JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 14 (2026) ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Leveraging MedlinePlus to Improve Health Information Access Among Patients and Caregivers: Systematic Literature Review (e79416)
Fei Yu, Xiaomeng Wang, Jia Liu, Tian Wang, Siobahn Grady, Lixin Song. . . . 12

Influencing Factors of Mobile Health Apps in Kidney Transplant Care: Systematic Review Using the Consolidated Framework for Implementation Research (e84139)
Yingtian Jia, Shaobo Guo, Xinran Yang, Xiaohong Lin, Jiaxin Fang, Lei Dong, Xiangru Li, Haiya Sun, Wanhui Yu, Hongxia Liu. . . . 36

Digital Public Reporting Systems for Evaluating Health Care Quality: Systematic Review (e80435)
Lanmei Du, Reima Suomi, Elorm Damalie. . . . 52

Machine Learning for Predicting Venous Thromboembolism After Joint Arthroplasty: Systematic Review of Clinical Applicability and Model Performance (e79886)
Junwei Ma, Huifeng Tang, Yunshan Zhang, Xuemei Yi, Tangsheng Zhong, Xinyun Li, Gang Wang. . . . 67

Using Health Information Systems to Support Behavioral Interventions in Local Contexts: Scoping Review (e77296)
Xi Yang, Martin Dijst, Joost Daams, Marc Suhrcke, Haoran Yang, Ronald Cornet. . . . 81

Core Elements, Development, and Implementation Strategies of the Nursing Minimum Data Set: Scoping Review (e84281)
Pascal Müller, Annabell Gangnus, Katrin Gayen, Uta Kirchner-Heklau, Patrick Jahn, Sebastian Hofstetter. . . . 109

Technological Strategies for the Patient Experience in Emergency Departments: Scoping Review (e79782)
Andrés Santafé, Juan Aranda, Wilmer Beltrán, William Guerrero, Maryory Guevara Lozano, Ingrid Bustos. . . . 119

Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review (e79039)
Dongliang Li, Syaheerah Lebai Lutfi. . . . 134

Causal Discovery in Observational Medical Research: Scoping Review (e82499)
Zuting Liu, Tian Luo, Hailin Ma, Jiali Mo, Xia Yang, Zhenglong Huang, Jingkun Li, Jie Kuang. . . . 154

XSL•FO
RenderX

Viewpoints

The Ethics of AI Scribes as Epistemic Agents (e88235)
Frank Ursin, Sabine Salloch. . . . 173

Developing Clinical Decision (Support) Systems Combining the Scientific and Regulatory Perspective: European Insights on Challenges, Requirements, and Practical Guidance (e72809)
Sanne Vrijlandt, Sanne Janssen, Annemarie van Rossum, Niek Achten, Rianne Oostenbrink. . . . 183

Dynamic Personalized Optimization: An AI Functionality Framework for Digital Therapeutics (e75256)
Dohyoung Rim. . . . 210

Viewpoint on the Consequences and Mitigation of Cognitive Bias in the Radiological Interpretation of Breast Cancer Imaging Using Artificial Intelligence (e78955)
Lorenzo Conti, Benedetta Capetti, Ottavia Battaglia, Roberto Grasso, Filippo Pesapane, Dario Monzani, Gabriella Pravettoni. . . . 216

Ethical Imperatives for Retrieval-Augmented Generation in Clinical Nursing: Viewpoint on Responsible AI Use (e79922)
Xinyi Tu, Chenghao Shi, Peilin Qian, Lizhu Wang. . . . 231

Improving Clinical Decision-Making in Treating Airway Diseases With an Expert System Built Upon the Free AI Tool Google NotebookLM (e78567)
Cheng-Hao Hsu, Ching-Li Hsu, Chih-Hsiang Tsou, Kuo-Fang Hsu, Hung-Yu Yang. . . . 239

Clinical Model Autophagy: The Risk of Interpretative Drift in Recursive Medical AI (e94813)
Pei Shih. . . . 271

Original Papers

Telepathology Development and Stakeholder Perspectives in China: Cross-Sectional Survey (e83514)
Linlin Wang, Fangfang Cui, Xianying He. . . . 192

Experiences With Integrating Medical Terminologies Into User Interfaces for a Decision Support System for Primary Care: Conceptual and Development Study (e74934)
Michaela Neff, Jannik Schaaf, Najia Ahmadi, Michele Zoch, Dania Schütze, Holger Storf. . . . 254

Developing a Suicide Risk Prediction Algorithm Using Electronic Health Record Data in Mental Health Care: Real-World Case Study (e74240)
Linda Hummel, Karin Lorenz-Artz, Joyce Bierbooms, Inge Bongers. . . . 277

Deep Learning Model Using Transfer Learning for Detecting Left Ventricular Systolic Dysfunction: Retrospective Algorithm Development and Validation Study (e83127)
Sungjae Lee, Jung-Woo Son, Sung-Ai Kim, Min-Soo Ahn, Sang Lee, Sang-Jin Han, Taehyun Joo, Yeongyeon Na, Sunghoon Joo, Hyun Ahn, Mineok Chang, Yeha Lee, Young Park. . . . 302

Classification of Cochrane Plain Language Summaries by Conclusiveness Using Transformer-Based Models and ChatGPT: Retrospective Observational Study (e72657)
Antonija Mijatovi , Luka Ursi , Nensi Brali , Ružica Bandi , Barbara a i , Ivan Buljan, Ana Maruši . . . 321

Rectal Cancer Radiotherapy Response Prediction: Retrospective Study of Development of a Deep Learning–Based Radiomics Model (e77313)
Yiqun Li, Hengchang Liu, Qiang Wei, Zhenqi Wang, Zhen Hu. . . . 333

XSL•FO
RenderX

Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study (e80163)
Jundan Chen, Zhihe Mao. . . . 354

Scalable and Privacy-Conscious End-to-End Processing of Large-Scale Clinical Data for Precision Medicine: Empirical Evaluation Study (e83487)
Jungwoo Lee, Sangwon Hwang, Kyu Lee. . . . 372

A Bilayer Feature Fusion Framework for Pan-Cancer Survival Prediction Based on Multihead Attention and Adaptive Differential Privacy: Model Development and Validation Study (e83743)
Yun Chen, Zhifang Deng, Lili Wang, Huanhuan Wang, Xiang Wu. . . . 386

Trust in AI-Supported Screening in General Practice Among Urban and Rural Citizens: Cross-Sectional Study (e69777)
Larisa Wewetzer, Katja Goetz, Soenke Freischmidt, Jost Steinhauser. . . . 406

Stakeholder Perspectives on Affinity Domains in Digital Health Interoperability: Qualitative Study (e83894)
Jan Bruthans, Petra Hospodková, Ilona Fejfarová. . . . 417

Current Status, Challenges, and Opportunities Associated With Implementing Clinical Data Interchange Standards Consortium Standards in Japanese Academic Medical Centers: Cross-Sectional Survey (e83774)
Kaori Nagai, Noriko Ito, Yuya Ikeda, Toshiki Saito. . . . 433

Automated ICD-10–Anchored Classification of Primary Care Text Data: Development and Evaluation of a Custom Multilabel Classifier (e86533)
Christina Haag, Thomas Grischott, Jakob Burgstaller, Stefan Markun, Oliver Senn, Viktor von Wyl. . . . 447

A Sentence Classification–Based Medical Status Extraction Pipeline for Electronic Health Records: Institutional Case Study (e77409)
Chuanming Dong, Boris Delange, Alex Poiron, Mohamed El Azzouzi, Clément François, Guillaume Bouzillé, Marc Cuggia, Sandie Cabon. . . 4 5 5

Iterative Large Language Model–Guided Sampling and Expert-Annotated Benchmark Corpus for Harmful Suicide Content Detection: Development and Validation Study (e73725)
Kyumin Park, Myung Baik, YeongJun Hwang, Yen Shin, HoJae Lee, Ruda Lee, Sang Lee, Je Sun, Ah Lee, Si Yoon, Dong-ho Lee, Jihyung Moon, JinYeong Bak, Kyunghyun Cho, Jong-Woo Paik, Sungjoon Park. . . . 477

Development and Evaluation of SNOMED CT Automated Mapping Tool: Advancing Terminology Standardization and Semantic Interoperability (e82670)
Youngsun Park, Hannah Kang, Jiwon Kim, Soo-Yong Shin, Dosang Cho, Sang Rhee, Hong Park, Kyung-Jae Lee, Sungchul Bae. . . . 501

Analysis Model for Infant Incubator Adverse Events Using Retrieval-Augmented Generation Combined With Dual-Adapter Fine-Tuning: Development and Evaluation Study (e83745)
Wenke Xia, Wanting Zhu, Tianchun Li, Li Wang, Weiqi Li, Peiming Zhang. . . . 517

Design Principles for Interactive Dashboards in Drug Safety Surveillance: Design Science Research (e75936)
Malwina Kotowicz, Mexhid Ferati, Soumitra Chowdhury, Cláudio Martins Pires. . . . 542

A Data-Centric Approach for Health Care and Research in a Health Knowledge Management Platform: Implementation and Requirement-Based Evaluation Study (e83608)
Björn Schreiweis, Benjamin Kinast, Hannes Ulrich, Tobias Bronsch, Ann-Kristin Kock-Schoppenhauer, Björn Bergh. . . . 575

Modeling Diabetes Risk and Progression With Public Health Data: Ontology-Guided, Simulation-Capable Digital Twin Study (e87374)
Qingrui Li, Kapileshwor Amat, Eric Johnson, Juan Li. . . . 596

XSL•FO
RenderX

Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making (e80553)
Adriano Fabi, Caroline Egli, Séverin Wendelspiess, Sebastian Griewing, Yvonne Haas, Laura De Pellegrin, Dirk Schaefer, Shan Qiu, Yves Harder, Elisabeth Kappos. . . . 615

Machine Learning Prediction of Progression to Dialysis in Patients With Polycystic Kidney Disease: Population-Based Retrospective Cohort Study (e80343)
Cheng-Hao Chang, Mingchih Chen, Ming-Hsien Tsai, Yen-Chun Huang, Hung-Hsiang Liou, Ben-Chang Shia, Chingying Liang, Yu-Wei Fang. 6 2 7

Mild Cognitive Impairment Detection System Based on Unstructured Spontaneous Speech: Longitudinal Dual-Modal Framework (e80883)
Yu-Shan Liao, Thiri Wai, Ting-Yun Liao, Ho-Ling Chang, Yu-Ling Chang, Li-Chen Fu. . . . 648

Enhancing Anesthetic Depth Assessment via Unsupervised Machine Learning in Processed Electroencephalography Analysis: Novel Methodological Study (e77830)
Po-Yu Huang, Wei-Lun Hong, Hui-Zen Hee, Wen-Kuei Chang, Ching-Hung Lee, Chien-Kun Ting. . . . 672

Concordance of Death Information Between Two Health Systems Serving the Same Region and the Social Security Administration Death Master File: Retrospective Observational Study (e85136)
Zhan Wang, Mahanaz Syed, Shweta Bansal, Meredith Zozus, Eric Eisenstein, Muayad Hamidi. . . . 684

Exploring Factors Associated With the Stalled Implementation of a Ground-Up Electronic Health Record System in South Africa: Qualitative Insights From the E-Tick Case Study Using the Consolidated Framework for Implementation Research (CFIR) (e73831)
Campion Zharima, Frances Griffiths, Jane Goudge. . . . 692

Development of Quality Indicators for the Correct Use of Electronic Medical Records in Primary Care: Modified Delphi Study (e80057)
Rico Paridaens, Steve Van den Bulck, Michel De Jonghe, Benjamin Fauquert, Liesbeth Meel, Willem Raat, Bert Vaes. . . . 708

Multi-Institutional Drug Use Patterns in Hospitalized Older Patients: Retrospective Cross-Sectional Study (e78353)
Chung Lee, Grace Kim, Suhyun Kim, Jee Hong, Won Hwang, Jong-Yeup Kim, Kye Lee, Kwangsoo Kim, Mingyu Kang, Ju Kim, Suehyun Lee. 7 1 9

Harmonizing Logical Observation Identifiers Names and Codes (LOINC) Codes and Units in Real-World Oncology Data: Method Development and Evaluation (e81254)
Parvati Naliyatthaliyazchayil, Travis Stenerson. . . . 739

Complete Blood Count–Derived Inflammation Indices to Predict 3-Year All-Cause Mortality in Patients With Diabetes and Acute Myocardial Infarction in Critical Care: Retrospective Cohort Study With Single-Center External Validation (e83328)
Tong Zhou, Kun Yang, Ying Yang, Song-Mei Liu. . . . 753

Neutrophil Percentage–to-Albumin Ratio as a Novel Prognostic Biomarker in Adult Diffuse Gliomas: Retrospective Study Integrating 3 Machine Learning Models and Cox Regression (e79945)
Congcong Zhu, Jiyang An, Lili Zhou. . . . 769

Predicting Pediatric Urological Surgery Duration Through Multimodal Patient-Physician Feature Fusion: Deep Learning Framework Incorporating Clinical Text Embedding (e82329)
Yonggen Zhao, Ruoge Lin, Yiying Sun, Lingdong Chen, Jian Huang, Guangjie Chen, Zhu Zhu, Gang Yu. . . . 793

XSL•FO
RenderX

Data-Efficient Language Model for Assessing Pulmonary Embolism Diagnostic Certainty From Radiology Reports: Model Development and Validation Study (e79972)
Feifan Liu, Ruofan Hu, Donghyuk Kim, Hao Lo, Yevgeniy Kharonov, Ansh Johri, Ben Gerber, Elke Rundensteiner, Lauren Westafer, Jeroan Allison, Catarina Kiefe, Alexander Bankier, Max Rosen. . . . 810

Methodological Development and Assessing Prescribing Determinants Through Cumulative Drug Exposure in Hospitalized Patients: Proof-of-Concept Retrospective Study (e76961)
Mathilde Bories, Guillaume Bouzillé, Marc Cuggia, Pascal Le Corre, Aurélie Bannay. . . . 827

Research on Risk Transfer Pathways for Lung Cancer Among Middle-Aged and Older Individuals Using Deep Reinforcement Learning: Retrospective Cohort Study (e74990)
Songjing Chen, Sizhu Wu. . . . 845

Real-Time Symptom Ratings Using Ecological Momentary Assessment Versus Traditional Questionnaires in Patients With Chronic Obstructive Pulmonary Disease: Observational Study (e79001)
Banchia Palmen, Maarten van Herck, Yvonne Goërtz, Zjala Ebadi, Qichen Deng, Melissa Thong, Chris Burtin, Jeannette Peters, Roy Sprooten, Erik Bischoff, Emiel Wouters, Jan Vercoulen, Sarah Houben-Wilke, Daisy Janssen, Martijn Spruit, Anouk Vaes. . . . 864

CT Radiomics–Based Machine Learning Model for Predicting Capsular and Neural Invasion in Thyroid Carcinoma: Diagnostic Accuracy Study (e77349)
Fang-fang Cong, Ke Tian, Qian Gao, Fulin Wang, Peng Sun, Nan Xu. . . . 880

Effects of Image Degradation on Deep Neural Network Classification of Scaphoid Fracture Radiographs: Comparison Study of Different Noise Types (e65596)
Chihung Lin, Alfred Yoon, Chien-Wei Wang, Tung Chao, Kevin Chung, Chang-Fu Kuo. . . . 901

Exploring Self-Management–Based Mobile Health User Typologies and Associations Between User Types and Satisfaction With Key Mobile Health Functions: Comparative Study of Various Fitness and Weight Management App User Types (e64860)
Tong Wang, Yiwen Fan, Zheng Li, Xiaoyi Jiao, Qichuan Fang, Junhao Ma, Jianbo Lei. . . . 921

Balancing Privacy and Utility in Child and Adolescent Mental Health Services Research: Retrospective Cohort Study on Synthetic Data Generation (e71819)
Mounir Haizoune, Bennett Leventhal, Dipendra Pant, Øystein Nytrø, Kaban Koochakpour, Roman Koposov, Lars Øhlckers, Norbert Skokauskas. . . . 941

Enhancing Model Generalizability in Medical Artificial Intelligence: Systematic Comparison of Categorical Encoding and Sampling Techniques for Imbalanced Data (e75655)
Chien-wei Chuang, Chung-Kuan Wu, Chao-Hsin Wu, Ben-Chang Shia, Mingchih Chen. . . . 958

Enterocutaneous Fistula–Associated Sepsis and Mortality: Development and Validation of a Multimodal Artificial Intelligence Prediction Model (e79985)
Hui Li, Jing Chen, Peijun Lin, Youmei Pan, Yawen Cao, Wenfeng Xie. . . . 981

Developing a Thai User Interface Terminology for Systematized Nomenclature of Medicine Clinical Terms Implementation in Primary Care: Cross-Sectional Content Coverage Analysis (e80039)
Nat Tangchitnob, Wanchana Ponthongmak, Boonchai Kijsanayotin, Oraluck Pattanaprateep, Sithakom Phusanti, Pongsakorn Atiksawedparit, Kamonporn Suwanthaweemeesuk, Jirayus Siangfu, Gareth McKay, John Attia, Ammarin Thakkinstian. . . . 1003

Large Language Models for Psychiatric Diagnosis Based on Multicenter Real-World Clinical Records: Comparative Study (e77699)
Maoqian Sun, Jia Yu, Zhuhong Long, Yun Yang, Tao Xiao, Jiaquan Liang, Jun Feng, Huaili Deng, Guoping Huang. . . . 1016

Acceptance of Artificial Intelligence in Clinical Practice Among Chinese Physicians: Nationwide Cross-Sectional Survey Using Extended Unified Theory of Acceptance and Use of Technology and Explainable Machine Learning (e85270)
Xuefei Shi, Zhanxiao Tian, Qi Guo, Boxuan Qiao, Xiaolong Wang. . . . 1030

XSL•FO
RenderX

An Intelligent Interactive Management Platform for Rheumatoid Arthritis Care: Real-World Observational Study (e90784)
Ziyun Zhang, Xinyue Zhang, Li Zhang, Mingyan Du, Lijuan Zhang, Siyu Yang, Xiaoli Cai, Shengchao Hou. 1045

Identifying and Analyzing Bot-Generated Responses in Online Health Care Surveys: Methodological Study (e73622)
Emily Hamovitch, Kaileah McKellar, Walter Wodchis. 1056

Chatbot Outreach in Value-Based Preventive Care: Retrospective Analysis (e81370)
Jincy Chacko, Yuane Jia, Kimon Stathakos, Kathleen Mazza, Doran Kim, Michael Diefenbach, Lorena Carlo. 1070

Development of a Practical Nomogram for Depression Risk Stratification in Older Adults With Hypertension and Diabetes: Retrospective Analysis of Data From the China Health and Retirement Longitudinal Study (e81529)
Ting Peng, Ying Zhang, Rujia Miao, Jiangang Wang. 1086

Deep Learning for Content-Based Medical Image Retrieval in Picture Archiving and Communication Systems for Brain Tumor Detection: Algorithm Development and Validation (e78300)
Chin-Lin Lee, Tzu-Hsuan Hsu, Yu-Te Wu, Wan-Yuo Guo, Woei-Chyn Chu, Chung-Yueh Lien. 1098

Two-Minute Deep Learning–Powered Brain Quantitative Mapping: Accelerating Clinical Imaging With Synthetic Magnetic Resonance Imaging (e79389)
Yawen Liu, Hongxia Yin, Zuofeng Zheng, Wenjuan Liu, Tingting Zhang, Linkun Cai, Haijun Niu, Han Lv, Zhenghan Yang, Zhenchang Wang, Pengling Ren. 1118

Nomograms Based on X-Ray Radiomics for Predicting Pain Progression in Knee Osteoarthritis Using Data From the Foundation for the National Institutes of Health: Development and Validation Study (e78338)
Yingwei Sun, Jing Liu, Chunbo Deng, Chengbao Peng, Shinong Pan, Xueyong Liu. 1134

Linking Electronic Health Records for Multiple Sclerosis Research: Comparative Study of Deterministic, Probabilistic, and Machine Learning Linkage Methods (e79869)
Ohoud Almadani, Yasser Albogami, Adel Alrwisan. 1147

Stress Hyperglycemia Ratio and the Risk of Sepsis in Patients With Heart Failure: Retrospective Cohort Study From Medical Information Mart for Intensive Care-IV (e81229)
Xianzhen Cai, Baoxin Yan, Jinhao Chen, Junjun Ye, Jiating Su, Shuangshuang Tong, Bin Xie, Bingna Zhang, Jilin Li. 1156

Knowledge-Guided Explainable Recommendation Tool for Cancer Risk Prediction Models Using Retrieval-Augmented Large Language Models: Development and Validation Study (e78519)
Shumin Ren, Xin Zheng, Jing Zhao, Jiale Du, Yuxin Zhang, Cheng Bi, Jie Song, Jinyi Zhang, Hongmei Lang, Fan Zhang, Bairong Shen. 1174

Impact of Age on Hospital Outcomes Following Minimally Invasive Posterior Lumbar Interbody Fusion: Retrospective Analysis of the Nationwide Inpatient Sample Database from 2016 to 2020 (e76424)
Yu-Jun Lin, Fu-Yuan Shih, Jin-Fu Huang, Chun-Wei Ting, Yu-Chin Tsai, Lin Chang, Yu-Hua Huang, Ming-Jung Chuang. 1190

Mortality Prediction Among People Living With HIV on Antiretroviral Therapy in Public Health Facilities in Gondar City Administration, Northwest Ethiopia: Machine Learning–Based Study (e78770)
Andualem Gedefaw, Getaye Biwota, Abraraw Mesele, Abraham Mengistu, Gizaw Teferi. 1201

Detection of Interpretable and Fine-Grained Brain Tumor Magnetic Resonance Imaging Based on Progressive Pruning: Machine Learning Model Development and Validation Study (e84095)
Yupeng Liu, Shuwei Song, Shibo Lian, Xiaochen Zhang. 1220

XSL•FO
RenderX

Comparative Performance of 3 Analytical Models in Identifying Associated Factors of Pulmonary Dysfunction–Depression Comorbidity: China Health and Retirement Longitudinal Study–Based Nationwide Cross-Sectional Study (e77940)
Qinglin Cheng, Qiancheng Cao, Weilin Teng, Ruoqi Dai, Qingjun Jia, Xuexin Bai, Qingchun Li, Yifei Wu, Yinyan Huang. 1249

Code-Based Versus AutoML Methods for Pill Recognition in Clinical Settings: Comparative Performance Study (e79160)
Amir Ashraf, Richárd Rádli, Zsolt Vörösházi, András Fittler. 1270

Bridging Population Patterns and Individual Prediction: Framework for Prospective Multimorbidity Study (e84261)
Qianyao Zhang, Runtong Zhang, Weiguang Ma, Butian Zhao, Xiaomin Zhu. 1289

Enhanced Prediction of Atrial Fibrillation in Patients With Ischemic Stroke Through Electronic Medical Records and Text Mining: Algorithm Development and Validation (e78117)
Yu-Wei Chen, Sheng-Feng Sung, Ya-Han Hu, Yu-Hsuan Yang. 1322

Machine Learning–Based Risk Prediction for Coronary Heart Disease Complicated by Hyperhomocysteinemia: Retrospective Study (e80809)
Ming-Yuan Du, Meng-Ke Lyu, Hai-long Liu, Yi-zhuo Li, Hai-feng Yan, Xiao-hui Li. 1339

Research on the Prediction of Coal Workers' Pneumoconiosis Based on Easily Detectable Clinical Data: Machine Learning Model Development and Validation Study (e80156)
Haiquan Li, Jiaqi Jia, Xu Shi, Yudie Dong, Songquan Wang, Yuming Cui, Wenlu Hang, Dekun Zhang. 1353

Assessing the Impact of Sociodemographic Factors on Artificial Intelligence Models in Predicting Dementia: Retrospective Cohort Study (e80405)
Xingyi Liu, Muskan Garg, Maria Vassilaki, Jennifer St Sauver, Ronald Petersen, Momin Malik, Chung Wi, Young Juhn, Sunghwan Sohn. . . .
1 3 6 8

Screening for Clinically Significant Nephrolithiasis Based on Simple Health Checkup Clinical and Urine Parameters in General Populations: Multicenter Machine Learning Study (e80764)
Hao-Wei Chen, Pei-Siou Wei, Yu-Chen Chen, Jeng-Yih Wu, Chia-I Lin, Yii-Her Chou, Yung-Shun Juan, Wen-Jeng Wu, Chung-Yao Kao, Jung-Ting Lee. 1386

Prediction of First and Multiple Antiretroviral Therapy Interruptions in People Living With HIV: Comparative Survival Analysis Using Cox and Explainable Machine Learning Models (e78964)
Donald Salami, Emily Koech, Janet Turan, Kristen Stafford, Lilly Nyagah, Stephen Ohakanu, Anthony Ngugi, Manhattan Charurat. 1404

Prospective Diagnostic Accuracy and Technical Feasibility of Artificial Intelligence-Assisted Rib Fracture Detection on Chest Radiographs: Observational Study (e77965)
Shu-Tien Huang, Liong-Rung Liu, Ming-Feng Tsai, Ming-Yuan Huang, Hung-Wen Chiu. 1424

Machine Learning Prediction of Pharmacogenetic Testing Uptake Among Opioid-Prescribed Patients Using Electronic Health Records: Retrospective Cohort Study (e81048)
Mohammad Yaseliani, Je-Won Hong, Jiang Bian, Larisa Cavallari, Julio Duarte, Danielle Nelson, Wei-Hsuan Lo-Ciganic, Khoa Nguyen, Md Hasan. 1440

Development of Venous Thromboembolism Risk Prediction Models Based on Whole Blood Gene Expression Profiling Using 20 Machine Learning Algorithms: Comprehensive Analysis Study (e75565)
Yedong Huang, Xiaoyun Chen, Guannan Bai, Yajun Zhao, Dapeng Kuang, Lin Zhang, Wei Lu. 1455

Automated Classification of Lymphoma Subtypes From Histopathological Images Using a U-Net Deep Learning Model: Comparative Evaluation Study (e72679)
Jin Zhao, Xiaolian Wen, Li Ma, Liping Su. 1466

XSL•FO
RenderX

Deep Learning for Dynamic Prognostic Prediction in Minimally Invasive Surgery for Intracerebral Hemorrhage: Model Development and Validation Study (e86327)
Jingxuan Wang, Jian Shi, Qing Ye, Danyang Chen, Yuhao Sun, Chao Pan, Yingxin Tang, Ping Zhang, Zhouping Tang. 1479

Ranking-Aware Multiple Instance Learning for Histopathology Slide Classification: Development and Validation Study (e84417)
Ho Kim, Gisu Hwang, Won Jeong, Young Ko. 1494

Early Prediction of Delirium in Postcardiac Surgery Patients: Machine Learning Model Development and External Validation (e73283)
Huixiu Hu, Yuxiang Wang, Houfeng Li, Qinglai Zang, Jing Huang, Ying Zhang, Jinjing Wu, Long Liu, Zhen Xing, Yaohua Yu. 1508

Ensemble Machine Learning Models for Predicting Patients With High Usage: Model Validation and Economic Impact Analysis (e77202)
Joshua Tan, Le Quan, Hao Tan, Su-Yen Goh, Julian Thumboo, Marianne Au, Yong Bee, Jen Tan. 1523

Enhanced Magnetic Resonance Imaging–Based Knee Cartilage Segmentation Using a Swin-UNet Conditional Generative Adversarial Network: Development and Validation Study (e86155)
Jun Park, Ji-Hoon Nam, Shakhboz Abdigapporov, Jong-Keun Kim, Yong-Gon Koh, Byung Cho, Hyuck Kwon, Kwan Park, Kyoung-Tak Kang. .
1 5 3 6

Leveraging Naturalistic Driving Digital Biomarkers for Early Mild Cognitive Impairment Detection: Deep Learning Strategies (e83622)
Firas Al-Hindawi, Teresa Wu, Yutong Wen, Peter Serhan, Erica Forzani, Francis Tsow, Yonas Geda. 1550

Explainable Machine Learning for Assessing Digital Health Literacy in Older Adults: Validation and Development of a Two-Stage Model Integrating Performance-Based and Self-Assessed Indicators (e86171)
Choonghee Park, Jiyeon Park, Seora Kim, Ye Bae, Jae-Heon Kang, Tae-Min Kim, Ji-Won Chun. 1571

Responsible AI for Predicting Delayed Hospital Discharge Among Older Adults: Development and Evaluation Study for Balancing Accuracy, Equity, and Explainability (e83244)
Somayeh Ghazalbash, Manaf Zargoush, Sara JT Guilcher, Kerry Kuluski. 1591

Comparison of Feature Selection Methods in Machine Learning Models of Cancer Information Seeking Among United States Adults: Cross-Sectional Study (e75862)
Ying Liu, Kesheng Wang. 1612

Patient Experiences With Outpatient Pharmacy Services in Hospitals Using Automated Pharmacy Systems: Cross-Sectional Study (e80963)
Ali Buabbas, Fatemah Alsaleh, Mariam Alsanafi, Maryam Safar, Ahmad Al-Sultan, Talal Zaman. 1641

Identification and Localization of Breast Tumor Components via a Convolutional Neural Network Based on High-Frequency Ultrasound Combined With Histopathologic Registration: Prospective Study (e81181)
Jia-Qian Yao, Wen-Wen Zhou, Zhi-Fei Chai, Fei Ren, Tong-Yi Huang, Tian-Tian Zhen, Hui-Juan Shi, Xiao-Yan Xie, Ze Zhao, Ming Xu. 1667

Automatic Recognition and Prognostic Prediction of Colorectal Liver Metastases Using a Multi-Scale Deep Learning Framework: Model Development and Validation Study (e73311)
Houwen Long, Feng Ying, Shi Wu, Long Li. 1679

Prompting and Fine-Tuning Large Language Models for Parkinson Disease Diagnosis: Comparative Evaluation Study Using the PPMI Structured Dataset (e77561)
Hyun-Ji Shin, Young Jeong, Sungmin Jun, Do-Young Kang. 1697

XSL•FO
RenderX

CO2 Laser Popularity in Germany: A Five-Year Google Trends Analysis (2020 2025) (e77651)
Michael Kirchberger, Andreas Eisenried. 1718

Applicability of Existing Gender Scores for German Clinical Research Data: Scoping Review and Data Mapping (e74162)
Lea Schindler, Hilke Beelich, Elpiniki Katsari, Daniele Liprandi, Sylvia Stracke, Dagmar Waltemath. 1726

Evaluation of AI Chatbot Responses to a Standardized Patient Query on Myelin Oligodendrocyte Glycoprotein Antibody–Associated Disease: Cross-Sectional Content Analysis (e81720)
Meryem Sönmez, Mehmet Yetkin, Duygu Mehdiyev, Nazlı Çelik, Merve Ercan, Pınar Öztürk, Ye im Akbo a, Emine Koç, Semra Mungan. . . . .
1 7 5 2

Large Language Model Automated Extraction of Clinical Signs and Symptoms From Emergency Department Reports for Machine Learning Prediction Models: Development and Validation Study (e81500)
Anoeska Schipper, Peter Belgers, Rory O'Connor, Lieke van de Wouw, Luc Builtjes, Joeran Bosma, Ron Kusters, Steef Kurstjens, Matthieu Rutten, Bram van Ginneken. 1767

Approval of AI-Based Medical Devices in China From 2020 to 2025: Retrospective Analysis (e85538)
Lingli Zhang, Jianzhou Yan. 1796

Tailoring Discharge Summaries to Health Care Providers' Needs (Part 1 of the Framework and Implementation of AI Tools Project): User-Centered Design Approach (e80613)
Mieke Deschepper, Helga Rogge, Mathias Syx, Kirsten Colpaert. 1809

Evaluating GPT-4 Responses on Scars or Keloids for Patient Education: Large Language Model Evaluation Study (e78838)
Mingjun Rao, Tang Xiujun, Wang Haoyu. 1821

Large Language Models in Patient Health Communication for Atherosclerotic Cardiovascular Disease: Pilot Cross-Sectional Comparative Analysis (e81422)
Pengfei Li, Yinfei Xu, Xiang Liu, Zhean Shen, Yi Wang, Xinyi Lv, Ziyi Lu, Hui Wu, Jiaqi Zhuang, Yan Chen. 1830

Multi-Evidence Clinical Reasoning With Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study (e82026)
Hang Wong, Tsz Wong. 1845

Machine Learning Algorithms to Predict Venous Thromboembolism in Patients With Sepsis in the Intensive Care Unit: Multicenter Retrospective Study (e80969)
Yan Zhang, Xia Ren, Luojie Liu, Junjie Zha, Yijie Gu, Hongwei Ye. 1864

Evaluation of Large Language Models for Radiologists' Support in Multidisciplinary Breast Cancer Teams: Comparative Study (e68182)
Hong Jiang, Chun Yang, Wenbin Zhou, Cheng-liang Yin, Shan Zhou, Rui He, Guanghui Ran, Wujie Wang, Meixian Wu, Juan Yu. 1885

Artificial Intelligence Models for Predicting Triage in Emergency Departments: Seven-Month Retrospective Comparative Study of Natural Language Processing, Large Language Model, and Joint Embedding Predictive Architectures (e83318)
Edouard Lansiaux, Ramy Azzouz, Emmanuel Chazard, Amélie Vromant, Eric Wiel. 1901

Application of a Large Visual Language Model on Tongue Image Description Generation and Physical Constitution Reasoning in Traditional Chinese Medicine (TongueVLM): Model Development and Validation Study (e87237)
Chengdong Peng, Jun Gao, Nuo Yang, Yong Wang, Renming Chen, Changwu Dong. 1922

Emotion Expression in Breast Cancer Support Seeking: Empirical Study of an Online Community (e83674)
Anqi Xu, Ahmad Aljanaideh, Jennifer Xu, Haijing Hao. 1935

XSL•FO
RenderX

Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study (e86474)
Sandra Taylor, Melissa Jost, Scott MacDonald, Yunyi Ren, Shelley Hilton, Sadie Davenport, Debbie Aizenberg, Bruce Hall, Courtney Lyles, Jason Adams. . . . 1950

Impact of an Ambient AI Scribe Among Clinicians and Patients: Real-World Prospective Observational Time-Motion Study (e85580)
Jonathan Tan, Iffat Rafi, Gerald Sng, Joshua Tung, Daniel Lim, Jasmine Ong, Gek Lim, Eileen Lew, Kuok Chia, Mark Heng. . . . 1958

A Bilingual Arabic-English Ambient AI Scribe for Clinical Documentation: Prospective Evaluation Study (e83335)
Umair Khan, Ammar Khan, Waleed Aljaadi, Razan Alhadlaq, Zahran Baqashmer, Yasin Alsafi, Yousef Alomran, Maha Al Rusaiyes, Muaddiyah Radif, Tahir Khan, Saleh Altamimi. . . . 1968

Large Language Model–Enabled Editing of Patient Audio Interviews From “This Is My Story” Conversations: Comparative Study (e80205)
Bikram Bains, Sampath Rapuri, Edgar Robitaille, Jonathan Wang, Arnav Khera, Catalina Gomez, Eduardo Reyes, Cole Perry, Jason Wilson, Elizabeth Tracey. . . . 1981

Implementation Report

Scaling Wireless Continuous Vital Sign Monitoring Across an 8-Hospital Health System: Digital Health Implementation Report (e78216)
Ngoc-Anh Nguyen, Grace Lee, Brendan Holderread, Terrie Holman, Sarah Pletcher, Roberta Schwartz. . . . 291

Letter to the Editor

Enhancing Predictive Accuracy of Mood Symptoms Using Wearable Devices and Machine Learning in Bipolar Disorder (e92172)
P Nair, Priyanka D'Souza. . . . 1631

Corrigenda and Addendas

Correction: Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study (e92491)
Jundan Chen, Zhihe Mao. . . . 1633

Correction: An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation (e94047)
He Xu, Bernard Maccari, Hervé Guillain, Julien Herzen, Fabio Agri, Jean Raisaro. . . . 1634

Correction: Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation (e88845)
JMIR Editorial Office. . . . 1640

XSL•FO
RenderX

Editorial

AI Scribes: Are We Measuring What Matters? (e89337)
Enrico Coiera, David Fraile-Navarro. . . . 1664

Research Letter

AI-Enabled Customer Relationship Management Platforms for Patient Services in Health Care, Early Lessons From Governance, and Program-Level Outcomes (e83564)
Anup Gupta. . . . 1748

Retractions

Retraction: Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study (e98661)
JMIR Publications Editorial Office. . . . 1993

Retraction: Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study (e98666)
JMIR Publications Editorial Office. . . . 1994

Retraction: Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation (e98667)
JMIR Publications Editorial Office. . . . 1995

Retraction: Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development (e98668)
JMIR Publications Editorial Office. . . . 1996

Retraction: Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development (e98669)
JMIR Publications Editorial Office. . . . 1997

Retraction: An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis (e98670)
JMIR Publications Editorial Office. . . . 1998

Retraction: Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring (e98672)
JMIR Publications Editorial Office. . . . 1999

Retraction: Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation (e98673)
JMIR Publications Editorial Office. . . . 2000

Retraction: Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study (e98674)
JMIR Publications Editorial Office. . . . 2001

XSL•FO
RenderX

Leveraging MedlinePlus to Improve Health Information Access Among Patients and Caregivers: Systematic Literature Review

Fei Yu[1], PhD, MPS; Xiaomeng Wang[2], MSIS; Jia Liu[3], PhD; Tian Wang[4], PhD; Siobahn Day Grady[5], PhD; Lixin Song[3], RN, PhD

[1]Carolina Health Informatics Program, School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive CB #3360, 100 Manning Hall, Chapel Hill, NC, United States

[2]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States

[3]University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[4]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, United States

[5]School of Library and Information Sciences, North Carolina Central University, Durham, NC, United States

Corresponding Author:
Fei Yu, PhD, MPS
Carolina Health Informatics Program, School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive CB #3360, 100 Manning Hall, Chapel Hill, NC, United States

Abstract

Background: MedlinePlus, developed by the National Library of Medicine (NLM) in the United States, is one of the most widely used, authoritative, consumer-grade health information resources on the web. Although extensively used and discussed in scholarly work for health literacy and patient education, it is unclear how MedlinePlus has been integrated into clinical care or embedded within health informatics applications.

Objective: This study aimed to understand how MedlinePlus has supported patients and caregivers by increasing access to health information for clinical care and illness management. The insights on this topic will inform the design and development of patient-facing digital health intervention tools for improved health communication, decision engagement, informed decision-making, and health outcomes.

Methods: We conducted a systematic literature review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. First, we developed a comprehensive literature search strategy, searched 9 citation databases, and aggregated and deduplicated search results before importing them into Covidence for manual screening using predefined inclusion and exclusion criteria. Second, reviewers independently assessed all studies at the title-abstract and full-text levels, resolving discrepancies through ongoing discussions. Third, we applied the PICO (problem/population, intervention, comparison, and outcome) and the Collaborative Chronic Care Model as guiding frameworks for data extraction and analysis. All included studies underwent quality assessment using the Mixed Methods Appraisal Tool.

Results: In total, 28 studies reported in 27 sources met our inclusion criteria. We categorized the extracted data into 4 areas. First, regarding bibliometrics, the studies were reported between 2004 and 2024, with 2010 having the highest number of studies. Of these studies, 25 were conducted in the United States, 2 were conducted in Iran, and 1 was conducted in Argentina. Health informatics journals and conference proceedings, as well as library science journals, were prominent publishing venues. The NLM funded half of the studies. Second, regarding participants, most studies focused on outpatients. Other participant roles included physicians, nurses, hospital staff, pharmacists, and librarians. Fewer than half of the studies addressed the social determinants of health. Third, regarding intervention, most studies implemented MedlinePlus information interventions within clinical settings. Other interventions occurred in community pharmacies, community organizations, libraries, online health platforms, or patient portals. Fourth, regarding outcome, only 4 studies assessed clinical outcomes, and the findings were mixed and inconsistent. However, 24 of 28 studies reported positive nonclinical outcomes, including improved attitudes toward and satisfaction with MedlinePlus and enhancements in patients' information-seeking behaviors, confidence, and willingness to engage in decision-making, physician-patient communication, self-management, and self-efficacy.

Conclusions: This systematic literature review is the first comprehensive examination of how MedlinePlus has been integrated into clinical care, supporting patients and caregivers with enhanced access to health information. Our findings offer evidence and insights through the Collaborative Chronic Care Model lens and can guide the development of digital health interventions to improve patient health.

(JMIR Med Inform 2026;14:e79416) doi:10.2196/79416

XSL•FO
RenderX

KEYWORDS

MedlinePlus; health information access; information intervention; information prescription; patients and caregivers; systematic literature review

Introduction

The National Library of Medicine (NLM) in the United States launched the MedlinePlus website in 1998 to provide the public with free access to "high-quality, relevant health and wellness information that is trusted, easy to understand, and free of advertising" [1]. Being available in both English and Spanish, MedlinePlus provides content covering more than 1000 diseases and health conditions, almost 300 medical test descriptions, more than 1500 brand-name and generic drugs, 1300 genetic conditions and 1400 genes, and dietary supplements and herbs, as well as more than 4000 medical encyclopedia articles and healthy recipes. In addition to textual content, MedlinePlus offers animated health videos produced by the NLM or selected from accredited external organizations and plentiful interactive health assessment tools, enabling people to assess their risk for diseases or health conditions and maintain healthy lifestyles [2].

Over the past two decades, the NLM has continuously expanded and enhanced MedlinePlus by adding new features and content to meet the evolving needs of its users. In particular, in the era of electronic health records (EHRs), when patients' and caregivers' engagement in shared decision-making is empowered by their access to medical records, the NLM activated MedlinePlus Connect [3]. This application programming interface service links MedlinePlus content to patient portals and EHRs, greatly extending the scope of MedlinePlus information availability, compatibility, and usage. For example, studies show that patient portals embed MedlinePlus to improve medical record comprehension [4,5].

MedlinePlus has become a widely used free health information resource, serving millions of users worldwide and supporting health care professionals, educators, and researchers in providing evidence-based health education and improving health literacy. In particular, it has been frequently cited in the scholarly literature on the concept and practice of information prescription (IP) because of its demonstrated utility in patient care. IP refers to the physician's prescription of current, personalized, evidence-based, and reliable health information to patients as part of their treatment plan to understand, manage, and control their health problems and to better engage in their health care [6-8]. Although IP can take paper or electronic format, the Information Rx (IRx) program, initiated in 2002 by the American College of Physicians Foundations and the NLM, provided physicians with prescription pads designed to guide patients to relevant content on the MedlinePlus website, tailored to their specific health conditions [9]. The IP and IRx initiatives positioned MedlinePlus as one of the key resources in disseminating consumer-grade health information. A recent scoping review confirmed that "the most commonly prescribed website to produce IP content is MedlinePlus," compared with 15 other identified IP contents [8].

Although researchers unanimously agree on the value of IP and IRx in patient education, self-management, and decision engagement, the integration of MedlinePlus into clinical practice, particularly embedding it within health care applications such as EHRs or digital health interventions, is unknown. No studies have systematically examined the extant research evidence about how MedlinePlus has been adopted, implemented, or integrated to improve patient care and health outcomes, especially in the context of designing and developing health information technology (Health IT). To address this gap, this study aimed to investigate the role of MedlinePlus in supporting patients and their caregivers by increasing access to health information pertinent to their care and illness management, with a focus on Health IT. A systematic mapping of the literature will inform the design and development of patient-centered digital health interventions that enhance health communication, decision engagement, informed decision-making, and health outcomes.

Methods

This study used a systematic literature review methodology, adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1) [10] for literature search, study selection, and data extraction.

Literature Search

A literature search strategy was first developed by connecting the 2 main keywords "MedlinePlus" and "Medline Plus" with the Boolean operator "OR" and then customizing the query based on the field codes of the chosen citation database. This strategy was executed across 9 citation databases, including PubMed, Embase, CINAHL, Global Health, PsycINFO, Cochrane Database of Systematic Reviews, Scopus, Web of Science, and Dimensions. The first 5 databases are well-known research article indexing resources in the health sciences and are primary sources for clinical studies, including Health IT and digital health interventions. We also searched the Cochrane reviews, the gold standard for systematic reviews in biomedical and health sciences, to identify relevant reviews and examine their cited empirical studies, ensuring comprehensive coverage of the literature. Furthermore, Scopus, Web of Science, and Dimensions were included because they index a broad range of peer-reviewed research articles across medicine and health sciences and are commonly used in systematic reviews and evidence-mapping studies. Our initial search was conducted in October 2021 and was updated in October 2022 and in September 2024 (Multimedia Appendix 1). In addition, we checked all references of the included articles (backward searching) and newer articles that cited the included studies (forward searching) in Scopus for an exhaustive literature search.

Study Selection

The search results from each database were exported, aggregated, deduplicated, and imported into Covidence [11] for

XSL•FO
RenderX

screening. Independent reviewers from our research team conducted the screening process, initially at the title-abstract level (FY, Erin Harvey, and Pengxuan Wang), followed by a full-text assessment (FY and XW). Discrepancies between reviewers were resolved through discussion to achieve agreement.

The studies were included if they (1) used MedlinePlus as the primary or one of the primary information or education resources for patients and/or their caregivers; (2) targeted patients or caregivers, as indicated by the recruited participants or by the research aims; (3) were English literature; and (4) were original research articles. The studies were excluded if they were (1) non-English literature, (2) did not use MedlinePlus as one of the information sources, (3) did not include patients or caregivers as direct information recipients (eg, only physicians or other health care professionals are information or training recipients), (4) were not original research articles (eg, perspectives, comments, editorials, news, bulletins, erratum, or abstract only), (5) were information readability or coverage assessment studies, or (6) were information retrieval, search behavior, or query analysis studies.

Data Extraction and Synthesis

To achieve the research aim, we adopted 2 frameworks to guide our data extraction and synthesis process. The first is the PICO (problem/population, intervention, comparison, and outcome) framework [12], widely recognized and used in evidence-based practice and research in health sciences. The PICO framework systematically addresses the targeted population, specific intervention or exposure, intervention comparison to an alternative (if applicable), and desired outcomes or end points. The second framework is the Collaborative Chronic Care Model (CCCM) [13], a refined version of the widely adopted original chronic care model. The CCCM emphasizes collaborative, continuous, evidence-based, and proactive care delivery for individuals with chronic, medical, or mental health conditions. The CCCM encompasses 6 major elements: health system organization, work role redesign, self-management support, decision support, information management, and community linkages. We created a customized data extraction framework (Table 1) and organized the data accordingly. Guided by the PICO and CCCM frameworks, our data extraction focused on the following 4 aspects: bibliometrics, participants, interventions, and outcomes.

Table . Data collection, extraction, and analysis framework.

Measure	Data extraction metrics	Data source	Analysis
Bibliometrics	Study distribution by year, source, country, and funding; authors and institutions; and citation and digital impact	Full-text article, Scopus, Altmetric Explorer	Quantitative
Participants	Participants' role, diagnosis, demographics (including SDOH[a]), sample size, and age range	Full-text article	Qualitative and quantitative
Interventions	Intervention features, information content, additional interventions, and setting	Full-text article	Qualitative
Outcomes	Clinical outcome, nonclinical outcome, outcome comparison, study duration, and follow-up method	Full-text article	Qualitative and quantitative

[a]SDOH: social determinants of health.

Bibliometrics

Bibliometrics capture the document details of the included studies, including reporting year, source title, funding sources, author names and affiliations, and citation impact. We assessed the citation impact of included studies using both raw citation counts (ie, how many times an article has been cited) and the field-weighted citation impact (FWCI), a field- and time-normalized citation impact metric [14] provided by the Elsevier Scopus database. An FWCI of 1 indicates that, compared with papers published in the same field and time, this article is cited at the global average level; a value greater than 1 suggests that it is cited more than the worldwide average and vice versa. Being an improved citation impact measure, FWCIs have been commonly adopted by research impact evaluators [15,16]. In addition, we used Altmetrics to capture the publication's digital attention [17], with scores obtained from Altmetric Explorer [18] through an institutional subscription. The composite Altmetric Attention Score (AAS) is a weighted and quantified measure of the online attention and engagement that a research output (eg, journal article) has received by tracking various sources such as public policy documents, mainstream news, social networks, blogs, and other platforms. Generally, an AAS of 20 or higher is considered good [19] and is often used as a benchmark in research evaluations [20].

Participants

The participant category captures participants' roles, patient diagnoses, CCCM-aligned demographics (including social determinants of health [SDOH]), sample size, and age range of patients and caregivers who used MedlinePlus. Participant roles include physicians, nurses, research associates, patients, and caregivers involved in using MedlinePlus as an information intervention. Demographics include SDOH, such as self-management skills, education and literacy levels, and economic and financial factors (eg, income, employment, and

XSL•FO
RenderX

health insurance). Reporting SDOH and demographic characteristics (eg, race, gender, and primary language) provides contextual information and aligns with CCCM's emphasis on whole-person care and linkage to community resources.

Interventions

Interventions describe how MedlinePlus was used as a health information intervention, including the type of content assessed, platform features used, and modes of delivery. Using the CCCM-aligned construct of intervention delivery, we systematically documented the key intervention attributes, including intervention features, health information content, any additional interventions beyond the MedlinePlus resource, and the settings where the interventions were implemented.

Outcomes

Outcomes cover clinical and nonclinical outcomes as well as outcome comparison and assessment, if applicable. Clinical outcomes are associated with "a measurable change in symptoms, overall health, ability to function, quality of life, or survival outcomes that result from giving care to patients" [21]. Nonclinical outcomes encompass a wide range of measures, including participants' attitudes toward health information resources, health information–seeking behavior, experience with health information interventions, information usage, and perceptions. We also examined the study durations and follow-up methods used to assess outcomes of MedlinePlus as an information intervention.

In total, 3 reviewers (FY, Pengxuan Wang, and XW) independently extracted data from the Scopus database, Altmetric Explorer, and the full-text articles of the included studies. We recorded data in a Microsoft Excel template for each included study based on our data extraction framework and discussed and reconciled data inconsistencies to reach consensus.

Quality Assessment

Two reviewers (FY and XW) independently assessed the quality of the included studies using the Mixed Methods Appraisal Tool (MMAT) [22] by following the MMAT user guide. MMAT has been widely used in systematic reviews to appraise qualitative studies, quantitative studies, randomized controlled trials (RCTs), nonrandomized studies, and mixed methods studies [23,24]. The MMAT tool offers a checklist. First, each study undergoes 2 screening questions to ensure that it is an empirical study (ie, "S1. Are there clear research questions?" and "S2. Do the collected data allow addressing the research questions?"). Second, reviewers select an appropriate study category for each included study. Third, based on the 5 criteria in the selected category, reviewers rate each criterion as "Yes," "No," or "Can't tell," documenting any observations in a "Comments" column. We rated "Yes" when a study provided sufficient information to fully comply with the criteria, "No" when the study failed to meet the criteria, and "Can't tell" when the information in the full-text article was insufficient, unclear, or partially met the criteria.

After discussing and resolving discrepancies between the 2 reviewers' independent ratings, the research team achieved consensus and presented the final rating results at the criteria level for all included studies, following the MMAT user guide (Multimedia Appendix 2).

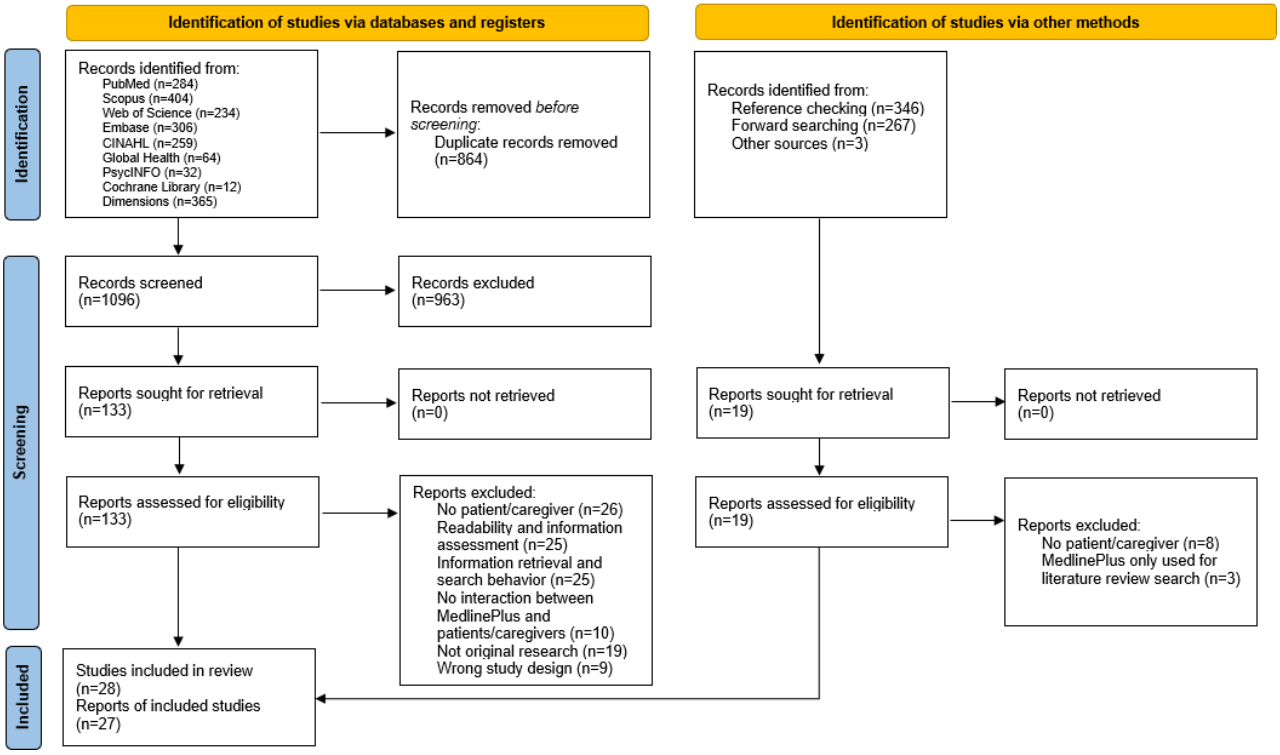
Results

Literature Search and Study Selection

In total, 28 studies reported in 27 sources met our inclusion criteria (Figure 1). Cohen κ measured the interreviewer agreement, which was 0.63 at the title-abstract level and 0.9 at the full-text level, indicating substantial agreement between reviewers. The results of all included studies are delineated by 4 categories of measures in Tables 2-5.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the systematic literature search, screening, and study selection process [10].



XSL•FO
RenderX

Table . Bibliometric characteristics of included studies.

Study	Source title	Country	Funding source	Citation count, n	FWCI[m]	AAS[n]
D'Alessandro et al, 2004 [25]	Archives of Pediatrics & Adolescent Medicine	USA	RWJ[c]	77	2.45	7
Siegel et al, 2006 [9]	Information Services & Use	USA	N/A[d]	43	2.61	N/A
Hailemeskel et al, 2007 [26]	Journal of the American Pharmacists Association	USA	NLM[e]	1	N/A	N/A
Leisey and Shipman, 2007 [27]	Journal of the Medical Library Association	USA	NLM	20	0.87	2
Smalligan et al, 2008 [28]	Journal of Investigative Medicine	USA	N/A	11	0.46	N/A
Zyskind et al, 2009 [29]	Information Services and Use	USA	NLM	7	N/A	N/A
McConnaughy and Wilson, 2010 [30]	Journal of Consumer Health on the Internet	USA	NNLM[f]	N/A	N/A	N/A
Coberly et al, 2010 [6]	Journal of the Medical Library Association	USA	NLM	25	1.84	3
Teolis et al, 2010 [31]	Journal of Consumer Health on the Internet	USA	NLM	15	0.5	3
Ulmer and Robishaw, 2010 [32]	Journal of Consumer Health on the Internet	USA	N/A	10	N/A	N/A
Lasky et al, 2011 [33]	Journal of Consumer Health on the Internet	USA	N/A	2	0.17	N/A
Coberly et al, 2012 [7]	Informatics in Primary Care	USA	NLM	6	0.6	1
Gavgani, 2012 [34]	Book Chapter in Evidence-Based Medicine	Iran	N/A	N/A	N/A	N/A
McCarthy et al, 2013 [35]	Annals of Emergency Medicine	USA	NLM and AHRQ[g]	21	1.73	63
Tarver et al, 2013 [36]	Medical Reference Services Quarterly	USA	NLM or NNLM	6	0.18	N/A
Ramesh et al, 2013 [37]	Studies in Health Technology and Informatics	USA	NCATS[h]	40	5.79	N/A
Borbolla et al, 2014 [4]	Studies in Health Technology and Informatics	Argentina	N/A	17	3.36	1
Koonce et al, 2015 [38]	Journal of the Medical Library Association	USA	IMLS[i]	20	1.44	1
Ancker et al, 2016 [5]	AMIA Annual Symposium proceedings	USA	AHRQ and NLM	15	0.95	5
Wilcox et al, 2016 [39]	Journal of the American Medical Informatics Association	USA	AHRQ, NLM, and NINR[j]	33	1.9	9

XSL•FO
RenderX

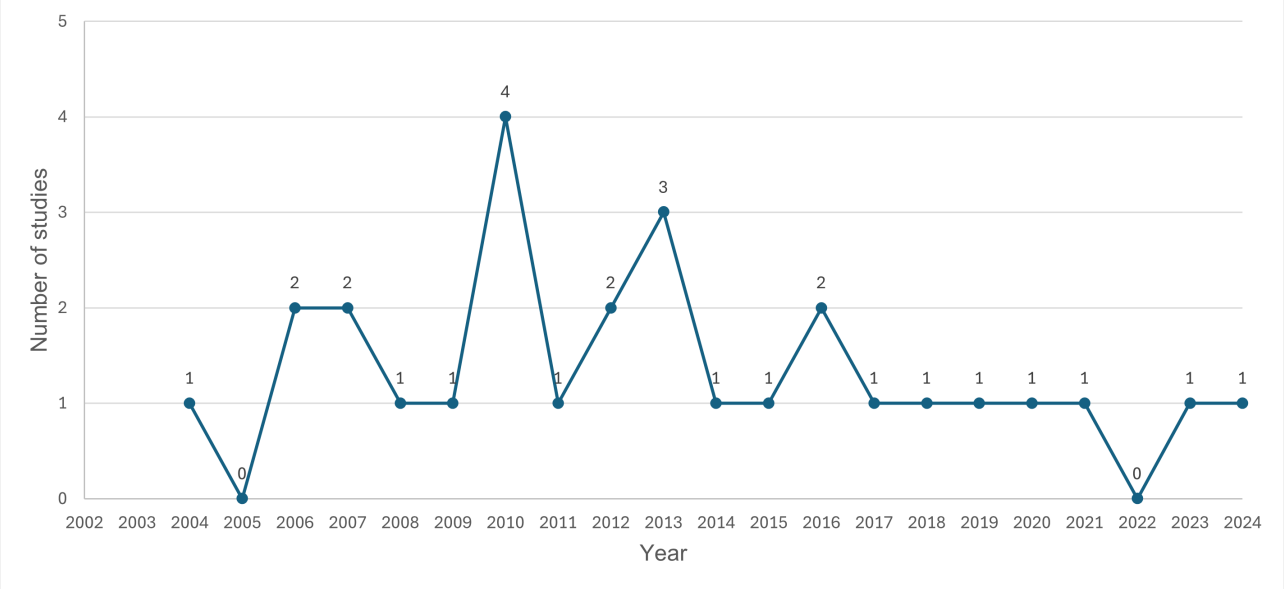
Study	Source title	Country	Funding source	Citation count, n	FWCI[m]	AAS[n]
Caufield-Noll and Gorman, 2017 [40]	Journal of Consumer Health on the Internet	USA	NLM	2	0.32	N/A
Sanders et al, 2018 [41]	Journal of the Medical Library Association	USA	NLM	2	0.11	1
Fenske, 2019 [42]	Journal of Hospital Librarianship	USA	NLM or NNLM	1	0.17	N/A
Fawcett, 2020 [43]	Dissertation	USA	N/A	N/A	N/A	N/A
Kazemi Majd et al, 2021 [44]	Health Informatics Journal	Iran	Tabriz University of Medical Sciences	5	0.33	5
Zhang et al, 2023 [45]	Journal of the American Medical Informatics Association	USA	NLM and NCATS	N/A	N/A	1
Yu et al, 2024 [46]	IEEE 12th International Conference on Healthcare Informatics	USA	DoD[k]	N/A	N/A	N/A

[a]FWCI: field-weighted citation impact.
[b]AAS: Altmetric Attention Score
[c]RWJ: Robert Wood Johnson Generalist Faculty Scholars Program.
[d]N/A: not reported in the full-text article.
[e]NLM: National Library of Medicine.
[f]NNLM: National Network of Libraries of Medicine.
[g]AHRQ: Agency for Healthcare Research and Quality.
[h]NCATS: National Center for Advancing Translational Sciences.
[i]IMLS: Institute of Museum and Library Services.
[j]NINR: US National Institute of Nursing Research.
[k]DoD: (US) Department of Defense.

Bibliometrics

The included studies were reported between 2004 and 2024 (Table 2). The year 2010 had the highest number of studies (n=4), followed by 2013 (n=3). The years 2006, 2007, 2012, and 2016 each had 2 studies, while all other years had either 1 or no studies (Figure 2).

Figure 2. Number of included studies by year (2004-2024).



XSL•FO
RenderX

Of the 28 studies, 25 were conducted in the United States, with 2 conducted in Iran [34,44] and 1 conducted in Argentina [4]. The document types varied and included 1 book chapter [34], 4 health informatics conference proceedings (ie, AMIA Annual Symposium [5], Studies in Health Technology & Informatics [4,37], and IEEE International Conference on Healthcare Informatics [46]), and 1 doctoral dissertation [43]. The remaining studies were published in 12 academic journals. The *Journal of Consumer Health on the Internet* was the most frequent publication venue, with 5 publications, followed by the *Journal of the Medical Library Association* (4 publications), the *Journal of the American Medical Informatics Association* (2 publications), and *Information Services & Use* (2 publications).

In total, 20 publications reported funding support from 9 agencies. The NLM was the primary funder, supporting 15 studies. Other funders included several US federal agencies (eg, Agency for Healthcare Research and Quality, National Center for Advancing Translational Sciences, National Institute of Nursing Research, and Department of Defense), 1 foundation program (ie, Robert Wood Johnson Generalist Faculty Scholars Program), and Tabriz University of Medical Sciences in Iran.

A total of 125 individuals from 65 organizations contributed to the included studies. Notable contributors included authors from the NLM (3 publications), George Washington University (2 publications), University of Missouri (2 publications), and Columbia University Medical Center (2 publications).

Of the 27 sources, 22 were cited in scholarly works. Eight publications had FWCI scores above 1.0, with the study by Ramesh et al [37] having the highest FWCI of 5.79 (Table 2), indicating that the publication has received more citations than expected for its field. Altmetric scores were available for 11 publications, with the highest score of 63 by McCarthy et al [35], indicating that it is widely shared, discussed, and referenced across multiple platforms. This paper is also the only one with an AAS above 20, indicating high engagement (frequent mentions across multiple sources, including news, policy reports, and mainstream discussions).

Participants

Of 28 studies, 18 involved outpatients, while 5 focused on inpatients and their family members. Notably, 2 studies designed digital information interventions incorporating MedlinePlus but recruited their participants for testing either through crowdsourcing [37] or from a local community [45]. Patients represented a range of clinical conditions, such as diabetes, cardiovascular or heart valve diseases, pediatric cancer, heart failure, and depression. More than half of the studies recruited patients with multiple or varied diagnoses (Table 3).

XSL•FO
RenderX

Table . Participant characteristics of included studies.

Study	Participants' role	Patient diagnosis	Demographics and SDOH[a]	Sample size, n	Age range (y)
D'Alessandro et al, 2004 [25]	Parents or legal guardians of outpatients and pediatricians	Pediatrics	Mostly White, medically insured, telephone access, English speaking, and some or finished graduate school	197	23-52
Siegel et al, 2006 [9]	T1: physicians and T2: outpatients[b]	Varied	N/A[c]	T1: 270; T2: 907	N/A
Hailemeskel et al, 2007 [26]	Outpatients, student pharmacists, and pharmacy residents	Varied	Predominantly African American women	92	18-35+
Leisey and Shipman, 2007 [27]	Outpatients and physicians	Varied	N/A	110	N/A
Smalligan et al, 2008 [28]	Outpatients and physicians	Varied	Limited internet access among some patients	893	N/A
Zyskind et al, 2009 [29]	Outpatients, physicians, and research assistants	Diabetes	Low socioeconomic status and education, predominantly underserved, and Spanish speaking	108	45-59[d]
McConnaughy and Wilson, 2010 [30]	Parents of children with special needs and librarians	Varied	Varying levels of internet computer literacy	324 visitors and 6 parents	N/A
Coberly et al, 2010 [6]	Outpatients, physicians, staff, and medical librarians	Varied	Majority White, highly educated, secure internet access, and insurance status	224	N/A
Teolis, 2010 [31]	Outpatients, health care professionals, and medical librarians	Varied	Majority unemployed without preventative health care, uninsured, indigent patients, and low health literacy	809	N/A
Ulmer and Robishaw, 2010 [32]	Inpatients, parents, librarians, physicians, and nurses	Varied	Minorities (Spanish-speaking patients) included and health literacy assessed	10	N/A
Lasky et al, 2011 [33]	Outpatients and pharmacy students	Varied	Majority White	198	46-65[d]
Coberly et al, 2012 [7]	Outpatients	Varied	Majority White and well-educated	592	51-52
Gavgani, 2012 [34]	Inpatients, cardiologists, information specialists, and research assistants	Cardiovascular disease	Persians with religion and rural population included	95	34-75[d]
McCarthy et al, 2013 [35]	ED[e] patients	Varied	Majority non-White, varied education, literacy, health insurance, prescription drug coverage, and computer or internet access	3386	35-54
Tarver et al, 2013 [36]	Outpatients, librarians, and hospital employees	Varied	N/A	5600	N/A
Ramesh et al, 2013 [37]	Participants from Amazon Mechanical Turk	N/A	Majority Asian and White and varied education and health literacy	59	N/A
Borbolla et al, 2014 [4]	Outpatients	Varied	Spanish speaking	N/A	N/A

XSL•FO
RenderX

Study	Participants' role	Patient diagnosis	Demographics and SDOH[a]	Sample size, n	Age range (y)
Koonce et al, 2015 [38]	Outpatients and physicians	Diabetes	Majority African American, high school education or less, household incomes ≤$20,000, employment status, and varied health literacy	160	53-54[d]
Ancker et al, 2016 [5]	Outpatients	Varied	Different genders, races, insurance, and ethnicity, including White, Black, and Latino populations	12,877	45-64
Wilcox et al, 2016 [39]	Inpatients, families, and pharmacists	Cardiothoracic surgery	Sex, age, and varied education levels	32	26-88
Caufield-Noll and Gorman, 2017 [40]	Outpatients, librarians, and physicians	Cardiology	N/A	234	N/A
Sanders et al, 2018 [41]	Outpatients, librarians, clinicians, and nurses	Varied	Minority, low education, and low income	302	Middle aged
Fenske, 2019 [42]	Inpatients, families, nurses, child life specialists, and librarians	Pediatric cancer and chronic illness	N/A	133	N/A
Fawcett, 2020 [43]	Outpatients	Diabetes	Majority college degrees or higher and insurance status	14	N/A
Kazemi Majd et al, 2021 [44]	Inpatients, cardiologists, and librarians	Heart failure	Literate patients or those cared for by a literate person	120	56-80
Zhang et al, 2023 [45]	Individuals from one community	Diabetes and depression	Majority immigrants, low income, low education levels, Spanish as preferred language, and secure internet access	134	18-80+
Yu et al, 2024 [46]	Outpatients	Localized prostate cancer	N/A	10	67-89

[a]SDOH: social determinants of health.

[b]T1: first evaluation study; and T2: second evaluation study.

[c]N/A: not reported in the full-text article.

[d] The majority of the participants in a study were within the specified age range or means.

[e]ED: emergency department.

In addition to patient participants, several studies also involved health care professionals: 12 involved clinicians with various backgrounds, 3 targeted nurses, 2 included research assistants, and 2 involved other hospital staff. Community-based participants comprised librarians from medical, health sciences, or public libraries (9 studies) and pharmacists or pharmacy students (3 studies).

SDOH factors were captured in the patients' demographic data. The studies explicitly addressed underserved and low socioeconomic status groups [29,38,41,45] and participants with low [31] or varied health literacy levels [37,38]. Several studies either focused on or included racial and ethnic populations, such as African American populations [26,38], Spanish-speaking underserved populations [29,32,45], indigent patients [31], Asian populations [37], Persian populations [34], or diverse race populations [5].

Health insurance status was detailed in 6 studies [5,6,25,31,35,43], with 3 indicating that partial participants were uninsured, while 2 reported that the vast majority were insured. While 1 study reported that 148 (75%) of 197 study participants had an annual income exceeding $50,000 [25], other studies reported low income or unemployment among their participants [29,31,38,41,45]. In addition, internet access was addressed in multiple studies. While 3 studies reported secure internet access [6,25,45], 2 studies indicated that participants had limited internet or computer access [28,31], highlighting persistent access barriers relevant to digital health implementation.

The sample sizes in these studies ranged from 10 to 12,877 participants. Most studies had sample sizes under 300 participants, while several studies reported larger samples exceeding 500 participants [5,7,28,31,35,36]. Among studies

XSL•FO
RenderX

reporting age, participants ranged from 17 to 88 years, with mean ages typically ranging between 35 and 65 years.

Intervention

Table 4 presents that 13 studies explicitly identified their main intervention feature as either “information prescription” or “information Rx.” Two of them [6,7] investigated the effectiveness of IPs delivered via paper or email. Additionally, 3 studies [31,33,41] investigated the use of kiosks to deliver multimodal MedlinePlus information. While most studies used MedlinePlus resources on desktops or laptops, 2 studies explored iPad or tablet use [41,42], and 6 studies embedded MedlinePlus information as hyperlinks in patient portals of EHRs [4,5,32,36,37,39].

XSL•FO
RenderX

Table . Intervention characteristics.

Study	Intervention feature	Information content	Additional interventions	Setting
D'Alessandro et al, 2004 [25]	Information prescription	Three preprinted websites (MedlinePlus, GeneralPediatrics.com, and the American Academy of Pediatrics website)	Hands-on training on internet health information search, handout about computer training, a list of community locations for computer access, and pediatrician training on internet patient education information	Clinical setting
Siegel et al, 2006 [9]	Information Rx	Physicians prescribed specific health information for patients to find on the MedlinePlus	Information Rx kit (pads, posters, and bookmarks) and referral to the MedlinePlus website	Clinical setting
Hailemeskel et al, 2007 [26]	Pharmacy-community partnership	One-on-one medication education and counseling	Student training classes (80 min), MedlinePlus materials, and written instructions	Pharmacy
Leisey and Shipman, 2007 [27]	Information prescription	Physician referral to MedlinePlus	A canvas gift bag containing MedlinePlus and women's health information	Clinical setting
Smalligan et al, 2008 [28]	Information Rx	Patients' encouragement to use MedlinePlus for a period	N/A[a]	Clinical setting
Zyskind et al, 2009 [29]	Information Rx	MedlinePlus X-plain diabetes tutorials on computer and given Information Rx	N/A	Clinical setting
McConnaughy and Wilson 2010 [30]	Information portal	InfoAble portal with links to health info from MedlinePlus	Other resources (ClinicalTrials.gov, PubMed, and Center for Disability Resource Library)	Library setting
Coberly et al, 2010 [6]	Email information prescription	Information prescription and condition-specific email linking to MedlinePlus topics	Physician and staff training on MedlinePlus by a medical librarian	Clinical setting
Teolis, 2010 [31]	Information Rx kiosk	48 selected MedlinePlus tutorials	MedlinePlus bookmark distribution, built-in printer, and dedicated phone line for librarian assistance	Clinical setting
Ulmer and Robishaw, 2010 [32]	Information Rx via EPIC	MedlinePlus and the library's subscription-based consumer health databases	Patient health education provided by a nurse	Clinical setting
Lasky et al, 2011 [33]	Pharmacy kiosk	MedlinePlus website	Pharmacist counseling, printed drug labeling and medication guides, and MedlinePlus bookmark and notepads	Pharmacy
Coberly et al, 2012 [7]	Paper versus email information prescription	Email or paper prescription containing 5 links to a study website with specific MedlinePlus web resources	Instructions for accessing a medical librarian for additional information	Clinical setting
Gavgani, 2012 [34]	Information prescription	Culturally adapted MedlinePlus information on coronary artery disease	Information selected from additional sources, translated into Persian, and simplified to grade 6 - 7 readability	Clinical setting

XSL•FO
RenderX

Study	Intervention feature	Information content	Additional interventions	Setting
McCarthy et al, 2013 [35]	Information prescription	MedlinePlus topics about the prescribed drug	Research assistant-guided review of MedlinePlus drug information, MedlinePlus URL magnet, prescription filling–related information, and medical librarian phone referral	Clinical setting
Tarver et al, 2013 [36]	MedlinePlus resources embedded in patient portals (MyChart)	Contextual information embedded in patient portals via MedlinePlus Connect and MedlinePlus search box	Printed materials, MyChart posters and cards with MedlinePlus resource links, library phone number, and waiting room computer	Public library setting or clinical setting
Ramesh et al, 2013 [37]	NoteAid system linking EHR[b] notes to patient education materials	EHR links to MedlinePlus content	EHR links to UMLS[c] and Wikipedia	Patient portal
Borbolla et al, 2014 [4]	Disease-specific information button on home screen and MedlinePlus links embedded in laboratory results screen	Infobutton within a PHR[d] system linking to MedlinePlus in Spanish	N/A	Patient portal
Koonce et al, 2015 [38]	Information prescription	Diabetes educational materials at 5th- and 8th-grade reading levels, visual, auditory, and kinesthetic formats, Spanish version	N/A	Clinical setting
Ancker et al, 2016 [5]	Hyperlinks containing ICD[e] codes associated with medical terms	Hyperlinks to MedlinePlus content on the patient portal account	N/A	Patient portal
Wilcox et al, 2016 [39]	Inpatient medication management and tracking	Medication names in patient portals linking to MedlinePlus	Training, demonstration, and observation sessions followed by an application use session with think-aloud protocol	Clinical setting
Caufield-Noll and Gorman 2017 [40]	Patient encounters with library staff for information consultation or instruction	Demonstration of MedlinePlus and AHA[f] website and informal conversation about a health topic	Patient handout and prepared materials from AHA and NLM[g]	Clinical setting
Sanders et al, 2018 [41]	Computer kiosks and EHR-embedded links	MedlinePlus bookmarks on kiosks and desktop computers in patient examination rooms	Computer classes and MedlinePlus training (group and one-on-one), waiting room demonstrations and recurring MedlinePlus video loops, clinician and nurse training with MedlinePlus integration into after-visit summary reviews, and patient access to iPads and laptops in examination rooms for MedlinePlus access and interactive videos	Clinical setting
Fenske, 2019 [42]	Tablets at bedside	MedlinePlus accessed from tablets	Genetics Home Reference, Genetic and Rare Disease Information Center, X-plain videos, and other NLM resources, nurse training, bookmarks, brochures, and pamphlets produced by NLM or NIH[h]	Clinical setting
Fawcett, 2020 [43]	Multimedia	MedlinePlus video	Educational videos from other resources	Clinical setting

XSL•FO
RenderX

Study	Intervention feature	Information content	Additional interventions	Setting
Kazemi Majd et al, 2021 [44]	Information prescription	Physician-prescribed MedlinePlus information and patient-asked additional information	Librarian prepared the health information across multiple sources for patients upon physician prescription	Clinical setting
Zhang et al, 2023 [45]	Dedicated bilingual web portal	Generic MedlinePlus information versus customized MedlinePlus content with localized community components	Community calendar, local service directory, social media integration for content sharing, commenting, and forming groups, Google Translate use, and webpages created by local organizations	Community setting
Yu et al, 2024 [46]	Dedicated web portal	Research team-curated information and MedlinePlus content retrieved via API[i]	Research staff online support available to facilitate prescribed tasks	Online setting via iPICS[j] platform

[a]N/A: not reported in the full-text article.

[b]EHR: electronic health record.

[c]UMLS: Unified Medical Language System.

[d]PHR: patient health record.

[e]ICD: International Classification of Diseases.

[f]AHA: American Heart Association.

[g]NLM: National Library of Medicine.

[h]NIH: National Institutes of Health.

[i]API: application programming interface.

[j]iPICS: Interactive prostate cancer information, communication, and support program

Primarily using MedlinePlus web resources, the interventions encompassed a broad range of health information topics, tutorials, and medical term explanations. Additional intervention practices included (1) hands-on computer and health information search training for patients [25,39,41,45]; (2) training for physicians, nurses, staff, and students [6,26,41,42]; (3) the use of other web-based health resources provided by health professionals or organizations [30,34,42,45]; (4) the distribution of supplementary MedlinePlus materials such as bookmarks, pamphlets, brochures, magnets, or waiting room TV loops featuring MedlinePlus videos [9,27,31,33,35,36,40-42]; (5) patient health education or counseling offered by health care professionals as part of the intervention [32,33,35]; and (6) access to librarian assistance, instruction, or training [7,31,35,40,44]. Two recent studies either customized MedlinePlus content with additional local community components (eg, recipes, infographics, and local resource listings) [45] or integrated MedlinePlus into a digital platform with the research team's self-curated health information [46].

Most studies implemented MedlinePlus interventions in a clinical setting, such as clinics or hospitals, while others used MedlinePlus in community pharmacies [26,33], a community health resource center [45], libraries [30,36], an online platform [46], or patient portals [4,5,37].

Outcome

Overall, studies reported more nonclinical outcomes than clinical outcomes (Table 5). Of 28 studies, 4 measured clinical outcomes [29,35,43,44]. Two of these studies reported significant differences between the intervention and control groups. For instance, Fawcett [43] reported significant changes in diabetic foot ulcer measurements ($F_{2,26}$=3.91; P=.03; and χ^2_2=18.5; P<.001). Kazemi Majd et al [44] found fewer patients in the intervention group died or were readmitted compared to the control group during 1 year (relative risk 0.47, 95% CI 0.20 - 1.06; P=.005), but not in the first 6-month follow-up (relative risk 0.6, 95% CI 0.15 - 2.40; P=.05). However, the other 2 studies [29,35] reported no significant differences in clinical outcomes between the intervention and control groups, although a difference in patient behavior was noted, with the intervention group being more likely to use the internet or fill their prescriptions.

XSL•FO
RenderX

Table . Clinical and nonclinical outcomes, comparisons, study duration, and follow-up method.

Study	Clinical outcome	Nonclinical outcome	Comparison	Study duration	Follow-up methods
D'Alessandro et al, 2004 [25]	N/A[a]	Participants' attitudes and behaviors about using internet health information resources	General internet search: I[b]>C[c] (*P*=.05); prescribed info search: I>C (*P*<.001); future IP[d] use intention: I>C (*P*=.02); and recommendation to others: I>C (*P*=.001)	3 mo	Phone survey
Siegel et al, 2006 [9]	N/A	T1[e]: physician attitudes and behavior and patient experience; and T2[f]: perceived impact on patient-physician relationships and perceived health outcome	Majority positive attitudes and experiences, improved referral behavior and medical communication, and better perceived health decisions and health outcomes (no *P* values reported)	T1: 24 mo and T2: 6 mo	T1: mail survey, telephone interview, or focus group; and T2: web survey
Hailemeskel et al, 2007 [26]	N/A	Familiarity with NLM[g] databases, including MedlinePlus and usage of MedlinePlus	Patients' familiarity and use of NLM databases and searching for health topics after intervention (*P*<.001); no significant difference in patients' interest in and confidence using MedlinePlus: (Fisher exact test, *P*=.40); NLM training perceived as beneficial by 91% of respondents	N/A	Phone interview
Leisey and Shipman, 2007 [27]	N/A	Physicians' unanimous agreement with IP benefits, no patient-reported recognition of the term, and only 14% patient recall of the recommendation	Discrepancy between physician-reported broad IP practice and no patient recognition of the term and between physician-reported encouragement and patient recall	6 mo	Phone interview
Smalligan et al, 2008 [28]	N/A	MedlinePlus usage; competency of accessing; trust, satisfaction, perceived ease to understand; better health decision-making; intent to reuse; and recommendation to others	N/A	N/A	Online or mail survey
Zyskind et al, 2009 [29]	Weight, BMI, blood pressure, HbA_{1c}[h], LDL[i] cholesterol, blood glucose, and total cholesterol	Patient knowledge about diabetes, verified awareness and use of online resources, including MedlinePlus for health and diabetes	No statistically significant differences between I and C groups for clinical measures; nonclinical measures (I>C): use internet to learn about health (*P*<.003) and diabetes (*P*<.003); and to have heard of MedlinePlus (*P*<.005)	12 mo	Survey

XSL•FO
RenderX

Study	Clinical outcome	Nonclinical outcome	Comparison	Study duration	Follow-up methods
McConnaughy and Wilson, 2010 [30]	N/A	Positive feedback on the InfoAble portal, usability improvement, and portal and engagement tracked via web use statistics	N/A	1 mo	N/A
Coberly et al, 2010 [6]	N/A	Health information–seeking behavior (ie, internet use to find health information, discussion of internet-sourced health information with a provider, heard of MedlinePlus, and MedlinePlus usage) and recall of receiving IP	No significant differences between the I and C groups before the clinical visit in (1) internet use for health information (P=.79), (2) discussing internet health information with a provider (P=.18), and (3) patient percentage who reported the use of MedlinePlus (15% vs 17%; P=.83); the patient percentage who had heard of MedlinePlus (I>C, 35% vs 20%, P=.02)	N/A	Mail survey
Teolis, 2010 [31]	N/A	Number of participants who used MedlinePlus for the first time, percentage of health professionals who used MedlinePlus for the first time, and perceived helpfulness and ease of understanding	N/A	8 mo	Ask for experience after using it
Ulmer and Robishaw, 2010 [32]	N/A	(1) Patients' understanding of condition or treatment, using information to ask physician questions, perceived helpfulness in overall treatment; (2) physicians' IP use and perceived impact on patients; and (3) the number of librarian-handled prescriptions	N/A	12 mo	Mailing survey for patients and online survey for physicians
Lasky et al, 2011 [33]	N/A	Health information–seeking behavior related to MedlinePlus, intention to return to the website, average encounter duration, and interest in returning to the website	Age-related differences in self-reported prior MedlinePlus usage and interest in returning to the website (P<.01)	2 mo	Kiosk prompted feedback soliciting
Coberly et al, 2012 [7]	N/A	Prescription filling rate, website engagement, health information search behavior, perception of information quality, and preference over email versus paper prescription	Higher filling rate for eHIP[j] patients than pHIP[k] patients (38% vs 23%; P<.001) and eHIP favored for perceived usefulness and ease of use	N/A	Phone and online survey
Gavgani, 2012 [34]	N/A	Total number of patients receiving IP, major topics, drivers, and barriers identification	N/A	4 mo	N/A

XSL•FO
RenderX

Study	Clinical outcome	Nonclinical outcome	Comparison	Study duration	Follow-up methods
McCarthy et al, 2013 [35]	Adverse effects and symptom improvement	Medication adherence, patient satisfaction with prescription drug information, ED[l] care, and subsequent health care use	Across all 3 sites, no significant differences between usual care and any interventions; at 1 site, subjects receiving practical prescription information or services or a combination of prescription information or service were more likely to fill their prescription than usual care (OR[m] 1.8, 95% CI 1.0-3.1); no clinically meaningful differences between groups in adverse-effect incidence or subsequent physician visits	>10 mo	Phone interview
Tarver et al, 2013 [36]	N/A	Patient use, interest, and feedback on contextual information links in MyChart	N/A	12 mo	N/A
Ramesh et al, 2013 [37]	N/A	Self-reported EHR[n] notes comprehension and readability levels	No statistical significance in improving users' comprehension of EHR notes with MedlinePlus-implemented NoteAid (P>.05) and self-reported EHR note comprehension scores ranged between 3 and 4 on a 5-point scale	N/A	N/A
Borbolla et al, 2014 [4]	N/A	Access frequency to MedlinePlus Connect via patient portal and percentage of access sessions on laboratory results	N/A	14 mo	N/A
Koonce et al, 2015 [38]	N/A	Diabetes knowledge scores at 2- and 6-week follow-ups	Diabetes knowledge scores at 2- and 6-week follow-ups: I>C (P<.001)	6 mo	Phone interview
Ancker et al, 2016 [5]	N/A	Frequency of patient portal access and usage frequency of MPC[o]	Black patients were more likely to use MPC than White patients (43% vs 40%; P<.001); Latino patients were more likely to use the resource than non-Latinos (43% vs 41%; P<.001); patients covered by Medicaid were more likely to use MPC than privately insured (44% vs 43%; P<.001)	36 mo	N/A
Wilcox et al, 2016 [39]	N/A	Usage frequency, usability, usefulness, acceptability, and audit logs	Perceived usefulness by both pharmacists and patients, preferring different features	19 wk	Interview

XSL•FO
RenderX

Study	Clinical outcome	Nonclinical outcome	Comparison	Study duration	Follow-up methods
Caufield-Noll and Gorman, 2017 [40]	N/A	Patient satisfaction	N/A	21 mo	Survey
Sanders et al, 2018 [41]	N/A	Patients' frequency of access to MedlinePlus, knowledge, attitude, and skills related to accessing online health resources and clinicians' use and recommendation of MedlinePlus	Patient use of MedlinePlus increased significantly from 2% before intervention to 6% after intervention (χ^2_1=5.2; *P*=.02); MedlinePlus recommendation from clinicians and nurses increased from 21% to 43% (χ^2_1=6.0; *P*<.001)	34 mo	Survey
Fenske, 2019 [42]	N/A	Number of patients and families served, understanding of diagnosis, procedures, and care; generating questions for providers; and confidence in searching MedlinePlus	Improved understanding of diagnosis, procedure, or care after video viewing	10 mo	Survey
Fawcett, 2020 [43]	DFU[p] area measurement	NAFF scores	No significant change in NAFF[q] scores after intervention (t_{13}=−0.28; *P*=.79) and significant change in DFU area ($F_{2,26}$=3.91; *P*=.03)	4 wk	NAFF self-evaluation and wound assessment
Kazemi Majd et al, 2021 [44]	Reduction in the hospital readmission and death rate	N/A	One-year mortality: fewer deaths in the intervention group than the control group (7 vs 15; RR[r] 0.47, 95% CI 0.20 - 1.06; *P*=.005); 6-month death or readmission: no statistically significant difference between groups (RR 0.6, 95% CI 0.15 - 2.40; *P*=.05)	12 mo	Phone interview
Zhang et al, 2023 [45]	N/A	PAM[s], disease knowledge, self-efficacy, behavior changes, ease of use and usage of the digital platform, and frequency and methods of using the resource	No statistically significant difference between groups in PAM, disease knowledge, or self-efficacy at 1 month; a significant improvement in PAM scores within intervention group (*P*=.048), but not in the control group (*P*=.06); both groups significantly improved self-efficacy scores on communication with providers, finding professional help, information, and knowledge (*P*<.03)	13 mo	Face-to-face survey

XSL•FO
RenderX

Study	Clinical outcome	Nonclinical outcome	Comparison	Study duration	Follow-up methods
Yu et al, 2024 [46]	N/A	Precision and recall of MedlinePlus search; task completion, time, and number of browsed content types	N/A	1 mo	N/A

[a]N/A: no relevant information reported in the full-text article.

[b]I: intervention.

[c]C: control.

[d]IP: information prescription.

[e]T1: first evaluation study.

[f]T2: second evaluation study.

[g]NLM: National Library of Medicine.

[h]HbA1c: hemoglobin A_{1c}.

[i]LDL: low-density lipoprotein.

[j]eHIP: email health information prescription.

[k]pHIP: paper health information prescription.

[l]ED: emergency department.

[m]OR: odds ratio.

[n]EHR: electronic health record.

[o]MPC: MedlinePlus Connect.

[p]DFU: diabetic foot ulcer area measurement.

[q]NAFF: Nottingham Assessment of Functional Footcare scores.

[r]RR: relative risk.

[s]PAM: Patient Activation Measure.

Most studies (24/28) reported positive nonclinical outcomes, focusing on (1) participants' attitudes, satisfaction, and behaviors related to online health information resource usage and information-seeking behaviors; (2) frequency and user experience of MedlinePlus usage; and (3) perceived usefulness, trust, ease of use, and competency in using MedlinePlus information for health decision-making, physician communication, self-management, and self-efficacy on finding professional help and disease knowledge.

The study durations varied, typically around 12 months, ranging from 1 to 34 months. Most studies used surveys or interviews for participant follow-ups.

Quality Assessment

All studies were empirical with clear research goals, objectives, or research questions, and they collected data to address their research questions, achieving MMAT S1 and S2. Among them, 2 qualitative studies were analyzed by MMAT criteria from 1.1 to 1.5, and only one criterion (ie, MMAT 1.1) was met in each study [34,36]. For 7 RCTs evaluated by MMAT 2.1 to 2.5, the number of criteria met ranged from 1 to 5, with many studies lacking sufficient details on either the outcome assessors blinded to the intervention (MMAT 2.4) or follow-up response rates of adherence to the intervention (MMAT 2.5). Each of the 4 nonrandomized studies met at least two criteria (eg, MMAT 3.2 and 3.5), and only one study [5] met MMAT 3.4 by addressing potential confounding variables adequately. Seven quantitative descriptive studies met 2 or 3 criteria (MMAT 4.3 and 4.5), and the majority rarely provided sufficient information about their sampling strategies or reduced nonresponse bias risk (MMAT 4.1, 4.2, and 4.4). Finally, among 8 mixed methods studies, 1 did not meet any criteria, while 6 met only 1 criterion. The study by Wilcox et al [39] met all 5 criteria. Most mixed methods studies lacked sufficient details on whether and how the qualitative and quantitative findings were integrated and interpreted (MMAT 5.3). Overall, when considering the average number of MMAT criteria met per study type, qualitative studies had the fewest met criteria (one per study), while nonrandomized studies had the most (ie, an average of 3.25 criteria per study). Detailed quality assessment results were documented in Multimedia Appendix 2.

Discussion

Bibliometrics

As the primary funding agency for half of the included studies, the US NLM has sustained MedlinePlus as a reliable and trusted health information resource. While most included studies took place in the United States, 3 studies were conducted in developing countries, demonstrating MedlinePlus's applicability and value across cultural, medical, and organizational contexts [4,34,44].

Despite nearly 25 years of availability, MedlinePlus has generated a relatively small body of empirical research examining its role in clinical care and chronic disease management. The limited number of identified studies (n=28), coupled with generally below-average citation impact (ie, FWCI scores <1.0) and below-benchmark Altmetric scores (<20), suggests that MedlinePlus has been underused and underexamined in health care practice and research. These gaps highlighted missed opportunities to rigorously document its

XSL•FO
RenderX

implementation effectiveness and potential contribution to patient- and caregiver-centered care.

Participants

Guided by the CCCM framework, this review highlights the engagement of diverse participant groups with MedlinePlus across coordinated and team-based care settings, including health care professionals, inpatient and outpatient contexts, patients with varied health conditions, family caregivers, pharmacists, and librarians. In particular, medical or public libraries often played a critical role in "physician-directed" information delivery [6], improving community health literacy [47], and serving as community partners to make MedlinePlus accessible to broader audiences. These findings underscore its potential role in integrated care delivery.

Although most studies reported patients' ages, races, and socioeconomic backgrounds, many lacked SDOH-related specifics such as health insurance, annual income, employment, or health literacy. This gap limits our understanding of how sociodemographics impacted MedlinePlus use, especially given the strong association between health literacy, internet access, and digital engagement [48].

Intervention

As an information intervention in clinical care, early studies on MedlinePlus [25,26,28] were directly influenced by the paper-based IP or Information Rx program launched by the NLM in 2002. However, later work reflected a growing interest in integrating MedlinePlus into EHR-enabled environments. For example, Ramesh et al [37] highlighted earlier efforts to embed MedlinePlus within Health IT systems. This trend continues, with more recent studies incorporating MedlinePlus into patient portals [4,5,39,41] and stand-alone digital health platforms for patients and caregivers [45,46], supporting their self-management and health information access beyond traditional clinical settings.

Several studies explored the use of kiosks in community pharmacies to deliver multimodal MedlinePlus information [31,33,41], aligning with the community linkage in the CCCM framework. Other studies also reported disseminating MedlinePlus information via iPad or tablet, or combining MedlinePlus content with additional resources, such as the American Heart Association patient materials, RxList, Mayo Clinic, and even religious texts (ie, the Holy Quran). These interventions were implemented across settings, from clinics to virtual environments, highlighting MedlinePlus' adaptability and versatility in supporting varied models of health information delivery.

Outcome

Clinical outcomes were reported in only 4 studies. Their results varied from no to significant differences in assessed measures, sometimes with mixed results even within the same study. These inconclusive results indicate the complex nature of directly linking MedlinePlus information interventions to patient health improvements. Researchers identified barriers, including inconsistent physician participation [27], small sample sizes [6], low computer literacy [29] or digital divide [5], limited evaluation periods [29], and information intervention design challenges [35]. Nevertheless, most studies reported nonclinical outcomes, including notable improvements in participants' attitudes, satisfaction, and behaviors related to online health information seeking. These nonclinical outcomes highlight the benefits of using MedlinePlus in enhancing patients' health engagement, communication, education, and self-efficacy, providing strong evidence to inform physician-directed health information interventions that complement routine clinical care.

Overall, in light of the identified research gaps, we recommend the following priorities for future research. First, additional studies are needed to rigorously examine the impact of MedlinePlus on patient- and caregiver-centered health outcomes and ensure transparent reporting and effective dissemination of findings within and beyond scholarly communities. Second, more rigorously designed RCTs are needed to evaluate its clinical efficacy. The existing RCTs missed or underreported several key methodological components. Thus, future studies should provide detailed information on allocation concealment, randomization implementation, and follow-up response rates regarding participants' adherence to the intervention. Third, digital solutions should seamlessly integrate MedlinePlus content into patients' health care journeys, such as through personalized apps or culturally adapted digital health platforms. Although the most recent 2 studies [45,46] suggest these burgeoning trends, future studies shall ensure that patients receive relevant information tailored to their specific needs, preferences, and literacy levels. Finally, future research shall consistently collect and report SDOH-related demographics, such as income, insurance, education, and health literacy, and support MedlinePlus as an equitable intervention responsive to the care needs of patients and caregivers of different backgrounds.

Despite its global importance, MedlinePlus faces challenges from rapid technological advances (eg, keeping pace with evolving web design and digital infrastructure), competition from commercial products (eg, Healthwise within EPIC EHRs), and uncertainties about the long-term sustainability of federal funding amid federal priority shifts. However, the synthesized evidence from this review demonstrated consistent nonclinical benefits and clinical potentials, validating MedlinePlus' enduring value and utility. The findings provide evidence to inform future Health IT innovations, investments, and policy decision-making.

Limitations

This study has several limitations. First, although we aimed to identify all relevant studies, we may have missed those published in languages other than English or not indexed in the 9 bibliographic databases we searched. Second, we excluded articles that reported the use of MedlinePlus for nonpatient or noncaregiver populations. While this exclusion helped us focus on MedlinePlus's role in clinical care, it might have limited the interpretations of MedlinePlus's real impact beyond clinical contexts. Third, the included studies are heterogeneous in their research designs, and many lack demographic information on socioeconomic status, constraining a comprehensive understanding of how SDOH-related demographics influence MedlinePlus utilization. Future studies shall address this gap

XSL•FO
RenderX

by ensuring more complete and consistent demographic data collection and reporting.

Conclusions

This systematic literature review is the first to examine how MedlinePlus has been integrated into clinical care, supporting patients and caregivers through increased health information access. Researchers agree that promoting MedlinePlus resources will help address the digital divide, particularly for individuals who are homebound, reside in rural areas, or are underserved. Although health care providers are the key stakeholders in advocating MedlinePlus, the effective delivery and optimization of MedlinePlus health information for patient care requires collective and coordinated community efforts. Our synthesis through the CCCM-guided lens will serve as a foundation for future research and practice in designing and developing accessible, equitable, and scalable digital health information solutions to improve patient engagement, health literacy, self-management, and health outcomes.

Acknowledgments

We would like to thank all team members, particularly our graduate students, Erin Harvey and Pengxuan Wang, for their assistance with study selection. This manuscript was prepared in Microsoft Word with a Grammarly plug-in. Text edits suggested by Grammarly were reviewed and cross-checked using ChatGPT 4o for grammar, word choices, and sentence structures. All editing recommendations by generative artificial intelligence tools were carefully evaluated before being accepted or dismissed.

Funding

The US Department of Defense supported this project through funding (W81XWH-21-1-0263, principal investigator: LS). However, the authors take full responsibility for the information in this study, which does not necessarily reflect the views of the US Department of Defense.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 16 KB - medinform_v14i1e79416_app1.docx]

Multimedia Appendix 2
Quality assessment results by Mixed Methods Appraisal Tool.
[DOCX File, 35 KB - medinform_v14i1e79416_app2.docx]

Checklist 1
PRISMA checklist.
[PDF File, 178 KB - medinform_v14i1e79416_app3.pdf]

References

1. About MedlinePlus. National Library of Medicine. URL: https://medlineplus.gov/about/ [accessed 2023-01-24]
2. Health check tools: MedlinePlus. National Library of Medicine. URL: https://medlineplus.gov/healthchecktools.html [accessed 2023-02-19]
3. Jentsch J. MedlinePlus connect: 10 years of linking electronic health records to consumer health information – NLM Musings from the Mezzanine. National Library of Medicine. 2021. URL: https://nlmdirector.nlm.nih.gov/2021/01/13/medlineplus-connect-10-years-of-linking-electronic-health-records-to-consumer-health-information/ [accessed 2023-02-19]
4. Borbolla D, Del Fiol G, Taliercio V, et al. Integrating personalized health information from MedlinePlus in a patient portal. Stud Health Technol Inform 2014;205:348-352. [Medline: 25160204]
5. Ancker JS, Mauer E, Hauser D, Calman N. Expanding access to high-quality plain-language patient education information through context-specific hyperlinks. AMIA Annu Symp Proc 2016;2016:277-284. [Medline: 28269821]
6. Coberly E, Boren SA, Davis JW, et al. Linking clinic patients to Internet-based, condition-specific information prescriptions. J Med Libr Assoc 2010 Apr;98(2):160-164. [doi: 10.3163/1536-5050.98.2.009] [Medline: 20428282]
7. Coberly E, Boren SA, Mittal M, et al. MedlinePlus-based health information prescriptions: a comparison of email vs paper delivery. Inform Prim Care 2012;20(3):197-205. [doi: 10.14236/jhi.v20i3.25] [Medline: 23710844]
8. Okhovati M, Bagherinezhad Z, Gavgani VZ, Safizadeh H. Physician-directed information prescription service. A scoping review. Patient Educ Couns 2022 Dec;105(12):3369-3380. [doi: 10.1016/j.pec.2022.08.002] [Medline: 35985907]

XSL•FO
RenderX

9. Siegel ER, Logan RA, Harnsberger RL, et al. Information Rx: evaluation of a new informatics tool for physicians, patients, and libraries. Inf Serv Use 2006;26(1):1-10. [doi: 10.3233/isu-2006-26101] [Medline: 17060946]
10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
11. Covidence. URL: https://www.covidence.org/ [accessed 2025-12-17]
12. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc 2006;2006:359-363. [Medline: 17238363]
13. Bauer MS, Stolzmann K, Miller CJ, Kim B, Connolly SL, Lew R. Implementing the collaborative chronic care model in mental health clinics: achieving and sustaining clinical effects. Psychiatr Serv 2021 May 1;72(5):586-589. [doi: 10.1176/appi.ps.202000117] [Medline: 33730885]
14. What is Field-Weighted Citation Impact (FWCI)? Scopus. URL: https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/~/what-is-field-weighted-citation-impact-%28fwci%29%3F/ [accessed 2021-06-02]
15. Yu F, Van AA, Patel T, et al. Bibliometrics approach to evaluating the research impact of CTSAs: a pilot study. J Clin Trans Sci 2020 Aug;4(4):336-344. [doi: 10.1017/cts.2020.29]
16. Qua K, Yu F, Patel T, Dave G, Cornelius K, Pelfrey CM. Scholarly productivity evaluation of KL2 scholars using bibliometrics and federal follow-on funding: cross-institution study. J Med Internet Res 2021 Sep 29;23(9):e29239. [doi: 10.2196/29239] [Medline: 34586077]
17. Yu F, Patel T, Carnegie A, Dave G. Evaluating the impact of a CTSA program from 2008 to 2021 through bibliometrics, social network analysis, and altmetrics. J Clin Transl Sci 2023;7(1):e44. [doi: 10.1017/cts.2022.530] [Medline: 36845314]
18. Introduction to the Altmetric Explorer. Altmetric. URL: https://help.altmetric.com/support/solutions/articles/6000240622-introduction-to-the-altmetric-explorer [accessed 2024-01-07]
19. Putting the Altmetric Attention Score in context. Altmetric. URL: https://help.altmetric.com/support/solutions/articles/6000233313-putting-the-altmetric-attention-score-in-context [accessed 2024-08-20]
20. Patel RB, Vaduganathan M, Bhatt DL, Bonow RO. Characterizing high-performing articles by Altmetric Score in Major Cardiovascular Journals. JAMA Cardiol 2018 Dec 1;3(12):1249-1251. [doi: 10.1001/jamacardio.2018.3823] [Medline: 30476952]
21. Clinical outcome – toolkit. Toolkit for Patient-Focused Therapy Development. URL: https://toolkit.ncats.nih.gov/glossary/clinical-outcome/ [accessed 2024-03-06]
22. Hong QN, Pluye P, Fàbregues S, et al. Mixed Methods Appraisal Tool (MMAT), Version 2018: User Guide. URL: https://mixedmethodsappraisaltoolpublic.pbworks.com [accessed 2024-03-06]
23. Sun X, Zhang J, Wang Y, et al. The impact of humor therapy on people suffering from depression or anxiety: an integrative literature review. Brain Behav 2023 Sep;13(9):e3108. [doi: 10.1002/brb3.3108] [Medline: 37340873]
24. McPhail L, Thornicroft G, Gronholm PC. Help-seeking processes related to targeted school-based mental health services: systematic review. BMC Public Health 2024 May 2;24(1):1217. [doi: 10.1186/s12889-024-18714-4] [Medline: 38698391]
25. D'Alessandro DM, Kreiter CD, Kinzer SL, Peterson MW. A randomized controlled trial of an information prescription for pediatric patient education on the internet. Arch Pediatr Adolesc Med 2004 Sep;158(9):857-862. [doi: 10.1001/archpedi.158.9.857] [Medline: 15351750]
26. Hailemeskel B, Dutta AP, Daftary M, et al. Campus-community pharmacy partnership to foster awareness and use of MedlinePlus. J Am Pharm Assoc (2003) 2007;47(3):385-389. [doi: 10.1331/JAPhA.2007.06067] [Medline: 17510035]
27. Leisey MR, Shipman JP. Information prescriptions: a barrier to fulfillment. J Med Libr Assoc 2007 Oct;95(4):435-438. [doi: 10.3163/1536-5050.95.4.435] [Medline: 17971891]
28. Smalligan RD, Campbell EO, Ismail HM. Patient experiences with MedlinePlus.gov: a survey of internal medicine patients. J Investig Med 2008 Dec;56(8):1019-1022. [doi: 10.2310/JIM.0b013e31818d9138] [Medline: 19105248]
29. Zyskind A, Jones KC, Pomerantz KL, LaFaye Barker A. Exploring the use of computer based patient education resources to enable diabetic patients from underserved populations to self-manage their disease. Inf Serv Use 2009 Sep 14;29(1):29-43. [doi: 10.3233/ISU-2009-0592]
30. McConnaughy RP, Wilson SP. InfoAble portal: developing a disability information portal for a support network for families. J Consum Health Internet 2010 Feb 26;14(1):1-11. [doi: 10.1080/02763860903542934]
31. Teolis MG. A MedlinePlus kiosk promoting health literacy. J Consum Health Internet 2010;14(2):126-137. [doi: 10.1080/15398281003780966] [Medline: 20808715]
32. Ulmer PA, Robishaw S. Information prescriptions: providing health information at the inpatient's point of medical need. J Consum Health Internet 2010 May 19;14(2):138-149. [doi: 10.1080/15398281003781006]
33. Lasky T, Kogut S, Campbell S, Risica PM. Computer kiosks to deliver medication information in the pharmacy. J Consum Health Internet 2011 Oct;15(4):347-360. [doi: 10.1080/15398285.2011.623579]
34. Gavgani VZ. Evidence based information prescription (IPs) in developing countries. In: Sitaras N, editor. Evidence-Based Medicine – Closer to Patients or Scientists?: InTechOpen; 2012. [doi: 10.5772/38238]
35. McCarthy ML, Ding R, Roderer NK, et al. Does providing prescription information or services improve medication adherence among patients discharged from the emergency department? A randomized controlled trial. Ann Emerg Med 2013 Sep;62(3):212-223. [doi: 10.1016/j.annemergmed.2013.02.002] [Medline: 23561463]

XSL•FO
RenderX

36. Tarver T, Jones DA, Adams M, Garcia A. The librarian's role in linking patients to their personal health data and contextual information. Med Ref Serv Q 2013;32(4):459-467. [doi: 10.1080/02763869.2013.837730] [Medline: 24180654]
37. Polepalli Ramesh B, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform 2013;192:714-718. [Medline: 23920650]
38. Koonce TY, Giuse NB, Kusnoor SV, Hurley S, Ye F. A personalized approach to deliver health care information to diabetic patients in community care clinics. J Med Libr Assoc 2015 Jul;103(3):123-130. [doi: 10.3163/1536-5050.103.3.004] [Medline: 26213503]
39. Wilcox L, Woollen J, Prey J, et al. Interactive tools for inpatient medication tracking: a multi-phase study with cardiothoracic surgery patients. J Am Med Inform Assoc 2016 Jan;23(1):144-158. [doi: 10.1093/jamia/ocv160] [Medline: 26744489]
40. Caufield-Noll C, Gorman L. Heart-healthy conversations: a librarian-physician partnership to promote disease management and wellness information to cardiology patients. J Consum Health Internet 2017 Apr 3;21(2):148-158. [doi: 10.1080/15398285.2017.1309169]
41. Sanders M, Bringley K, Thomas M, Boyd M, Farah S, Fiscella K. Promoting MedlinePlus utilization in a federally qualified health center using a multimodal approach. J Med Libr Assoc 2018 Jul;106(3):361-369. [doi: 10.5195/jmla.2018.216] [Medline: 29962914]
42. Fenske RF. Tablet computer use at the bedside in a new patient/family education program. J Hosp Librariansh 2019;19(2):110-128. [doi: 10.1080/15323269.2019.1586286] [Medline: 31708714]
43. Fawcett KA. The effectiveness of multimedia education to improve diabetic foot care knowledge, practice, and wound healing [Dissertation]. : Grand Canyon University; 2020.
44. Kazemi Majd F, Gavgani VZ, Golmohammadi A, Jafari-Khounigh A. Effect of physician prescribed information on hospital readmission and death after discharge among patients with health failure: a randomized controlled trial. Health Informatics J 2021;27(1):1460458221996409. [doi: 10.1177/1460458221996409] [Medline: 33657912]
45. Zhang TM, Millery M, Aguirre AN, Kukafka R. A randomized controlled trial of generic and localized MedlinePlus-based information resources for hard-to-reach urban Hispanic community. J Am Med Inform Assoc 2023 Jan 18;30(2):282-291. [doi: 10.1093/jamia/ocac213] [Medline: 36350567]
46. Yu F, Neidre D, Rakib NA, et al. Black-box testing of the interactive prostate cancer information, communication, and support program to ensure reliability for patients and caregivers. Presented at: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI); Jun 3-6, 2024. [doi: 10.1109/ICHI61247.2024.00065]
47. Ottosen T, Mani NS, Fratta MN. Health information literacy awareness and capacity building: present and future. IFLA Journal 2019 Oct;45(3):207-215. [doi: 10.1177/0340035219857441]
48. Estacio EV, Whittle R, Protheroe J. The digital divide: examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019 Oct;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]

Abbreviations

AAS: Altmetric Attention Score
CCCM: Collaborative Chronic Care Model
EHR: electronic health record
FWCI: field-weighted citation impact
IP: information prescription
IRx: Information Rx
MMAT: Mixed Methods Appraisal Tool
NLM: National Library of Medicine
PICO: problem/population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SDOH: social determinants of health

Edited by A Benis; submitted 20.Jun.2025; peer-reviewed by S Ashraf, Y Chi; revised version received 28.Dec.2025; accepted 17.Feb.2026; published 27.Apr.2026.

Please cite as:
Yu F, Wang X, Liu J, Wang T, Grady SD, Song L
Leveraging MedlinePlus to Improve Health Information Access Among Patients and Caregivers: Systematic Literature Review
JMIR Med Inform 2026;14:e79416
URL: https://medinform.jmir.org/2026/1/e79416
doi:10.2196/79416

XSL•FO
RenderX

© Fei Yu, Xiaomeng Wang, Jia Liu, Tian Wang, Siobahn Day Grady, Lixin Song. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Influencing Factors of Mobile Health Apps in Kidney Transplant Care: Systematic Review Using the Consolidated Framework for Implementation Research

Yingtian Jia[1*], BSN; Shaobo Guo[1*], MSN; Xinran Yang[1], BSN; Xiaohong Lin[2], MSN; Jiaxin Fang[2], MSN; Lei Dong[2], MSN; Xiangru Li[1], MSN; Haiya Sun[3], PhD; Wanhui Yu[3], MSN; Hongxia Liu[1], PhD

[1]School of Nursing, Beijing University of Chinese Medicine, No. 11 Beisanhuan East Road, Chaoyang District, Beijing, China

[2]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China

[3]School of Nursing, Jining Medical College, Jining, China

[*]these authors contributed equally

Corresponding Author:
Hongxia Liu, PhD
School of Nursing, Beijing University of Chinese Medicine, No. 11 Beisanhuan East Road, Chaoyang District, Beijing, China

Abstract

Background: Kidney transplant recipients require lifelong self-management and follow-up care to maintain allograft function. Mobile health (mHealth) effectively improves self-management behaviors and clinical indicators, consequently enhancing nursing care quality. However, these apps commonly face challenges, including low adoption rates and high discontinuation. Although researchers have explored associated facilitators and barriers from various perspectives, a systematic review of these influencing factors is lacking.

Objective: The objective of this study was to systematically review the influencing factors of mHealth apps in kidney transplant care and to provide evidence for developing targeted interventions.

Methods: The systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, and the protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews, CRD420251091361). PubMed, Web of Science, Embase, MEDLINE, and Chinese databases, including China National Knowledge Infrastructure, Wanfang Data, China Science and Technology Journal Database, and SinoMed, were searched from inception to March 2025. The Mixed Methods Appraisal Tool was used for quality assessment given its suitability for appraising diverse study designs. Influencing factors were identified and coded according to the Consolidated Framework for Implementation Research due to its utility in systematically identifying multilevel implementation factors.

Results: A total of 19 studies (all English-language publications) were included, comprising 9 qualitative studies, 5 mixed methods studies, and 5 quantitative studies, involving 1265 kidney transplant recipients and 34 health care providers. A total of 16 facilitators and 14 barriers were identified and categorized into 5 domains: intervention characteristics, outer setting, inner setting, characteristics of individuals, and process.

Conclusions: The use of mHealth apps in kidney transplant care is influenced by multidimensional factors, with intervention characteristics constituting the most prominent domain, while the outer setting and process domains are relatively underrepresented. Future research should investigate these influencing factors and implement multidimensional strategies to optimize mHealth apps in kidney transplant care.

(JMIR Med Inform 2026;14:e84139) doi:10.2196/84139

KEYWORDS

mobile health; mHealth; kidney transplant; nursing care; influencing factors; systematic review; Consolidated Framework for Implementation Research; CFIR

Introduction

Kidney transplantation is currently the most effective treatment for end-stage renal disease, effectively alleviating clinical symptoms and prolonging patient survival [1,2]. Research indicates that the 1-year survival rate for kidney transplant recipients has reached 97.8%, with the 5-year survival rate at 88.1% and the 10-year survival rate at approximately 60% [2]. However, due to the risks of graft rejection and long-term immunosuppressant use, kidney transplant recipients require lifelong self-management and follow-up care to maintain allograft function [3].

XSL•FO
RenderX

The World Health Organization defines mobile health (mHealth) as "medical and public health practice supported by mobile devices, including mobile phones, patient monitoring devices, personal digital assistants, and other wireless technologies" [4]. Currently, mHealth has been applied across multiple domains of kidney transplant care, including medication adherence reminders [5], weight management [6], and postoperative self-monitoring [7]. Studies demonstrate that mHealth interventions can effectively enhance self-management behaviors, improve treatment compliance, optimize clinical outcomes, and elevate nursing care quality [8-10]. Additionally, mHealth apps can reduce transportation costs and time burdens for recipients, mitigate disease exposure risks, and decrease nosocomial infection rates [11,12]. Nevertheless, challenges persist, including low software adoption rates, high discontinuation rates, and short user retention periods [13]. Existing research has explored facilitators and barriers from the perspectives of device functionality, user experience, and management protocols [11,14,15], yet it lacks a comprehensive synthesis of these influencing factors.

The Consolidated Framework for Implementation Research (CFIR) was developed by Damschroder et al [16]. As a widely recognized and extensively cited framework in implementation science, it is applicable to all implementation stages, including preimplementation, active implementation, and sustainability [17]. This theoretical framework encompasses 5 domains (intervention characteristics, outer setting, inner setting, characteristics of individuals, and process) and 39 underlying constructs. It enables systematic identification of multidimensional and multilevel influences during intervention implementation processes. Guided by the CFIR framework, this study aims to systematically evaluate the influencing factors of mHealth apps in kidney transplant care, identify key facilitators and barriers, and provide evidence-based recommendations for developing targeted intervention strategies.

Methods

Search Strategy

A comprehensive computerized search was conducted across 8 English and Chinese databases: PubMed, Web of Science, Embase, MEDLINE, China National Knowledge Infrastructure, Wanfang Data, China Science and Technology Journal Database, and SinoMed, covering literature from inception to March 2025. The search focused on identifying influencing factors of mHealth apps in kidney transplant care. Search terms were structured using a combination of Medical Subject Headings and keywords. English search terms: “kidney transplant* OR renal transplant* OR kidney graft* OR renal graft* OR kidney allograft* OR renal allograft* OR transplanted kidney OR kidney homotransplantation* OR renal homotransplantation*,” “tele?medicine OR telemedicine OR mobile health OR m?Health OR mHealth OR tele?health OR telehealth OR e?Health OR eHealth OR tele?rehabilitation OR telerehabilitation OR tele?monitoring OR telemonitoring OR video?conferencing OR videoconferencing OR online OR technolog* OR mobile information technolog* OR electronic OR digital OR Internet-based OR app* OR mobile app* OR software OR wearable electronic device* OR WeChat OR platform* OR smartphone*,” “motivator* OR motivation* OR enabler* OR promote OR drive OR encourage OR facilitator* OR barrier* OR obstacle* OR challenge* OR difficult* OR issue* OR influencing factor* OR impact factor* OR relative factor* OR experience* OR perception* OR opinion* OR feeling* OR need* OR attitude*.” The search strategy was adjusted according to different databases. Key adaptations included the use of database-specific subject headings (eg, Medical Subject Headings in PubMed, Emtree in Embase) and adjustments in field tags (eg, [Title/Abstract] in PubMed; ab, kw, ti in Embase). Taking the PubMed database as an example, the search query is provided in Multimedia Appendix 1. To ensure comprehensive reporting, the PRISMA checklist is provided in Checklist 1.

Inclusion and Exclusion Criteria

Inclusion criteria:

1. Study subjects: kidney transplant recipients and stakeholders involved in mHealth apps (eg, health care providers, caregivers);
2. Study content: facilitators, barriers, or influencing factors related to the application of mHealth, with mHealth functions relevant to transplant care;
3. Study design: qualitative, quantitative, or mixed methods studies.

Exclusion criteria:

1. Conference abstracts, unpublished data, or inaccessible full texts;
2. Secondary literature (reviews, systematic reviews) or duplicate publications;
3. Non-Chinese or non-English publications.

Literature Screening and Data Extraction

Two researchers trained in evidence-based nursing systematically and independently conducted the screening process using predefined protocols. Duplicates were removed using NoteExpress 3.8 software. Initial exclusion was performed through title and abstract screening, followed by full-text assessment. Discrepancies were resolved through consultation with a third researcher.

Data extraction was independently performed by 2 researchers. To ensure accuracy and completeness, all extracted data were cross-checked against the original full-text articles by a third researcher. Any disagreements or uncertainties were resolved through team discussion. The data extraction form was independently developed by the researchers, subsequently refined through research team discussions, and ultimately finalized for implementation. Extracted data included authors, publication year, country, study design, sample size, study population, sampling method, data collection methods or tools, names and functions of mHealth apps, and Mixed Methods Appraisal Tool (MMAT) rating.

Quality Assessment

Given the heterogeneity of study designs, 2 researchers independently evaluated the methodological quality of included studies using the MMAT, 2018 edition [18]. Discrepancies were

XSL•FO
RenderX

resolved through consensus discussions with a third researcher. The MMAT was developed by a research team at McGill University, Canada, in 2006 and updated in 2018 [18]. It is designed to appraise the quality of qualitative, quantitative, and mixed methods research studies. The tool comprises 2 screening questions and 15 appraisal criteria. After addressing the 2 screening questions, the appropriate study category is selected for appraisal. Qualitative and quantitative studies require the evaluation of 5 criteria each. Each item was rated as "Yes," "No," or "Unclear." A rating of 5***** or 100% indicates fulfillment of all quality criteria, 4**** or 80% indicates fulfillment of 80% of the criteria, and so forth. For mixed methods studies, all 15 criteria (instead of 5) must be appraised. The overall quality rating adheres to the principle that the quality of the whole is determined by its weakest component.

Data Synthesis

This study conducted data analysis anchored in the 5 domains of the CFIR [16] and adhered to the Joanna Briggs Institute guidelines [19] for mixed methods systematic reviews. The analytical process involved: (1) integrating findings from distinct research methodologies separately; (2) translating quantitatively synthesized results into qualitative narratives and merging them with qualitatively synthesized evidence; and (3) finally establishing integration categories, inducing thematic patterns, and deriving consolidated outcomes.

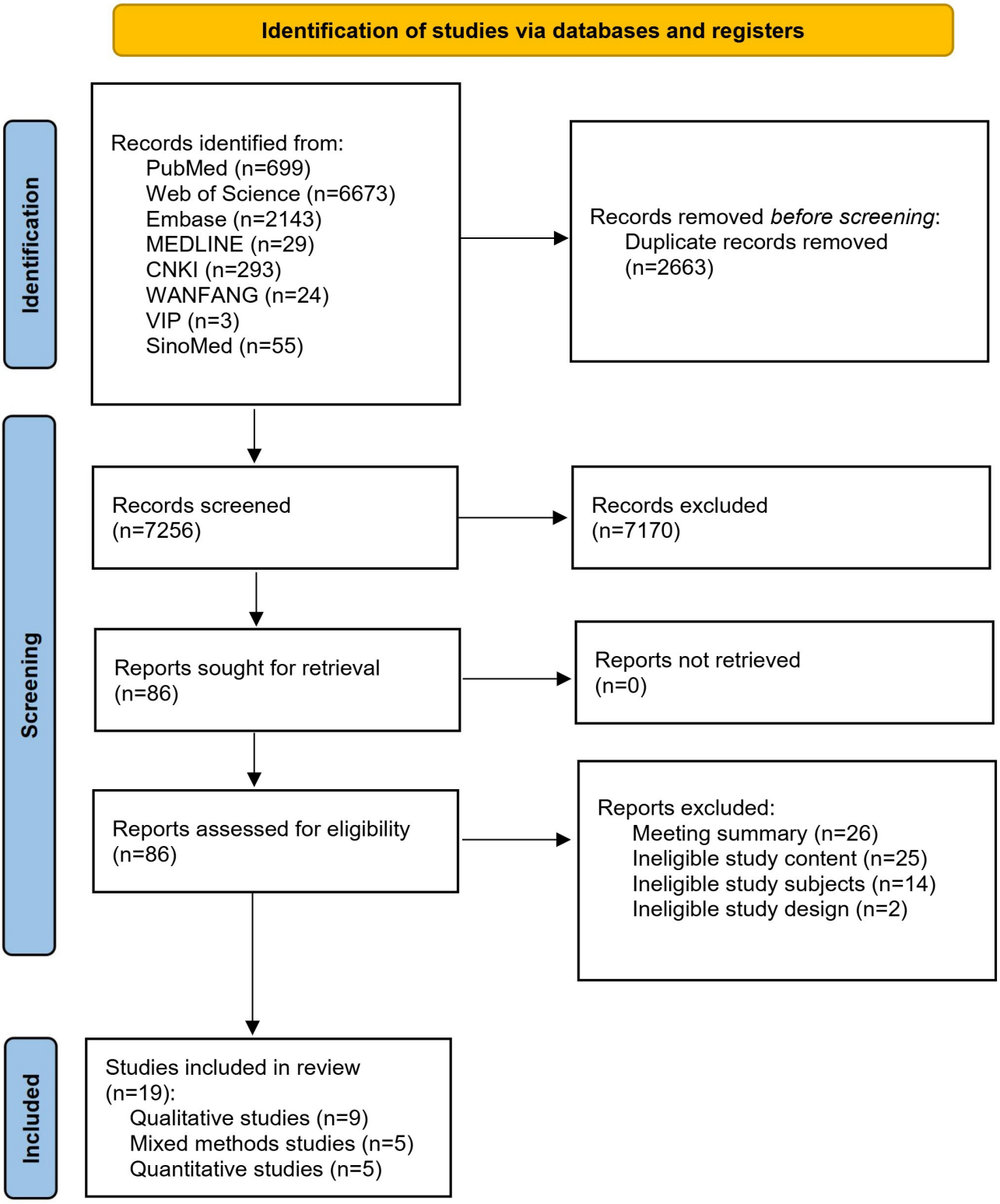
Results

Literature Search Results

A total of 9919 relevant studies were retrieved from the databases. After removing duplicates, 7256 studies remained. Following preliminary screening, 86 studies were retained. Ultimately, 19 studies were included after full-text review. The literature selection process and the list of included studies are presented in Figure 1 and Table 1, respectively.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. CNKI: China National Knowledge Infrastructure; VIP: China Science and Technology Journal Database.



XSL•FO
RenderX

Table . Characteristics of the included studies.[a]

Author(s)	Year	Country	Study design	Sample size (n)	Study population	Sampling method	Data collection methods/tools	Names/functions of mHealth	MMAT rating
Esayed et al [11]	2025	United States	Qualitative	20	Adult kidney transplant recipients	Purposive sampling	Semistructured interviews	A health care delivery platform using a live video visits for post-transplant follow-up	5*****
Malo et al [20]	2024	Canada	Qualitative	11	Kidney transplant recipients participating	Purposive sampling	Semistructured interviews	KEeP ACTIVe Club for increasing physical activity levels and reducing loneliness in kidney transplant recipients	5*****
Tang et al [21]	2023	Australia	Mixed methods	91	Adult kidney transplant recipients	Multistage sampling	Questionnaires (Self-developed questionnaire, eHealth Literacy Scale); semistructured interviews	eHealth	1*
Tang et al [22]	2022	Australia	Qualitative	30	Adult kidney transplant recipients	Purposive sampling	Semistructured interviews	eHealth	5*****
Castle et al [23]	2022	United Kingdom	Mixed methods	n_1=17 (quantitative); n_2=13 (qualitative)	Adult kidney transplant recipients	Unreported (quantitative); Purposive sampling (qualitative)	Questionnaires (General Practice Physical Activity Questionnaire, Nutrition Self-Efficacy Scale, Physical Exercise Self-Efficacy Scale, EuroQol 5-Dimension-5 Level, Chalder Fatigue Scale, 6-min walk test); body composition and parameter measurements; semistructured interviews	ExeRTion for prevention of post–kidney transplantation weight gain	2**
Huuskes et al [24]	2021	Australia	Qualitative	34	Adult kidney transplant recipients	Unreported	Focus group interviews	Telehealth	5*****

XSL•FO
RenderX

Author(s)	Year	Country	Study design	Sample size (n)	Study population	Sampling method	Data collection methods/tools	Names/functions of mHealth	MMAT rating
Castle et al [15]	2021	United Kingdom	Qualitative	17	Kidney transplant recipients (n=11), transplant-related health care providers (n=6)	Purposive sampling	Think-aloud method; semistructured interviews	ExeRTion for prevention of post–kidney transplantation weight gain	5*****
Krause et al [25]	2021	Germany	Mixed methods	55	Adult kidney transplant recipients	Convenience sampling	Questionnaires (4-item Basel Assessment of Adherence to Immunosuppressive Medication Scale, Self-developed questionnaire); blood assays; semistructured interviews	SimpleMed+ for estimating medication adherence in kidney transplant patients	1*
O'Brien et al [26]	2020	United States	Quantitative	53	Older kidney transplant recipients (≥60 y)	Convenience sampling	Questionnaires (Self-developed questionnaire)	Fitbit Charge 2 for collecting physical activity data in real time	4****
Nielsen et al [27]	2020	Denmark	Qualitative	36	Adult kidney transplant recipients (n=16), clinicians (n=16), nurses (n=4)	Purposive sampling	Semistructured interviews; focus group interviews	Telehealth solution consisted of a mobile app and a workflow for posttransplant follow-up	5*****
O'Brien et al [28]	2020	United States	Qualitative	20	Adult kidney transplant recipients	Purposive sampling	Semistructured interviews	mHealth apps	5*****
Côté et al [29]	2019	Canada	Qualitative	10	Adult kidney transplant recipients	Purposive sampling	Semistructured interviews	Web-based tailored virtual nursing intervention for promoting medication adherence and supporting self-management among kidney transplant recipients	5*****

XSL•FO
RenderX

Author(s)	Year	Country	Study design	Sample size (n)	Study population	Sampling method	Data collection methods/tools	Names/functions of mHealth	MMAT rating
Wedd et al [14]	2019	United States	Quantitative	710	Adult kidney transplant recipients (n=455) and liver transplant recipients (n=255)	Unreported	Medical records and database extraction	Web-based patient portal systems	3***
O'Brien et al [30]	2018	United States	Quantitative	165	Adult kidney transplant recipients	Convenience sampling	Questionnaires (Self-developed questionnaire)	Mobile apps for self-management of care among kidney transplant recipients	4****
Wang et al [31]	2017	Netherlands	Quantitative	46	Adult kidney transplant recipients	Unreported	Questionnaires (Self-developed questionnaire)	Self-management support systems (SMSS) for increasing the level of self- management among kidney transplant recipients	4****
Van Lint et al [32]	2015	Netherlands	Mixed methods	30	Adult kidney transplant recipients	Unreported	Questionnaires (Self-developed questionnaire, Worry Scale, Transplant Effects Questionnaire, Self-efficacy Scale, Health Care Climate Questionnaire); semistructured interviews	StatSensor Xpress for creatinine self-measurement; Microlife WatchBP Home for blood pressure self-measurement; disease-management system (DMS) for registering measurement results	5*****
McGillicuddy et al [33]	2013	United States	Quantitative	99	Adult kidney transplant recipients	Convenience sampling	Questionnaires (Self-developed questionnaire, Perceived Stress Scale, 7-item Morisky Medication Adherence Scale)	A mHealth remote monitoring system for monitoring medication adherence and physiological parameters	5*****
Schäfer-Keller et al [34]	2009	Switzerland	Qualitative	22	Adult kidney transplant recipients (n=14), clinicians (n=8)	Purposive sampling	Structured interviews	OTIS for increasing the level of self-management among kidney transplant recipients	5*****

XSL•FO
RenderX

Author(s)	Year	Country	Study design	Sample size (n)	Study population	Sampling method	Data collection methods/tools	Names/functions of mHealth	MMAT rating
Russell et al [35]	2009	United States	Mixed methods	85	Older kidney transplant recipients (>55 y)	Convenience sampling	Questionnaires (Self-developed questionnaire); semistructured interviews	MEMS for improving medication adherence among kidney transplant recipients	1*

[a]The Mixed Methods Appraisal Tool (MMAT) uses the following rating system: 5***** indicates full compliance with 100% of the quality assessment criteria, 4**** corresponds to meeting 80% of the criteria, 3*** signifies adherence to 60%, 2** represents 40%, and 1* denotes meeting 20% of the standards.

Characteristics of the Included Studies

All included studies were published in English, with publication years spanning from 2009 to 2025. These studies originated from 8 countries, with the United States contributing the highest number of publications [11,14,26,28,30,33,35]. The research designs comprised 9 qualitative studies, 5 mixed methods studies, and 5 quantitative studies. The study populations included 1265 kidney transplant recipients and 34 health care professionals. The basic characteristics of the included studies are summarized in Table 1.

Quality Assessment of the Included Studies

A total of 9 qualitative studies exhibited high overall quality, each scoring 5*****. Among the 5 quantitative studies, 1 [14] did not clearly describe whether the sampling method was appropriate for addressing the research questions, another [14] failed to specify the representativeness of the sample relative to the target population, 1 [30] lacked clarity on the risk of nonresponse bias, 1 [26] did not explicitly address confounding factors in the study design or analysis, and 1 [31] did not confirm whether interventions were implemented as intended during the research period. The 5 mixed methods studies demonstrated lower overall quality: 4 studies [21,23,25,35] did not justify the rationale for adopting a mixed methods design to address the research questions, 4 [21,23,25,35] inadequately explained the integration of qualitative and quantitative results, 4 [21,23,25,35] insufficiently addressed discrepancies or heterogeneity between quantitative and qualitative findings, and 3 [21,25,35] failed to cohesively integrate distinct components of their research. The MMAT scores for all included studies are presented in Table 1.

Factors Influencing the Application of mHealth in Kidney Transplant Care

This study identified 16 facilitators and 14 barriers influencing the application of mHealth in kidney transplant care. Based on the CFIR, both facilitators and barriers were categorized into 5 domains, as detailed in Table 2.

XSL•FO
RenderX

Table . Factors influencing the application of mHealth in kidney transplant care.

Category	Facilitators	Barriers
Intervention characteristics	Evidence strength and quality: • Authoritative content [21,22] Relative advantage: • Convenience and flexibility (time-, effort-, and cost-saving) [11,21,22,24,27,29,35] • Reduced infection risk [11,22,24,31] • Improved comfort [11] Design quality and packaging: • User-friendly design (easy to understand, operate, and humane) [22,28,29,35] • Personalized design [15,22,28]	Evidence strength and quality: • Lack of authoritative content [34] Complexity: • Communication/interaction limitations [11,20,22,24,27] • Privacy/security risks [22,23] • Negative online content [22] Design quality and packaging: • Poor design (complex operation, unclear content, inappropriate sizing) [15,21,25,26,28,30,34,35]
Outer setting	Patient needs and resources: • No additional costs [33]	Patient needs and resources: • External environmental interference [23,24] • Extra costs [30]
Inner setting	Networks and communications: • Professional support [23,32] • Peer support [15,20,21]	Structural characteristics: • Unstable network/equipment [24] Networks and communications: • Lack of multidisciplinary team support [24] Compatibility: • Time conflicts [20] • Conflicts with user habits [25,35]
Characteristics of individuals	Knowledge and beliefs about the intervention: • mHealth literacy [14,21] Self-efficacy: • Self-efficacy [23,32] Individual identification with organization: • Perceived benefits [20,21,29,31,34] Other personal attributes: • Intrinsic motivation (curiosity, altruism, responsibility) [20,23,26,29] • Demographics (race) [33]	Knowledge and beliefs about the intervention: • Lack of mHealth literacy [21,22,24,30] Other personal attributes: • Poor physical condition [24,30]
Process	Engaging: • Training on device functions [15,30] • Executing: SMS/email reminders [15]	Engaging: • Lack of training on device functions [27]

Facilitators

Intervention Characteristics

Intervention characteristics are critical predictors of the successful application of interventions. The authoritative content of mHealth platforms can effectively alleviate user concerns and enhance their confidence in using these tools (“These sorts of webpages would need to carry some sort of certification so that we, the individuals, know we are getting professional advice as opposed to some blogger.” [21]). Compared to traditional face-to-face care models, mHealth offers advantages such as convenience and flexibility ("Now [with telehealth] it’s very convenient for me because I can even take calls while I’m working in the office or whether I’m home.” [24]; “I think video visits might actually help with compliance, especially for socioeconomically disadvantaged or older patients that don’t drive.” [11]), reduced infection risks (“I prefer [eHealth], I’m so scared of COVID. I don’t want to get sick, so I’m happy to do the online, video calls and things like that.” [22]), and enhanced comfort ("I was comfortable sharing personal information since, like I said, I’m in my own house. So, it’s easier to actually sometimes talk to the doctor when you’re at home.” [11]). These advantages serve as key motivators for users to persist with mHealth. Additionally, some users emphasized that designs featuring clear language (“Level of language, get the language to a layperson’s point of view. Doctors would understand what it was and so would the person who is reading it… [or else] they would go away from it.” [22]), ease of use (“You know, it’s user-friendly, going to it isn’t complicated. […] It’s pretty easy to figure out. […] You quickly get the hang of it.” [29]), human-centered care ("The nurse came across as very professional […] as someone who cares about your health […] Me, I found her to be warm. Even though it was a video, it’s. it’s not. it can seem a little cold, but it’s someone who’s talking to us, deep down.” [29]), and personalization (“If something was going wrong, yes, rather

XSL•FO
RenderX

than just getting an automatic message that was just saying 'keep going for your goals!' and you're like 'well I haven't been on 3 weeks'. I would prefer something more personable." [15]) would make mobile devices more appealing.

Outer Setting

Users' needs and resource accessibility are critical external factors determining the adoption of mHealth. If mHealth solutions require no additional financial burden on users, their willingness to adopt is significantly enhanced (79% of users endorsed this view [15]).

Inner Setting

Multidisciplinary collaboration and communication are central to the successful application of mHealth. Regular communication and professional guidance from health care providers motivate patients to adhere to care plans ("I think, um, it will be helpful [for my adherence] because I spoke to my physio quite a bit. She used to call me and, um, she would—I'd tell her sometimes—and she'd be like, you know what, you know keep busy, do this and do that and stuff like that. She would give me advice." [23]). Some patients highlighted that sharing experiences and emotional support with peers fosters a sense of connection, alleviates loneliness, and strengthens their resolve to persist ("When I had a question, I could send a text and they [peers] responded right away, right away. So it was really good, you know." [20]; "Being together, that always motivates more than when we're alone." [20]).

Characteristics of Individuals

Individual characteristics are pivotal factors influencing the adoption of mHealth. An individual's mHealth literacy directly impacts the effectiveness of mHealth use. Those with higher mHealth literacy exhibit stronger technical proficiency, information evaluation skills, and greater acceptance of mHealth ("Factors associated with mHealth use include higher eHEALS scores and higher education, with odds ratios [ORs] of 1.21, 95% CI 1.06 - 1.38 and 7.78, 95% CI 2.19 - 27.7, respectively." [21]). Self-efficacy reinforces individuals' confidence in using mHealth ("I was just [pause] following the program through. Um, but that was just my personal thing, just because I have—you know—I have the knowledge and the confidence to do my own thing." [23]). Some patients noted that using mHealth offers benefits such as enhanced self-management ("Patient portal containing access to all of the above suggestions in one unified system, where my blood results, education, and app monitoring can all be tracked by myself at a click of a button. This will ultimately increase self-management among patients and allow for more detailed 'healthier life' conversations with health professionals." [21]) and disease-related knowledge ("All important things are included [in this platform]: where I have to take precautions, what I need to do. Everything is well explained." [34]). Intrinsic motivations such as curiosity ("It was out of curiosity. I'm naturally curious. So I really wanted to see how far it went..." [29]), altruism ("Well, I did it for you… it was to help you out, because you help me in a way. Your aim… the reason that you're doing this, I imagine, is to help me." [29]), and a sense of responsibility ("So, my point there is in terms of being accountable to something. Even though it's not a human being, you are being accountable to a system…" [23]) drive patients to persist. Additionally, race influences attitudes toward mHealth ("Black individuals exhibited more positive attitudes than White individuals [mean 4.25, SD 0.88 vs mean 3.76, SD 1.07; *P*=.02]." [33]).

Process

Supportive strategies during the implementation process are critical factors for ensuring the sustainability of mHealth interventions. Participants emphasized that receiving training prior to using mHealth builds operational confidence and foundational knowledge, reducing adherence decline or misuse caused by technical unfamiliarity ("I think the first steps you need it face-to-face to start with… then you can do it on your own at home." [15]). Additionally, reminder features in devices play a role in encouraging adherence ("If there's a 12-week program, what would be useful would be a text reminder or an email reminder." [15]).

Barriers

Intervention Characteristics

Inherent limitations of mHealth may hinder its application. When mHealth platforms provide content lacking official certification or conflicting with clinical recommendations, patients may question its reliability ("My doctor advised taking tacrolimus between meals, but the platform instructed me to take it 1 h before or 1.5 - 2 h after meals." [34]). User-unfriendly designs, such as overly lengthy interfaces ("Instead of scrolling through all the different things because that's [the excessively long pages] what drives me crazy with apps." [28]) or bulky devices ("The mobile health device [MEMS] is too large to carry around or travel with." [35]), foster resistance and reduce willingness to continue use. The virtual nature of mHealth limits nonverbal interaction ("I like face - to - face because my doctor tells me if he thinks that I'm putting on weight or I'm not doing the right thing." [24]; "Telehealth can't take your blood pressure… and they can't take a look at your wound or check you or examine you or weigh you. Sitting down, I could have gained 25 kg, but no one would know." [22]). Participants also expressed concerns about privacy breaches ("Hearing about hackers, and you don't want your records and things like that when you subscribe to these sites and things. You're putting yourself at risk and your whole medical history at risk. There are things in my medical history that I don't want people to know." [22]). Additionally, exposure to negative online content during use imposed psychological burdens ("Sometimes it got a bit depressing… it was like really a bit of a mental burden as well, just like constantly getting notifications about this message and that message and someone dying. Sometimes when you're in a difficult headspace, it can be really hard to be confronted with that." [22]).

Outer Setting

The inability of external resources to meet user needs can hinder the application of mHealth. The use of mHealth may be interfered with by external environments, thereby negatively impacting its effectiveness ("Because I think face - to - face, you're in the room, you're kind of in the moment, with

XSL•FO
RenderX

telehealth, you can get distracted. You can get a bit confused" [24]; "If you're doing telehealth, then you might miss something like the dog barking in the background or the kids screaming." [24]). Additionally, participants noted that additional costs associated with mHealth reduce willingness to use such services, as these expenses impose financial burdens [30].

Inner Setting

Structural barriers within mHealth systems represent core constraints affecting their application. Participants reported that unstable network equipment compromises their user experience ("Like when the internet's down, when it's slow, which it is often at my place. Even my phones aren't that good. The landline, even… [Whether I can use mHealth normally] just depends on if everything's working." [24]). Compared to traditional face-to-face clinical models, mHealth struggles to ensure patients receive simultaneous support from multidisciplinary teams ("When I do telehealth, you become sort of like missing out on all the other aspects of allied health." [24]; "Nurses are a part of my training team as a nephrologist, and I feel that connect with the nurses is completely cut off [with telehealth]." [24]). Some mHealth services require advance appointments, yet users often fail to attend due to personal scheduling conflicts ("The schedule was a problem, it [mHealth] just didn't work for me at all." [20]). Additionally, some participants expressed difficulty adapting to new self-management models via mHealth due to entrenched habits (eg, medication or exercise routines) ("When the situation differs from my routine, it [mHealth] gets in the way and is hard to accommodate." [35]).

Characteristics of Individuals

User characteristics directly influence the effectiveness of mHealth adoption. For users with limited mHealth literacy, such as older adults or those with lower education levels, technical challenges encountered during mHealth use often lead to abandonment of the technology ("Not being a tech-savvy person, I find trying to navigate and download apps very stressful." [21]; "The problem is that when you get older, it's harder for you to understand and pick up if you are not constantly on the computer. I don't use it much." [22]). Additionally, users' physical conditions impact mHealth utilization, with patients reporting that visual impairments and limited hand dexterity hinder their engagement with mHealth platforms [30].

Process

Insufficient training on mHealth use results in ongoing challenges for users during later implementation ("I would say it is incredibly complex in the sense that when the doctor says it, it sounds logical and totally correct, but when you come home, then what? Because then if I make a slight deviation, is it the same, or is it a new situation?" [27]).

Discussion

Principal Findings

An important finding of this review is that, despite the comprehensive 5-domain analytical structure offered by CFIR, a notable imbalance exists in the focus of existing research. Influencing factors are predominantly concentrated within the domains of intervention characteristics, characteristics of individuals, and the inner setting. In contrast, factors related to the outer setting and process domains, such as policy support and stakeholder collaboration, remain significantly underrepresented.

This pattern may be explained by 2 primary reasons. First, the majority of included primary studies focused on technology acceptance and user experience. Their participants were predominantly kidney transplant recipients, who are generally not in a position to provide in-depth perspectives on macrolevel issues such as health care policies, regional resource distribution, or multistakeholder coordination. Second, evidence concerning the complex factors of the outer setting and process is more likely to reside in what is known as "grey literature," including project evaluation reports, dissertations, or specific policy studies, which was not captured by our predefined academic database search.

Optimizing mHealth Design and Performance to Enhance User Experience

The findings of this study indicate that the application of mHealth is closely tied to its inherent characteristics. The domain of intervention characteristics contains the highest number of influencing factors, which aligns with prior studies [36]. For successful integration into the workflow, the intervention needs to align with the prevailing value system of key stakeholders.

After enduring the hardships of dialysis, most kidney transplant recipients highly value their transplanted kidney and exhibit heightened caution in health management [37]. The authoritative nature of mHealth content helps build trust in the platform and strengthens users' confidence. Studies [38,39] reveal that approximately 30% of kidney disease–related apps on the market lack clinically validated content, posing risks such as incorrect medication advice or deviations from the latest guidelines, which significantly undermine patient trust and hinder sustained use. Therefore, mHealth development should be grounded in evidence-based medicine and clinical guidelines, with regular updates. Prior to releasing content, experts in kidney transplantation should review materials to ensure scientific rigor. The design and performance of mHealth directly impact user experience. Kidney transplant recipients vary widely in age, educational background, and health literacy. With global population aging, the proportion of older recipients is rising [40,41]. Older users may struggle with complex medical terminology or operational workflows due to cognitive decline or limited technical adaptability. Thus, mHealth design should prioritize accessibility, using plain language and intuitive interfaces to lower barriers and improve acceptance. Kidney transplant recipients commonly face psychological challenges such as anxiety and depression [42]. Research [43] shows that integrating emotional support features into mHealth can alleviate patient anxiety and enhance long-term engagement. Additionally, lifelong self-management is essential for recipients. If apps lack personalized feedback, users may abandon them due to perceived low value. To address this, personalized features such as achievement badges to motivate

XSL•FO
RenderX

users or tailored notifications to boost engagement could be incorporated.

mHealth is a double-edged sword. While offering advantages such as convenience, flexibility, and reduced infection risks, it also has limitations, including security risks, hindered clinician-patient interaction, and exposure to negative content, which diminish user willingness to adopt it. Studies [44,45] highlight widespread patient concerns about data security in mHealth, particularly for long-term conditions such as kidney transplantation, where sensitivity to privacy breaches is heightened [46]. Thus, health care providers should strengthen awareness of data protection and regularly debug and update software to prevent cyberattacks. Beyond safeguarding privacy, stakeholders could enhance user experience by improving content moderation and integrating wearable devices to mitigate mHealth's shortcomings.

Strengthening Multidimensional Support Systems to Improve Resource Accessibility

The findings of this study reveal that the adoption of mHealth is influenced by multiple internal and external factors. Additionally, outer setting factors are relatively few among all influencing factors. This aligns with a prior scoping review by Rangachari et al [47].

Peer, professional, and multidisciplinary team support significantly impact kidney transplant recipients' engagement with mHealth, suggesting that health care providers should establish multidimensional support strategies. Peer support fosters emotional resonance and experience sharing among patients with similar backgrounds, reducing loneliness and enhancing willingness to use mHealth. Therefore, integrating peer communication modules into mHealth platforms is recommended; however, medical advice or medication sharing should be explicitly prohibited to ensure compliance. Guidance and feedback from health care professionals can alleviate patients' technological apprehensions. Studies [48] indicate that lack of clinician support is a barrier to mHealth adoption, aligning with this study's conclusions. Thus, remote follow-ups by professionals to guide patients, alongside accessible channels for reporting issues, are proposed to build trust and sustain engagement. The development and implementation of mHealth require collaboration and guidance from multidisciplinary teams. Without such support, unresolved technical or medical challenges during use may lead to unmet needs and reduced compliance [49]. Future initiatives should establish mHealth teams comprising physicians, specialist nurses, rehabilitation therapists, and IT engineers to optimize application effectiveness. Additionally, internal and external factors such as additional costs and unstable network infrastructure hinder mHealth accessibility. Expenses related to data consumption or smart device procurement may impose financial burdens on patients. Research [50] reports that 23% of patients in resource-limited settings abandon mHealth due to unaffordable device costs. To address this, governments and hospitals should establish subsidy mechanisms to reduce user expenses. Developing low-data or offline versions of mHealth platforms could simultaneously mitigate data consumption and connectivity issues, thereby improving resource accessibility.

Stimulating Individuals' Intrinsic Motivation, Enhancing Perceived Benefits, and eHealth Literacy

The findings of this study indicate that the application of mHealth is influenced by individual characteristics, such as intrinsic motivation and perceived benefits. Notably, factors at the individual level are relatively numerous among all identified influences, which is similar to the findings of the previous review by Neil-Sztramko et al [51].

Intrinsic motivation refers to an individual's inherent tendency to seek novel challenges, develop competencies, and engage in exploratory learning driven by personal interest or enjoyment of the activity itself, which is more likely to sustain long-term behavioral adherence [52]. For users lacking intrinsic motivation, motivational interviewing techniques such as goal setting, reflective practices, verbal persuasion, and action planning can be used to initiate and reinforce their motivation, thereby promoting behavioral persistence [53]. This study also reveals that perceived benefits can enhance users' sustained engagement with mHealth. Research demonstrates [54] that perceived benefits indirectly influence behavioral adoption through the mediating role of self-efficacy. Health care professionals should acknowledge and encourage users' perceived benefits, reinforce positive experiences through multiple channels, and further enhance self-efficacy to facilitate long-term mHealth adherence. When users fail to perceive benefits, their behavioral persistence may be compromised [55]. In such cases, practitioners should promptly adjust intervention strategies through telephone follow-ups, video demonstrations, and other methods to strengthen perceived benefits and improve self-efficacy. Furthermore, this study identifies that insufficient eHealth literacy constitutes a barrier to mHealth utilization. Some users experience difficulties in independently operating smartphones or applications due to age, educational background, or limited digital skills, leading to resistance or abandonment of mHealth tools. Therefore, beyond simplifying the usability of mHealth systems during development, health care providers should strengthen preimplementation training on device functionalities. Personalized training programs tailored to users' varying health literacy levels should be developed to compensate for eHealth literacy deficiencies.

Limitations

Limitations of the Included Studies

The inclusion of studies with varied methodological quality, while ensuring comprehensive analysis, may have introduced potential bias. The lower overall ratings for some mixed methods studies primarily reflected limitations in their integrative components rather than in the qualitative or quantitative parts from which data were extracted. Future studies should be more methodologically rigorous and better designed to enhance the overall quality and reliability of the evidence base in this field. All included studies were published in English and conducted in developed countries, introducing both linguistic and geographic bias. Consequently, the findings primarily reflect contexts with abundant resources and well-developed digital infrastructure, which may limit the generalizability of the conclusions. It is imperative to conduct research focusing on developing countries and non–English-speaking regions. No

XSL•FO
RenderX

subgroup analysis was conducted for key variables (eg, age, mHealth type). This was primarily because most studies did not stratify outcomes by age or treated mHealth as a broad concept without specifying functional types, hindering the extraction of subgroup-specific factors. Publication bias was not formally assessed. Due to the nature of this review, which included diverse study designs and focused on synthesizing influencing factors rather than quantifying a single effect size, standard graphical tests for publication bias (eg, funnel plots) were not applicable or were not conducted.

Limitations of the Review Methodology

The transformation from quantitative evidence to qualitative descriptions may have resulted in inevitable information loss. The lack of evaluation of literature in languages other than Chinese and English may have led to the omission of some influencing factors. Although 2 researchers independently performed literature screening and data extraction to minimize subjective bias, the results of consistency testing were not reported. This shortcoming represents a methodological limitation regarding the transparency of the selection process.

Conclusion

Guided by the CFIR, this study systematically analyzed the influencing factors of mHealth apps in kidney transplant care, which were categorized into 5 domains: intervention characteristics, outer setting, inner setting, characteristics of individuals, and process. The findings revealed that intervention characteristics constituted the most prominent influencing domain, while the outer setting and process demonstrated relatively fewer associated elements. Future research should further investigate the impact of these identified factors on mHealth apps in kidney transplant care. Additionally, health care providers should adopt multidimensional strategies targeting these determinants to optimize mHealth integration into clinical practice.

Acknowledgments

The authors express their sincere gratitude to all authors whose papers were included in this meta-synthesis.

Data Availability

The data are from published academic research literature and can be found in the original article.

Authors' Contributions

Conceptualization: YJ, HL
Data curation: XL, JF
Formal analysis: YJ, SG, XY
Investigation: YJ, SG, XY
Supervision: HL
Writing – original draft: YJ, SG
Writing – review & editing: XL, JF, LD, HS, WY, HL
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy in PubMed.
[DOCX File, 13 KB - medinform_v14i1e84139_app1.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 28 KB - medinform_v14i1e84139_app2.docx]

References

1. Eltayeb HHH, Rawat A, Salazar González JF, et al. Exploring the impact of diabetes on kidney transplant: patient outcomes and management strategies. Cureus 2025 Mar;17(3):e80843. [doi: 10.7759/cureus.80843] [Medline: 40255815]
2. Poggio ED, Augustine JJ, Arrigain S, Brennan DC, Schold JD. Long-term kidney transplant graft survival-making progress when most needed. Am J Transplant 2021 Aug;21(8):2824-2832. [doi: 10.1111/ajt.16463] [Medline: 33346917]
3. Neuberger JM, Bechstein WO, Kuypers DRJ, et al. Practical recommendations for long-term management of modifiable risks in kidney and liver transplant recipients: a guidance report and clinical checklist by the Consensus on Managing Modifiable Risk in Transplantation (COMMIT) group. Transplantation 2017 Apr;101(4S Suppl 2):S1-S56. [doi: 10.1097/TP.0000000000001651] [Medline: 28328734]

XSL•FO
RenderX

4. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. : World Health Organization; 2011 URL: https://iris.who.int/server/api/core/bitstreams/ad1b13c0-7c82-47b4-8dd5-f0a26c3a3cc3/content [accessed 2026-02-28]
5. Beck DK, Tielen M, Rechards M, et al. Comparison of a home-based (multi) systemic intervention to promoting Medication AdheRence and Self-management among kidney transplant recipients with care-as-usual: the MARS randomized controlled trial protocol. BMC Nephrol 2020 Aug 28;21(1):374. [doi: 10.1186/s12882-020-02008-z] [Medline: 32859157]
6. Massierer D, Sapir-Pichhadze R, Bouchard V, et al. Web-based self-management guide for kidney transplant recipients (The Getting on With Your Life With a Transplanted Kidney Study): protocol for development and preliminary testing. JMIR Res Protoc 2019 Jun 24;8(6):e13420. [doi: 10.2196/13420] [Medline: 31237243]
7. van Lint C, Wang W, van Dijk S, et al. Self-monitoring kidney function post transplantation: reliability of patient-reported data. J Med Internet Res 2017 Sep 26;19(9):e316. [doi: 10.2196/jmir.7542] [Medline: 28951385]
8. Pollock MD, Stauffer N, Lee HJ, et al. MyKidneyCoach, patient activation, and clinical outcomes in diverse kidney transplant recipients: a randomized control pilot trial. Transplant Direct 2023 Apr;9(4):e1462. [doi: 10.1097/TXD.0000000000001462] [Medline: 36935874]
9. Xie X, Wang X, Li A, et al. A study of the effectiveness of mobile health application in a self-management intervention for kidney transplant patients. Iran J Kidney Dis 2023 Sep;17(5):263-270. [doi: 10.52547/ijkd.7693] [Medline: 37838936]
10. Kuang M, Yao Z, Chen J, Yan Y, Li Z. The effectiveness of remote interventions based on digital health technology in kidney transplant recipients: a systematic review. World J Urol 2025 Mar 27;43(1):193. [doi: 10.1007/s00345-025-05582-9] [Medline: 40148521]
11. Esayed S, Kim E, Sung HC, et al. Hybrid telemedicine and in‐person care for kidney transplant follow‐up: a qualitative study. Clin Transplant 2025 Feb;39(2):e70106. [doi: 10.1111/ctr.70106] [Medline: 39945199]
12. Ghezzo A, Pascale G, Dew MA, et al. Perspectives on conducting behavioral intervention with adolescents using a virtual platform: a thematic analysis from the viewpoint of study coordinators. Pediatr Transplant 2025 Feb;29(1):e70006. [doi: 10.1111/petr.70006] [Medline: 39776082]
13. Zhou L, Wan J. Research progress on the application of telemedicine technology in medication adherence of kidney transplant recipients [Artcile in Chinese]. Chin Nurs Res 2024 Aug;38(19):3508-3512 [FREE Full text]
14. Wedd J, Basu M, Curtis LM, et al. Racial, ethnic, and socioeconomic disparities in web-based patient portal usage among kidney and liver transplant recipients: cross-sectional study. J Med Internet Res 2019 Apr 22;21(4):e11864. [doi: 10.2196/11864] [Medline: 31008707]
15. Castle EM, Greenwood J, Chilcot J, Greenwood SA. Usability and experience testing to refine an online intervention to prevent weight gain in new kidney transplant recipients. Br J Health Psychol 2021 Feb;26(1):232-255. [doi: 10.1111/bjhp.12471] [Medline: 32931645]
16. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
17. Luo M, Li Y, Zhang S, et al. Application of the consolidated framework for implementation research in the field of evidence-based nursing practice: a scoping review [Article in Chinese]. J Nurs Sci 2025 Jun;40(12):119-124 [FREE Full text]
18. Liao X, Hu R, Li B, et al. An introduction to Mixed Methods Appraisal Tool. Chin Gen Pract 2021 Sep;24(31):4015-4020. [doi: 10.12114/j.issn.1007-9572.2021.00.215]
19. Stern C, Lizarondo L, Carrier J, et al. Methodological guidance for the conduct of mixed methods systematic reviews. JBI Evid Synth 2020 Oct;18(10):2108-2118. [doi: 10.11124/JBISRIR-D-19-00169] [Medline: 32813460]
20. Malo MF, Janaudis-Ferreira T, Affdal A, et al. KEeP active club study: kidney transplant recipients' experiences of a physical activity and social interaction virtual group. Can J Kidney Health Dis 2024 Feb;11:20543581241229254. [doi: 10.1177/20543581241229254] [Medline: 38344312]
21. Tang J, Howell M, Lee VW, et al. Patients' perspectives, factors, and patterns of ehealth use in kidney transplant recipients. Kidney Int Rep 2023 Apr;8(4):727-736. [doi: 10.1016/j.ekir.2023.01.007] [Medline: 37069969]
22. Tang J, Howell M, Roger S, Wong G, Tong A. Perspectives of kidney transplant recipients on eHealth: semistructured interviews. Transplant Direct 2022 Dec;8(12):e1404. [doi: 10.1097/TXD.0000000000001404] [Medline: 36382131]
23. Castle EM, Dijk G, Asgari E, et al. The feasibility and user-experience of a digital health intervention designed to prevent weight gain in new kidney transplant recipients—the ExeRTiOn2 trial. Front Nutr 2022 Mar;9:887580. [doi: 10.3389/fnut.2022.887580] [Medline: 35677553]
24. Huuskes BM, Scholes-Robertson N, Guha C, et al. Kidney transplant recipient perspectives on telehealth during the COVID-19 pandemic. Transpl Int 2021 Aug;34(8):1517-1529. [doi: 10.1111/tri.13934] [Medline: 34115918]
25. Krause AV, Bertram A, Nöhre M, Bauer-Hohmann M, Schiffer M, de Zwaan M. Use of an electronic medication monitoring device to estimate medication adherence in kidney transplant patients. Transl Behav Med 2021 Apr 7;11(3):842-851. [doi: 10.1093/tbm/ibaa122] [Medline: 33710349]
26. O'Brien T, Meyer T. A feasibility study for teaching older kidney transplant recipients how to wear and use an activity tracker to promote daily physical activity. Nephrol Nurs J 2020 Jan;47(1):47-51. [Medline: 32083436]

XSL•FO
RenderX

27. Nielsen C, Agerskov H, Bistrup C, Clemensen J. Evaluation of a telehealth solution developed to improve follow-up after kidney Transplantation. J Clin Nurs 2020 Apr;29(7-8):1053-1063. [doi: 10.1111/jocn.15178] [Medline: 31889357]
28. O'Brien T, Rosenthal A. Preferred features in mobile health applications for kidney transplant recipients: a qualitative approach. Nephrol Nurs J 2020 Nov;47(6):529-536. [Medline: 33377753]
29. Côté J, Fortin MC, Auger P, et al. Web-based tailored nursing intervention to support medication self-management: a qualitative study of the experience of kidney transplant recipients. Comput Inform Nurs 2019 Nov;37(11):564-572. [doi: 10.1097/CIN.0000000000000572] [Medline: 31599744]
30. O'Brien T, Russell CL, Tan A, Washington M, Hathaway D. An exploratory correlational study in the use of mobile technology among adult kidney transplant recipients. Prog Transpl 2018 Dec;28(4):368-375. [doi: 10.1177/1526924818800051]
31. Wang W, van Lint CL, Brinkman WP, et al. Renal transplant patient acceptance of a self-management support system. BMC Med Inform Decis Mak 2017 May 8;17(1):58. [doi: 10.1186/s12911-017-0456-y] [Medline: 28482896]
32. van Lint CL, van der Boog PJ, Wang W, et al. Patient experiences with self-monitoring renal function after renal transplantation: results from a single-center prospective pilot study. Patient Prefer Adherence 2015 Dec;9:1721-1731. [doi: 10.2147/PPA.S92108] [Medline: 26673985]
33. McGillicuddy JW, Weiland AK, Frenzel RM, et al. Patient attitudes toward mobile phone-based health monitoring: questionnaire study among kidney transplant recipients. J Med Internet Res 2013 Jan 8;15(1):e6. [doi: 10.2196/jmir.2284] [Medline: 23305649]
34. Schäfer-Keller P, Dickenmann M, Berry DL, Steiger J, Bock A, De Geest S. Computerized patient education in kidney transplantation: testing the content validity and usability of the Organ Transplant Information System (OTIS). Patient Educ Couns 2009 Jan;74(1):110-117. [doi: 10.1016/j.pec.2008.03.027] [Medline: 18515034]
35. Russell CL, Owens S, Hamburger KQ, et al. Medication adherence and older renal transplant patients' perceptions of electronic medication monitoring. J Gerontol Nurs 2009 Oct;35(10):17-21. [doi: 10.3928/00989134-20090903-06] [Medline: 19772223]
36. Liverpool S, Mota CP, Sales CMD, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317. [doi: 10.2196/16317] [Medline: 32442160]
37. Li XR, Liu HX, Zhao Y, et al. Path analysis of the factors influencing self-management among kidney transplant recipients based on the health promotion model. Chin Nurs Manag 2025 Apr;25(4):562-567. [doi: 10.3969/j.issn.1672-1756.2025.04.015]
38. Paneerselvam GS, Lua PL, Chooi WH, Rehman IU, Goh KW, Ming LC. Effectiveness of mobile apps in improving medication adherence among chronic kidney disease patients: systematic review. J Med Internet Res 2025 Apr 16;27:e53144. [doi: 10.2196/53144] [Medline: 40239197]
39. Kelly JT, Jegatheesan DK, Dawson J, et al. Are digital health technologies and models of nutrition care the future of chronic kidney disease management? J Ren Nutr 2023 Nov;33(6S):S80-S87. [doi: 10.1053/j.jrn.2023.02.004] [Medline: 36965753]
40. Boenink R, Kramer A, Masoud S, et al. International comparison and time trends of first kidney transplant recipient characteristics across Europe: an ERA Registry study. Nephrol Dial Transplant 2024 Mar 27;39(4):648-658. [doi: 10.1093/ndt/gfad189] [Medline: 37653455]
41. Betjes MGH, Kho MML, Roodnat J, de Weerd AE. Transplant candidates of 70+ years have superior survival if receiving pre-emptively a living donor kidney. J Clin Med 2024 Mar 23;13(7):1853. [doi: 10.3390/jcm13071853] [Medline: 38610618]
42. Temür BN, Aksoy N. Psychosocial problems in recipients of kidney transplant. Exp Clin Transplant 2024 Jan;22(Suppl 1):187-191. [doi: 10.6002/ect.MESOT2023.P18] [Medline: 38385395]
43. Gibson CA, Gupta A, Naik A, et al. Developing a healthy lifestyle program for recent kidney transplant recipients. Prog Transplant 2023 Sep;33(3):193-200. [doi: 10.1177/15269248231189878] [Medline: 37469164]
44. Messner EM, Sturm N, Terhorst Y, et al. Mobile apps for the management of gastrointestinal diseases: systematic search and evaluation within app stores. J Med Internet Res 2022 Oct 5;24(10):e37497. [doi: 10.2196/37497] [Medline: 36197717]
45. Steubl LS, Reimann J, Simon L, et al. A systematic quality rating of available mobile health apps for borderline personality disorder. Borderline Personal Disord Emot Dysregul 2022 Jun 1;9(1):17. [doi: 10.1186/s40479-022-00186-w] [Medline: 35642024]
46. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024 May 31;26:e50715. [doi: 10.2196/50715] [Medline: 38820572]
47. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022 Nov 30;22(1):1450. [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
48. Qin R, Ma L, Xing H. Systematic review and meta integration of qualitative research on the use of mobile medical apps for chronic disease management diabetes patients. China Health Ind 2020 Jan;17(3):8-14 [FREE Full text]
49. O'Brien T, Rose K, Focht B, Al Kahlout N, Zynda L, Daloul R. Lessons learned for training a research team and teaching kidney transplant recipients how to use a mobile health technology for dietary intake and physical activity via a virtual platform. Nephrol Nurs J 2022 Nov;46(6):527-530. [Medline: 36645363]

XSL•FO
RenderX

50. Shukla R, Kapur A. Methods and lessons from costing a large mHealth intervention at scale in India. Front Public Health 2022 Jun;10:894390. [doi: 10.3389/fpubh.2022.894390] [Medline: 35719673]
51. Neil-Sztramko SE, Smith-Turchyn J, Fong A, Kauffeldt K, Tomasone JR. Community-based exercise programs for cancer survivors: a scoping review of program characteristics using the Consolidated Framework for Implementation Research. Arch Phys Med Rehabil 2022 Mar;103(3):542-558. [doi: 10.1016/j.apmr.2021.06.026] [Medline: 34375631]
52. Ryan RM, Deci EL. A self-determination theory approach to psychotherapy: the motivational basis for effective change. Can Psych 2008 Aug;49(3):186-193. [doi: 10.1037/a0012753]
53. Yu Y, Yu G, Xiong X, et al. Overview of behavior change wheel theory and progress in clinical nursing application. J Nurs Chin 2023 Mar;30(4):38-43. [doi: 10.16460/j.issn1008-9969.2023.04.038]
54. Li J, He J, Yang G, Cao Z, Zhong X. Influencing factors of HCV testing willingness among men who have sex with men in China: a structural equation modeling analysis. PLoS One 2025 Apr;20(4):e0321469. [doi: 10.1371/journal.pone.0321469]
55. Bechard LE, Beaton D, McGilton KS, Tartaglia MC, Black SE. Physical activity perceptions, experiences, and beliefs of older adults with mild cognitive impairment or Alzheimer's disease and their care partners. Appl Physiol Nutr Metab 2020 Nov;45(11):1216-1224. [doi: 10.1139/apnm-2019-0730] [Medline: 32369706]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CNKI: China National Knowledge Infrastructure
mHealth: Mobile Health
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews

Edited by A Benis; submitted 15.Sep.2025; peer-reviewed by J Lyu, M Galani; revised version received 21.Jan.2026; accepted 17.Feb.2026; published 24.Mar.2026.

Please cite as:
Jia Y, Guo S, Yang X, Lin X, Fang J, Dong L, Li X, Sun H, Yu W, Liu H
Influencing Factors of Mobile Health Apps in Kidney Transplant Care: Systematic Review Using the Consolidated Framework for Implementation Research
JMIR Med Inform 2026;14:e84139
URL: https://medinform.jmir.org/2026/1/e84139
doi:10.2196/84139

© Yingtian Jia, Shaobo Guo, Xinran Yang, Xiaohong Lin, Jiaxin Fang, Lei Dong, Xiangru Li, Haiya Sun, Wanhui Yu, Hongxia Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Public Reporting Systems for Evaluating Health Care Quality: Systematic Review

Lanmei Du, MSc; Reima Suomi, PhD; Elorm Damalie, MSc

Department of Management and Entrepreneurship, Turku School of Economics, University of Turku, Rehtorinpellonkatu 3, Turku, Finland

Corresponding Author:
Lanmei Du, MSc
Department of Management and Entrepreneurship, Turku School of Economics, University of Turku, Rehtorinpellonkatu 3, Turku, Finland

Abstract

Background: Health care public reporting (PR) refers to making information about the quality and performance of health care providers available to the public. The primary targeted use of PR is the selection of a health care provider. Previous studies suggest that PR has improved health care quality; however, the overall adoption rate of PR systems remains low. Misalignment between PR information and users' actual needs can explain this gap.

Objective: This study conducts a systematic literature review of PR systems in health care, aiming to explore how data encoding and presentation influence user utilization, with particular attention to the impact of data inconsistencies and inaccuracies on actual use.

Methods: A literature search was conducted on 5 electronic databases (Web of Science, CINAHL, Embase, PsycINFO, and PubMed), focusing on studies providing information in health care from an individual perspective. Subsequently, the quality of the included studies was assessed using the Mixed Methods Appraisal Tool (McGill University). Finally, a total of 25 empirical studies were included in this study.

Results: Among the 25 studies, 40% (10/25) investigated website presentation formats, 20% (5/25) explored dashboard-based data presentation formats in PR systems, 48% (12/25) examined data quality, and 20% (5/25) addressed user heterogeneity. Quantitative methods were used in 80% (20/25) of studies, while qualitative and mixed method designs accounted for 16% (4/25) and 4% (1/25), respectively. Findings suggest that standardized website and dashboard guidelines improve data reliability and user comprehension and that user heterogeneity mediates the effects of structural and process indicators on outcomes. Data collection was completed in February 2025, and the study was projected to be fully completed by September 2025.

Conclusions: This study provides an integrated structure-process-outcome–based framework for PR in health care. By presenting data through dashboards, inconsistencies and inaccuracies in data across different web pages can be mitigated, thereby decreasing conflicting information and improving comprehension in PR systems. Based on the literature findings and identified knowledge gaps, this study also proposes future research directions for online PR.

(JMIR Med Inform 2026;14:e80435) doi:10.2196/80435

KEYWORDS

digital health; public reporting systems; data and information process; health literacy; quality indicators

Introduction

Background

Digital health refers to the development and use of digital technologies to enhance health-related knowledge and practices [1]. It is not only a technological concept but also profoundly shapes how people access health information. For instance, a 2023 report by Digitalis Medical indicates that 7% of all daily Google searches are health-related, equating to approximately 70,000 health care searches every minute [2].

The core function of public reporting (PR) systems is to assist individuals in selecting health care providers and services [3]. Improving the use of the PR system requires understanding how data are encoded and decoded in digital health. This understanding is crucial for individuals who use the PR system to proactively manage their health and engage in recommended behaviors [4].

Previous studies have shown that inconsistencies or errors in PR data can affect trust and reduce users' confidence in the reporting system [5]. Such inconsistencies not only challenge data reliability but also influence how users understand PR information. However, empirical research on these data inconsistencies remains fragmented. For example, some studies suggest that information presentation affects users' actual use of PR systems [6,7]. Moreover, these studies largely remain focused on information presentation formats, while research on

XSL•FO
RenderX

data heterogeneity within systems remains limited, underscoring the complexity of PR data usage.

Patient activation measure (PAM) refers to an individual's knowledge, skills, and confidence essential for managing their own health and health care [8,9]. Although patients show strong interest in accessing PR information to make informed health care decisions, the overall adoption of the PR system remains low, largely due to information presentations that do not match patients' comprehension levels [10-12]. Therefore, exploring data generated in real-world physician-patient co-management contexts may help alleviate data communication challenges and support informed decision-making [13,14].

To address this gap, this study conducts a systematic literature review to provide an overview of current research on PR in digital health information systems. To achieve this objective, we address the following research questions: (1) How do system-level structural indicators influence process indicators, and subsequently influence outcome indicators? and (2) How do decoding errors in PR systems and user-system interactions influence users' information use?

Accordingly, we systematically identified and reviewed 25 empirical studies on information presentation in PR in health care contexts. Drawing upon the structure-process-outcome (SPO) theoretical framework, we categorized and synthesized the structural indicators (website guidelines and dashboard guidelines), process indicators (data quality guidelines and data process criteria), and outcome indicators (user satisfaction and surgical mortality). Meanwhile, we position user heterogeneity as a mediating variable that links process indicators to outcome indicators. Specifically, variations in contextual and individual-level factors (age, gender, political affiliation, employment status, income, and area of residence) can influence the interactions of users' engagement with information.

The remainder of this paper is structured as follows: the "Methods" section describes the methodology for planning and conducting the literature review, focusing on recent research; the "Results" section presents the findings in response to the 2 research questions; and the "Discussion" section discusses the progress and identifies challenges in current literature, concluding with recommendations for future research. Checklist 1 presents the PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) 2020 checklist which outlines recommended items to address in a systematic review protocol. Multimedia Appendix 1 contains the search strategies used in each database. Multimedia Appendix 2 shows the Mixed Methods Appraisal Tool (MMAT; McGill University) quality assessment report. Multimedia Appendix 3 contains the key characteristics of the included studies. Multimedia Appendix 4 includes additional details of the included studies.

Theoretical Background

Overview

The presentation of information in public reports significantly influences user engagement with these systems [15]. To gain an understanding of individual users' attitudes toward the PR system and how this forms health information adoption behavior, it is necessary to explore the information presented in the PR systems. While information is composed of data, the way data is structured determines how information is understood [16]. Based on this premise, this study assesses whether existing PR data align with users' informational needs and examines how such alignment influences users' adoption of PR systems.

PR and Quality Indicators

PR in health care refers to the disclosure of performance-related information about providers, organizations, or systems to the general public. Such information is typically nonanonymous, identifying specific professionals or institutions, thereby enabling direct comparison and accountability [17,18]. By disclosing the performance data, individuals (patients and nonpatients) can choose higher-performing providers and make health plan decisions [17,19].

The assessment of PR systems, rather than direct health care delivery, has increasingly drawn academic attention. Donabedian [20] addresses this issue by conceptualizing health care quality through the SPO framework. In health care, structure indicators refer to the system's resources and organizational arrangements, and process indicators include all service delivery activities such as diagnosis, treatment, nursing, and patient education. Finally, outcome indicators reflect the resulting health effects, including patient health improvements, satisfaction, and mortality rates. By applying the SPO framework in this manner, we shift the focus from assessing health care delivery quality to evaluating how PR system structures and processes influence user outcomes.

In recent times, we have also seen that PR in health care is demanded from the perspective of environmental, social, and governance (ESG) reporting, which involves public disclosure of an organization's performance and its impacts on environmental, social, and governance factors [21]. Currently, PR seems to be moving toward health care processes beyond basic clinical activity, such as PR on clinical trials [22]. In the era of the current artificial intelligence revolution, the impact of artificial intelligence on PR is, of course, an important topic [23,24].

Although defining and measuring quality in health care remains challenging, extensive evidence supports the essential role of the SPO framework in measuring health care quality [20,25]. Within the structural dimension, PR systems are an important component [25]. The health care system is depicted through structural indicators, which in turn influence process indicators, ultimately affecting outcome indicators [20]. PR has to inform its users of all these health care components. Previous research on process indicators in PR primarily focused on operational standardization, consistency, and service standard implementation [15,26]. However, few studies have examined the structural dimension from an individual-system interaction perspective.

At the process level, patient-physician collaboration (PPC) generates substantial user data [27]. During interactions between health care organizations and users, systems serve as supporting tools that can enhance treatment efficiency. The data generated in this process holds significant value for collection and use

XSL•FO
RenderX

[28]. The fine details of PPC, however, remain a difficult item to document in PR.

Existing research has paid insufficient attention to patient-generated data from the PPC process, particularly in the digital era, where the processes of data collection, organization, and use remain inadequate compared to the opportunities that digital data processing could offer. Such gaps may lead to unintended consequences, including users' mistrust of PR systems.

Finally, outcome indicators include subjective or functional metrics, such as in our sample, patient satisfaction and mortality rate [29,30]. Online PR websites enable individuals to interact directly with information, whereas individuals often rely on their own knowledge to evaluate content [17]. In contrast, in offline settings, user interaction is typically mediated by health care institutions, resulting in a more indirect experience [31].

Previous research indicates that attitudes formed through direct experience are more effective in predicting behavior than those formed through indirect experience [32]. However, studies examining interactions between individuals and system information in online environments in a PR context remain relatively limited. Taken together, these gaps highlight the need for a framework that integrates system information processing into the SPO dimensions of PR.

Sociotechnical Theory

According to Bostrom and Heinen [33], sociotechnical theory suggests that information systems should be designed so that the individual and technology operate as an integrated whole. In other words, the design of information systems should not only focus on data collection, processing, and transmission; the interaction between humans and information should also be emphasized. Within the context of PR, this principle is embodied through the interaction between PPC, user heterogeneity, and system-level demand for data reliability and standardization.

The operational goal of any PR system is to serve as a channel through which health care organizations and governments convey PR information to the public. This process enables individuals to filter health care information via the selection pathway and make health care choices suited to their needs [12]. PR information selection is both receptive and selective, and at the system stage, it primarily involves information processing.

In the context of PR, effective information processing is critical because it shapes how users understand and use performance data to make informed decisions. The information processing in PR systems involves a series of systematic steps that transform raw health data into actionable information for diverse stakeholders such as policymakers, providers, and the public [34]. On the PR website, data are symbols that represent the properties of objects and events [35].

Information consists of processed data, with the processing directed at increasing its usefulness [16]. System information processing involves integrating data from multiple sources into a central database, which is then retrieved and displayed on websites [36]. Regarding quality indicators, many individuals still experience significant analytical challenges, particularly when information is not clearly presented [37]. Thus, it is necessary for the PR system to reduce information asymmetries and empower patients, families, and payers to select high-quality providers [38].

These studies provide insights into the interaction between PR information and individuals, particularly regarding how end users use PR information. Previous research has primarily focused on issues of data inconsistency and low accuracy in PR [39]. However, previous studies have provided insufficient investigation into how data inconsistency issues can be effectively addressed within the field of information systems. Specifically, it is equally important to consider the objective quality of the data, incorporating data guidelines to address data inconsistency.

These findings underscore the critical role of data quality in shaping the adoption of PR systems. However, extant research demonstrates several notable limitations. First, the analytical depth remains constrained. Most studies focus primarily on information presentation as the primary mode of data delivery; however, they often overlook how data processing within systems fundamentally influences website presentation and user interpretation.

Second, investigations in the PR domain have predominantly adopted an awareness-to-behavior framework, emphasizing the role of awareness in guiding user actions [17]. Nevertheless, the complex interrelations between data quality and adoption remain underexplored, resulting in substantive gaps between PR systems and users' actual attitudes. Consequently, there is a pressing need to examine online PR systems through a lens that integrates the system and individuals.

Methods

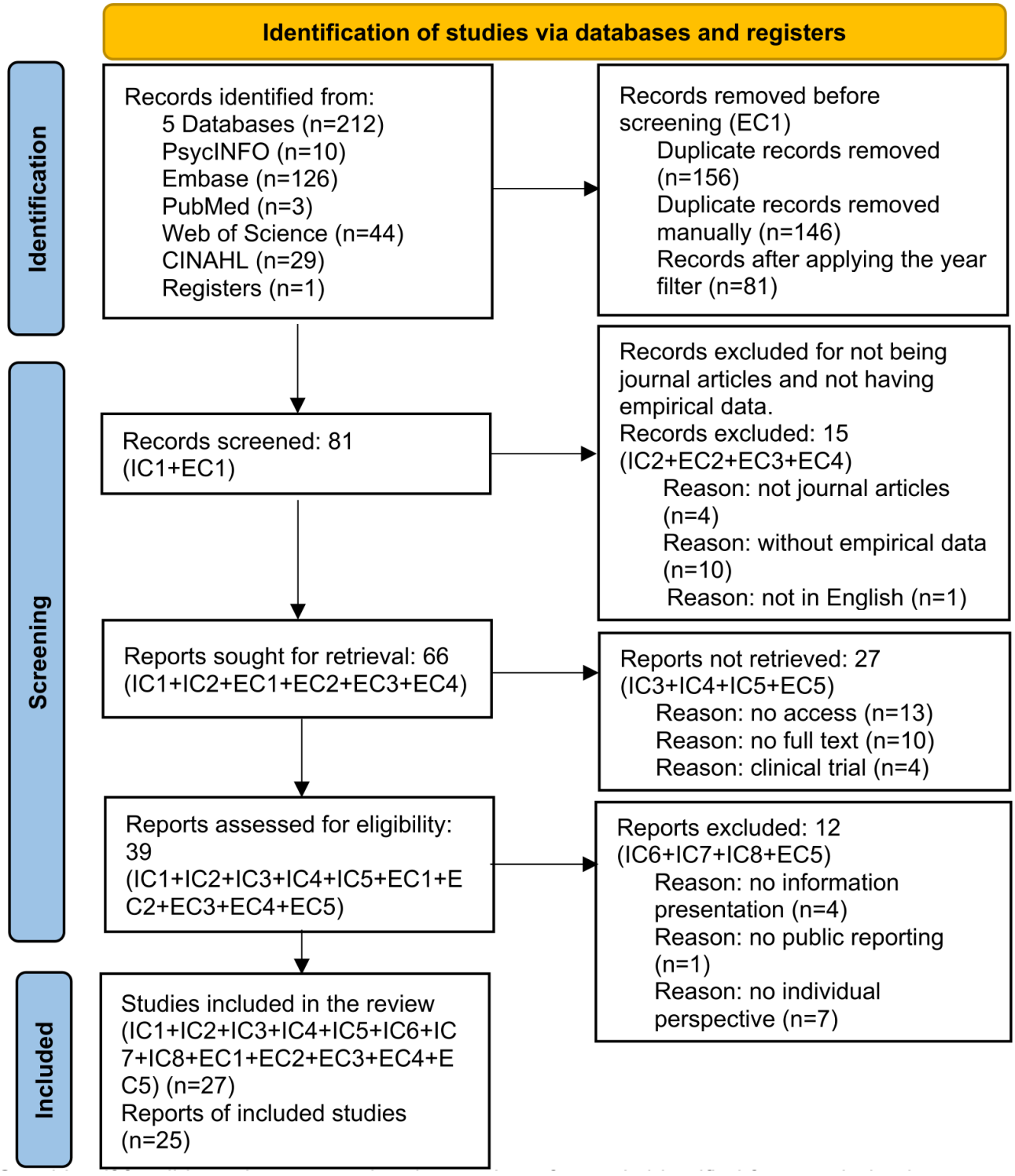
Overview

A systematic literature review requires clearly defined research questions, a transparent search process, data extraction, and data presentation [17,40]. Typical reviews cover a span of approximately 10 years. In this study, the time frame was set from 2016 to 2025. The rationale for selecting studies after 2015 is that, from this year onward, PR increasingly transitioned to internet-based platforms. The system reached its peak in July 2015, but then declined rapidly and has remained at a relatively low level since [41]. Additionally, the Hospital Consumer Assessment of Healthcare Providers and Systems star ratings system was launched, enhancing patients' visual understanding of hospital quality information [42].

We used the guidelines of Okoli and Schabram [43] and the framework proposed by Webster and Watson [44] to conduct our literature review. This review adopts a structured approach to the literature. As shown in Figure 1, the complete selection and synthesis process is summarized in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the selection process for public reporting systems retrieved from 5 databases between 2016 and 2025. EC: exclusion criteria; IC: inclusion criteria.



First, when designing the search strategy, efforts were made to capture all studies that met the inclusion criteria [45]. This research aims to explore the presentation of information in PR systems. Therefore, the search term was set to information; meanwhile, the research mainly focuses on PR, and the keyword includes "PR." The search term is provided in Multimedia Appendix 1, and the search date was September 14, 2025.

Second, PR has evolved as an interdisciplinary field, drawing from health policy and behavioral science. Conducting a review can be helpful to fill in the gap of data inconsistency and identify future directions from the information system perspective [46]. Access to these databases was obtained via Volter (Library of the University of Turku), ensuring that the user was logged in to minimize the number of ignored studies due to access issues. The data sources were chosen from Web of Science (multidisciplinary and broad literature search), CINAHL (health and public health information systems), Embase (health and public health information systems), PsycINFO (cognitive science, decision-making, and behavioral economics), and PubMed (biomedicine literature database).

XSL•FO
RenderX

Third, using tools to assess the risk of bias in studies is an important step described in systematic review protocols, as judgments made during this process may influence the analysis and interpretation of the review results [45]. Meanwhile, applying objective and reproducible methods helps ensure the appropriateness and fairness of the study findings [43]. The quality assessment protocol covers study methodology, level of structuring, study outcomes, and research topics. After the deduplication of studies (n=146) in EndNote (Clarivate) combined with manual checking, 81 studies were included in the final scoring pool.

Among the included studies, the methodologies involved both qualitative and quantitative approaches. The scoring table referenced from Hong et al [47] is provided in Multimedia Appendix 2.

Screening Process of the Systematic Review

In the screening process, 81 studies were initially included in the research pool. Review or descriptive studies that were not journal publications and did not contain empirical data were excluded. A total of 66 articles were retrieved, and full-text searches were conducted for these studies. During the full-text review, studies without accessible full text and clinical trial papers were excluded. Ultimately, 39 studies were included in the quality assessment.

Some of these studies examined PR systems from an organizational perspective. This study, however, focuses primarily on the individual perspective, specifically the alignment between individual users and the information presented in the PR system. Studies that explored system processes at the organizational level, which do not directly reflect the information presentation encountered by individual users, were excluded. Additionally, since this study focuses on PR systems in the health care domain, studies from non–health care fields were also excluded.

In Figure 1, abbreviated labels are used to represent inclusion criteria (IC) and exclusion criteria (EC).

The selection was conducted independently and double-checked manually. For clarity, the IC correspond to the following criteria: IC1 indicates studies published between 2016 and 2025, IC2 indicates empirical studies, IC3 indicates studies with open access, IC4 indicates studies with full text availability, IC5 indicates studies published after peer review, IC6 indicates studies focused on information presentation, IC7 indicates studies focused on the individual level, and IC8 indicates studies conducted in health care contexts. The EC corresponds to the following criteria: EC1 indicates studies published outside the 2016 to 2025 timeframe, EC2 indicates nonjournal studies, EC3 indicates nonempirical studies, EC4 indicates studies not published in English, and EC5 indicates clinical trials and other studies not focused on PR systems or individual-level information use.

After the initial full‐text screening, 27 studies were eligible for quality appraisal. The methodological quality of the included studies was assessed using the MMAT (2018 version) [47], a validated instrument specifically developed for appraising empirical studies in mixed-study systematic reviews. The MMAT evaluates 5 categories of study designs—qualitative research, randomized controlled trials, nonrandomized quantitative studies, quantitative descriptive studies, and mixed methods studies—each using 5 methodological criteria.

Following MMAT guidance, studies can be classified as low (1‐2 criteria met), moderate (3 criteria met), or high quality (4‐5 criteria met). All assessments were performed independently by the first and third authors, and discrepancies were resolved through discussion. Two studies with MMAT scores of 1/5 and 2/5 were excluded from the analysis, resulting in a final set of 25 studies for article synthesis.

Literature Synthesis

Thematic and Narrative Synthesis

We conducted a thematic synthesis of the included studies to extract key findings, which were then grouped into distinct themes [45]. Within each theme, dimensional indicators were identified and presented using an integration framework to address the first research question. This approach enables the qualitative integration of evidence across studies, highlighting patterns in structure indicators, process indicators, and outcome indicators.

Given the fragmented and broad scope of research in the field of PR, a narrative synthesis approach was adopted [48]. To enhance the transparency and rigor of the synthesis, descriptive statistics were also used to summarize the number of studies within each category and their methodological characteristics.

Regarding the research questions, the theme under the SPO framework is reported quantitatively, whereas subthemes (eg, website, dashboard, and data) are explored using a narrative approach. During synthesis, studies were grouped based on the specific dimensions of the SPO framework and their relevance to the research questions. Furthermore, a subset of studies investigated how structural indicators and process indicators influence the outcome indicators. Examining findings on the same aspects allows them to extend one another, which can be viewed as a form of confirmation [48].

SPO Framework–Based Results Mapping

Each subtheme in the “Results” section corresponds to a specific type of data extracted from the literature: (1) structure (website and dashboard): indicators describing the system design, presentation, and functionality were collected and summarized; (2) process (data quality and data process): studies reporting on data accuracy, completeness, accessibility, and workflow processes were analyzed; (3) outcome (patient satisfaction and surgical mortality): measures were extracted from studies evaluating the effect of PR systems on users’ decision-making, satisfaction, and clinical outcomes; and (4) user heterogeneity: information on demographic characteristics, health literacy, and contextual factors influencing user interaction with PR systems was also extracted and synthesized.

Quality Assurance and Integration

First, we conducted an initial synthesis using thematic analysis, presenting the study results in Multimedia Appendices 3 and 4. Subsequently, 1 author summarized and categorized the included

XSL•FO
RenderX

studies within an integrative framework. Based on this synthesis, an integrative framework was developed to organize the identified themes, which are further elaborated in the "Results" section.

Results

Overview

Among the 25 studies included in this review, 40% (10/25) explored PR from the perspective of information presentation formats, while 20% (5/25) of the studies focused on the dashboard in online PR. Additionally, 48% (12/25) of the studies examined PR systems from the perspective of data presentation, and 20% (5/25) addressed user heterogeneity. Finally, 12% (3/25) of the studies investigated PR from the perspective of patient satisfaction and surgical mortality.

Meanwhile, 80% (20/25) adopted quantitative methods, 16% (4/25) used qualitative designs, and 4% (1/25) used a mixed methods approach. Regarding data sources, 60% (15/25) collected secondary data (eg, web content analyses), while 40% focused on primary data through surveys, interviews, or focus groups. Data collection was conducted between February 2025 and September 2025.

Structure

Website

Among the included studies, 40% (10/25) of the studies examined PR from the perspective of information presentation formats on websites. For instance, studies found that the format of information presentation on PR websites can significantly influence public understanding and trust, policymaking, and pandemic response [49].

Specifically, the findings show that standardized data guidelines help improve the quality of PR websites, including accuracy, completeness, technical elements, design and aesthetics, readability, usability, and accessibility [50]. For example, challenges in information presentation may also result from mismatches between recommended reading levels for patient education materials (typically up to sixth grade) and the actual readability of websites, which seldom meet the needs of most patients [50]. Such mismatches create a gap between the health information users need and what is actually provided. Appropriate readability facilitates comprehension, enabling better disease management and adoption of healthy lifestyles [51].

Moreover, in terms of presentation formats, we found that online ratings may serve as a valuable source of information for both consumers and policymakers [52]. It is worth noting that storytelling techniques that convey narrative trends on dashboards can help users better understand the information presented in PR [53]. Additionally, supplementing this with characteristics of professional care institutions, which are often under-represented, can enhance the comprehensiveness of PR information [54].

Furthermore, the widespread use of aggregated ratings, such as star systems, may further strengthen consumer responses [55]. Interestingly, research has shown that "breadcrumb navigation" (forward-backward topic navigation) helps maintain the continuity of information, enabling users to quickly review previously read content and efficiently locate needed information [50]. Meanwhile, studies indicate that participants accessing "drill-down scores" tend to engage more deeply with the provided data (manifested as increased interactions on the website), but their impact on physician selection is lower than that of participants accessing "roll-up scores" [56].

Additionally, studies have found that over 90% of websites do not include accessibility features for individuals with disabilities, such as hearing, cognitive, neurological, motor, language, or visual impairments [51,57,58]. Therefore, considering accessibility benefits not only for disabled users but also the general population (eg, those using mobile devices, with slow internet, or limited bandwidth), it provides equal access to information for elderly individuals or residents in remote areas [51,57,58].

Dashboard

In our sample, a lot of discussion revolved around dashboards. We see them as a subcategory of web-based information, offering more dynamic and versatile analysis possibilities than static websites.

Overall, 20% (5/25) of studies examined the presentation of information on dashboards. Relevant and high-quality data are crucial for constructing dashboards, and research indicates that data sources are key factors in dashboard development [59,60]. Dashboards are especially useful in presenting epidemic and pandemic trends, such as those of COVID-19 [61], but can also be used in classical resource selection performed by PR users.

In terms of content presentation, dashboards should serve as tools that clearly link current trends with past policy decisions and individual behaviors, thereby enhancing the use of indicators [61]. However, evidence suggests that known and validated presentation techniques were overall underused, especially explanatory narratives [61]. To enhance usability, effective dashboards include: (1) understanding the audience and their information needs; (2) managing the type, quantity, and flow of displayed information; (3) clearly reporting data sources and methods; (4) linking time trends with policy decisions; (5) providing "close-to-home" data; (6) segmenting populations into relevant subgroups; and (7) using storytelling and visual cues [61].

Among these studies, the predominant approach involved data provision (specificity of geographic breakdowns, range of indicators reported, and explanations of data sources or calculations) and advancements enabled by the technologies employed (customization of time trends and interactive or visual chart elements) [53]. Meanwhile, studies also note that the clarity of data sources and methods has improved in dashboard development [53]. Teams differ in how they explore and define the purposes of dashboards. Some view dashboards as tools for presenting raw data to the public, while others attempt to provide explanations through narrative or visualization [59].

Moreover, 8% (2/25) of studies viewed the purpose of dashboards as solely presenting data (raw numbers) for the

XSL•FO
RenderX

public to interpret independently, whereas others sought to provide explanations through narrative or visual methods [53]. Additionally, incorporating storytelling elements and visual cues within dashboards has been shown to enhance users' ability to process information effectively [53]. For example, the development of dashboards that allow users to create profiles and receive personalized scores enables consumers to "see themselves in the data" [62].

Process

Data Quality

Overall, 48% (12/25) of studies examined PR from the perspective of website data. Notably, 32% (8/25) of studies indicated that standardized data guidelines help improve the quality of PR websites [15,45-47,58,59]. Additionally, some research described implementing check types within a data quality framework covering conformance, completeness, and plausibility, including both verification and validation [63].

Data Process

In the context of PPC, the generation of data encompasses multiple sources and activities. By using the data quality framework, data were collected across 4 domains, multiple measures, and time periods to examine access and equity; efficiency and sustainability; quality, safety, and patient orientation; and employee engagement [60].

Among these studies, 28% (7/25) investigated chart information on PR websites. The use of publicly reported data by consumers, policymakers, and researchers has grown substantially [64]. Most websites use charts, graphical elements, and quantitative descriptions as tools to convey differences between hospitals [65]. In addition, some research suggests adding charts to improve navigation [15]. For example, appropriate disease-related images and illustrations help patients better understand the information [50].

Meanwhile, the trend toward simplified composite rating systems may be reasonable to improve consumer responses to PR [55]. Previous research identified seven key elements that enhance user understanding [7]. The elements include unmarked bar charts, bar charts with simple symbols, bar charts with traffic light indicators, bar charts with thumbs-up symbols, provider rankings based on performance, clearly indicating whether higher values represent better performance, and implying quality levels through performance ranges without explicitly marking scale directions [7].

Furthermore, the transparent communication of standardized databases provides a crucial supplement to the missing components commonly observed in observational research [63]. This transparency can enhance the effectiveness of PR tools, as consumers may question the reliability of the quality information provided and consequently avoid using report cards [66].

Within the PPC process, presenting data in a structured format on dashboards can facilitate clear communication between health care providers and patients [67]. Meanwhile, PR websites can also serve as supplementary materials recommended by physicians, thereby reducing the need for face-to-face education and consultation [50].

Moderating Variable

User Heterogeneity

Overall, 20% (5/25) of the studies examined user heterogeneity, suggesting that themes such as simplicity, trust, collaboration, software and data integration, and adaptability have facilitated the development of PR systems [67]. However, the use of identical data across different websites may result in confusion or even contradictions regarding the same provider [15].

Moreover, research has highlighted that improvements in data timeliness, the level of detail available, and the categorization of comparable population groups can enhance the quality and effectiveness of PR [60]. For instance, individuals with higher health literacy tend to spend more time reviewing PR information than those with lower health literacy [68].

Individual Characteristics and Contextual Factors

Similarly, individual characteristics and contextual factors, such as age, political orientation, employment status, income level, and residential area, significantly influence users' adoption of PR information [38]. Younger adults are generally more likely than older adults to trust information from all sources [38]. In addition, political orientation, employment status, income level, and residential location are correlated with trust in medical quality information from companies, government agencies, or family and social network sources [38]. Collectively, these findings indicate that both individual characteristics and contextual factors are critical determinants of the perception of PR information.

Outcome

Patient Satisfaction

Overall, although patients generally prefer to obtain health information from medical professionals, the media remains an important source of information [69]. Within the PPC process, patient satisfaction primarily depends on the quality of care and service delivery. Specifically, patient satisfaction appears to depend largely on postoperative pain levels and access to timely appointments [70].

Surgical Mortality

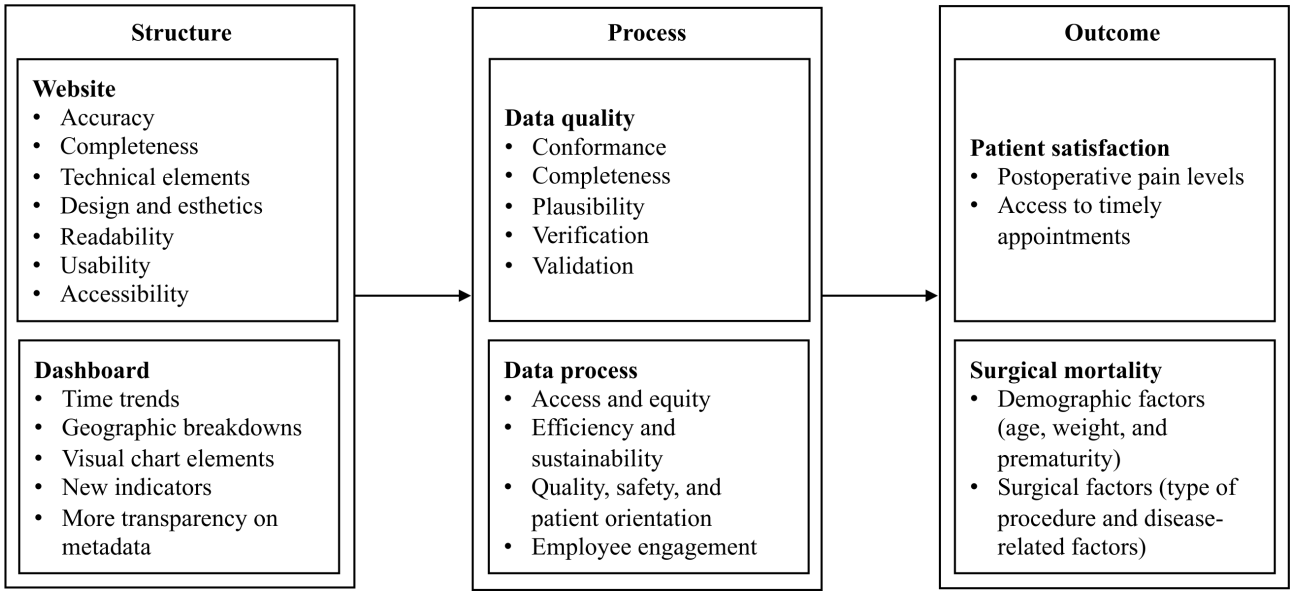
In addition, studies have pointed out that the variables influencing patient risk levels in surgical mortality include demographic factors such as age, weight, and prematurity, surgical factors such as the type of procedure, as well as disease-related factors [71].

A Framework to Understand the PR Quality Indicators in Health Care

The framework used in this study is based on the Donabedian quality model [20]. As shown in Figure 2, the system constitutes a key component of structural indicators [20]. User heterogeneity functions as a mediating mechanism linking the structural indicators and process indicators to outcome indicators. Specifically, the structural indicators and process indicators features of PR systems jointly shape users' perception of the information [38].

XSL•FO
RenderX

Figure 2. Integrative structure-process-outcome (SPO) framework for public reporting systems in health care.



The presentation format of PR websites refers to the channel through which users directly interact with the displayed content of the PR systems. Adherence to the website quality dimension and the dashboard quality dimension in the presentation format facilitates users' understanding of the system in a structured context.

Previous studies have indicated that data inconsistencies negatively affect the actual use of the system, thereby influencing users' adoption of the system [72]. The reviewed study indicates that the dashboards present data that have been processed and integrated within the underlying data architecture, resulting in higher accuracy and consistency. This, in turn, helps users use information that is more reliable and precise.

While structural indicators define the foundational quality of PR websites, process indicators determine how these systems function and generate reliable data. In terms of process indicators, data quality encompasses conformance, completeness, plausibility, verification, and validity. Data processes include accessibility and equity, efficiency and sustainability, quality, safety, patient-centered orientation, as well as employee engagement.

Building on the research gap regarding data inconsistency in existing PR, this study examines the dimensional indicators of data generated within process metrics. Standardizing data generated through PPC processes facilitates process-level data handling, thereby supporting the further use of the data. This aligns with prior findings from reviewed studies indicating that standardized data guidelines help users make better-informed health decisions [51].

User heterogeneity consists of individual characteristics and contextual design features [37]. Prior research suggests that dashboard design often fails to account for distinct user groups, leading to limited engagement [59]. Integrating contextual factors, health literacy levels, and individual characteristics into the design of PR systems allows for a more targeted presentation of information, thereby enhancing user comprehension and decision-making [37,59,68].

Previous research has shown that demographic variations influence information perception [73]. In line with this evidence, the reviewed studies identify multiple factors associated with users' interactions with PR systems, including age, gender, political orientation, employment status, and region of residence. Beyond demographic variations, health literacy and contextual design features jointly shape how users interpret and use PR data [37].

The reviewed studies have demonstrated that audiovisual and narrative information can positively influence satisfaction measures; contextual features such as storytelling techniques, interactivity, and visual chart elements enhance user understanding [37]. Health literacy plays a critical role in users' ability to interpret PR data. The reviewed studies found that individuals with higher health literacy are better able to comprehend PR information, which may explain their longer visual attention to such content [68]. This suggests that health literacy influences both the depth and quality of patient engagement in PPC processes. To address this, design features such as explanatory narratives [59], temporal trends and interactive functions [57], and appropriate visual chart elements can assist users with varying levels of health literacy in better understanding PR information [68].

This study further notes that within PR, patient satisfaction is primarily affected by postoperative pain levels (higher pain correlates with lower satisfaction) and access to timely appointments [70], consistent with previous findings linking health status to user satisfaction. Meanwhile, demographic factors (age, weight, and prematurity) and surgical factors (type of procedure and disease-related factors) influence surgical mortality [71].

By mapping the reviewed studies to the SPO framework, we illustrate that system data are no longer confined solely to structural indicators; they also play a role in process and outcome indicators. Accordingly, this study identifies website guidelines and dashboard guidelines under structural indicators, data quality and data processes under process indicators, and patient satisfaction and surgical mortality under outcome

XSL•FO
RenderX

indicators. Furthermore, the underlying database sources are displayed within dashboards, helping to address inconsistencies between databases across different systems.

Discussion

Principal Findings

Health technology assessment ensures that medical innovations genuinely improve patient outcomes rather than merely increasing the burden [74]. In this literature review, we synthesized the findings of 25 empirical studies by categorizing them under structural indicators, process indicators, and outcome indicators [20]. Our main findings indicate that structural indicators (website and dashboard guidelines) and process indicators (data quality and data processes) jointly influence user outcomes, such as satisfaction and surgical mortality, with user heterogeneity mediating these effects.

This study adopts a dual perspective of the individual and the system in selection pathways to systematically examine the mechanisms through which information presentation and user interaction operate within PR systems [19]. Existing literature indicates that the presentation of information within information systems not only supports patients in selecting health care providers and using report cards [68] but also influences the quality of user decision-making [55]. Consistent with prior research, the mode of information presentation, the level of aggregation, and report personalization significantly affect the extent of user engagement in decision-making [7,62]. However, this study finds that the operability and data-display characteristics of dashboards play a unique role in supporting individualized decisions [53,61] and that the supplementation of online information can enhance users' understanding of the data [51]. Additionally, the personal and contextual characteristics of health care information sources affect users' adoption in PR systems [37].

Regarding consistency, most studies emphasize the importance of PR dashboards for user decision-making [67], while user health literacy and individual characteristics modulate actual information use [68]. This study further highlights that the standardized design of data and websites can enhance decision-making effectiveness [60]. Unified dashboard guidelines, along with aggregation and search functionalities, can reduce information conflicts and increase user comprehension and trust [59,68].

In terms of heterogeneity, studies have examined the impact of different information types (eg, online reviews, report cards, and dashboards during the pandemic) and diverse user groups (classified by age and health status) on decision-making [54]. Specifically, users differ in their focus on data operability versus information sources [37,61]. Common challenges in PR include missing information, delayed updates, complex formats, lack of standardization, and limited accessibility for users with low health literacy [51,68].

Regarding the user selection pathway, the system is considered a critical avenue for decision-making. Dashboards can mitigate existing issues in PR, such as missing data, delayed updates, and information confusion. As Donabedian [20] noted, while quality indicators are difficult to measure, their establishment contributes to improving actual health care quality. Standardized data guidelines and structured information presentation can support more effective decisions within structural and process indicators, while reducing information asymmetry and enhancing the quality of health services [55].

In summary, by referencing different PR scenarios and information types, this study proposes practical system optimization recommendations and categorizes differences among user groups as user-consciousness heterogeneity, which serves as a moderating variable. It emphasizes the role of standardized guidelines in facilitating the flow of information.

Future Research Agenda

Overview

In the era of digital health, the adoption of PR has evolved. From the perspective of availability, it is possible to examine the innovativeness of technologies in shaping users' perceptions of what they can accomplish [75]. Information systems, as a critical component of structural indicators [20], also play a significant role in enhancing patient awareness [15]. Based on the classification of diverse user groups and the multiple roles involved in the PPC process, incorporating dashboards and other visualization tools into system design can facilitate the adoption of PR information.

Based on the analysis of the studies included in this review, the public's understanding and use of PR remain at a preliminary stage across different population groups [76,77]. To advance the practical use of PR information, future studies should adopt diverse methodological and theoretical approaches to explore how enhancing awareness can foster trust between health care organizations and various demographic groups. Research on this topic should be further expanded along four key dimensions: conceptual understanding, thematic exploration, methodological innovation, and technological development.

Conceptual Agenda

User attitudes are fundamental determinants of behavioral intentions and actual behaviors [17]. Underlying these attitudes are user cognitions, particularly awareness and understanding, which align with levels 1 and 2 of the PAM. While existing research highlights the role of health literacy in PR and its connection to PAM, little attention has been paid to its functional and interactive dimensions [78].

Although system design and information presentation can foster awareness and literacy, the conceptual link between the two remains underexplored. From a systems perspective, perception marks the initial awareness stage; subsequent information processing drives decision-making and system adoption. Information framing, as shown in media studies, shapes comprehension and interpretation [79]. Empirical evidence shows that most patients remain at level 1 of PAM [10]. Enhancing health literacy may help users progress from basic awareness to initial engagement.

XSL•FO
RenderX

Thematic Agenda

This study notes that participants with more severe health conditions tend to seek care more frequently than healthier individuals, making the process of choosing a physician particularly important for them. Notably, significant differences in decision quality are observed among these participants [56]. Previous research on the design of PR websites has largely focused on conditions such as hypertension, diabetes, and infectious diseases [15]. The future PR system should pay greater attention to the needs of patients with complex conditions, such as those with multiple comorbidities.

Moreover, this study points out that contextual factors, health literacy, and individual characteristics affect an individual's adoption. Future research should explore how individuals' emotional states and literacy capacities interact with and interpret PR information. For instance, Shaller [80] proposed a context-based consumer choice framework that emphasizes the integration of users' decision contexts, emotional states, and literacy capacities.

Methodological Agenda

Primary data collection was primarily used in surveys, experiments, and semistructured interviews, whereas qualitative approaches, such as interviews, focus groups, and thematic content analysis, remained less common. Mixed method studies were rare, typically combining expert evaluation with descriptive and thematic analyses.

Methodologically combining natural language processing, causal inference, and experimental designs could advance the understanding of how system-level indicators interact with user heterogeneity, health literacy, and decision-making behaviors in the PR system.

Technological Agenda

In health care organizations, the PPC extends beyond physicians and patients to include family members, friends, caregivers, and other stakeholders. Future system designs that integrate and accommodate these multiple roles could enhance communication between actual users and the system.

Previous studies have indicated that patients with poorer health conditions exhibit more pronounced differences in decision-making [56]. Population segmentation can be categorized into low-health-demand, multiple comorbidities, medically complex, and frail groups [77]. Future research could explore the heterogeneity of responses across these groups in relation to information presentation. For example, web design could incorporate switchable versions to accommodate the health information needs of patients with different health conditions.

Limitations

First, this study retrieved literature through keyword searches within selected databases. This approach may have limited the scope of study selection and potentially excluded relevant studies from other databases.

Second, this study explored PR systems from the perspective of individuals in health care. Even though employees act as individual users of the PR system, broadening the search term to include employees of health care organizations in their dual roles can bring new insights. It can be helpful to explore the interaction between employees in the organization and individuals within the sociotechnical theory. Additionally, given that qualitative research methods are interpretive, potential subjective biases may exist when evaluating study findings.

Conclusions

This study provides the following contributions to the existing literature. First, we synthesized within the SPO framework. By systematically analyzing current research on the use of PR information presentation, this study provides an overview of the website and dashboard dimensions in the structure indicators, data quality and data process in the process indicators, and patient satisfaction and surgical mortality in the outcome indicators. Moreover, this study contributes to the literature by identifying contextual factors, such as health literacy, individual characteristics, political affiliation, employment status, income, and area of residence that influence users' perceptions.

Second, this study builds a connection between data appearing during PPC and the PR systems in the structure quality dimension. Finally, this research outlines future research directions from conceptual, methodological, thematic, and information system perspectives. These insights provide valuable guidance for advancing research in this field, bridging existing gaps, and improving the overall quality of academic inquiry.

Acknowledgments

The authors thank Professor Hongxiu Li for her guidance in conceptualizing the study and providing comments on the manuscript. We also acknowledge the librarians at the University of Turku for their assistance with refining the search strategies. Finally, the authors are grateful to the anonymous reviewers for their comments on the conceptual framing and comments regarding the search strategy.

Funding

LD was financially supported by the Chinese Scholarship Council (grant 202306770017). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript writing.

XSL•FO
RenderX

Data Availability

All extracted data, coding notes, and synthesis outputs generated in this study are publicly available in the Open Science Framework repository at [81]. This includes structured data files (ciw/ris/nbib). Access is unrestricted. Researchers can download and use these materials for noncommercial purposes. During the preparation of this manuscript, generative artificial intelligence (AI; ChatGPT) was used solely for translation and language polishing purposes. All outputs generated by the AI were carefully reviewed and verified by the authors to ensure accuracy and appropriateness. The AI was not involved in study design, data collection, analysis, interpretation, or in writing the scientific content of the manuscript.

Authors' Contributions

Conceptualization: LD, RS
Data curation: LD
Formal analysis: LD, ED
Investigation: LD, ED
Methodology: LD
Resources: RS
Supervision: RS
Validation: LD, ED
Visualization: LD
Writing–original draft: LD
Writing–review & editing: LD, ED

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies in each database.
[DOCX File, 17 KB - medinform_v14i1e80435_app1.docx]

Multimedia Appendix 2
Mixed Methods Appraisal Tool (MMAT) quality assessment report.
[DOCX File, 54 KB - medinform_v14i1e80435_app2.docx]

Multimedia Appendix 3
Key characteristics of the included studies.
[DOCX File, 50 KB - medinform_v14i1e80435_app3.docx]

Multimedia Appendix 4
Details of the included studies.
[DOCX File, 40 KB - medinform_v14i1e80435_app4.docx]

Checklist 1
PRISMA-P checklist.
[DOCX File, 27 KB - medinform_v14i1e80435_app5.docx]

References

1. Evans L, Evans J, Pagliari C, Källander K. Scoping review: exploring the equity impact of current digital health design practices. Oxf Open Digit Health 2023 Jan 1;1(8):qad006. [doi: 10.1093/oodh/oqad006]
2. All the healthcare marketing statistics to pay close attention to in 2023. Digitalis Medical. 2021. URL: https://digitalismedical.com/blog/healthcare-marketing-statistics/ [accessed 2026-03-13]
3. Salampessy BH, Portrait FRM, van der Hijden E, Klink A, Koolman X. On the correlation between outcome indicators and the structure and process indicators used to proxy them in public health care reporting. Eur J Health Econ 2021 Nov;22(8):1239-1251. [doi: 10.1007/s10198-021-01333-w] [Medline: 34191196]
4. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012 May;27(5):520-526. [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]

XSL•FO
RenderX

5. Knopik C. Evaluating data integrity and reporting challenges in public health: lessons from COVID-19 data collection in Washington state. : Science, Public Health Policy and the Law; 2024 URL: https://publichealthpolicyjournal.com/wp-content/uploads/2024/10/Knopik_SciencePublicHealthPolicyAndTheLaw_v5.2019-2024.Oct_2024.pdf [accessed 2026-02-27]
6. Kurtzman ET, Greene J. Effective presentation of health care performance information for consumer decision making: a systematic review. Patient Educ Couns 2016 Jan;99(1):36-43. [doi: 10.1016/j.pec.2015.07.030] [Medline: 26277826]
7. Sander U, Emmert M, Dickel J, Meszmer N, Kolb B. Information presentation features and comprehensibility of hospital report cards: design analysis and online survey among users. J Med Internet Res 2015 Mar 16;17(3):e68. [doi: 10.2196/jmir.3414] [Medline: 25782186]
8. Alemu W, Girma E, Mulugeta T. Patient awareness and role in attaining healthcare quality: a qualitative, exploratory study. Int J Afr Nurs Sci 2021;14:100278. [doi: 10.1016/j.ijans.2021.100278]
9. Janamian T, Greco M, Cosgriff D, Baker L, Dawda P. Activating people to partner in health and self-care: use of the patient activation measure. Med J Aust 2022 Jun 6;216 Suppl 10(Suppl 10):S5-S8. [doi: 10.5694/mja2.51535] [Medline: 35665937]
10. Chen M, Zhang X, Liu C, Chen H, Wang D, Liu C. Effects of public reporting of prescription indicators on patient choices: evidence from propensity scores matching. Front Pharmacol 2023 Oct 9;14:1110653. [doi: 10.3389/fphar.2023.1110653]
11. Chimonas S, Fortier E, Li DG, Lipitz-Snyderman A. Facts and fears in public reporting: patients' information needs and priorities when selecting a hospital for cancer care. Med Decis Making 2019 Aug;39(6):632-641. [doi: 10.1177/0272989X19855050] [Medline: 31226909]
12. Cacace M, Geraedt M. Public reporting as a quality strategy. Eur J Public Health 2017 Nov 1;27(suppl_3):ckx187-ckx751. [doi: 10.1093/eurpub/ckx187.751]
13. Schlesinger M, Grob R, Shaller D. Using patient‐reported information to improve clinical practice. Health Serv Res 2015 Dec;50(Suppl 2):2116-2154. [doi: 10.1111/1475-6773.12420] [Medline: 26573890]
14. Greene SM, Tuzzio L, Cherkin D. A framework for making patient-centered care front and center. Perm J 2012 Sep;16(3):49-53. [doi: 10.7812/TPP/12-025]
15. Sapin M, Ehlig D, Geissler A, Vogel J. Public reporting in five health care areas: a comparative content analysis across nine countries. Health Policy 2025 Feb;152:105222. [doi: 10.1016/j.healthpol.2024.105222] [Medline: 39700634]
16. Ackoff R. Ackoff's Best: His Classic Writings on Management: Wiley; 1999.
17. Cacace M, Geraedts M, Berger E. Public reporting as a quality strategy. In: Reinhard B, Niek K, Dimitra P, Wilm Q, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies: WHO Regional Office for Europe (with OECD); 2019:331-355.
18. Totten AM, Wagner J, Tiwari A, O'Haire C, Griffin J, Walker M. Closing the quality gap: revisiting the state of the science (vol. 5: public reporting as a quality improvement strategy). Evid Rep Technol Assess (Full Rep) 2012 Jul(208.5):1-645. [Medline: 24422977]
19. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. Med Care 2003 Jan;41(1 Suppl):I30-I38. [doi: 10.1097/00005650-200301001-00004] [Medline: 12544814]
20. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q 1966 Jul;44(3):166-206. [Medline: 5338568]
21. Pratici L, Salvatore FP, Fanelli S, Zangrandi A, Milone M. Using ESG paradigm as a basis for social reporting in nonprofit organizations: evidence from cases in healthcare. Manag Decis 2025 Feb 13;63(2):586-609. [doi: 10.1108/MD-10-2023-2012]
22. Chan AW, Karam G, Pymento J, et al. Reporting summary results in clinical trial registries: updated guidance from WHO. Lancet Glob Health 2025 Apr;13(4):e759-e768. [doi: 10.1016/S2214-109X(24)00514-X] [Medline: 40155113]
23. Agha RA, Mathew G, Rashid R, et al. Revised Strengthening the Reporting of Cohort, Cross-Sectional and Case-Control Studies in Surgery (STROCSS) guideline: an update for the age of artificial intelligence. Prem J Sci 2025. [doi: 10.70389/PJS.100081]
24. Non LR, Marra AR, Ince D. Rise of the machines - artificial intelligence in healthcare epidemiology. Curr Infect Dis Rep 2025 Dec;27(1):4. [doi: 10.1007/s11908-024-00854-8]
25. Donabedian A. The quality of care. How can it be assessed? JAMA 1988;260(12):1743-1748. [doi: 10.1001/jama.260.12.1743] [Medline: 3045356]
26. Quality improvement and monitoring at your fingertips. Agency for Healthcare Research and Quality. 2016. URL: https://qualityindicators.ahrq.gov/ [accessed 2026-03-05]
27. Staehelin D, Dolata M, Stöckli L, Schwabe G. How patient-generated data enhance patient-provider communication in chronic care: field study in design science research. JMIR Med Inform 2024 Sep 10;12(1):e57406. [doi: 10.2196/57406] [Medline: 39255481]
28. Fiske A, Degelsegger-Márquez A, Marsteurer B, Prainsack B. Value-creation in the health data domain: a typology of what health data help us do. Biosocieties 2022 Apr 12:1-25. [doi: 10.1057/s41292-022-00276-6] [Medline: 35432575]
29. Tang Y, Liu C, Zhang X. Performance associated effect variations of public reporting in promoting antibiotic prescribing practice: a cluster randomized-controlled trial in primary healthcare settings. Prim Health Care Res Dev 2017 Sep;18(5):482-491. [doi: 10.1017/S1463423617000329] [Medline: 28606190]
30. Yang L, Liu C, Wang L, Yin X, Zhang X. Public reporting improves antibiotic prescribing for upper respiratory tract infections in primary care: a matched-pair cluster-randomized trial in China. Health Res Policy Syst 2014 Oct 10;12:61. [doi: 10.1186/1478-4505-12-61] [Medline: 25304996]

XSL•FO
RenderX

31. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021 Sep 3;20(1):158. [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
32. Fazio RH, Zanna MP. Direct experience and attitude-behavior consistency. Adv Exp Soc Psychol 1981;14:161-202. [doi: 10.1016/S0065-2601(08)60372-X]
33. Bostrom RP, Heinen JS. MIS problems and failures: a socio-technical perspective. MIS Q 1977 Sep 1;1(3):17-32. [doi: 10.2307/248710]
34. Prang KH, Maritz R, Sabanovic H, Dunt D, Kelaher M. Mechanisms and impact of public reporting on physicians and hospitals' performance: a systematic review (2000-2020). PLoS ONE 2021;16(2):e0247297. [doi: 10.1371/journal.pone.0247297] [Medline: 33626055]
35. Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956 Mar;63(2):81-97. [Medline: 13310704]
36. Sivarajkumar S, Mohammad HA, Oniani D, et al. Clinical information retrieval: a literature review. J Healthc Inform Res 2024 Jun;8(2):313-352. [doi: 10.1007/s41666-024-00159-4] [Medline: 38681755]
37. Moore AR, Hudson C, Amey F, Chumbler N. Trusting sources of information on quality of physician care. Inquiry 2020;57:46958020952912. [doi: 10.1177/0046958020952912] [Medline: 32830580]
38. Pross C, Averdunk LH, Stjepanovic J, Busse R, Geissler A. Health care public reporting utilization - user clusters, web trails, and usage barriers on Germany's public reporting portal Weisse-Liste.de. BMC Med Inform Decis Mak 2017 Apr 21;17(1):48. [doi: 10.1186/s12911-017-0440-6] [Medline: 28431546]
39. Zhang Y, Callaghan-Koru JA, Koru G. The challenges and opportunities of continuous data quality improvement for healthcare administration data. JAMIA Open 2024 Oct;7(3):ooae058. [doi: 10.1093/jamiaopen/ooae058] [Medline: 39091510]
40. Kitchenham B, Pearl Brereton O, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering—a systematic literature review. Inf Softw Technol 2009 Jan;51(1):7-15. [doi: 10.1016/j.infsof.2008.09.009]
41. Li J. Public reporting and consumer demand in the home health sector. Med Care Res Rev 2024 Apr;81(2):122-132. [doi: 10.1177/10775587231221852] [Medline: 38235538]
42. Beckett MK, Quigley DD, Cohea CW, et al. Trends in HCAHPS survey scores, 2008–2019: a quality improvement perspective. Med Care 2024 Jun 1;62(6):416-422. [doi: 10.1097/MLR.0000000000002001] [Medline: 38728680]
43. Okoli C, Schabram K. A guide to conducting a systematic literature review of information systems research. SSRN Electron J 2010. [doi: 10.2139/ssrn.1954824]
44. Webster J, Watson RT. Analyzing the past to prepare for the future: writing a literature review. MIS Q 2002 Jun 1;26(2):xiii-xxiii. [doi: 10.2307/4132319]
45. Higgins JPT, Thomas J, Chandler J, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions: Wiley; 2019. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119536604 [accessed 2026-02-27] [doi: 10.1002/9781119536604]
46. Jasperson J(, Carte TA, Saunders CS, et al. Review: power and information technology research: a metatriangulation review. MIS Q 2002 Dec 1;26(4):397-459. [doi: 10.2307/4132315]
47. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Nov 1;34(4):285-291. [doi: 10.3233/EFI-180221]
48. Sandelowski M, Voils CI, Barroso J. Defining and designing mixed research synthesis studies. Res Sch 2006;13(1):29. [Medline: 20098638]
49. Hennessee I, Clennon JA, Waller LA, Kitron U, Bryan JM. Considerations for improving reporting and analysis of date-based COVID-19 surveillance data by public health agencies. Am J Public Health 2021 Dec;111(12):2127-2132. [doi: 10.2105/AJPH.2021.306520] [Medline: 34878867]
50. Siddhanamatha HR, Heung E, de los Angeles Lopez-Olivo M, et al. Quality assessment of websites providing educational content for patients with rheumatoid arthritis. Semin Arthritis Rheum 2017 Jun;46(6):715-723. [doi: 10.1016/j.semarthrit.2017.01.006] [Medline: 28258768]
51. Lopez-Olivo MA, des Bordes JKA, Syed MN, et al. Quality appraisal of educational websites about osteoporosis and bone health. Arch Osteoporos 2021 Feb 10;16(1):28. [doi: 10.1007/s11657-021-00877-x] [Medline: 33566216]
52. Temkin-Greener H, Mao Y, McGarry B. Online customer reviews of assisted living communities: association with community, county, and state factors. J Am Med Dir Assoc 2023 Jun;24(6):841-845.e3. [doi: 10.1016/j.jamda.2023.02.007] [Medline: 36934775]
53. Barbazza E, Ivanković D, Wang S, et al. Exploring changes to the actionability of COVID-19 dashboards over the course of 2020 in the Canadian context: descriptive assessment and expert appraisal study. J Med Internet Res 2021 Aug 6;23(8):e30200. [doi: 10.2196/30200] [Medline: 34280120]
54. Ryskina KL, Andy AU, Manges KA, Foley KA, Werner RM, Merchant RM. Association of online consumer reviews of skilled nursing facilities with patient rehospitalization rates. JAMA Netw Open 2020 May 1;3(5):e204682. [doi: 10.1001/jamanetworkopen.2020.4682] [Medline: 32407501]
55. Perraillon MC, Konetzka RT, He D, Werner RM. Consumer response to composite ratings of nursing home quality. Am J Health Econ 2019;5(2):165-190. [doi: 10.1162/ajhe_a_00115] [Medline: 31579236]

XSL•FO
RenderX

56. Cerully JL, Parker AM, Rybowski L, et al. Improving patients' choice of clinician by including roll-up measures in public healthcare quality reports: an online experiment. J Gen Intern Med 2019 Feb;34(2):243-249. [doi: 10.1007/s11606-018-4725-y] [Medline: 30446909]
57. Reid RO, Deb P, Howell BL, Conway PH, Shrank WH. The roles of cost and quality information in Medicare Advantage plan enrollment decisions: an observational study. J Gen Intern Med 2016 Feb;31(2):234-241. [doi: 10.1007/s11606-015-3467-3] [Medline: 26282952]
58. Emmert M, Wiener M. What factors determine the intention to use hospital report cards? The perspectives of users and non-users. Patient Educ Couns 2017 Jul;100(7):1394-1401. [doi: 10.1016/j.pec.2017.01.021] [Medline: 28189471]
59. Barbazza E, Ivanković D, Davtyan K, et al. The experiences of 33 national COVID-19 dashboard teams during the first year of the pandemic in the World Health Organization European Region: a qualitative study. Digit Health 2022;8:20552076221121154. [doi: 10.1177/20552076221121154] [Medline: 36060614]
60. Lloyd S, Cliff C, FitzGerald G, Collie J. Can publicly reported data be used to understand performance in an Australian rural hospital? Health Inf Manag 2021;50(1-2):35-46. [doi: 10.1177/1833358320948559] [Medline: 32935590]
61. Ivanković D, Barbazza E, Bos V, et al. Features constituting actionable COVID-19 dashboards: descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. J Med Internet Res 2021 Feb 24;23(2):e25682. [doi: 10.2196/25682] [Medline: 33577467]
62. Smith MA, Bednarz L, Nordby PA, et al. Increasing consumer engagement by tailoring a public reporting website on the quality of diabetes care: a qualitative study. J Med Internet Res 2016 Dec 21;18(12):e332. [doi: 10.2196/jmir.6555] [Medline: 28003173]
63. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021 Sep 18;28(10):2251-2257. [doi: 10.1093/jamia/ocab132] [Medline: 34313749]
64. Hsu SH, Hung P, Wang SY. Factors associated with hospices' nonparticipation in medicare's hospice compare public reporting program. Med Care 2019 Jan;57(1):28-35. [doi: 10.1097/MLR.0000000000001016] [Medline: 30489545]
65. Downing NS, Cloninger A, Venkatesh AK, et al. Describing the performance of U.S. hospitals by applying big data analytics. PLoS One 2017;12(6):e0179603. [doi: 10.1371/journal.pone.0179603] [Medline: 28662045]
66. Kast K, Otten SM, Konopik J, Maier CB. Web-based public reporting as a decision-making tool for consumers of long-term care in the United States and the United Kingdom: systematic analysis of report cards. JMIR Form Res 2023 Dec 14;7:e44382. [doi: 10.2196/44382] [Medline: 38096004]
67. Gephart SM, Tolentino DA, Quinn MC, Wyles C. Neonatal intensive care workflow analysis informing NEC-Zero clinical decision support design. Comput Inform Nurs 2023 Feb 1;41(2):94-101. [doi: 10.1097/CIN.0000000000000929] [Medline: 35470306]
68. Yılmaz NG, Timmermans DR, Van Weert JC, Damman OC. Breast cancer patients' visual attention to information in hospital report cards: an eye-tracking study on differences between younger and older female patients. Health Informatics J 2023;29(1):14604582231155279. [doi: 10.1177/14604582231155279] [Medline: 36840473]
69. Walsh-Childers K, Braddock J, Rabaza C, Schwitzer G. One step forward, one step back: changes in news coverage of medical interventions. Health Commun 2018 Feb;33(2):174-187. [doi: 10.1080/10410236.2016.1250706] [Medline: 27983868]
70. Shah NS, Umeda Y, Newyear B, et al. Patients with higher postoperative pain after ambulatory shoulder surgery reported lower satisfaction: a prospective observational study. AME Surg J 2022 Dec 1;2:32-32. [doi: 10.21037/asj-22-11]
71. Jacobs JP. The society of thoracic surgeons congenital heart surgery database public reporting initiative. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 2017 Jan;20:43-48. [doi: 10.1053/j.pcsu.2016.09.008] [Medline: 28007064]
72. Rambur B, Vallett C, Cohen JA, Tarule JM. Metric-driven harm: an exploration of unintended consequences of performance measurement. Appl Nurs Res 2013 Nov;26(4):269-272. [doi: 10.1016/j.apnr.2013.09.001] [Medline: 24206760]
73. Wong YK, Low KL, Pooke TG. Factors associated with dimensions of patients' trust in chiropractic doctors in the international medical university healthcare chiropractic center: an exploratory study. J Chiropr Med 2022 Jun;21(2):83-96. [doi: 10.1016/j.jcm.2022.02.003] [Medline: 35774631]
74. Facey K, Henshall C, Sampietro-Colom L, Thomas S. Improving the effectiveness and efficiency of evidence production for health technology assessment. Int J Technol Assess Health Care 2015 Jan;31(4):201-206. [doi: 10.1017/S0266462315000355] [Medline: 26646858]
75. Zheng H, Xia W, Ma J. Affordance types and their determinants in patient mobile health portals use: an empirical investigation. J Med Internet Res 2025 Aug 27;27(2025):e74351. [doi: 10.2196/74351] [Medline: 40865120]
76. Kumpunen S, Trigg LJ, Rodrigues R. Public reporting in health and long-term care to facilitate provider choice. : World Health Organization; 2014 URL: https://iris.who.int/server/api/core/bitstreams/ca6534be-7003-4d72-892a-45023ddbbf01/content [accessed 2026-02-27]
77. Langton JM, Wong ST, Burge F, et al. Population segments as a tool for health care performance reporting: an exploratory study in the Canadian province of British Columbia. BMC Fam Pract 2020 May 31;21(1):98. [doi: 10.1186/s12875-020-01141-w] [Medline: 32475339]
78. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000 Sep 1;15(3):259-267. [doi: 10.1093/heapro/15.3.259]

XSL•FO
RenderX

79. de Vreese CH, Boomgaarden HG, Semetko HA. (In)direct framing effects: the effects of news media framing on public support for Turkish membership in the European Union. Communic Res 2011 Apr;38(2):179-205. [doi: 10.1177/0093650210384934]
80. Shaller D, Kanouse DE, Schlesinger M. Context-based strategies for engaging consumers with public reports about health care providers. Med Care Res Rev 2014 Oct;71(5 Suppl):17S-37S. [doi: 10.1177/1077558713493118] [Medline: 23819945]
81. Understanding user behavior toward public reporting in digital health: a systematic review. OSF. URL: https://osf.io/62bdk/overview [accessed 2026-03-13]

Abbreviations

EC: exclusion criteria
ESG: environmental, social, and governance
IC: inclusion criteria
MMAT: Mixed Methods Appraisal Tool
PAM: patient activation measure
PPC: patient-physician collaboration
PR: public reporting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analyses Protocols
SPO: structure-process-outcome

Edited by A Benis; submitted 10.Jul.2025; peer-reviewed by H Maheshwari, Y Oyende, Z Guo; revised version received 09.Feb.2026; accepted 17.Feb.2026; published 18.Mar.2026.

Please cite as:
Du L, Suomi R, Damalie E
Digital Public Reporting Systems for Evaluating Health Care Quality: Systematic Review
JMIR Med Inform 2026;14:e80435
URL: https://medinform.jmir.org/2026/1/e80435
doi:10.2196/80435

© Lanmei Du, Reima Suomi, Elorm Damalie. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning for Predicting Venous Thromboembolism After Joint Arthroplasty: Systematic Review of Clinical Applicability and Model Performance

Junwei Ma[1], BSc; Huifeng Tang[2], MSN; Yunshan Zhang[2], BSc; Xuemei Yi[2], BSc; Tangsheng Zhong[3], DN; Xinyun Li[1], BSc; Gang Wang[2], MSN

[1]School of Nursing, Jilin University, Changchun, China

[2]The Second Operation Room, First Hospital of Jilin University, No. 71 Xinmin Street, Chaoyang District, Changchun, China

[3]Nursing Department, First Hospital of Jilin University, Changchun, China

Corresponding Author:
Gang Wang, MSN
The Second Operation Room, First Hospital of Jilin University, No. 71 Xinmin Street, Chaoyang District, Changchun, China

Abstract

Background: There is increasing research on machine learning in predicting venous thromboembolism after joint arthroplasty, but the quality and clinical applicability of these models remain uncertain.

Objective: This systematic review aims to evaluate the predictive performance and methodological quality of machine learning models for venous thromboembolism risk after joint replacement surgery.

Methods: Web of Science, Embase, Scopus, CNKI, Wanfang, Vipro, and PubMed were searched until December 15, 2024. The risk of bias and applicability were evaluated using the PROBAST (Prediction Model Risk of Bias Assessment Tool) checklist. A qualitative comprehensive analysis was conducted to extract and describe the data related to the model's characteristics and performance.

Results: This review encompassed 34 prediction models from 9 studies. The most frequently used machine learning models were extreme gradient boosting and logistic regression. The results showed that all studies had significant heterogeneity and high risk of bias. Although some models reported nearly flawless area under the curve (>0.9), they lacked external validation and may have overfitted. The models tested on large external datasets demonstrated more conservative performance.

Conclusions: The predictive performance of machine learning models varied greatly. Although the reported area under the curve values indicated that some models have good discriminative ability, this performance varied greatly and was inconsistent among the included studies. These models have a high risk of bias, and it is necessary to take this into account when they are used in clinical practice. Future studies should adopt a prospective study design, ensure appropriate data handling, and use external validation to improve model robustness and applicability.

Trial Registration: PROSPERO CRD42024625842; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024625842

(JMIR Med Inform 2026;14:e79886) doi:10.2196/79886

KEYWORDS

joint arthroplasty; venous thromboembolism; machine learning; meta-analysis; systematic review

Introduction

Joint replacement surgery is one of the main treatment methods for severe joint injuries, which can alleviate patients' pain and improve joint function [1,2]. In recent years, due to the aging population and the advancement of medical technology, more and more patients have undergone joint replacement surgery. But surgical complications are a big risk that can change how well joint replacement surgery works. Venous thromboembolism (VTE), including deep vein thrombosis (DVT) and pulmonary embolism (PE), is a serious postoperative complication that significantly affects the prognosis of patients [3,4].

Studies have shown that for patients who did not receive preventive medication before undergoing major orthopedic surgeries, such as total hip arthroplasty (THA) or total knee arthroplasty (TKA), the incidence of postoperative VTE can be as high as 26.6% to 60.8% [5]. Asymptomatic VTE may not present any obvious clinical symptoms but can be detected through ultrasound examination or other diagnostic techniques. Some patients with VTE may experience discomfort and swelling in the affected limb, and their functional condition may be poor, which may even affect their postoperative recovery [6,7]. Furthermore, PE also has a significant impact on patients'

XSL•FO
RenderX

lives and prognosis. Studies have shown that up to 34% of postoperative deaths may be caused by PE [8,9].

Being able to accurately identify which patients are at high risk of VTE is of great significance for the prognosis of these patients. Accurate predictions can maximize the effectiveness of anticoagulant drugs and reduce side effects, such as bleeding [10,11]. Through the analysis of vast volumes of high-dimensional data, machine learning (ML) may uncover hidden patterns and intricate nonlinear relationships, aiding illness early warning [12,13]. In recent studies, ML models have been developed to predict the incidence of postoperative VTE. The most commonly used algorithms include extreme gradient boosting (XGB), random forest (RF), and support vector machine (SVM) [14]. In the prediction of VTE, ML technology exhibits many advantages. For instance, compared to traditional approaches, it can provide superior predictions by combining several types of information, such as imaging data, biomarkers, and clinical data [15,16]. Although these studies provide new methods for the field, there are some uncertainties in the consistency and reliability of the results in the existing literature owing to the single data source, small sample size, and the model's limited capacity for generalization. In addition, there are justified concerns about bias, explainability, and applicability [17]. Therefore, this study aimed to assess the effectiveness and methodological quality of ML models in predicting VTE following joint replacement through a systematic review in order to provide evidence-based support for their clinical application, identify current research gaps, and offer guidance for future high-quality studies.

Methods

Registration of the Study

This study adheres to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Checklist 1) guidelines for conducting systematic reviews [18]. The study protocol was registered on PROSPERO (International Prospective Register of Systematic Reviews; registration number: CRD42024625842).

Data Sources and Search Strategy

A comprehensive search was conducted in both Chinese and English databases, including Web of Science, Embase, Scopus, CNKI, Wanfang, Vipro, and PubMed, from the time of their formation until December 15, 2024. The following search terms were applied: "joint arthroplasty," "venous thromboembolism," "machine learning," "risk prediction," "prediction models," and their synonyms and variant terms. A manual search of reference lists containing studies to find related publications was conducted. For detailed search strategies, see Multimedia Appendix 1, which includes keyword combinations and database settings.

Inclusion or Exclusion Criteria

The following were the inclusion criteria: (1) studies on patients after joint replacement; (2) observational studies, case-control studies, cohort studies, and randomized controlled trials; (3) postoperative VTE risk was predicted using an ML model; (4) studies that provide performance evaluation metrics (eg, accuracy, sensitivity, specificity, area under the curve [AUC]) for the model and give clear statistical analysis results; and (5) studies published in English or Chinese.

The following were the exclusion criteria: (1) review articles, case reports, opinion pieces, commentaries, conference abstracts, and other nonoriginal research; (2) studies that do not mention particular kinds of ML models or algorithms; (3) animal experiments, simulation studies, and other nonclinical research; and (4) studies for which the entire text was not retrievable.

Study Selection and Data Synthesis

The retrieved results were imported into EndNote 21, and duplicates were eliminated. Two researchers trained in systematic evaluation techniques independently conducted the study selection process based on preestablished eligibility criteria. First, the titles and abstracts of the literature were screened, and the studies failing to meet the eligibility criteria were initially excluded. Second, it was filtered through full-text reading, and if 2 researchers disagreed, a third researcher stepped in to reach an agreement.

The data extraction process was also performed independently by 2 researchers following CHARMS (Critical Appraisal and Data Extraction for Systematic Reviews of Prediction) guidelines [19], and a third researcher settled any discrepancies. There are 2 groups into which the data taken from the chosen studies are separated: (1) basic information: author, year of publication, country, study design, participants, data source, clinical outcome, and sample size and (2) model information: missing data handling, candidate predictors, screening variable methods, prediction timepoint, prediction horizon, model development, internal verification, external verification, model performance, number of variates, calibration, and method of visualization.

Risk of Bias Assessment

The included studies' applicability and bias risk were evaluated using the PROBAST (Prediction Model Risk of Bias Assessment Tool) tool. The assessment has 20 questions, including 4 basic areas: participants, predictors, outcomes, and analysis. There are answers to the questions in each domain: "yes/probably" (low risk of bias), "no/probably not" (high risk of bias), and "no information" (unclear). A domain was deemed to have a high risk of bias if its answer contained "no/probably not." If every question had a "yes/probably yes" response, the domain was deemed to have a low risk of bias. When there is "no information," it is deemed to be highly biased if it cannot be explained by the original or additional material. If each domain was considered low risk, then the total bias risk was considered low. Furthermore, the total risk of bias was deemed high if any domains were deemed to be at high risk.

The bias risk in each model is assessed independently by 2 researchers, and if there are differences, they will be discussed to reach an agreement. Due to the majority of research using various models or having several outcomes, the bias risk is evaluated independently for each model and outcome. Finally, since there are no differences in the risk of bias between models in a single study, each study only shows the result once.

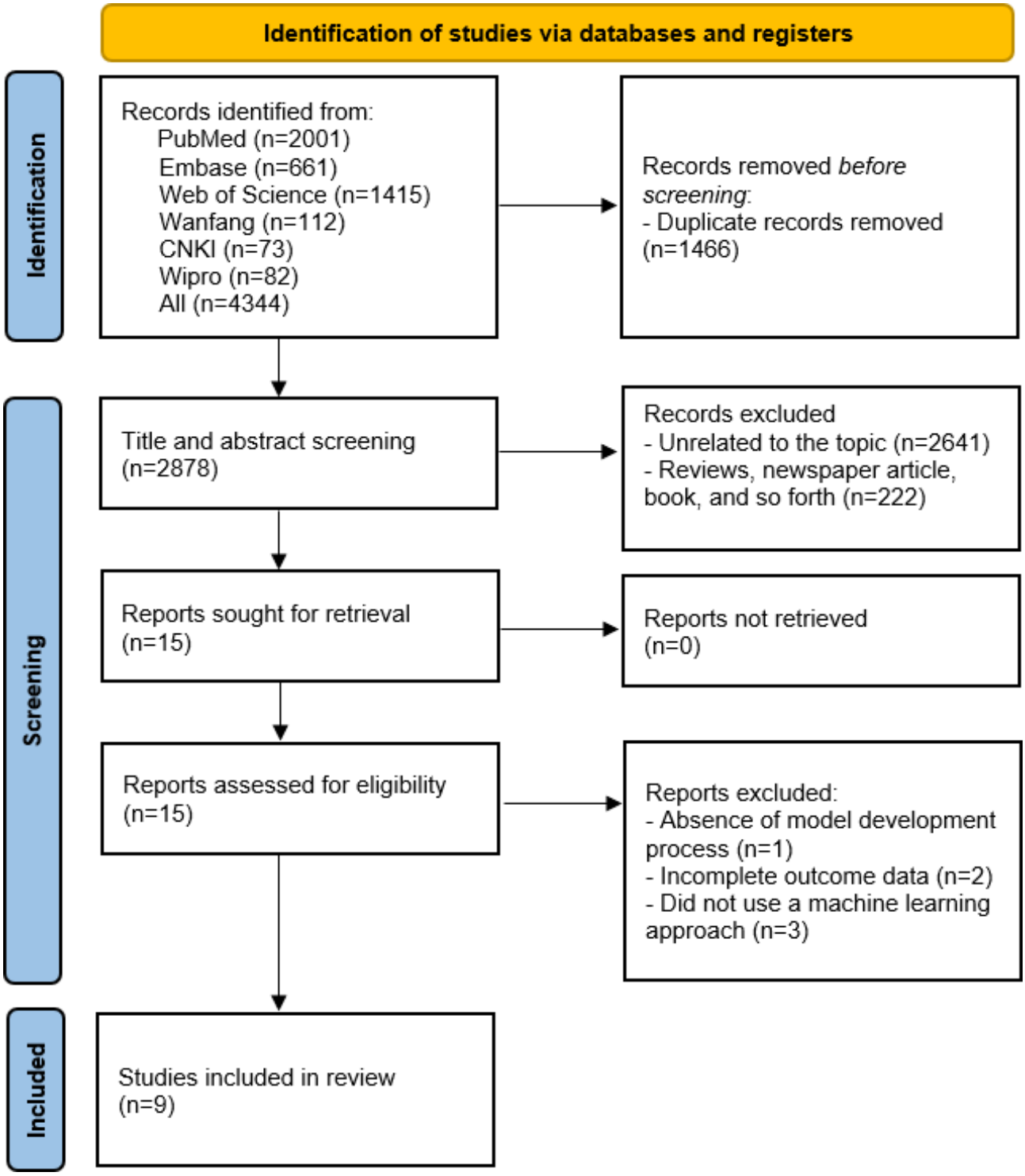
XSL•FO
RenderX

Results

Results of Literature Screening

The study selection process is detailed in the PRISMA flow diagram (Figure 1). A total of 4344 records were initially identified from 6 electronic databases (PubMed, n=2001; Embase, n=661; Web of Science, n=1415; Wanfang, n=112; CNKI, n=73; and Wipro, n=82). After removing 1466 duplicates, 2878 unique records remained for screening. During the title and abstract screening phase, 2863 records were excluded because they were either unrelated to the topic (n=2,641) or were nonoriginal research, such as reviews and newspaper articles. This left 15 reports to be sought for full-text retrieval. The subsequent critical appraisal excluded 6 studies based on predefined criteria: 3 did not use an ML approach, 2 had incomplete outcome data, and 1 did not involve a model development process. Ultimately, 9 studies met the full inclusion criteria and were included in this systematic review.

Figure 1. Literature screening process. CNKI: China National Knowledge Infrastructure.



XSL•FO
RenderX

General Study Characteristics

Table 1 shows the basic characteristics of the 9 studies. These studies were published from 2018 to 2024 and included 1 prospective study [3] and 8 retrospective studies [20-27]. Among them, 4 studies [3,20,21,25] were conducted in China; 3 studies [22-24] were from the United States; 1 was conducted in the Netherlands; and 1 transnational study used data from the Netherlands, the United Kingdom, and Denmark. In terms of study population, 2 studies [3,20] focused on patients with THA, 1 study [21] targeted patients with TKA, and the remaining studies [22-27] focused on both THA and TKA. In terms of clinical outcomes, 2 studies [20,21] concentrated solely on DVT, while the remaining 7 studies [3,22-27] evaluated both DVT and PE. In addition, the studies' sample sizes ranged from 60 to 3,92,661. Cases of VTE ranged from 24 to 4042, with an incidence of 0.9% to 40%.

Table . Basic characteristics of the included studies.

Author (year)	Country	Research type	Data source	Participants	Clinical outcome	Cases/sample size
Ding et al (2023) [3]	China	Prospective study	Shanghai Changzheng Hospital	THA[a]	VTE[b]	76/1481
Chen and Jiang (2022) [21]	China	Retrospective study	Clinical database of a hospital in Chongqing	TKA[c]	DVT[d]	24/60
Harris et al (2019) [22]	United States	Retrospective study	VASQIP data, VA Corporate Data Warehouse	TKA, THA	VTE	566/70,569
Rasouli Dezfouli et al (2022) [23]	United States	Retrospective study	The Cerner Health Facts data warehouse	TKA, THA	VTE	4042/392,661
Shohat et al (2023) [24]	United States	Retrospective study	Rothman Orthopedic Institute at Thomas Jefferson University	TKA, THA	DVT, PE[e]	308/35,963
Wang et al (2023) [25]	China	Retrospective study	A health system	TKA, THA	VTE	1161/6897
Xu et al (2024) [20]	China	Retrospective study	A hospital in Wenzhou, China	THA	DVT	92/333
Sweerts et al (2022) [26]	Netherlands	Retrospective study	Multiple databases from the Dutch	TKA, THA	VTE	26/3776
Nemeth et al (2024) [27]	Netherlands, Denmark, England	Retrospective study	Multiple databases from the United Kingdom, the Netherlands, and Denmark	TKA, THA	VTE	897/64,032

[a]THA: total hip arthroplasty.

[b]VTE: venous thromboembolism.

[c]TKA: total knee arthroplasty.

[d]DVT: deep vein thrombosis.

[e]PE: pulmonary embolism.

Characteristics and Performance of ML Models

Table 2 provides a detailed summary of the methodological design and data characteristics of the included studies. Table 3 presents a summary of the performance of the model development, validation, and inclusion of the studies. Among the 9 studies included, 5 studies [3,20,23-25] evaluated multiple ML algorithms, and 4 studies [21,22,26,27] used a single algorithm. Across these studies, 15 kinds of ML algorithms were used to develop 34 predictive models, and each study developed 1-5 models. Logistic regression (LR) and XGB were the most common algorithms, appearing in 5 studies (5/9, 56%) each. This is followed by RF (3/9, 33%) [23-25] and SVM (3/9, 33%) [20,24,25]. In particular, the XGB model showed strong performance in predicting DVT and PE, with an AUC range of 0.71 - 0.982. Figure 2 details the frequency of the model used.

XSL•FO
RenderX

Table . Summary of methodological design and data characteristics of the included studies.

Author (year)	Missing data handling	Candidate predictors	Screening variable methods	Prediction timepoint	Prediction horizon	Number of variates
Singh et al (2019) [1]	Data were excluded, or mean imputation, or mode imputation	67	LASSO[a] regression	Preoperative	During postoperative hospitalization	27
Yau et al (2022) [2]	NR[b]	NR	Single-factor analysis	Postoperative	Within 3-5 days after surgery	8
Ding et al (2023) [3]	Treats missing data as a separate category	NR	NR	Preoperative	Within 30 days after surgery	22
Simon et al (2023) [4]	NR	NR	Genetic algorithm	Preoperative	Within 30 days after surgery	25
Lee et al (2013) [5]	NR	56	Univariate analysis	Postoperative	Within 90 days after surgery	27
Malcolm et al (2020) [6]	NR	66	NR	Preoperative	Within 5 weeks after surgery	66
Lewis et al (2019) [7]	missForest	26	Recursive feature elimination	Postoperative	During postoperative hospitalization	10
Meng et al (2021) [8]	Multiple imputation	5	Literature, clinical reasoning, eyeballing	Preoperative	Within 1 year after surgery	5
Runner et al (2021) [9]	No missing data	39	LR[c]	Preoperative	Within 90 days after surgery	12

[a]LASSO: least absolute shrinkage and selection operator regression.

[b]NR: not reported.

[c]LR: logistic regression.

XSL•FO
RenderX

Table . Summary of model development, validation, and performance of included studies.

Author (year)	Model development	Internal verification	External verification	AUC[a] (95% CI)	Calibration	Method of visualization
Singh et al (2019) [1]	MLP[b], XGB[c], AdaBoost[d], GBC[e], KNN[f], LR[g]	10-fold cross-validation, random split validation	None	0.955 (0.917 - 0.993); 0.982 (0.954 - 1.000); 0.980 (0.944 - 1.000); 0.978 (0.953 - 1.000); 0.837 (0.743 - 0.931); 0.944 (0.878 - 1.000)	NR[h]	ROC[i] curve, DCA[j] curve, Shapley additive explanations, nomogram
Yau et al (2022) [2]	XGB	Random split validation	None	0.832 (0.748 - 0.916)	NR	ROC curve
Ding et al (2023) [3]	LASSO[k]	Internal cross-validation	Yes	0.613 (0.608 - 0.617)	Calibration plot	Calibration curve
Simon et al (2023) [4]	RF[l], GBT[m], TE[n], FCDNN[o]	Random split validation	None	0.698; 0.638; 0.635; 0.780	NR	ROC curve
Lee et al (2013) [5]	XGB, RF, LASSO, SVM[p]	Random split validation, repeated cross-validation	None	DVT[q]: 0.759 (0.667 - 0.881); 0.71; 0.68; 0.64; PE: 0.774 (0.683 - 0.835); 0.8; 0.8; 0.67	Calibration plot	ROC curve, PR[r] curve, calibration curve
Malcolm et al (2020) [6]	XGB, RF, SVM, BPNN[s], LR, EM[t]	5-fold cross-validation, GridSearchCV	None	0.914 (0.894 - 0.935); 0.907 (0.883 - 0.931); 0.903 (0.874 - 0.930); 0.910 (0.8949 - 0.926); 0.885 (0.868 - 0.902); 0.921 (0.896 - 0.936)	NR	ROC curve
Lewis et al (2019) [7]	XGB, KNN, SVM, NB[u], MLP[v], LR	Random split validation, 10-fold cross-validation	None	0.800 (0.674 - 0.927); 0.755 (0.634 - 0.874); 0.753 (0.627 - 0.878); 0.681 (0.571 - 0.788); 0.739 (0.611 - 0.846); 0.579 (0.450 - 0.696)	Brier score	Shapley additive explanations, feature importance
Meng et al (2021) [8]	Multiple imputation	LR	None	0.663 (0.627 - 0.699)	Hosmer-Lemeshow, calibration plot, Brier score	ROC curve, calibration curve
Runner et al (2021) [9]	No missing data	LR	Yes	0.65 (0.63 - 0.67) external validation: 0.64 (0.61 - 0.67)	Calibration plot	Calibration curve

[a]AUC: area under the curve.

[b]MLP: multilayer perceptron.

[c]XGB: extreme gradient boosting.

[d]AdaBoost: adaptive boosting.

[e]GBC: gradient boosting classifier.

[f]KNN: k-nearest neighbors.

[g]LR: logistic regression.

[h]NR: not reported.

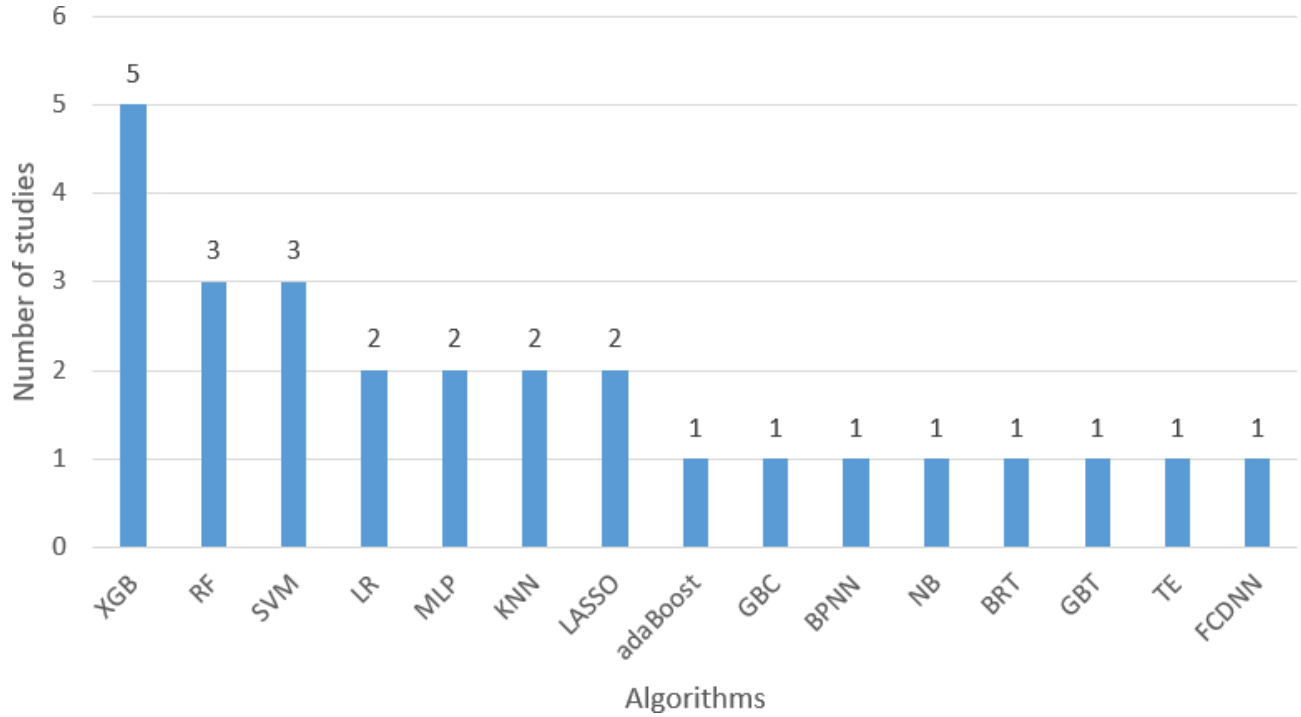
[i]ROC: receiver operating characteristic.

[j]DCA: decision curve analysis.

XSL•FO
RenderX

[k]LASSO: least absolute shrinkage and selection operator regression.
[l]RF: random forest.
[m]GBT: gradient boosting tree.
[n]TE: tree-based ensemble methods.
[o]FCDNN: fully connected deep neural network.
[p]SVM: support vector machines.
[q]DVT: deep vein thrombosis.
[r]PR: precision-recall.
[s]BPNN: back propagation neural network.
[t]EM: ensemble method.
[u]NB: naive Bayes.
[v]MLP: multilayer perceptron.

Figure 2. Frequency of machine learning models. adaBoost: adaptive boosting; BPNN: back propagation neural network; EM: ensemble method; FCDNN: fully connected deep neural network; GBT: gradient boosting tree; GBC: gradient boosting classifier; KNN: k-nearest neighbors; LASSO: least absolute shrinkage and selection operator regression; LR: logistic regression; MLP: multilayer perceptron; NB: naive Bayes; RF: random forest; SVM: support vector machine; TE: tree-based ensemble methods; XGB: extreme gradient boosting.



The included studies used a variety of metrics to evaluate ML models, of which AUC was the most common (8/9, 89%). The predicted AUC range for DVT and PE was 0.579 - 0.982. One study did not report AUC but reported C-statistic [22]. XGB, popularized by Chen and Guestrin [28] in 2016, is a widely adopted ML algorithm on the basis of gradient boosting [29,30]. The included XGB studies [3,20,21,24,25] reported AUC ranging from 0.71 to 0.982, demonstrating strong model performance. LR estimates the association of one or more independent variables with binary dependent variables [31,32]. The included LR studies [3,20,25-27] reported AUC ranging from 0.579 to 0.944. RF is a widely used ML algorithm that helps mitigate overfitting and improves the model's stability and accuracy by aggregating the outcomes from various decision trees [33]. It functions as a modeling algorithm as well as a variable selection technique. Three studies [23-25] used RF, with reported AUC ranging from 0.698 to 0.907. In addition, calibration plots were provided in 3 studies [22,24,27]; Brier scores were provided in 1 study [20]; and 1 study [26] also provided Hosmer-Lemeshow, calibration plot, and Brier score. The other 4 studies did not include data on calibration performance. Additional reported metrics encompass accuracy, sensitivity, and specificity.

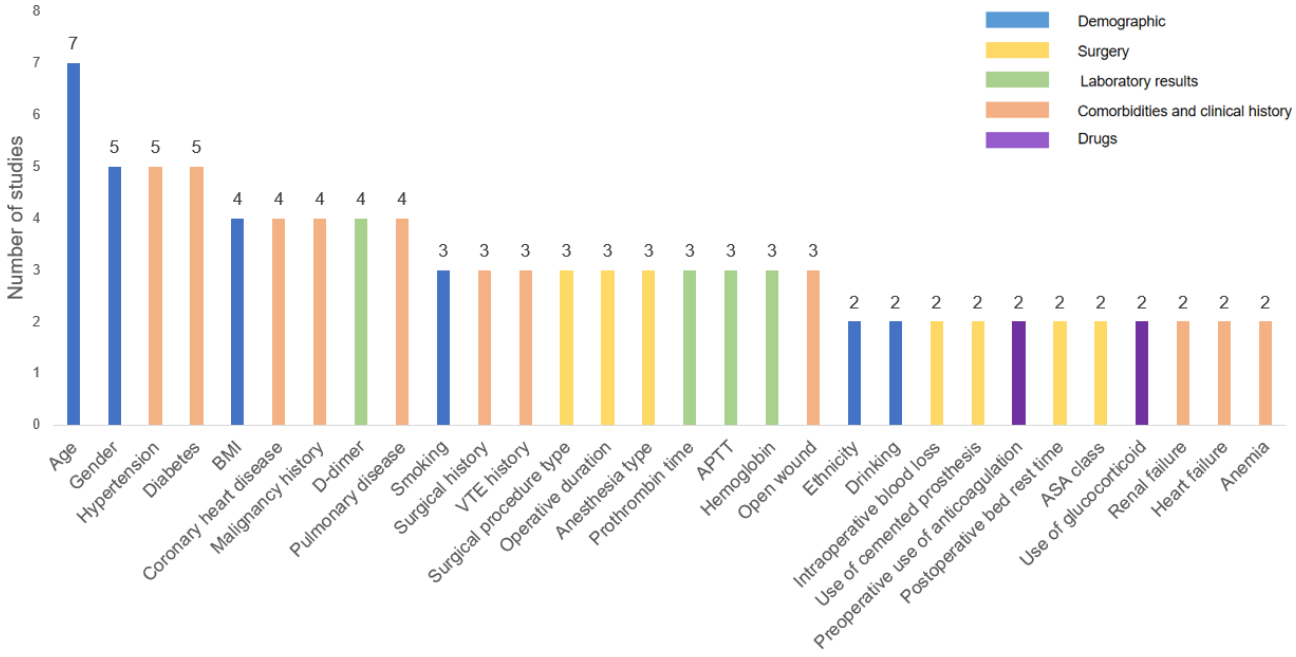
These 9 studies differ in terms of the optimal timing for risk assessment. Six studies have created preoperative prediction models [3,22,23,25-27] to identify high-risk individuals before surgery. The last 3 studies focused on postoperative prediction, assessing immediate risks using postoperative data [20,21,24]. Additionally, there is diversity in the risk assessment's predictive horizon. Its duration ranges from a brief period following the procedure, such as "within 3 - 5 days after surgery" or "during the postoperative hospital stay" [3,20], to a lengthy risk assessment that might continue for up to a year following the procedure [26]. Common time horizons include 30 days [22,23] and 90 days [24,27].

From a low of 5 [26] to a high of 66 [25], the number of predictor variables that ultimately made it into the model varied

XSL•FO
RenderX

widely. The descriptive analysis of the predictive factors used in each study (as shown in Figure 3) shows that a constant core set of variables is regularly utilized, despite the wide variations in the number of predictive factors used in each investigation. These variables can be broadly classified into four categories: (1) demographic and biological characteristics of the patients (such as age, gender, BMI); (2) major comorbidities (such as hypertension, diabetes, coronary heart disease, and malignant tumors); (3) specific VTE risk histories (such as previous VTE); and (4) key laboratory indicators (such as D-dimer).

Figure 3. Frequency of the predictors. APTT: activated partial thrombin time; ASA: American Society of Anesthesiologists; VTE: venous thromboembolism.



Model Validation

All of the models in the included studies underwent internal validation; however, only 2 studies [22,27] used external validation and reported the AUC value of external validation. The main methods used for internal validation were k-fold cross-validation and random-split validation. External validation plays a key role in confirming whether a model can maintain good performance in different populations and clinical settings. The absence of external validation will limit the generalization and clinical application of these models.

Assessing Risk of Bias

According to the PROBAST evaluation, all models had a high risk of bias (Table 4). The assessment results showed that the main sources of bias were concentrated in the “participants” and “analysis” domains. In the participants domain, 8 studies had a high risk of bias, mainly because these studies relied on retrospective data. In the analysis domain, all models had a high risk of bias. The PROBAST assessment pointed out several specific methodological issues: 4 studies [21,23-25] did not perform model calibration, 4 studies [3,21,23,25] did not report the handling of missing data, and 2 studies [21,23] were insufficient in measures to prevent model overfitting. In contrast, the outcome domain was rated as having a low bias risk level in all studies. In terms of applicability assessment, only 1 study [25] had poor applicability because the model was overly complex and required 66 predictor variables.

XSL•FO
RenderX

Table . Prediction Model Risk of Bias Assessment Tool (PROBAST) results of the included studies.

Author (year)	Risk of bias				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Risk of bias	Applicability
Ding et al (2023) [3]	+[a]	+	+	–[b]	+	+	+	–	+
Chen et al (2023) [21]	–	+	+	–	+	+	+	–	+
Harris et al (2019) [22]	–	+	+	–	+	+	+	–	+
Rasouli et al (2022) [23]	–	+	+	–	+	+	+	–	+
Shohat et al (2022) [24]	–	+	+	–	+	+	+	–	+
Wang et al (2023) [25]	–	–	+	–	+	–	+	–	–
Xu et al (2024) [20]	–	+	+	–	+	+	+	–	+
Sweerts et al (2022) [26]	–	+	+	–	+	+	+	–	+
Nemeth et al (2024) [27]	–	+	+	+	+	+	+	–	+

[a]Low risk of bias.

[b]High risk of bias.

Discussion

Principal Findings

VTE is a common and serious postoperative complication that is preventable [10]. Risk assessment and management-based strategies can help prevent or reduce the occurrence of VTE [34]. Therefore, selecting an appropriate predictive model for VTE is crucial. This systematic review examined the methodological quality and clinical applicability of ML methods in predicting VTE after joint replacement surgery. Among the reviewed studies, 15 ML algorithms were utilized to construct 34 predictive models. XGB and LR were the most used model types, followed by RF and SVM. The AUC values of the models ranged from 0.579 to 0.982, indicating that the discrimination ability of the models varied greatly, and some of the models showed excellent performance. Two research [1,6] reports demonstrated extremely high discrimination ability, with their AUC values exceeding 0.9. Such high performance often raises concerns about overfitting, where the model learns specific statistical noise specific to the training data rather than the generalizable underlying biological signals, leading to an optimistically biased performance estimate. Moreover, these models have not yet undergone external validation. This deficiency is a major limitation. In the clinical environment of joint replacement, there may be significant differences among patients, and surgical techniques as well as VTE prevention protocols may also vary among different medical institutions. Models trained based on data from a single hospital are likely to learn the specific treatment patterns of that hospital rather than a universally applicable VTE prediction signature. This may lead to unreliable prediction results when the model is applied to different patient groups. The risk of overfitting and the lack of reliable external validation seriously undermine the clinical value of these highly performing models. These models can only be considered for practical application after undergoing extensive external validation.

Another finding is that the XGB model demonstrates superior predictive performance compared to other models. For instance, 2 studies [3,20] compared the XGB and SVM models with the LR models, and the results showed that the XGB and SVM models outperformed the LR model. This supremacy is probably due to the smart way the algorithm was made. First, XGB has strong built-in features (such as regularization) that keep the model from overfitting when working with small datasets, making it stable even when the data are noisy. Second, XGB is very good at identifying the most important risk factors on its own. It can also capture complicated nonlinear interaction relationships, which typical linear models cannot do. Furthermore, the fact that XGB functions so effectively most likely indicates that the research team spent a great deal of time and energy optimizing the hyperparameters and determining the ideal configurations for their particular dataset. However, research has indicated that LR models may be more successful and easier to apply in some situations, whereas XGB and SVM models may not necessarily perform better. LR's fundamental structure makes it simple to comprehend and analyze. Additionally, it produces dependable findings and has strong convergence, particularly when used for small datasets [35]. The Caprini score is a clinically prevalent instrument utilized to evaluate VTE risk by analyzing many clinical risk factors, including age, surgical history, malignancy, and obesity [36]. One study [25] compared the Caprini Score to ML models and showed that the ML models were better at predicting VTE risk. Although the predictive performance of most models ranged

XSL•FO
RenderX

from moderate to good, the risk of bias in all studies restricted the applicability of these models in clinical practice.

Related Work

The PROBAST assessment results indicate that all the models have a high risk of bias, suggesting that their predictive performance in practical applications may be lower than the reported levels. This widespread bias is mainly due to the flaws in the research design and the statistical analysis methods adopted by the models. In terms of the research design, most studies rely on retrospective data, which increases the risk of selecting an unrepresentative patient population, thereby weakening the generalization ability of the models beyond the original dataset. Additionally, there are numerous flaws in the analysis aspects of the model construction and testing processes. The main issues are that some studies did not check the calibration process. This makes it impossible for us to determine whether the "30% risk of venous thrombosis" predicted by the model truly means a 30% probability in reality, thus preventing its use in clinical decision-making. Some studies also failed to provide enough protections against the "overfitting" problem, in which the model just remembered the noise in the training data instead of identifying the true risk patterns, while others did not provide solutions for handling missing patient data. These analytical flaws make these models likely to perform poorly when applied to new patients. Furthermore, the actual usability is also an issue. One model used an excessive 66 predictive variables, making it overly complex and unsuitable for clinical use. To raise the accuracy and clinical application of predictive models, future studies should apply more rigorous methodological methods. Moreover, the significant bias risk indicates inconsistency in the development and validation of current models, which could influence their widespread application. Future studies should give priority to enhancements in a number of important domains, including the research design (ie, cohort or case-control study design), the number of events, the method for predictor selection, the processing of complicated data, and the processes of model calibration and fitting.

One serious and widespread flaw in these studies is the absence of model calibration reports. Although most of the included studies reported high discrimination metrics such as AUC, the discrimination does not guarantee that the predicted probabilities are accurate. Calibration addresses a more fundamental clinical question: Does the predicted "30% risk of venous thrombosis" really mean that there will be a 30% incidence rate among a group of similar patients? Inaccurate calibration might have very harmful consequences. For example, a model that consistently overestimates risks may incorrectly identify patients with low risks as having high risks [37]. This may lead to the improper use of potent anticoagulant drugs, thereby exposing patients to a serious risk of bleeding, a risk that could have been avoided. However, if the model underestimates the risks, it may prevent doctors from taking necessary preventive measures to ensure the safety of patients with high risk. This could lead to potentially fatal thrombosis incidents [38]. Consequently, owing to the absence of calibration data, most of these models are unreliable for informing individual patient decisions, irrespective of their quoted AUC values. Future research must prioritize calibration as the principal measure of model performance. We strongly urge that studies include both visual calibration plots and quantitative indicators (such as the Brier Score) at the same time to make sure their predictions are accurate in the actual world. Additionally, most of the included studies did not provide reports on the sensitivity and specificity of their models. This affects the overall assessment of model performance, as the reported AUC values may not fully reflect its clinical applicability. Future research should follow standardized reporting guidelines to evaluate its model performance.

External validation is critical to confirm the generalizability of the model's performance and its applicability to different populations [39,40]. However, all the studies used internal validation methods, such as K-fold cross-validation or random split validation, and only 2 studies conducted external validation. Most of these models lack external validation, further increasing the challenge of clinical implementation. In such a sensitive sector as health care, where predictive models are finally meant to be employed in real-world situations, these ML models may not accurately reflect their applicability and dependability without outside validation [41]. Future studies should incorporate external validation where possible, such as in different hospitals or using publicly accessible databases, to optimize the model generalization through data synthesis and larger datasets [42]. Additionally, we advise temporal validation to assess the model's resilience at different points in time.

The ML model's predictive variables found in this study have a number of important effects on clinical practice. First, the predictive factors frequently used in these models play a very important role in current clinical practice and future research. For instance, D-dimer is a well-known clinically validated biomarker for VTE. Studies have shown that it is highly sensitive in predicting the risk of VTE. Prothrombin time, activated partial thromboplastin time, and international normalized ratio are a few other markers that may assist physicians better understand the risk of blood clots and quickly identify and treat patients with high risk [43]. Another established risk factor for VTE is age. Research has indicated that the risk of VTE rises with age. People over the age of 60 years account for around 70% of occurrences of VTE [44]. Considering age while evaluating patients before surgery may enhance outcomes and diminish complication rates. Common chronic disorders, such as diabetes and hypertension, were also among recurring factors. Research has shown that these disorders can raise VTE incidence. A personalized anticoagulant treatment strategy might help those with past histories of these disorders lower their VTE risk even more [45]. Second, the results of the 9 studies on the predictors can guide the development of future predictive models and contribute to later studies on related risk elements. For instance, 1 study [20] included bone cement prosthesis as a predictor in the model, although further study is required to confirm its association with VTE. However, variance in the choosing and reporting of predictors in these studies may affect the comparability and practical utility of the models in different clinical settings.

Limitations

To our knowledge, this is the first study to evaluate ML in predicting VTE in patients undergoing joint replacement

XSL•FO
RenderX

surgery. Nevertheless, the study had certain limitations. First, most of the included studies have been conducted in China and the United States. This geographical concentration restricts the general applicability of the model to different areas. Second, numerous studies focused solely on developing predictive models without considering external validation or practical application of the model. The absence of external validation may affect the credibility of our pooled estimates. Third, since the traditional methods used to assess publication bias (such as funnel plots and Egger or Begg tests) are not applicable to AUC data, no quantitative assessment of publication bias was conducted for the included studies. However, the risk of publication bias remains: “positive” studies (eg, high AUC values) are more likely to be published than “negative” or inconclusive ones. Therefore, the studies included in this review may overestimate the true performance of these models. Finally, the included studies exhibited a high risk of bias and insufficient transparency. These methodological shortcomings could affect the accuracy of model predictions, therefore reducing the dependability and application of the conclusions.

Conclusions

The study identified 9 studies with a total of 34 ML models for predicting VTE after joint replacement. Most studies have modest to good discriminative ability from their AUC values. However, among these 9 studies, only 2 of them underwent effective external validation. The widespread neglect of external validation has left the generalization capabilities of these models, which performed well on the original dataset, completely unknown. In addition, these studies also show a high risk of bias and significant heterogeneity. Most of the studies had significant methodological shortcomings, including a lack of rigorous study design and absence of calibration measurements. These deficiencies will affect the reliability of the model in clinical applications and reduce the promotion value of the model. To improve the prediction capabilities of ML models, future research must closely follow the PROBAST criteria, which place a strong emphasis on meticulous study design and quality control. External validation is also necessary to improve the generalization and applicability of ML models in clinical settings.

Funding

This work was supported by the Jilin Provincial Department of Science and Technology (grant number 20210203091SF).

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JM
Data curation: HT, YZ
Formal analysis: HT, JM, XY
Funding acquisition: GW
Investigation: GW
Methodology: GW, JM, XY
Project administration: GW
Resources: HT
Software: HT
Supervision: GW
Validation: TZ, XY, YZ
Visualization: JM, XL
Writing – original draft: JM
Writing – review & editing: XY, YZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of the search strategy.
[DOCX File, 24 KB - medinform_v14i1e79886_app1.docx]

Checklist 1
PRISMA checklist.
[PDF File, 206 KB - medinform_v14i1e79886_app2.pdf]

References

XSL•FO
RenderX

1. Singh JA, Yu S, Chen L, Cleveland JD. Rates of total joint replacement in the United States: future projections to 2020-2040 using the national inpatient sample. J Rheumatol 2019 Sep;46(9):1134-1140. [doi: 10.3899/jrheum.170990] [Medline: 30988126]
2. Yau LK, Henry FU, Man Hong C, et al. Swelling assessment after total knee arthroplasty. J Orthop Surg (Hong Kong) 2022;30(3):10225536221127668. [doi: 10.1177/10225536221127668] [Medline: 36122907]
3. Ding R, Ding Y, Zheng D, et al. Machine learning-based screening of risk factors and prediction of deep vein thrombosis and pulmonary embolism after hip arthroplasty. Clin Appl Thromb Hemost 2023;29:10760296231186145. [doi: 10.1177/10760296231186145] [Medline: 37394825]
4. Simon SJ, Patell R, Zwicker JI, Kazi DS, Hollenbeck BL. Venous thromboembolism in total hip and total knee arthroplasty. JAMA Netw Open 2023 Dec 1;6(12):e2345883. [doi: 10.1001/jamanetworkopen.2023.45883] [Medline: 38039005]
5. Lee WS, Kim KI, Lee HJ, Kyung HS, Seo SS. The incidence of pulmonary embolism and deep vein thrombosis after knee arthroplasty in Asians remains low: a meta-analysis. Clin Orthop Relat Res 2013 May;471(5):1523-1532. [doi: 10.1007/s11999-012-2758-9] [Medline: 23264001]
6. Malcolm TL, Knezevic NN, Zouki CC, Tharian AR. Pulmonary complications after hip and knee arthroplasty in the United States, 2004-2014. Anesth Analg 2020 Apr;130(4):917-924. [doi: 10.1213/ANE.0000000000004265] [Medline: 31206434]
7. Lewis S, Glen J, Dawoud D, et al. Venous thromboembolism prophylaxis strategies for people undergoing elective total knee replacement: a systematic review and network meta-analysis. Lancet Haematol 2019 Oct;6(10):e530-e539. [doi: 10.1016/S2352-3026(19)30155-3] [Medline: 31444124]
8. Meng H, Zhu Y, Zhang J, et al. Incidence and risk factor for preoperative deep vein thrombosis (DVT) in isolated calcaneal fracture, a prospective cohort study. Foot Ankle Surg 2021 Jul;27(5):510-514. [doi: 10.1016/j.fas.2020.06.007] [Medline: 32595018]
9. Runner RP, Shau DN, Staley CA, Roberson JR. Utilization patterns, efficacy, and complications of venous thromboembolism prophylaxis strategies in revision hip and knee arthroplasty as reported by American Board of Orthopaedic Surgery Part II candidates. J Arthroplasty 2021 Jul;36(7):2364-2370. [doi: 10.1016/j.arth.2021.01.072] [Medline: 33674164]
10. Chong J, Curtain C, Gad F, et al. Development and implementation of venous thromboembolism stewardship across a hospital network. Int J Med Inform 2021 Nov;155:104575. [doi: 10.1016/j.ijmedinf.2021.104575] [Medline: 34560489]
11. Agarwal AR, Das A, Harris A, Campbell JC, Golladay GJ, Thakkar SC. Trends of venous thromboembolism after total hip arthroplasty in the United States: analysis from 2011 to 2019. J Am Acad Orthop Surg 2023 Apr 1;31(7):e376-e384. [doi: 10.5435/JAAOS-D-22-00708] [Medline: 36727960]
12. Wei C, Wang J, Yu P, et al. Comparison of different machine learning classification models for predicting deep vein thrombosis in lower extremity fractures. Sci Rep 2024 Mar 22;14(1):6901. [doi: 10.1038/s41598-024-57711-w] [Medline: 38519523]
13. Khera R, Haimovich J, Hurley NC, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol 2021 Jun 1;6(6):633-641. [doi: 10.1001/jamacardio.2021.0122] [Medline: 33688915]
14. Ryan L, Mataraso S, Siefkas A, et al. A machine learning approach to predict deep venous thrombosis among hospitalized patients. Clin Appl Thromb Hemost 2021;27:1076029621991185. [doi: 10.1177/1076029621991185] [Medline: 33625875]
15. Guan C, Ma F, Chang S, Zhang J. Interpretable machine learning models for predicting venous thromboembolism in the intensive care unit: an analysis based on data from 207 centers. Crit Care 2023 Oct 24;27(1):406. [doi: 10.1186/s13054-023-04683-4] [Medline: 37875995]
16. Hou T, Qiao W, Song S, et al. The use of machine learning techniques to predict deep vein thrombosis in rehabilitation inpatients. Clin Appl Thromb Hemost 2023;29:10760296231179438. [doi: 10.1177/10760296231179438] [Medline: 37365805]
17. Arina P, Kaczorek MR, Hofmaenner DA, et al. Prediction of complications and prognostication in perioperative medicine: a systematic review and PROBAST assessment of machine learning tools. Anesthesiology 2024 Jan 1;140(1):85-101. [doi: 10.1097/ALN.0000000000004764] [Medline: 37944114]
18. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
19. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
20. Xu Q, Yu B, Zhou P, Dai S, Dong X. Construction of interpretable prediction model for lower limb deep vein thrombosis in patients undergoing total hip arthroplasty based on machine learning and SHAP. Chin J Hosp Stat 2024;31(1) [FREE Full text]
21. Chen Y, Jiang Y. Construction of prediction model of deep vein thrombosis risk after total knee arthroplasty based on XGBoost algorithm. Comput Math Methods Med 2022 Jan 25;2022:3452348. [doi: 10.1155/2022/3452348] [Medline: 35116072]

XSL•FO
RenderX

22. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin Orthop Relat Res 2019 Feb;477(2):452-460. [doi: 10.1097/CORR.0000000000000601] [Medline: 30624314]
23. Rasouli Dezfouli E, Delen D, Zhao H, Davazdahemami B. A machine learning framework for assessing the risk of venous thromboembolism in patients undergoing hip or knee replacement. J Healthc Inform Res 2022;6(4):423-441. [doi: 10.1007/s41666-022-00121-2] [Medline: 36744082]
24. Shohat N, Ludwick L, Sherman MB, Fillingham Y, Parvizi J. Using machine learning to predict venous thromboembolism and major bleeding events following total joint arthroplasty. Sci Rep 2023 Feb 7;13(1):2197. [doi: 10.1038/s41598-022-26032-1] [Medline: 36750656]
25. Wang X, Xi H, Geng X, et al. Artificial intelligence-based prediction of lower extremity deep vein thrombosis risk after knee/hip arthroplasty. Clin Appl Thromb Hemost 2023;29:10760296221139263. [doi: 10.1177/10760296221139263] [Medline: 36596268]
26. Sweerts L, Hoogeboom TJ, van Wessel T, van der Wees PJ, van de Groes SAW. Development of prediction models for complications after primary total hip and knee arthroplasty: a single-centre retrospective cohort study in the Netherlands. BMJ Open 2022 Aug 24;12(8):e062065. [doi: 10.1136/bmjopen-2022-062065] [Medline: 36002218]
27. Nemeth B, Smeets M, Pedersen AB, et al. Development and validation of a clinical prediction model for 90-day venous thromboembolism risk following total hip and total knee arthroplasty: a multinational study. J Thromb Haemost 2024 Jan;22(1):238-248. [doi: 10.1016/j.jtha.2023.09.033] [Medline: 38030547]
28. Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); Aug 13-17, 2016. [doi: 10.1145/2939672.2939785]
29. Liang D, Wang L, Liu S, et al. Global incidence of diarrheal diseases—an update using an interpretable predictive model based on XGBoost and SHAP: a systematic analysis. Nutrients 2024 Sep 23;16(18):3217. [doi: 10.3390/nu16183217] [Medline: 39339819]
30. Davagdorj K, Pham VH, Theera-Umpon N, Ryu KH. XGBoost-based framework for smoking-induced noncommunicable disease prediction. Int J Environ Res Public Health 2020 Sep 7;17(18):6513. [doi: 10.3390/ijerph17186513] [Medline: 32906777]
31. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA 2016 Aug 2;316(5):533-534. [doi: 10.1001/jama.2016.7653] [Medline: 27483067]
32. Moore A, Bell M. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. Clin Med Insights Cardiol 2022;16:11795468221133611. [doi: 10.1177/11795468221133611] [Medline: 36386405]
33. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 2009 Dec;14(4):323-348. [doi: 10.1037/a0016973] [Medline: 19968396]
34. Raskob GE, Angchaisuksiri P, Blanco AN, et al. Thrombosis: a major contributor to global disease burden. Semin Thromb Hemost 2014 Oct;40(7):724-735. [doi: 10.1055/s-0034-1390325] [Medline: 25302681]
35. Chen X, He L, Shi K, Wu Y, Lin S, Fang Y. Interpretable machine learning for fall prediction among older adults in China. Am J Prev Med 2023 Oct;65(4):579-586. [doi: 10.1016/j.amepre.2023.04.006] [Medline: 37087076]
36. Caprini JA. Thrombosis risk assessment as a guide to quality patient care. Dis Mon 2005;51(2-3):70-78. [doi: 10.1016/j.disamonth.2005.02.003] [Medline: 15900257]
37. Ghadri JR, Cammann VL, Jurisic S, et al. A novel clinical score (InterTAK Diagnostic Score) to differentiate takotsubo syndrome from acute coronary syndrome: results from the International Takotsubo Registry. Eur J Heart Fail 2017 Aug;19(8):1036-1042. [doi: 10.1002/ejhf.683] [Medline: 27928880]
38. Zaka A, Mutahar D, Gorcilov J, et al. Machine learning approaches for risk prediction after percutaneous coronary intervention: a systematic review and meta-analysis. Eur Heart J Digit Health 2024 Oct;6(1):23-44. [doi: 10.1093/ehjdh/ztae074] [Medline: 39846069]
39. Looi JL, Chan C, Pemberton J, et al. External validation of a clinical score to differentiate takotsubo syndrome from non-ST-elevation myocardial infarction in women. Heart Lung Circ 2023 Jun;32(6):696-701. [doi: 10.1016/j.hlc.2023.04.002] [Medline: 37121882]
40. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016 Jun 22;353:i3140. [doi: 10.1136/bmj.i3140] [Medline: 27334381]
41. Zobeiri A, Rezaee A, Hajati F, Argha A, Alinejad-Rokny H. Post-cardiac arrest outcome prediction using machine learning: a systematic review and meta-analysis. Int J Med Inform 2025 Jan;193:105659. [doi: 10.1016/j.ijmedinf.2024.105659] [Medline: 39481177]
42. Mao A, Su J, Ren M, Chen S, Zhang H. Risk prediction models for falls in hospitalized older patients: a systematic review and meta-analysis. BMC Geriatr 2025 Jan 14;25(1):29. [doi: 10.1186/s12877-025-05688-0] [Medline: 39810076]
43. Alhabibi AM, Eldewi DM, Wahab MAA, Farouk N, El-Hagrasy HA, Saleh OI. Platelet-derived growth factor-beta as a new marker of deep venous thrombosis. J Res Med Sci 2019;24:48. [doi: 10.4103/jrms.JRMS_965_18] [Medline: 31160915]

XSL•FO
RenderX

44. Wang H, Rosendaal FR, Cushman M, van Hylckama Vlieg A. Procoagulant factor levels and risk of venous thrombosis in the elderly. J Thromb Haemost 2021 Jan;19(1):186-193. [doi: 10.1111/jth.15127] [Medline: 33090602]
45. Anghele AD, Marina V, Dragomir L, Moscu CA, Anghele M, Anghel C. Predicting deep venous thrombosis using artificial intelligence: a clinical data approach. Bioengineering (Basel) 2024 Oct 25;11(11):1067. [doi: 10.3390/bioengineering11111067] [Medline: 39593727]

Abbreviations

AUC: area under the curve
CHARMS: Critical Appraisal and Data Extraction for Systematic Reviews of Prediction
DVT: deep vein thrombosis
LR: logistic regression
ML: machine learning
PE: pulmonary embolism
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
RF: random forest
SVM: support vector machine
THA: total hip arthroplasty
TKA: total knee arthroplasty
VTE: venous thromboembolism
XGB: extreme gradient boosting

Edited by A Benis; submitted 30.Jun.2025; peer-reviewed by S Mohanadas, T Shi; revised version received 18.Nov.2025; accepted 29.Dec.2025; published 12.Feb.2026.

Please cite as:
Ma J, Tang H, Zhang Y, Yi X, Zhong T, Li X, Wang G
Machine Learning for Predicting Venous Thromboembolism After Joint Arthroplasty: Systematic Review of Clinical Applicability and Model Performance
JMIR Med Inform 2026;14:e79886
URL: https://medinform.jmir.org/2026/1/e79886
doi:10.2196/79886

© Junwei Ma, Huifeng Tang, Yunshan Zhang, Xuemei Yi, Tangsheng Zhong, Xinyun Li, Gang Wang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Health Information Systems to Support Behavioral Interventions in Local Contexts: Scoping Review

Xi Yang[1,2], MArch; Martin Dijst[2,3], Prof Dr; Joost G Daams[4], MA; Marc Suhrcke[5,6], Prof Dr; Haoran Yang[7,8], Prof Dr; Ronald Cornet[1], Prof Dr

[1]Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Meibergdreef 15, Amsterdam, The Netherlands

[2]Department of Urban Development and Mobility, Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg

[3]University of Luxembourg, Esch-sur-Alzette, Luxembourg

[4]Medical Library, Research Support, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[5]Department of Living Conditions, Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg

[6]Centre for Health Economics, University of York, York, United Kingdom

[7]School of Geographic Sciences, East China Normal University, Shanghai, China

[8]Liwa Institute of Skin Health, East China Normal University, Shanghai, China

Corresponding Author:
Ronald Cornet, Prof Dr
Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Meibergdreef 15, Amsterdam, The Netherlands

Abstract

Background: Individual-level behavioral interventions are designed to improve health behaviors and manage noncommunicable diseases. Neighborhood geo-referenced contexts (NGRCs) significantly impact the success of these interventions. Integrating NGRC data into health information systems (HISs), including electronic medical records (EMRs), electronic health records (EHRs), and personal health records (PHRs), can enhance personalized NGRC-focused behavioral interventions and improve health outcomes. Despite the potential benefits, there is a notable gap in the literature about NGRC-focused behavioral interventions using HISs.

Objective: This scoping review aims to review the current status and stakeholder insights of NGRC-focused behavioral interventions using HISs.

Methods: Two reviewers examined publications indexed by MEDLINE (Ovid) and Scopus. Publications reporting on NGRC-focused behavioral interventions using HISs were included. We extracted data on study characteristics, population attributes, diseases, health setting characteristics, HISs, NGRCs, behavioral interventions, outcomes, and conclusions, with suggestions for future research.

Results: The literature search identified 24 studies for inclusion. Of these, EHRs or EMRs were mostly used in 20 studies, with limited focus on PHRs. HISs were not used in the entire process of NGRC-focused behavioral interventions in 23 studies. Behavioral interventions in 21 studies focused exclusively on social needs interventions, with comparatively limited attention given to lifestyle change interventions. NGRCs mainly focused on built environments (n=24), with less attention to natural (n=1) and social (n=4) environments. The spatiotemporal characteristics of NGRCs were described in limited detail. “Destination accessibility” was the main focus within built environments. Other characteristics (“density,” “design,” “diversity,” and “distance”) were less studied. Stakeholders’ perspectives highlighted the value, convenience, patient autonomy, and privacy of integrating content (NGRCs) and/or technology (HISs) into behavioral interventions, despite the persistence of certain social concerns. We summarized 15 recommendations in terms of NGRC-focused behavioral interventions using HISs.

Conclusions: PHRs have the potential to actively integrate more diverse, dynamic, and/or high-resolution NGRCs. Additionally, PHRs can participate in the entire process of NGRC-focused behavioral interventions, particularly in outpatient lifestyle change interventions. Health care providers can support these interventions to help achieve public health goals and advance health equity.

(JMIR Med Inform 2026;14:e77296) doi:10.2196/77296

KEYWORDS

health information system; behavioral intervention; environments; geo-referenced contexts; neighborhood geo-referenced contexts

XSL•FO
RenderX

Introduction

Behavioral interventions are strategies and techniques to influence health-related behaviors at the personal, community, and public policy levels [1]. They may encourage individuals to quit smoking or to exercise regularly. At the community level, behavioral interventions aim to change behaviors by modifying the environments that support them [1]. For example, enhancing existing sidewalks with improved lighting, repainting bike lane markings, and adding benches can create a more enjoyable experience that encourages walking and cycling. At the public policy level, the government or private groups can share health information to encourage behavioral change [1]. The absence of individual-level behavioral interventions poses a significant challenge in managing noncommunicable diseases (NCDs), which account for 70% of global deaths annually [2], as effective interventions for NCDs require long-term engagement and multiple follow-ups to detect relapses or late effects as soon as possible [3].

Neighborhood geo-referenced contexts (NGRCs) have been identified as barriers or facilitators to the acceptance of, participation in, and adherence to individual-level behavioral interventions [4]. NGRCs refer to the physical and social environments of a location or locations along one's daily activities, and they influence people's health-related behaviors [5-8]. NGRCs are a spatial manifestation of the social determinants of health (SDoH), which refer to the conditions in which people are born, grow, live, work, and age, as well as the broader systems and forces that shape these daily life conditions [9]. Beyond the social and natural environments in NGRCs, the built environment is characterized using the 5D framework, which includes *density*, *diversity*, *design*, *distance*, and *destination accessibility*. The failure to address NGRCs is evident from the focus on limited prevention programs or resources in health care settings aimed at addressing individual attitudes and behaviors, along with the limited resources for preventive care [10]. NGRC-focused behavioral interventions allow health care providers to tailor interventions to specific contexts, offering a more personalized approach. These interventions can be categorized into social needs interventions (eg, arranging transportation) and individual lifestyle change interventions (eg, walking in a nearby park). To achieve NGRC-focused behavioral interventions, studying the interaction between NGRCs and modifiable health-related behaviors is essential [11-13].

Health information system (HIS) data have been used to investigate the relationships between NGRCs and health behaviors among health providers, HIS vendors, payers, and researchers [14-24]. A HIS is a system for collecting and processing data, information, and knowledge in health care settings [25]. The category includes (1) provider-centric electronic medical record (EMR) systems used within a single health care setting, (2) electronic health record (EHR) systems spanning multiple settings, and (3) patient-centric personal health record (PHR) systems. Among PHRs, standalone PHRs are not integrated with EMRs or EHRs, whereas tethered and interconnected PHRs are linked to EMRs or EHRs [25-27]. These systems enable various stakeholders, such as health providers, patients, and caregivers, to access and process multimodal data, such as diagnoses, medication use, treatments, NGRCs, and behavioral information.

At the individual level, HISs can enhance NGRC-focused behavioral intervention via comprehensive NGRC data capture and effective intervention prescription, monitoring, and assessment [3]. HISs can help collect NGRC data during routine clinical care, supplemented by patient-reported information, or through linkage with publicly available external data sources like geographic information systems. At the community and policy levels, HISs can support cost-effective, large-volume, and time-efficient assessments of large-scale public health efforts, such as environmental interventions to support behavioral changes, compared with traditional surveys [28].

Although integrating NGRCs with HISs is not new [29], many reviews on the functionalities and features of EHRs, EMRs, and PHRs lack clarity regarding NGRCs [30-44]. Additionally, several systematic reviews have investigated the linkage between NGRC data and clinical data sources (eg, EHRs) for population health research, but they lack specifics regarding behavioral interventions [45,46]. Recent systematic reviews on SDoH screening, linkage, extraction, analysis, and interventions within HISs, especially EHRs, similarly show limited attention to NGRCs and/or intervention-based approaches [47-50]. Given the increasing interest and existing research gap in using HISs to prepare, prescribe, monitor, and assess NGRC-focused behavioral interventions, we conducted, to our knowledge, the first scoping review of the literature in this field to distinguish it from broader SDoH-focused reviews and those lacking an intervention-based approach [46,50]. This review excluded standalone PHRs, as they are independent of provider-centric EHRs or EMRs.

This scoping review aimed to answer the following questions:

1. To what extent and what types of HISs have been used in NGRC data capture, intervention prescriptions, monitoring, and assessments of NGRC-focused behavioral interventions, and across the 4 stages?
2. What types, spatiotemporal characteristics, and databases of NGRCs have been utilized in each type of behavioral intervention supported by HISs?
3. What are stakeholders' insights regarding the use of HISs and/or NGRCs in behavioral interventions?
4. What are the gaps in research on the use of NGRCs and HISs in behavioral interventions?

Methods

Checklist

The analysis adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Checklist 1) [51].

Search for Relevant Research

An information specialist from the Amsterdam Medical Center library (JGD) was consulted to help define key concepts and establish the evaluation framework, which included selecting the appropriate databases, relevant keywords, and MeSH terms.

XSL•FO
RenderX

MEDLINE (Ovid) and Scopus were searched from inception on April 29, 2024, to include studies up to the search date.

The search strategy could be conceptually summarized as follows: ([GRCs] AND (([neighborhoods] AND ([chronic care model] OR ([HISs] AND [health behaviors]))) OR ([community health facilities] AND [health behaviors]))), involving the following four elements related to our research questions: (1) HISs, (2) geo-referenced contexts (GRCs), (3) neighborhoods, and (4) health behaviors. Detailed information regarding the search terms in Scopus is shown in Textbox 1. The inclusion and exclusion criteria are presented in Textbox 2. For search terms in MEDLINE, refer to Table S2 in Multimedia Appendix 1.

Textbox 1. Search terms in Scopus.

[GRCs] AND (([neighborhoods] AND ([chronic care model] OR ([HISs] AND [health behaviors]))) OR ([community health facilities] AND [health behaviors]))

(((TITLE-ABS-KEY((("electronic health" or personal or digital or medical or patient or participant) W/1 record) OR (("electronic health" or personal or digital or medical or patient or participant or subject or adult) W/2 data) OR ehr or referral OR ((screen or screening) W/4 ("social determinant" or "social need" or sdoh))) OR ABS("clinical data" or "clinical information" or "clinical record")) AND TITLE-ABS-KEY(((socio or socioeconomic or social or psychosocial) W/0 (risk or stress or vulnerable or vulnerability or disadvantage or status)) or (((socio or socioeconomic or social or psychosocial) W/1 (need or environment)) or "health need") or ((socio or socioeconomic or social or psychosocial) W/2 (determinant or exposure or factor or problem)) or sdoh or hrsn or health equity or patient contextual or "life circumstance" or (socioeconomic W/0 (factor or deprivation or status)) or (socio W/1 (factor or deprivation or status)) or (community and (prevalence or risk)) or (community W/1 (type or context or factor or health*)) or (environment* PRE/0 (factor or design)) or ((community or environment*) W/2 exposure)) AND (TITLE-ABS-KEY(("residence characteristic" or "place of birth" or "birth place" or domicile or "health service area" or "catchment area" or neighbourhood or housing or census or urban or rural or county or counties or household) OR geocode or geocoding or "geo-code" or "geo-coding" OR (geograph W/2 (code or coding or data or divers or distribute or distribution or disparity or inequity or information or measure or measurement or record or residence)) OR ((patient W/0 context) or (contextual W/1 data)) OR ((demograhic W/2 (data or information or record)) or census) OR ((area or community) W/2 (factor or resource or research or impact or level or link or linked or linking or linkage)) OR (social W/2 information)) or ABS((demograhic W/3 factor) OR ((regional or region) W/9 patient) OR ((personal or individual) W/1 (sdoh or "social determinant")) OR "family centered" OR (social W/2 (risk or problem or factor)) OR geography or geographies)) OR (TITLE-ABS-KEY(((socio or socioeconomic or social or psychosocial) W/0 (risk or stress or vulnerable or vulnerability or disadvantage or status)) or (((socio or socioeconomic or social or psychosocial) W/1 (need or environment)) or "health need") or ((socio or socioeconomic or social or psychosocial) W/2 (determinant or exposure or factor or problem)) or sdoh or hrsn or health equity or patient contextual or "life circumstance" or (socioeconomic W/0 (factor or deprivation or status)) or (socio W/1 (factor or deprivation or status)) or (community and (prevalence or risk)) or (community W/1 (type or context or factor or health*)) or (environment* PRE/0 (factor or design)) or ((community or environment*) W/2 exposure)) AND TITLE-ABS-KEY((community W/1 (register or registry or data or record or program or provider)) or "community health center" or "community mental health center"))) AND (TITLE-ABS-KEY(((health or unhealth) PRE/1 behavior) or ((healthy or unhealthy) PRE/1 behavior) or smoking or tobacco or diet or food or nutrition or lifestyle or alcohol or "physical activit" or sedentary or (("social determinant" or SDOH) W/0 (screen or screening))))) OR ((TITLE-ABS-KEY ("chronic care model")) AND (TITLE-ABS-KEY (("residence characteristic" OR "place of birth" OR "birth place" OR domicile OR "health service area" OR "catchment area" OR neighbourhood OR housing OR census OR urban OR rural OR county OR counties OR household) OR geocode OR geocoding OR "geo-code" OR "geo-coding" OR (geograph W/2 (code OR coding OR data OR divers OR distribute OR distribution OR disparity OR inequity OR information OR measure OR measurement OR record OR residence)) OR ((patient W/0 context) OR (contextual W/1 data)) OR ((demograhic W/2 (data OR information OR record)) OR census) OR ((area OR community) W/2 (factor OR resource OR research OR impact OR level OR link OR linked OR linking OR linkage)) OR (social W/2 information)) OR ABS ((demograhic W/3 factor) OR ((regional OR region) W/9 patient) OR ((personal OR individual) W/1 (sdoh OR "social determinant")) OR "family centered" OR (social W/2 (risk OR problem OR factor)) OR geography OR geographies)) AND (TITLE-ABS-KEY (((socio OR socioeconomic OR social OR psychosocial) W/0 (risk OR stress OR vulnerable OR vulnerability OR disadvantage OR status)) OR (((socio OR socioeconomic OR social OR psychosocial) W/1 (need OR environment)) OR "health need") OR ((socio OR socioeconomic OR social OR psychosocial) W/2 (determinant OR exposure OR factor OR problem)) OR sdoh OR hrsn OR health AND equity OR patient AND contextual OR "life circumstance" OR (socioeconomic W/0 (factor OR deprivation OR status)) OR (socio W/1 (factor OR deprivation OR status)) OR (community AND (prevalence OR risk)) OR (community W/1 (type OR context OR factor OR health*)) OR (environment* PRE/0 (factor OR design)) OR ((community OR environment*) W/2 exposure))))

XSL•FO
RenderX

Textbox 2. Inclusion and exclusion criteria.

Inclusion criteria

- Study type: Original empirical research studies.
- Health information systems (HISs) and behavioral interventions: Used HISs to prepare, prescribe, monitor, and/or assess behavioral interventions.
- Neighborhood geo-referenced contexts (NGRCs): Used one or more NGRC attributes about the outside world to assist and/or adjust behavioral interventions.
- NGRCs and behavioral interventions: Studies in which NGRC and intervention data pertain to concurrent or longitudinal periods. Studies in which NGRC attributes and intervention attributes use either objective or subjective measurements.

Exclusion criteria

- Study type: Study protocols without substantive findings and studies without results (eg, experimental descriptions and commentaries).
- HISs: Studies for which it cannot be confirmed whether the records are part of an HIS due to an unspecified format (electronic or paper), or it is not explicitly stated whether these data are connected to HISs. Studies on standalone personal health records managed solely by patients and/or caregivers without the integration of data in medical settings (eg, hospitals). Studies using an HIS that was not focused on interventions but was used as a database to identify participants for empirical studies.
- NGRCs: Studies focusing only on broad geographic categories (eg, county, city, region, nation, and urban/rural), using sensors without specific attributes, NGRCs about health services, or economic factors (eg, transportation reimbursement). Studies for which it cannot be determined whether NGRCs were involved due to a lack of specific details on social determinants of health screening tools and interventions.

Results

Search Term Coverage and Rationale

GRCs include (1) built environments from the 5D dimensions (density, diversity, design, distance, and destination accessibility); (2) natural environments; and (3) social environments, which encompass neighborhood socioeconomic status and neighborhood social relationships. Neighborhoods include (1) static administrative areas, (2) various delineations surrounding each participant's address, (3) perceived neighborhoods, and (4) geographic areas encompassing all locations visited by an individual in their daily activities. For the definitions and examples of NGRCs and neighborhoods, see Table S1 in Multimedia Appendix 1 [1,9,26,52-76]. We used 3 different search strategies because some concepts can be merged and replaced by related ones. For example, HISs and health behaviors are 2 critical elements of the chronic care model, which is a valuable framework for patient empowerment, supporting self-management and enhancing health outcomes [53-56]. Consequently, substituting HISs and health behaviors with search terms pertinent to the chronic care model while connecting them with GRCs and neighborhoods constituted the second part of the search strategy. Additionally, community health facility workers may use HISs to address neighborhood-related issues. Therefore, replacing HISs and neighborhoods with search terms relevant to community health facilities (eg, community health centers and community data) while linking these terms with GRCs and health behaviors formed the third part of our search strategy.

Eligibility Criteria

All original peer-reviewed scientific studies in English indexed up to April 29, 2024, were included, while non–peer-reviewed studies and gray literature were excluded. Textbox 2 outlines the inclusion and exclusion criteria.

Selection Process

First, we screened titles and abstracts and excluded studies according to the inclusion and exclusion criteria. When uncertainty arose, studies were retained for the full-text screening stage. Second, XY and MD independently conducted a full-text review and extracted the relevant NGRCs, HISs, and behavioral interventions from the potentially included papers. Any disagreements were first addressed by referring to a predefined inclusion/exclusion checklist. If consensus was not reached using the checklist, the reviewers discussed each case, and, if necessary, a third reviewer (RC) made the final decision. Although interrater reliability was not calculated, this structured approach ensured consistent and transparent decision-making.

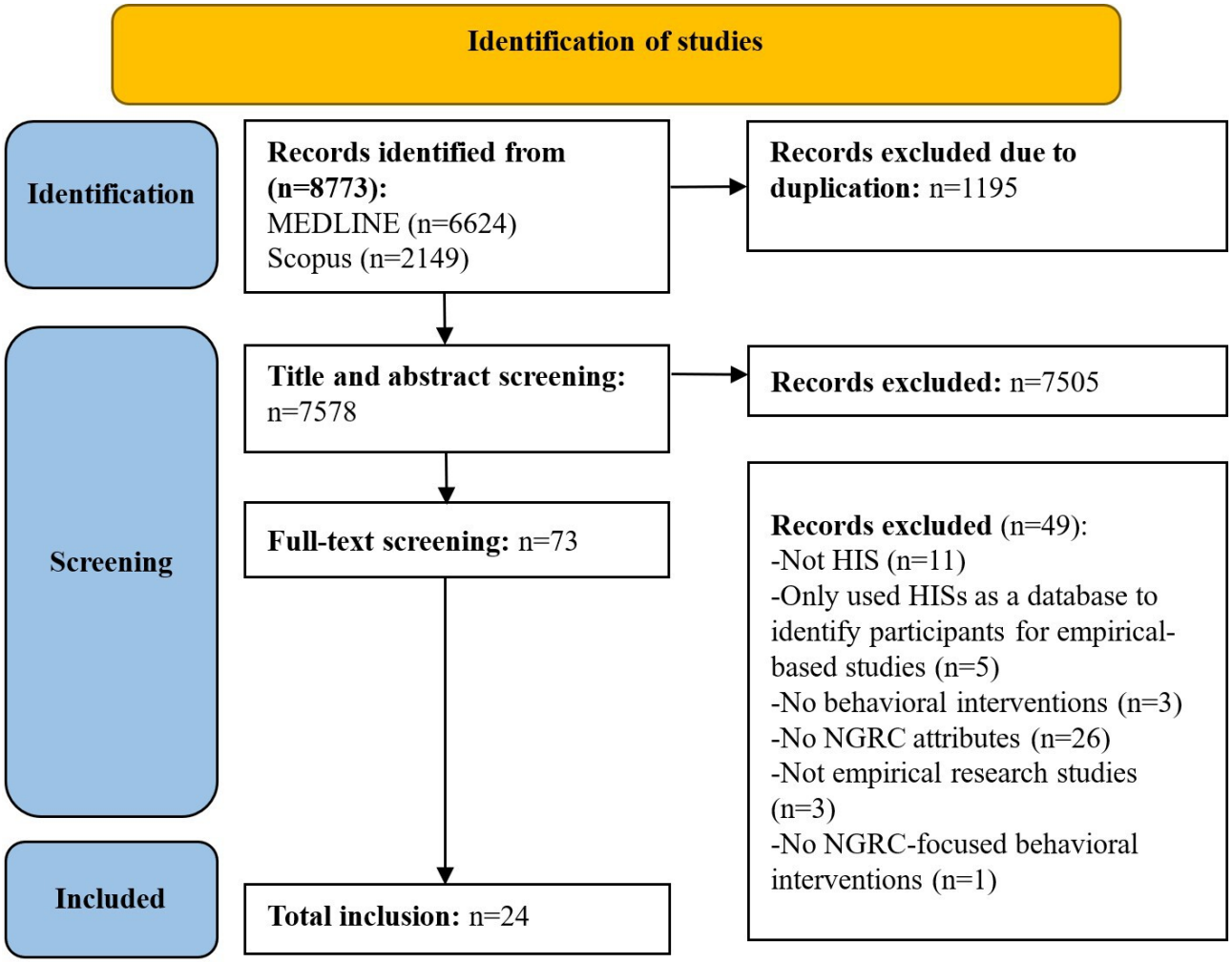
Data Extraction

We extracted data from the included studies using a 7-item table. These seven items included: (1) study attributes; (2) population attributes, diseases, and health setting characteristics; (3) HISs; (4) NGRCs; (5) behavioral interventions; (6) results; and (7) conclusions and future studies. Table S3 in Multimedia Appendix 1 provides further details.

Figure 1 presents the PRISMA flowchart, which includes the identification, screening, and selection process. A total of 24 peer-reviewed studies were finally included.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart. HIS: health information system; NGRC: neighborhood geo-referenced context.



Study Characteristics

We reviewed the descriptive characteristics of all 24 included papers. Basic characteristics of the studies are presented in Table S4 in Multimedia Appendix 1, including information on (1) the type, number, gender, age distributions, and language of participants, and (2) the study type and year of publication. The characteristics with appropriate properties for summarization include (1) the type, NCD, and health and computer literacy of participants, and (2) health settings. All studies were conducted in the United States. Details are provided in Table S5 in Multimedia Appendix 1.

We divided the studies by participant type as follows: 2 studies [29,77] focused solely on providers, 4 studies concentrated on caregivers [78-81], 13 studies examined patients only [10,82-93], and 5 studies involved both health providers and patients [93-97]. Among patients, most studies focused on adults.

Summarization of the Results

Overview

In this section, we aim to summarize the findings from the 24 included articles to address the first 3 research questions. Each of these themes is explored in detail in the subsections that follow.

Question 1: Question Regarding HISs

We summarized whether HISs were used and the types of HISs used. Table 1 presents HIS usage across all stages, while Table S5 in Multimedia Appendix 1 lists the types of HISs. Table S4 in Multimedia Appendix 1 provides the specific names of the HISs.

Table 2 details the stage 1 NGRC data capture, including the settings (eg, health care settings, such as hospitals, and non–health care settings, such as homes), those collecting data (eg, patients, providers, and systems), and those documenting it (systems or providers).

XSL•FO
RenderX

Table . The role of HISs[a] in the entire process of NGRC[b]-focused behavioral interventions, including situations and actors.

Group and study	NGRC data capture (n=18)	Prescription (n=13)				Monitoring (n=2)	Assessment (n=11)
		Digital prescriptions (n=6)		Prescriptions by HPs[c] and documented in HISs (n=7)			
		Tools	Interventions	Types of HPs	Interventions		
Social needs interventions							
Buitron de la Vega et al [82]	HISs[d]	EHR[d,e]	Referral guides[d]	—[f]	—	—	—
Cordova-Ramos et al [78]	HISs[d]	—	—	Nurses[d]	—	—	—
Fortin et al [79]	HISs[d]	—	—	Social workers[d]	EHR-based resource maps[d]	—	—
Lynch et al [94]	HISs[d]	—	—	—	—	Manual documentation[d]	—
Palakshappa et al [95]	HISs[d]	A tablet-based patient portal[d]	CRLs[g] and TIAtHPs[d,h]	—	—	—	—
Percac-Lima et al [85]	HISs[d]	—	—	—	—	—	The percentage of CRC[i] screening[d]
Russell et al [84]	HISs[d]	—	HISs[d]	Nurses[d]	—	—	—
Wallace et al [86]	—	—	—	—	—	—	HCUD[d,j]
Wright et al [96]	HISs[d]	—	HISs[d]	Social workers[d]	—	—	—
Wu et al [97]	HISs[d]	—	—	—	—	—	HCUD and medication adherence[d]
Roth et al [88]	—	—	—	—	—	—	HCUD, HbA_{1c} levels, and blood pressure[d]
Tanumihardjo et al [87]	HISs[d]	—	—	—	—	—	HbA_{1c}, LDL[k], and blood pressure[d]
Holt et al [29]	HISs[d]	Web-based PatientWisdom patient portal[d]	Resource linkages and TIAtHPs[d]	—	—	—	—
Bechtel et al [89]	—	—	—	—	—	—	HCUD[d]
Taber et al [90]	HISs[d]	—	—	—	—	—	—
Stark et al [80]	HISs[d]	Healthy Planet app[d]	CRLs[d]	—	—	—	—
Gupta et al [91]	HISs[d]	NowPow embedded in EMR[d,l]	Resource linkages[d]	—	—	Automatic documentation[d]	HCUD[d]
Britto et al [92]	HISs[d]	—	—	Researchers[d]	HISs[d]	—	The percentage of patients with well-controlled asthma[d]

XSL•FO
RenderX

Group and study	NGRC data capture (n=18)	Prescription (n=13)				Monitoring (n=2)	Assessment (n=11)
		Digital prescriptions (n=6)		Prescriptions by HPs[c] and documented in HISs (n=7)			
		Tools	Interventions	Types of HPs	Interventions		
Glasser et al [81]		EHR-integrated CommunityRx[d]	Resource linkages[d]	—	HISs[d]	—	—
Loo et al [93]	HISs[d]	—	—	EDCS[d,m]	HISs[d]	—	—
Percac-Lima et al [98]	HISs[d]	—	—	Navigators[d]	HISs[d]	—	HCUD[d]
Lifestyle change interventions							
Jilcott et al [10]	—	—	—	—	—	—	HCUD, blood lipids, glucose, pressure, BMI, and body composition[d]
Johnston et al [83]	—	—	—	—	—	—	HCUD[d]
The entire care process							
Kane et al [77]	HISs[d]	—	—	—	—	—	—

[a]HIS: health information system.

[b]NGRC: neighborhood geo-referenced context.

[c]HP: health provider.

[d]HISs have functioned in this category.

[e]EHR: electronic health record.

[f]Not applicable.

[g]CRL: community resource list.

[h]TIAtHP: timely intervention alert to health provider.

[i]CRC: colorectal cancer.

[j]HCUD: health care utilization data.

[k]LDL: low-density lipoprotein.

[l]EMR: emergency medical record.

[m]EDCS: emergency department clinical staff.

XSL•FO
RenderX

Table . The role of HISs[a] in NGRC[b] data capture.

Group and study	NGRC preparation (n=18)								
	H/NH[c]	NGRC data collection (n=18)						NGRC data documentation (n=18)	
		Patient-administered electronic collection (n=5)		Patient-administered paper-based collection (HPs[d] to assist if needed) (n=2)	HP-administered collection (n=10)		System-administered collection (n=1)	System (n=5)	HPs (n=13)
		Tools	Types of HPs		V/E[e]	Types of HPs			
Social needs interventions									
Buitron de la Vega et al [82]	H	—[f]	—	Medical assistants[g]	—	—	—	—	Medical assistants[g]
Cordova-Ramos et al [78]	H	—	—	—	V[g]	Nurses[g]	—	—	Nurses[g]
Fortin et al [79]	H	REDCap[g]	—	—	—	—	—	—	CTAs[g,h]
Lynch et al [94]	H	—	—	—	E[g]	Medical assistants[g]	—	—	Medical assistants[g]
Palakshappa et al [95]	H	Tablet-based patient portal[g]	—	—	—	—	—	HISs[g]	—
Percac-Lima et al [85]	H	—	—	—	V[g]	Navigators[g]	—	HISs[g]	Navigators[g]
Russell et al [84]	H	eScreening platform[g]	CTAs[g]	—	—	—	—	—	—
Wright et al [96]	H	—	—	Nurses[g]	—	—	—	—	Nurses[g]
Wu et al [97]	H	—	—	—	V[g]	Pharmacists[g]	—	—	Pharmacists[g]
Tanumihardjo et al [87]	H	—	—	—	V[g]	CHWs[g,i]	—	—	CHWs[g]
Holt et al [29]	NH	Web-based PatientWisdom patient portal[g]	—	—	—	—	—	HISs[g]	—
Taber et al [90]	H	—	—	—	V[g]	Pharmacists, nurses, social workers, and CHWs[g]	—	—	Pharmacists, nurses, social workers, and CHWs[g]
Stark et al [80]	NH	Healthy Planet app[g]	—	—	—	—	—	HISs[g]	—
Gupta et al [91]	H	—	—	—	E[g]	Without specificity[g]	—	—	Without specificity[g]
Britto et al [92]	H	—	—	—	V[g]	Researchers[g]	—	—	Researchers[g]

XSL•FO
RenderX

Group and study	NGRC preparation (n=18)								
	H/NH[c]	NGRC data collection (n=18)						NGRC data documentation (n=18)	
		Patient-administered electronic collection (n=5)		Patient-administered paper-based collection (HPs[d] to assist if needed) (n=2)	HP-administered collection (n=10)		System-administered collection (n=1)	System (n=5)	HPs (n=13)
		Tools	Types of HPs		V/E[e]	Types of HPs			
Loo et al [93]	H	—	—	—	V[g]	Nurses and clinicians[g]	—	—	Nurses[g]
Percac-Lima et al [85]	H	—	—	—	V[g]	Navigators[g]	—	—	Navigators[g]
Entire care process									
Kane et al [77]	H	—	—	—	—	—	Publicly available datasets[g]	HISs[g]	—

[a]HIS: health information system.

[b]NGRC: neighborhood geo-referenced context.

[c]H/NH: physical health settings (H), and non–health care settings (NH).

[d]HP: health provider.

[e]V/E: verbal surveys embedded in electronic health records (EHRs) or electronic medical records (EMRs) (E), or surveys not embedded in EHRs or EMRs (V).

[f]Not applicable.

[g]HISs have functioned in this category.

[h]CTA: care team assistant.

[i]CHW: community health worker.

Question 2: Question Regarding NGRCs

We summarized the reviewed studies in terms of behavioral intervention types as follows: (1) social needs interventions, (2) lifestyle change interventions, and (3) whole care without specifying the behavioral intervention type. This categorization agrees with the information in Table S1 in Multimedia Appendix 1. Table 3 summarizes the types and spatiotemporal characteristics of NGRC attributes.

XSL•FO
RenderX

Table . The types and spatiotemporal characteristics of NGRCs[a] related to social needs interventions, lifestyle interventions, and the entire care process.

Group and NGRC	B/N/S[b]	Neighborhood type (address)	Data source[c]	SDoH[d] screening	Ob/Su[e]	I/O[f]	Li/So/E[g]	Reference
Social needs interventions (n=21)								
Transportation barriers (inpatient)	B	PNs[h]	Sc	Yes	Su	I	So	[78,82,84,92,94,96,98]
Transportation barriers (outpatient)	B	PNs	Sc	Yes	Su	O	So	[29,80,81,93,97]
Neighborhood safety (individuals' fears and feelings of insecurity in their environment)	S	PNs	Sc	Yes	Su	I	So	[90]
Lifestyle interventions (n=2)								
Design: sidewalk/bike lane shortage	B	PNs (RDs[i])	—[j]	Yes	Su	I	So	[10]
Density: intensive fast foods, insufficient healthy food options (eg, food stores, restaurants, farmer's markets, and fruit stands), and healthy breakfast access	B	PNs (RDs)	—	Yes	Su	I	So	[10]
Density: recreational facility scarcity for walking (eg, parks, gyms, and tracks), affordable exercise facilities, and suitable physical activity programs	B	PNs (RDs)	—	Yes	Su	I	So	[10]
Traffic	B	PNs (RDs)	—	Yes	Su	I	So	[10]
Design: lack of streetlights	B	PNs (RDs)	—	Yes	Su	I	So	[10]
Distance to a high-risk location, such as a bar or supplier the individual used to frequent	B	Daily routes	L	No	Ob	O	Li	[83]

XSL•FO
RenderX

Group and NGRC	B/N/S[b]	Neighborhood type (address)	Data source[c]	SDoH[d] screening	Ob/Su[e]	I/O[f]	Li/So/E[g]	Reference
Living in an environment where alcohol use is prevalent or normalized	S	PNs (RDs)	Sc	No	Su	O	Li	[83]
Community crime	S	PNs (RDs)	Sc	No	Su	I	So	[10]
Verbal abuse from people on the street and speeding drivers	S	PNs (RDs)	Sc	No	Su	I	So	[10]
Air pollution from cars or factories	N	PNs (RDs)	Sc	No	Su	I	So	[10]
The entire care process (n=1)								
Food desert status (yes/no) and low food access census tract (yes/no)	B	Census tracts (RDs)	P	No	Ob	I	E	[77]
Location type (urban/rural)	B	Census tracts (RDs)	P	No	Ob	I	E	[77]
Total population, low-income population, children aged 0-17 years, seniors aged 65 years or older, and housing units lacking vehicle access among populations with limited access, defined by how far the closest supermarket is (0.5, 1, 10, and 20 miles away)	S	Census tracts (RDs)	P	No	Ob	I	E	[77]

XSL•FO
RenderX

Group and NGRC	B/N/S[b]	Neighborhood type (address)	Data source[c]	SDoH[d] screening	Ob/Su[e]	I/O[f]	Li/So/E[g]	Reference
Community-level socioeconomic and demographic data, including age, race, ethnicity, nationality distribution, median age, median family and household income, education level distribution, median home value and gross rent, renter-occupied household distribution, total population, total housing units, and low-income census tract (yes/no)	S	Census tracts (RDs)	P	No	Ob	I	E	[77]

[a]NGRC: neighborhood geo-referenced context.

[b]B/N/S: built environment (B), natural environment (N), or social environment (S).

[c]Source: NGRC screening (Sc), location-aware devices (L), or public datasets (P).

[d]SDoH: social determinants of health.

[e]Ob/Su: objective (Ob) or subjective (Su).

[f]I/O: inpatient (I) or outpatient (O).

[g]Li/So/E: lifestyle change interventions (Li), social needs interventions (So), or entire care process (E).

[h]PN: perceived neighborhood.

[i]RD: residential address.

[j]Not available.

We specified substructures for the categorization of social needs interventions, lifestyle change interventions, and whole care. Social needs interventions used actively collected, subjective, and static measures of perceived neighborhood social and built environments. Lifestyle change interventions included (1) actively collected, subjective, and static measures of perceived neighborhood social, natural, and built environments at residential addresses, and (2) passively collected, objective, real-time, and location-based distances to high-risk areas (eg, bars). Whole care included passively collected, objective, and static social and built environment indicators derived from the census tract of residential addresses.

In social needs interventions, actively collected, subjective, and static measures of perceived transportation barriers and neighborhood safety were the only 2 assessed NGRCs (21/24, 88%). These were assessed using electronic or paper-based surveys [29,78-82,84-98]. The focus on transportation barriers (21/24, 88%) may reflect their relative ease of mitigation through interventions (eg, door-to-door transportation to medical appointments [8,92]). These interventions were shaped by specific health service settings and available community resources. Similarly, transportation barriers were assessed using a variety of question formats, time frames, destination types, and SDoH screening tools. Despite these efforts, detailed information, such as transport mode, distance, travel time, and the locations of perceived neighborhoods, was generally rarely collected. Additional details are presented in Table 4.

XSL•FO
RenderX

Table . Additional details regarding transportation barriers.

Questions about NGRCs[a]	Timespan	Survey frequency	SDoH[b] screening tools	Reference
Transportation barriers to attending medical appointments (eg, CRC[c] screening) or obtaining medications (n=10)				
Do you have trouble getting transportation to medical appointments?	—[d]	—	Tool for Health and Resilience In Vulnerable Environments Social Determinants of Health (THRIVE)	[82,93]
Do you have trouble with transportation to or from NICUs[e]? For example, paying for parking or getting a ride to the hospital.	—	—	Tool for Health and Resilience In Vulnerable Environments Social Determinants of Health (THRIVE)	[78]
Query regarding transportation barriers to CRC screening	—	—	—	[85]
Query regarding transportation barriers to CT[f] screening tests (eg, lung cancer and chest)	—	—	—	[98]
Query regarding transportation to appointments	—	—	—	[92]
Have you missed any medical appointments because of difficulties with transportation?	Past 6 and 12 months	Annually	Health Risk Assessment (HRA)	[87]
Have you ever not seen a doctor because you did not have a way to get to the clinic or hospital?	Past 12 months	—	HealthLeads	[86]
Has a lack of transportation kept you from medical appointments or from getting medications? (query 1)	Past 12 months	—	Intensifying Community Referrals for Health (SINCERE)	[80]
Has a lack of transportation kept you from medical appointments or from getting medications? (query 2)	Past 3 months	—	Monteliore's 10-question SDoH survey	[91]
Transportation barriers to attending medical appointments, meetings, or work, or getting things needed for daily living (n=9)				
Has a lack of transportation kept you from medical appointments, meetings, or work, or from getting things needed for daily living? (query 1)	Past 12 months	—	Health-Related Social Needs (HRSN)	[79,95]
Has a lack of transportation kept you from medical appointments, meetings, or work, or from getting things needed for daily living? (query 2)	Past 12 months	—	Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE)	[87]
Has a lack of transportation kept you from medical appointments, meetings, or work, or from getting things needed for daily living? (query 3)	—	—	Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE)	[88,94]
Has a lack of transportation kept you from medical appointments, meetings, or work, or from getting things needed for daily living? (query 4)	—	—	Assessing Circumstances and Offering Resources for Needs	[84]
Has a lack of transportation kept you from work, attending medical appointments including dialysis treatments, or getting things you need for your daily living?	—	Monthly [96]; — [89]	The Core 5 Determinants of Health Screening Tool	[89,96]
Query regarding challenges in finding affordable, reliable, or handicap-accessible transportation to obtain resources such as medical care, food, or medications	—	—	—	[90]
Transportation barriers without specifying which destinations were affected (n=3)				
Query regarding transport	—	—	—	[29,81]

XSL•FO
RenderX

Questions about NGRCs[a]	Timespan	Survey frequency	SDoH[b] screening tools	Reference
By the way, we have this new resource team that can help with things not related to health care, like food, housing, and transportation. Is this something you want to talk about?	—	Weekly	—	[97]

[a]NGRC: neighborhood geo-referenced context.

[b]SDoH: social determinants of health.

[c]CRC: colorectal cancer.

[d]Information not available.

[e]NICU: neonatal intensive care unit.

[f]CT: computed tomography.

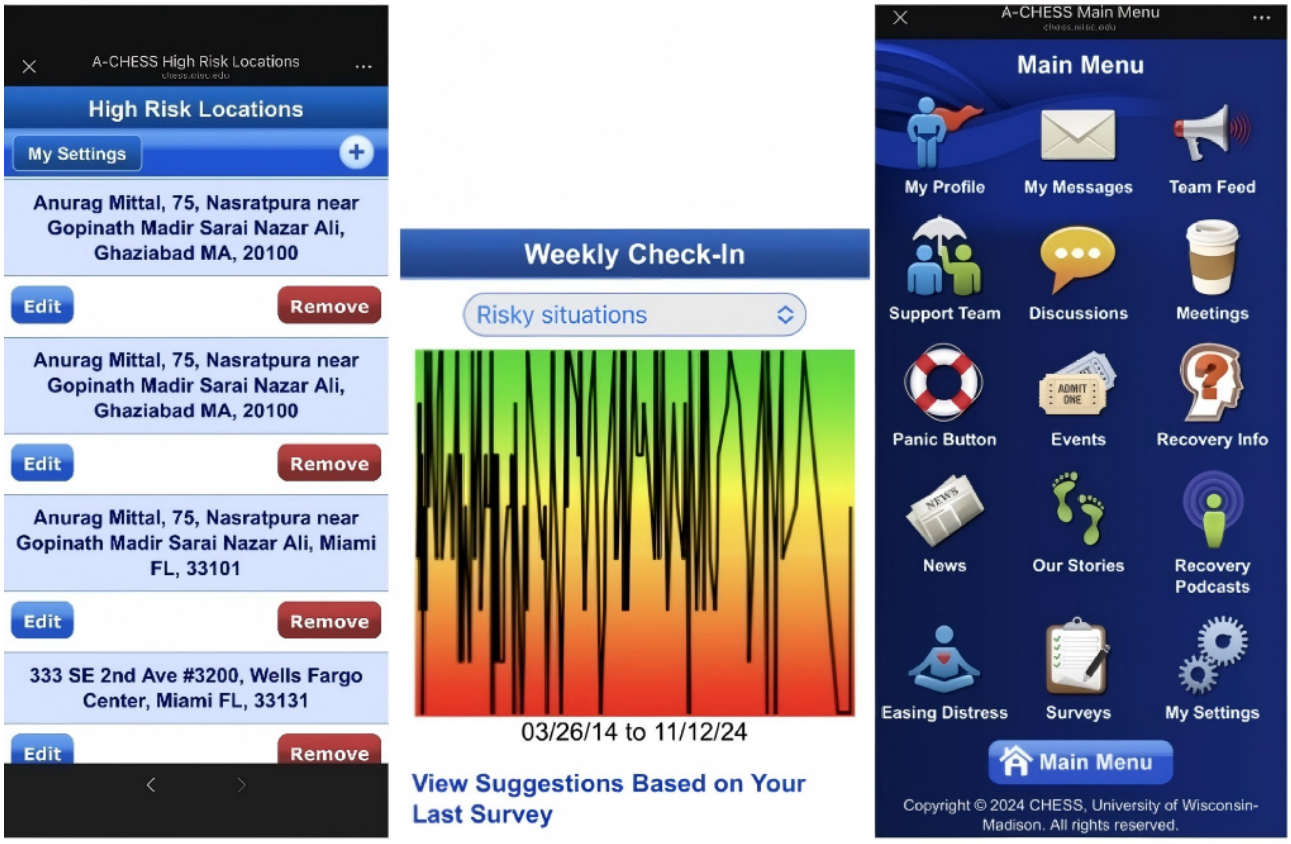
Lifestyle change interventions assessed actively collected, subjective, and static measures of social, natural, and built environments at participants' perceived residential neighborhoods. These were assessed with "neighborhood assessment" during follow-up visits [10]. These assessments showed the multifaceted interplay between NGRCs and lifestyle behaviors. The environmental barriers to healthy eating and physical activity included air quality, food environment density, physical activity facility density, streetlight design, and neighborhood safety (Table 3). The absence of subsequent similar studies highlights the practical challenges of consistently partnering with community groups and agencies to address environmental barriers to lifestyle change.

In addition, passively collected, objective, real-time, and location-based distances to high-risk areas (eg, bars) were assessed using GPS data from the A-CHESS (Addiction Comprehensive Health Enhancement Support System) app [83] (Figure 2). Unlike approaches that depend on community-level resources of residential addresses, A-CHESS optimizes users' daily routes. This suggests that leveraging existing environments along these routes—not just around residential areas—can be practical without requiring extensive organizational collaboration. A-CHESS also supports weekly surveys to actively assess participants' perceived alcohol use in their residential neighborhoods [83]. While transportation barriers are relatively stable, neighborhood social cultures, such as drinking behaviors, fluctuate rapidly, requiring active and higher-resolution data collection to capture their health effects.

In whole care, passively collected, objective, and static social and built environment indicators based on the census tract level of the residential address were displayed in the Envirome Web Service [77]. Integrated with the Cerner Millennium EHR, this service displays socioeconomic data, food desert status, and contextual information at the census tract level using public datasets (eg, the US Department of Agriculture and the 2010 Decennial Census). Compared with A-CHESS, the Envirome Web Service offers broader, population-level coverage of social and built environment factors through passive data collection but lacks dynamic, route-specific, or nonresidential contextual information.

XSL•FO
RenderX

Figure 2. The A-CHESS (Addiction Comprehensive Health Enhancement Support System) app [83] supports recovery from alcohol use disorder by providing 24/7 smartphone access to services such as cognitive-behavioral therapy, addiction-related resources, and alcohol-free events. The left panel displays a list of identified high-risk locations, the middle panel summarizes risky situations based on weekly surveys, and the right panel shows the app's homepage. Using GPS, A-CHESS monitors when users approach predesignated high-risk locations, such as bars, and sends an alert asking whether they intend to be there. No alert is issued if the user is not near a high-risk location. The screenshots were captured from the app link [99].



Question 3: Question Regarding Stakeholders' Insights

We summarized stakeholders' insights on the use of NGRCs and/or HISs in social needs interventions and whole care from 10 included studies. The details are shown in Table S6 in Multimedia Appendix 1.

Seven studies mentioned attitudes toward NGRC integration. Six out of the seven studies on social needs interventions found that patients, caregivers, and providers valued addressing transportation barriers for comfort and convenience [84,93,96,97]. However, challenges remain in addressing patients' concerns about potential discrimination, such as reduced courtesy or respect, as well as providers' discomfort with stigmatizing questions [81,86,93,96,97]. At the core, the challenge lies in balancing actionable social needs interventions and sensitive care. The last study [77] examined providers' views on NGRC data needs, clinical usefulness, and data visualizations in whole care. Providers emphasized the need for clear definitions and simple visualizations to make complex data more actionable in practice [77].

Nine studies examined HIS integrations from 3 aspects: value and convenience, patient autonomy, and privacy. The efficiency and convenience of electronic NGRC capture tools were recognized by providers, patients, and caregivers, particularly when patients completed them before visits [29,77,86,95-97]. However, providers also expressed concerns about increased workload, time constraints, unfitness to the workflow, and the possibility of raising unmet patient expectations [29,86,94-96]. These concerns underscore the need to integrate NGRC-focused behavioral interventions seamlessly into clinical workflows to avoid a higher administrative burden. Patient autonomy and engagement were enhanced [77,86,95-97]. Improved autonomy can support person-centered and team-based care. However, the absence of data on patients' digital literacy could lead to misinterpretation of patients' attitudes (eg, acceptability, adoption, and usage of HISs), especially for those with low literacy levels, as higher digital literacy can alleviate social concerns [29,84]. Privacy was improved by reducing patients' feelings of judgment and promoting greater honesty in disclosure [84,86,95]. Improved privacy suggests that well-designed digital systems can mitigate stigma associated with sensitive NGRCs.

Discussion

Principal Findings and Recommendations

In this scoping review, we reviewed the current state of NGRC-focused behavioral interventions using HISs in order to propose recommendations for future studies. We included 24 studies, and nearly half of the studies (11/24, 46%) were conducted in 2023 or later (Table S4 in Multimedia Appendix 1). Due to unstructured reporting in many of the included studies, reliably extracting data was difficult. We organized our findings and recommendations according to the 4 research questions.

XSL•FO
RenderX

Question 1: Question Regarding HISs

The predominant reliance on provider-centric EHRs/EMRs and the limited and inconsistently described use of HISs across the entire NGRC-focused behavioral intervention cycle significantly constrain the integration of NGRCs into behavioral interventions. Apart from the unspecified type of HIS in 1 study [10], most included studies relied on provider-centric EHRs/EMRs (20/24, 83%), with limited use of more patient-empowered PHRs. Only 3 studies used tethered PHRs, and none utilized interconnected PHRs (Table S5 in Multimedia Appendix 1). The reliance on provider-centric EHRs/EMRs may lead to difficulties in monitoring and assessment. Therefore, most studies (23/24, 96%) did not apply HISs in the entire process (NGRC data capture, intervention prescription, monitoring, and assessment). HISs were mainly used for (1) NGRC data capture (18/24, 75%), with limited patient-led (5/24, 21%) or system-led (1/24, 4%) cases; (2) intervention prescriptions (13/24, 54%); and (3) assessments (11/24, 46%) focusing on health care utilization, with minimal use in monitoring (2/24, 8%). Limited monitoring and a narrow focus on health care utilization data for assessments leave the sustainability of behavior change largely unaddressed. Given its critical role in improving long-term health outcomes, future research should aim to establish integrated frameworks that embed tethered or interconnected PHRs consistently across the 4 phases. Furthermore, the lack of details on HIS functionalities limits the ability to synthesize evidence and derive practical design recommendations for future interventions. Future studies should provide sufficient details to ensure reproducibility.

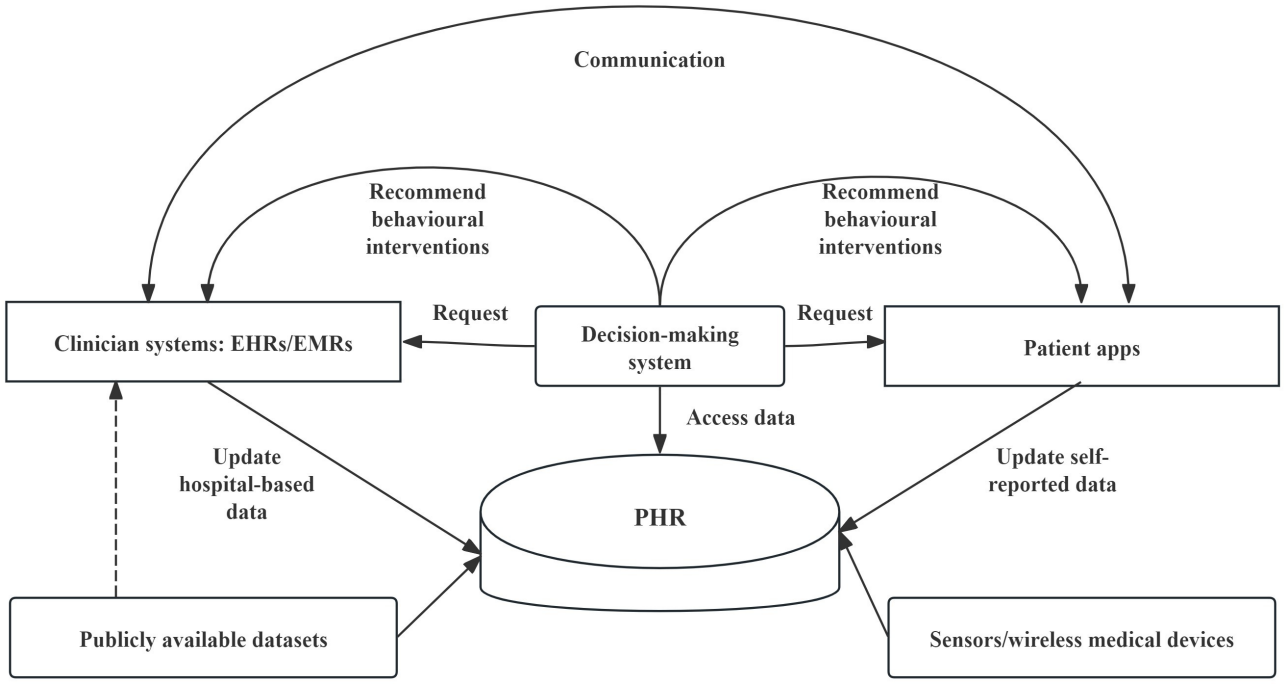
Although HISs have been used frequently in NGRC data capture, the dominance of active collection (eg, patient-administered or provider-administered surveys; 22/24, 92%) and the underuse of passive collection (eg, linkage with public datasets; 1/24, 4%) and GPS data (1/24, 4%) have left existing data underutilized, and active collection may cause user information to be unclear. On one hand, the majority of studies (17/24, 71%) relied on manual NGRC collection and documentation into EHRs/EMRs by stakeholders (Table S6 in Multimedia Appendix 1). Active data collection likely contributes to negative provider attitudes due to increased workload and limited consultation time. This potentially hinders the integration of NGRCs into routine care. Self-completed assessments using tethered or interconnected PHRs may address providers' concerns, and both patients and providers hold positive attitudes toward them (Table S6 in Multimedia Appendix 1). However, tethered or interconnected PHRs were underutilized in the included studies. The gap between positive stakeholder attitudes and practice suggests missed opportunities to reduce provider burden and engage patients more actively in their care using PHRs. On the other hand, although the A-CHESS app achieved passive data collection, it relied solely on HISs as assessment tools and had limited provider access to behavioral data, which constrained the understanding of how HISs and providers influence behavior change and its sustainability. Similarly, the Envirome Web Service [77] passively collected NGRC data via public dataset linkages. However, the lack of detailed descriptions of the use of HISs in subsequent steps (eg, monitoring and assessment) further limited insights into their role in supporting sustained behavior change.

The complexity of provider types and roles in the entire process of NGRC-focused behavioral interventions (Table 1) further complicates implementation in clinical settings. Future studies should clearly define the responsibilities of every provider to enhance coordination across care teams, thereby improving intervention fidelity and consistency.

In summary, implementing tethered or interconnected PHRs throughout the intervention process, particularly in outpatient settings, is recommended (Figure 3). Connection with EHRs/EMRs can also achieve provider support. By ensuring privacy, security, and controlled access, PHRs can enhance user acceptance and usage [100]. Secure data exchange among patient apps, public datasets, sensors, and EMRs/EHRs can enable tailored interventions, timely issue detection, improved communication, greater engagement and adherence, better care coordination, higher satisfaction, and reduced treatment burden [86,101]. We have summarized 3 recommendations in Textbox 3.

XSL•FO
RenderX

Figure 3. Tethered and interconnected advantages of PHRs in NGRC-focused behavioral interventions. The image has been modified [3]. EHR: electronic health record; EMR: electronic medical record; NGRC: neighborhood geo-referenced context; PHR: personal health record.



XSL•FO
RenderX

Textbox 3. Recommendations.

Recommendations regarding the types and roles of health information systems (HISs) in the entire process of neighborhood geo-referenced context (NGRC)–focused behavioral interventions

1. Consider tethered or interconnected personal health records (PHRs), especially in outpatient settings, in the entire process of NGRC-focused behavioral interventions.
    - Use tethered or interconnected PHRs to actively or passively collect multiple NGRCs and generate a more comprehensive view of patients' NGRCs [84]. PHRs can passively obtain location-related NGRCs through connections with electronic health records (EHRs)/electronic medical records (EMRs), public datasets (eg, government data and cohorts), sensors, and wireless medical devices. The locations can include not only the residential address but also regularly visited places or GPS-derived data. On the other hand, PHRs can also embed patient- or caregiver-completed surveys to actively collect perceived NGRCs. Determining the optimal timing and frequency of data capture is essential to balance user burden and data quality [82,84,102,103].
    - Clinical decision support systems (CDSSs) can use NGRC data from PHRs to personalize behavioral interventions, delivering recommendations to patients via apps and to providers via EHRs/EMRs. For instance, real-time CDSS alerts sent at optimal intervals can support providers in adjusting NGRC-focused interventions.
    - Consider integrating PHRs with other portable devices (eg, patient apps, sensors, and wireless medical devices) to allow real-time monitoring through surveys, GPS, and sensors.
    - Integrating PHRs with EHRs/EMRs and other portable devices can provide sufficient long-term follow-up data. CDSSs can then use these data to assess and adapt NGRC-focused behavioral interventions [46,85,104], allowing the evaluation of the sustainability of outcomes and the identification of interventions that produce lasting effects.
2. Incorporate the support of health care providers throughout NGRC-focused behavioral interventions. This support can be provided through (1) providers' guidance on CDSS recommendations and (2) direct patient-provider communication.
    - Explore the roles of various health care providers in the entire process of NGRC-focused behavioral interventions.
3. Balance active and passive NGRC data collection to achieve higher acceptability, feasibility, and effectiveness of NGRC-focused behavioral interventions [24,84,86].

Recommendations regarding the types, spatiotemporal characteristics, and datasets of NGRCs in each type of behavioral intervention

1. Accommodate NGRC-focused social needs interventions with lifestyle change interventions.
2. Include a broader range of NGRCs with detailed spatiotemporal information for different behavioral interventions.
    - Expand the focus beyond influential built environments to also include key natural and social environments for specific behaviors.
    - Incorporate a comprehensive 5D (density, diversity, design, distance, and destination accessibility) framework of built environments.
    - Optimize NGRC spatiotemporal resolution using an evolutionary approach. Apply high-resolution or real-time data when available for precise interventions; otherwise, use static or coarse data (eg, administrative datasets) to capture neighborhood influences. Combining dynamic and static NGRCs can mitigate challenges such as the modifiable areal unit problem, the uncertain geographic context problem, and neighborhood effect averaging.
    - Prioritize key subjective and objective NGRCs for specific behaviors (eg, cycling and walking) and standardize their definitions, spatiotemporal attributes, and measurements through multidisciplinary collaboration. Such efforts should leverage and expand existing terminologies (eg, "community details" in Systematized Nomenclature of Medicine—Clinical Terms [SNOMED CT] and Logical Observation Identifiers Names and Codes [LOINC]) and ontologies (Social Determinants of Health Ontology [SOHO] [105] and social determinants of health ontology [SDoHO] [106]) [24].
    - Assess the availability of influential objective NGRCs in existing datasets and explore opportunities to collect missing objective and subjective data through passive methods (eg, sensors). When certain NGRCs cannot be captured passively, develop self-administered questionnaires to obtain the remaining information. Finally, evaluate the quality, usability, and improvement potential of patient-completed data.
3. Address both technical and social considerations (eg, ethical issues) in NGRC data collection [92].

Recommendations regarding stakeholders' insights on the use of HISs and/or NGRCs in behavioral interventions

1. Assess stakeholder acceptance, attitude, and preferred interactive features for various NGRC-focused behavioral interventions using HISs, particularly those targeting lifestyle changes. For example, evaluate patients' willingness and desired features in PHR-based self-management tools, and design EHR/EMR features that fit providers' workflows to mitigate concerns like time constraints and workload.
2. Develop and iteratively test efficient, user-friendly data visualization tools for stakeholders [33,107-111].
3. Explore stakeholders' perspectives on integrating diverse and/or real-time NGRCs into behavioral interventions, extending beyond transportation barriers to broader applications. Future research should identify which NGRCs stakeholders consider most actionable and relevant for specific interventions, and determine effective, nonstigmatizing data collection methods. For example, assess patient willingness to share GPS data when they can control access to their information.
4. Consider stakeholders' computer and health literacy when evaluating their attitudes toward NGRC-focused behavioral interventions using HISs.

XSL•FO
RenderX

5. Iteratively build support systems, responding to negative feedback by aligning patient expectations and addressing stakeholders' social concerns.

Recommendations regarding research designs

1. Adopt standardized and detailed descriptions of NGRCs [24], HIS functionalities, intervention designs, and relevant disease contexts—ideally guided by common metadata standards or consensus-based frameworks—to enhance transparency, comparability, and collaboration across studies.
2. Evaluate the feasibility, acceptability, adoption, sustainability, effectiveness, cost-effectiveness, ethics issues, and mechanisms of change of NGRC-focused behavioral interventions using HISs, with detailed descriptions of research design for reproduction, testing, and comparison. For example, in research design, the only difference between the intervention and control groups should be whether the intervention is NGRC-focused, in order to isolate the effect of NGRCs.
3. Initiate comparative studies examining how different NGRCs and HIS functionalities influence specific intervention effectiveness, which would further inform the design of evidence-based interventions.

Generalizability

1. Explore the unique challenges and context-specific solutions relevant to countries outside the United States.

Question 2: Question Regarding NGRCs

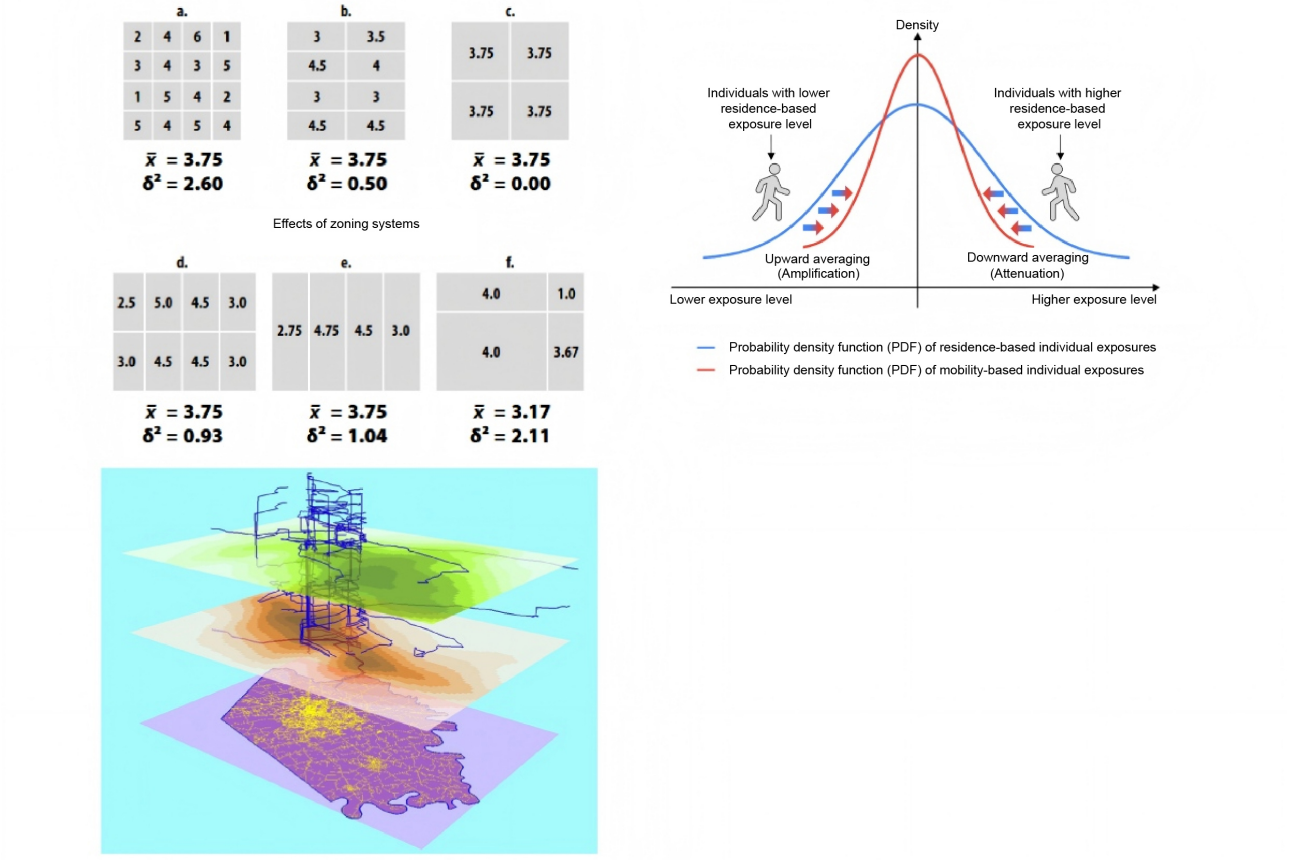
The dominance of structured, measurable social needs interventions in the involved studies likely reflects that health care prioritizes short-term outcomes. In contrast, long-term lifestyle change interventions are overlooked. This may be due to operational challenges, including reimbursement constraints, lower adherence, long-term evaluation, etc [112-115]. However, lifestyle change interventions are crucial for preventing NCDs. To raise attention on lifestyle change interventions, incorporating NGRCs is beneficial for improving adherence and effectiveness. Future studies should focus on leveraging NGRC data to design and evaluate personalized, NGRC-focused lifestyle interventions aiming at advancing patient-centered and preventive care.

Current research has largely overlooked NGRCs beyond transportation barriers within the "destination accessibility" dimension of built environments. Neglecting other 4D elements, as well as social (eg, community crime, neighborhood safety, and socioeconomic data; included in 4 studies) and natural environments (specifically air pollution; included in 1 study), may limit our understanding of how specific NGRCs can inform behavioral interventions. Within the built environment domain, the exploration of "density," "distance," and "design" in the "5D" framework was limited, and "diversity" was not addressed. Future studies should systematically explore other less-studied NGRCs, such as social, natural, and other built NGRCs, to inform more holistic solutions of behavioral interventions using HISs.

Insufficient attention to objective, dynamic, and/or high-resolution NGRCs has left several key methodological challenges largely unaddressed. These include the modifiable areal unit problem, uncertain geographic context problem, neighborhood effect averaging problem, and problem of temporal variations in NGRCs (Figure 4) [60,116-120]. One likely reason is that 22 studies (92%) relied on actively collected, subjective, and static NGRC data, which require substantial participant effort and are therefore difficult to scale for capturing dynamic and/or high-resolution NGRCs. Passive data collection methods (eg, GPS tracking and linkage with public datasets) could help overcome these limitations by enabling more dynamic and/or high-resolution NGRC measurement. However, passive approaches alone cannot capture all real-time and/or high-resolution NGRCs, particularly those related to individual perceptions. Therefore, active data collection remains essential for assessing perceived NGRCs. While some studies have focused exclusively on perceived measures, others have underscored the added value of combining subjective and objective NGRCs to more comprehensively understand their associations with behavioral outcomes [121-124]. When collecting perceived dynamic and/or high-resolution NGRCs, issues, such as user vagueness and the quality of self-reported data, must be carefully considered. Future work should prioritize developing optimized and balanced approaches that integrate active and passive data collection, thereby leveraging dynamic and/or high-resolution NGRC measures to advance behavioral intervention research.

XSL•FO
RenderX

Figure 4. The top-left figure displays the modifiable areal unit problem. The mean value does not change with aggregation, but the variance declines [118,119]. The top-right figure presents how the neighborhood effect averaging problem operates in individual air pollution exposures [116,120]. The bottom figure shows the uncertain geographic context problem, which arises because of the spatial uncertainty in the actual areas that exert contextual influences on the individuals being studied, and the temporal uncertainty in the timing and duration in which individuals experienced these contextual influences [60,117].



Furthermore, all involved studies illustrated the ongoing challenges of balancing granularity and comprehensiveness in assessing NGRCs. On the one hand, actively collected, subjective NGRC measures provide individualized insights but are limited in granularity and comprehensiveness. On the other hand, passively collected, objective indicators offer broad coverage but often lack individual insights and granularity, potentially due to existing dataset constraints. This trade-off underscores the need for integrated approaches that combine active and passive data to capture the complex, dynamic interplay between NGRCs and health behaviors.

To improve the granularity and comprehensiveness of NGRCs in various behavioral interventions, especially lifestyle change interventions, HISs—preferably tethered or interconnected PHRs in outpatient settings—should be considered. We have summarized 3 recommendations in Textbox 3.

Question 3: Question Regarding Stakeholders' Insights

Improved perceived value, convenience, patient autonomy, and privacy were consistently reported as outcomes of integrating technological (HISs) and content (NGRCs) innovations into behavioral interventions in all 10 included studies. On the other hand, social concerns persisted, including provider workload and the potential for discrimination. These benefits and concerns could be influenced by participants' computer literacy, which was rarely reported in the studies. Given this gap, future research should explore how to balance technological and content innovations with unresolved social concerns, accounting for variations in computer literacy.

The NGRC information needs of providers, the clinical usefulness of NGRC data, and the usefulness of NGRC-related data visualizations are not yet well supported by existing evidence. Future studies should investigate which NGRCs should be presented in HISs, how they should be used in practice, and which interactive features best support their use.

Stakeholders' insights into the integration of NGRCs and/or HISs into lifestyle interventions were not explored in the included studies. This highlights a critical gap in both technological and content innovations, as well as in addressing social concerns and in the integration of HISs and NGRCs into lifestyle change interventions.

We have summarized 5 recommendations focusing on enhancing NGRC and/or HIS integrations in behavioral interventions in Textbox 3.

Question 4: Question Regarding Research Design

We found limited coverage and inconsistent health outcomes of NGRC-focused behavioral interventions using HISs. Health outcomes focused mainly on health care utilization (eg, emergency visits), with limited coverage of clinical indicators, including medication adherence [97], asthma control [92], cancer screening [98], and blood pressure [87]. This limited coverage of clinical health outcomes limits the understanding of long-term

XSL•FO
RenderX

behavior change and highlights the need for more comprehensive assessments. Moreover, the inconsistent health care utilization outcomes [88,91] may be caused by the inconsistent and/or insufficient reporting of 5 aspects. These aspects include (1) NGRC attributes (eg, definitions and spatiotemporal characteristics), (2) HIS types and functionalities, (3) assessment methods (eg, clinical metrics [eg, HbA_{1c}] and health care utilization), (4) stakeholders (eg, only HISs or various health providers), and (5) targeted diseases. Such heterogeneity causes low interstudy comparability and reproducibility. It also limits the ability to determine which behavioral interventions, leveraging specific NGRCs via particular HIS functions, are most effective, thereby complicating the translation of findings into evidence-based practice.

No studies have addressed key issues such as cost-effectiveness and ethical considerations. Future research should examine these critical factors to ensure successful implementation.

To develop evidence, we have summarized 3 recommendations in Textbox 3.

Generalizability of the Results

Our objective was to synthesize global literature on NGRC-focused behavioral interventions supported by HISs. However, we were unable to identify published, peer-reviewed studies conducted outside of the United States. Initiatives, such as the European Health Data Space (EHDS) [125] and China's Healthy China Initiative [126], reflect growing efforts to integrate SDoH, including NGRCs, into HISs. On one hand, contextual differences do not preclude transferability. The contextual differences include cultural norms, socioeconomic conditions, and geographic contexts. With appropriate localization and system-level adaptation, the core principles identified in United States–based studies can be effectively applied in diverse settings. However, the limited diversity of study settings may obscure unique challenges and solutions relevant to low-resource or non-Western countries. Addressing these gaps is essential, and future research should ensure that context-specific NGRC-focused behavioral interventions using HISs are inclusive, adaptable, and globally relevant (Table 3). On the other hand, technological and political differences across the world do not influence the generalizability of our study. These differences include (1) HIS infrastructure, architecture, and integration capabilities, and (2) data standards, privacy regulations (eg, the European General Data Protection Regulation [127]), and governance policies. The commonalities of these differences help support the generalizability of the findings [128].

Strengths and Limitations

Our review has 4 key strengths. First, this article reviews the current state of NGRC-focused behavioral interventions using HISs. This provides a comprehensive analysis from 3 perspectives: technology, NGRC content, and stakeholders' insights. Second, we searched a wide range of databases, leading to publications from diverse disciplines, including medical journals, environmental journals, and computer science conferences. Third, we consulted a library information specialist to develop the search strategy, ensuring its comprehensiveness and accuracy, which thus enhanced the research's quality and credibility. Finally, to provide readers with sufficient context and allow them to assess the reliability of the evidence, we extracted and reported descriptive information, including study type, publication year, and sample size (Table S4 in Multimedia Appendix 1). We did not conduct a critical appraisal of the included studies, as it is not mandatory in scoping reviews. According to the Joanna Briggs Institute (JBI) methodology for scoping reviews [129], a critical appraisal is generally unnecessary unless it is directly relevant to the review objectives. Additionally, most quality assessment tools, including the Mixed Methods Appraisal Tool (MMAT) [130], primarily evaluate study design, data collection, and methodological rigor. Therefore, applying such tools would not provide meaningful information for our objectives (eg, HIS functionalities and NGRC indicators).

Our study also has several limitations that should be considered. First, NGRCs are ambiguous, suggesting that they are open to multiple interpretations or are vague. For example, in various studies, neighborhood safety may refer to different things, but there are no clear details of this concept in the studies. However, resolving this ambiguity is challenging because of the lack of standardized definitions and the context-dependent characteristics of NGRCs. Therefore, future studies should adopt detailed and consistent definitions of NGRCs, ideally guided by standardized terminologies or ontologies. These terminologies or ontologies can leverage and expand existing terminologies (eg, "community details" in Systematized Nomenclature of Medicine—Clinical Terms [SNOMED CT] and Logical Observation Identifiers Names and Codes [LOINC]) and SDoH ontologies [24].

Possible Mechanisms and Implications for Clinicians or Policymakers

HISs are often seen as passive recorders of health data. However, tethered or interconnected PHRs can actively participate in the entire process of NGRC-focused behavioral interventions, especially in outpatient lifestyle change interventions, with the relatively limited help of health care providers. For example, platforms like the iManageCancer platform for cancer patients and the iManageMyHealth app for their physicians provide personalized content from reliable public resources on cancer. However, these platforms do not currently consider NGRC factors and focus on behavioral interventions [3]. This underscores the importance for providers and policymakers to utilize tethered or interconnected PHRs to collect comprehensive, dynamic, and static NGRCs; deliver timely behavioral interventions in outpatient settings; monitor behaviors and the health status; and assess interventions. By doing so, they can achieve public health goals and promote health equity.

Conclusion

In this study, we found that most NGRC-focused behavioral interventions utilized transportation barriers to support social needs interventions. Meanwhile, HISs were involved only in specific aspects of the NGRC-focused behavioral intervention process. To achieve comprehensive dynamic and/or

XSL•FO
RenderX

high-resolution NGRC considerations and the full integration of HISs throughout the behavioral intervention process, given the limited and heterogeneous evidence, further research is needed to determine how tethered or interconnected PHRs can effectively support NGRC-focused interventions, especially lifestyle change interventions. During implementation, it is essential to address social concerns alongside technological (eg, HIS integration) and content innovations (eg, NGRC integration). We compiled a list of 15 recommendations for future research. These recommendations are likely to accelerate the integration of tethered or interconnected PHRs in NGRC-focused behavioral interventions.

Acknowledgments

Generative artificial intelligence (AI) tools were used for writing assistance. Generative AI tools (eg, ChatGPT) were used for language editing and minor improvements in phrasing after the original writing was completed. We would like to thank Valerie Moran from Luxembourg Institute of Socio-Economic Research (LISER) for shaping the thematic focus and structure of this scoping review.

Funding

The study was funded by the China Scholarship Council (CSC) (file number: 202208140027).

Authors' Contributions

XY, MD, and RC are the guarantors of the integrity of the entire study. All authors contributed to the study's conceptualization and design. JGD performed the literature search. XY, MD, and RC contributed to study selection and interpretation. XY wrote the manuscript. MD, RC, HY, MS, and JGD reviewed and edited the manuscript. All authors read and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information to support the scoping review.
[DOCX File, 62 KB - medinform_v14i1e77296_app1.docx]

Checklist 1
PRISMA-ScR checklist
[PDF File, 151 KB - medinform_v14i1e77296_app2.pdf]

References

1. Anderson NB, Bulatao RA, Cohen B, National Research Council (US) Panel on Race, Ethnicity,and Health in Later Life. Critical Perspectives on Racial and Ethnic Differences in Health in Late Life: National Academies Press (US); 2004. [doi: 10.17226/11086] [Medline: 20669464]
2. Lunde P, Nilsson BB, Bergland A, Kværner KJ, Bye A. The effectiveness of smartphone apps for lifestyle improvement in noncommunicable diseases: systematic review and meta-analyses. J Med Internet Res 2018 May 4;20(5):e162. [doi: 10.2196/jmir.9751] [Medline: 29728346]
3. Schera F, Schäfer M, Bucur A, et al. iManageMyHealth and iSupportMyPatients: mobile decision support and health management apps for cancer patients and their doctors. Ecancermedicalscience 2018;12:848. [doi: 10.3332/ecancer.2018.848] [Medline: 30079110]
4. Burgess E, Hassmén P, Pumpa KL. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. Clin Obes 2017 Jun;7(3):123-135. [doi: 10.1111/cob.12183] [Medline: 28296261]
5. Cummings KM, Fong GT, Borland R. Environmental influences on tobacco use: evidence from societal and community influences on tobacco use and dependence. Annu Rev Clin Psychol 2009;5:433-458. [doi: 10.1146/annurev.clinpsy.032408.153607] [Medline: 19327036]
6. Basta LA, Richmond TS, Wiebe DJ. Neighborhoods, daily activities, and measuring health risks experienced in urban environments. Soc Sci Med 2010 Dec;71(11):1943-1950. [doi: 10.1016/j.socscimed.2010.09.008]
7. Kepper MM, Myers CA, Denstel KD, Hunter RF, Guan W, Broyles ST. The neighborhood social environment and physical activity: a systematic scoping review. Int J Behav Nutr Phys Act 2019 Dec 9;16(1):124. [doi: 10.1186/s12966-019-0873-7] [Medline: 31815626]
8. Popkin B, Duffey K, Gordonlarsen P. Environmental influences on food choice, physical activity and energy balance. Physiol Behav 2005 Dec 15;86(5):603-613. [doi: 10.1016/j.physbeh.2005.08.051]

XSL•FO
RenderX

9. Healthy People 2030 - social determinants of health. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. URL: https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health [accessed 2025-10-15]
10. Jilcott SB, Keyserling TC, Samuel-Hodge CD, et al. Linking clinical care to community resources for cardiovascular disease prevention: the North Carolina Enhanced WISEWOMAN project. J Womens Health (Larchmt) 2006 Jun;15(5):569-583. [doi: 10.1089/jwh.2006.15.569] [Medline: 16796484]
11. Christian BJ. Translational research - the intersection between sociocultural and environmental factors and the health of children and families. J Pediatr Nurs 2017;37:127-131. [doi: 10.1016/j.pedn.2017.09.013] [Medline: 28988617]
12. Diez Roux AV, Mair C. Neighborhoods and health. Ann N Y Acad Sci 2010 Feb;1186(1):125-145. [doi: 10.1111/j.1749-6632.2009.05333.x] [Medline: 20201871]
13. Samuels-Kalow ME, Camargo CA. The use of geographic data to improve asthma care delivery and population health. Clin Chest Med 2019 Mar;40(1):209-225. [doi: 10.1016/j.ccm.2018.10.012] [Medline: 30691713]
14. Alharthy A, Hosler AS, Leckman-Westin E, Kammer JR. Association between tobacco retailer density and smoking among adults with diabetes and serious mental illness in New York state. Prev Chronic Dis 2022 Jan 6;19:E01. [doi: 10.5888/pcd19.210270] [Medline: 34990338]
15. Cottrell EK, Hendricks M, Dambrun K, et al. Comparison of community-level and patient-level social risk data in a network of community health centers. JAMA Netw Open 2020 Oct 1;3(10):e2016852. [doi: 10.1001/jamanetworkopen.2020.16852] [Medline: 33119102]
16. Edmonds AT, Rhew IC, Jones-Smith J, et al. Neighborhood disadvantage, patterns of unhealthy alcohol use, and differential associations by gender, race/ethnicity, and rurality: a study of Veterans Health Administration patients. J Stud Alcohol Drugs 2022 Nov;83(6):867-878. [doi: 10.15288/jsad.21-00110] [Medline: 36484584]
17. García R, Muñoz MA, Navas E, Vinyoles E, Verdú-Rotellar JM, Del Val JL. Variability in cardiovascular risk factor control in patients with heart failure according to gender and socioeconomic status. J Womens Health (Larchmt) 2022 May;31(5):690-697. [doi: 10.1089/jwh.2021.0404] [Medline: 35041531]
18. Howell NA, Tu JV, Moineddin R, Chu A, Booth GL. Association between neighborhood walkability and predicted 10-year cardiovascular disease risk: the CANHEART (Cardiovascular Health in Ambulatory Care Research Team) cohort. J Am Heart Assoc 2019 Nov 5;8(21):e013146. [doi: 10.1161/JAHA.119.013146] [Medline: 31665997]
19. Kurani SS, Lampman MA, Funni SA, et al. Association between area-level socioeconomic deprivation and diabetes care quality in US primary care practices. JAMA Netw Open 2021 Dec 1;4(12):e2138438. [doi: 10.1001/jamanetworkopen.2021.38438] [Medline: 34964856]
20. Mathur R, Schofield P, Smith D, Gilkes A, White P, Hull S. Is individual smoking behaviour influenced by area-level ethnic density? A cross-sectional electronic health database study of inner south-east London. ERJ Open Res 2017 Jan;3(1):00130-2016. [doi: 10.1183/23120541.00130-2016] [Medline: 28435842]
21. Miranda O, Fan P, Qi X, et al. DeepBiomarker2: prediction of alcohol and substance use disorder risk in post-traumatic stress disorder patients using electronic medical records and multiple social determinants of health. J Pers Med 2024 Jan 14;14(1):94. [doi: 10.3390/jpm14010094] [Medline: 38248795]
22. Rodriguez-Mireles S, Lopez-Valcarcel BG, Galdos-Arias P, Perez-Diaz E, Serra-Majem L. Socioeconomic disparities in diet and physical activity in children: evidence from well-child visit electronic health records in the Canary Islands, Spain. J Epidemiol Community Health 2024 May 9;78(6):354-359. [doi: 10.1136/jech-2023-220335] [Medline: 38458631]
23. To D, Joyce C, Kulshrestha S, et al. The addition of United States census-tract data does not improve the prediction of substance misuse. AMIA Annu Symp Proc 2021;2021:1149-1158. [Medline: 35308901]
24. DesRoches CM, Wachenheim D, Harcourt K, Garcia A, Shah R. Collecting, documenting, and using social determinants of health data: patient and clinician views. : OpenNotes; 2025 URL: https://www.opennotes.org/wp-content/uploads/2025/02/ASTP_OpenNotes_SDOH-FinalReport_508.pdf [accessed 2026-02-06]
25. Haux R. Health information systems - past, present, future. Int J Med Inform 2006;75(3-4):268-281. [doi: 10.1016/j.ijmedinf.2005.08.002] [Medline: 16169771]
26. Kim JW, Ryu B, Cho S, et al. Impact of personal health records and wearables on health outcomes and patient response: three-arm randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 4;7(1):e12070. [doi: 10.2196/12070] [Medline: 30609978]
27. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126. [doi: 10.1197/jamia.M2025] [Medline: 16357345]
28. Cottrell EK, O'Malley JP, Dambrun K, et al. The impact of social and clinical complexity on diabetes control measures. J Am Board Fam Med 2020;33(4):600-610. [doi: 10.3122/jabfm.2020.04.190367] [Medline: 32675271]
29. Holt JM, Cusatis R, Asan O, et al. Incorporating patient-generated contextual data into care: clinician perspectives using the consolidated framework for implementation science. Healthcare (Basel) 2020 Mar;8(1):100369. [doi: 10.1016/j.hjdsi.2019.100369]

XSL•FO
RenderX

30. Abd-Alrazaq AA, Bewick BM, Farragher T, Gardner P. Factors that affect the use of electronic personal health records among patients: a systematic review. Int J Med Inform 2019 Jun;126:164-175. [doi: 10.1016/j.ijmedinf.2019.03.014] [Medline: 31029258]
31. Andrikopoulou E, Scott P, Herrera H, Good A. What are the important design features of personal health records to improve medication adherence for patients with long-term conditions? A systematic literature review. BMJ Open 2019 Sep 26;9(9):e028628. [doi: 10.1136/bmjopen-2018-028628] [Medline: 31558449]
32. Aslani N, Ahmadi M, Samadbeik M. A systematic review of the attributes of electronic personal health records for patients with multiple sclerosis. Health Technol 2020 May;10(3):587-599. [doi: 10.1007/s12553-019-00387-4]
33. Azizi A, Aboutorabi R, Mazloum-Khorasani Z, Hoseini B, Tara M. Diabetic personal health record: a systematic review article. Iran J Public Health 2016 Nov;45(11):1388-1398. [Medline: 28032056]
34. Chiesa V, Chiarenza A, Mosca D, Rechel B. Health records for migrants and refugees: a systematic review. Health Policy 2019 Sep;123(9):888-900. [doi: 10.1016/j.healthpol.2019.07.018] [Medline: 31439455]
35. Crameri KA, Maher L, Van Dam P, Prior S. Personal electronic healthcare records: what influences consumers to engage with their clinical data online? A literature review. HIM J 2022 Jan;51(1):3-12. [doi: 10.1177/1833358319895369]
36. Eastway J, Lizarondo L. Experiences of adult patients with chronic non-communicable disease using electronic personal health records for self-management: a qualitative systematic review protocol. JBI Database System Rev Implement Rep 2019 Nov;17(11):2334-2342. [doi: 10.11124/JBISRIR-2017-003941] [Medline: 31232888]
37. Fernández-Alemán JL, Seva-Llor CL, Toval A, Ouhbi S, Fernández-Luque L. Free web-based personal health records: an analysis of functionality. J Med Syst 2013 Dec;37(6):9990. [doi: 10.1007/s10916-013-9990-z] [Medline: 24221916]
38. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Evaluation of personal health record systems through the lenses of EC research projects. Comput Biol Med 2015 Apr;59:175-185. [doi: 10.1016/j.compbiomed.2013.11.004] [Medline: 24315661]
39. Harahap NC, Handayani PW, Hidayanto AN. Functionalities and issues in the implementation of personal health records: systematic review. J Med Internet Res 2021 Jul 21;23(7):e26236. [doi: 10.2196/26236] [Medline: 34287210]
40. Jabour A, Jones JF. Facilitators and barriers to patients' engagements with personal health records: systematic review. In: Stephanidis C, Antona M, editors. Universal Access in Human-Computer Interaction Applications and Services for Quality of Life UAHCI 2013 Lecture Notes in Computer Science: Springer; 2013, Vol. 8011:472-481. [doi: 10.1007/978-3-642-39194-1_55]
41. Kharbouch M, Idri A, Fernandez-Aleman JL, Redman L, Toval A, Stelate Y. Reviewing the features and functionalities of contraception mPHRs. Health Policy Technol 2022 Sep;11(3):100633. [doi: 10.1016/j.hlpt.2022.100633]
42. Kneale L, Demiris G. Lack of diversity in personal health record evaluations with older adult participants: a systematic review of literature. J Innov Health Inform 2017 Jan 15;23(4):881. [doi: 10.14236/jhi.v23i4.881] [Medline: 28346131]
43. Thompson MJ, Reilly JD, Valdez RS. Work system barriers to patient, provider, and caregiver use of personal health records: a systematic review. Appl Ergon 2016 May;54:218-242. [doi: 10.1016/j.apergo.2015.10.010] [Medline: 26851482]
44. Zhao JY, Song B, Anand E, et al. Barriers, facilitators, and solutions to optimal patient portal and personal health record use: a systematic review of the literature. AMIA Annu Symp Proc 2017;2017:1913-1922. [Medline: 29854263]
45. Golembiewski E, Allen KS, Blackmon AM, Hinrichs RJ, Vest JR. Combining nonclinical determinants of health and clinical data for research and evaluation: rapid review. JMIR Public Health Surveill 2019 Oct 7;5(4):e12846. [doi: 10.2196/12846] [Medline: 31593550]
46. Schinasi LH, Auchincloss AH, Forrest CB, Diez Roux AV. Using electronic health record data for environmental and place based population health research: a systematic review. Ann Epidemiol 2018 Jul;28(7):493-502. [doi: 10.1016/j.annepidem.2018.03.008] [Medline: 29628285]
47. Chishtie JA, Tea J, Ester M, et al. Integration of screening and referral tools for social determinants of health and modifiable lifestyle factors in the Epic electronic health record system: scoping review. J Med Internet Res 2025 Sep 15;27:e73615. [doi: 10.2196/73615] [Medline: 40953441]
48. Arroyave Caicedo NM, Parry E, Arslan N, Park S. Integration of social determinants of health information within the primary care electronic health record: a systematic review of patient perspectives and experiences. BJGP Open 2024 Apr;8(1):BJGPO.2023.0155. [doi: 10.3399/BJGPO.2023.0155] [Medline: 37673433]
49. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. J Am Med Inform Assoc 2020 Nov 1;27(11):1764-1773. [doi: 10.1093/jamia/ocaa143] [Medline: 33202021]
50. Li C, Mowery DL, Ma X, et al. Realizing the potential of social determinants data in EHR systems: a scoping review of approaches for screening, linkage, extraction, analysis, and interventions. J Clin Trans Sci 2024;8(1). [doi: 10.1017/cts.2024.571]
51. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
52. Health informatics — electronic health record — definition, scope and context. : ISO (the International Organization for Standardization); 2005 ISO/TR 20514:2005(en) URL: https://www.iso.org/obp/ui/en/#iso:std:iso:tr:20514:ed-1:v1:en [accessed 2026-02-06]

XSL•FO
RenderX

53. Coleman K, Austin BT, Brach C, Wagner EH. Evidence on the chronic care model in the new millennium. Health Aff (Millwood) 2009;28(1):75-85. [doi: 10.1377/hlthaff.28.1.75] [Medline: 19124857]
54. Siminerio LM, Piatt GA, Emerson S, et al. Deploying the chronic care model to implement and sustain diabetes self-management training programs. Diabetes Educ 2006;32(2):253-260. [doi: 10.1177/0145721706287156] [Medline: 16554429]
55. Siminerio LM, Piatt G, Zgibor JC. Implementing the chronic care model for improvements in diabetes care and education in a rural primary care practice. Diabetes Educ 2005;31(2):225-234. [doi: 10.1177/0145721705275325] [Medline: 15797851]
56. Stellefson M, Dipnarine K, Stopka C. The chronic care model and diabetes management in US primary care settings: a systematic review. Prev Chronic Dis 2013;10:E26. [doi: 10.5888/pcd10.120180] [Medline: 23428085]
57. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 6;10(1):231. [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
58. Tang PC, Lee TH. Your doctor's office or the internet? Two paths to personal health records. N Engl J Med 2009 Mar 26;360(13):1276-1278. [doi: 10.1056/NEJMp0810264] [Medline: 19321866]
59. Stimson RJ, Golledge RG. Spatial Behavior: A Geographic Perspective: Guilford Press; 1996.
60. Kwan MP. The uncertain geographic context problem. Annals of the Association of American Geographers 2012 Sep;102(5):958-968. [doi: 10.1080/00045608.2012.687349]
61. Rodríguez DA, Evenson KR, Diez Roux AV, Brines SJ. Land use, residential density, and walking. Am J Prev Med 2009 Nov;37(5):397-404. [doi: 10.1016/j.amepre.2009.07.008]
62. Giles-Corti B, Donovan RJ. Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment. Prev Med 2002 Dec;35(6):601-611. [doi: 10.1006/pmed.2002.1115] [Medline: 12460528]
63. Giles-Corti B, Donovan RJ. Relative influences of individual, social environmental, and physical environmental correlates of walking. Am J Public Health 2003 Sep;93(9):1583-1589. [doi: 10.2105/ajph.93.9.1583] [Medline: 12948984]
64. Crawford D, Cleland V, Timperio A, et al. The longitudinal influence of home and neighbourhood environments on children's body mass index and physical activity over 5 years: the CLAN study. Int J Obes 2010 Jul;34(7):1177-1187. [doi: 10.1038/ijo.2010.57]
65. Zhang R, Wulff H, Duan Y, Wagner P. Associations between the physical environment and park-based physical activity: a systematic review. J Sport Health Sci 2019 Sep;8(5):412-421. [doi: 10.1016/j.jshs.2018.11.002] [Medline: 31534816]
66. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. Chronobiol Int 2013 Dec;30(10):1223-1230. [doi: 10.3109/07420528.2013.813528] [Medline: 23971909]
67. Feinglass J, Lee J, Semanik P, Song J, Dunlop D, Chang R. The effects of daily weather on accelerometer-measured physical activity. J Phys Act Health 2011 Sep;8(7):934-943. [doi: 10.1123/jpah.8.7.934] [Medline: 21885884]
68. Ferguson T, Curtis R, Fraysse F, et al. Weather associations with physical activity, sedentary behaviour and sleep patterns of Australian adults: a longitudinal study with implications for climate change. Int J Behav Nutr Phys Act 2023 Mar 14;20(1):30. [doi: 10.1186/s12966-023-01414-4] [Medline: 36918954]
69. Hu X, Knibbs LD, Zhou Y, Ou Y, Dong GH, Dong H. The role of lifestyle in the association between long-term ambient air pollution exposure and cardiovascular disease: a national cohort study in China. BMC Med 2024 Mar 5;22(1):93. [doi: 10.1186/s12916-024-03316-z] [Medline: 38439026]
70. Strak M, Janssen N, Beelen R, et al. Associations between lifestyle and air pollution exposure: potential for confounding in large administrative data cohorts. Environ Res 2017 Jul;156:364-373. [doi: 10.1016/j.envres.2017.03.050] [Medline: 28395240]
71. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
72. Nair BR, Browne W. Chronic disease management: seizing the moment in medical education. Pharm Med 2008 Nov;22(6):351-354. [doi: 10.1007/BF03256731]
73. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148. [doi: 10.2196/jmir.4255] [Medline: 26104044]
74. Kamiński M, Kręgielska-Narożna M, Bogdański P. Seasonal variation in lifestyle behavior in Poland: Google searches and market sales analysis. BMC Public Health 2021 Aug 6;21(1):1516. [doi: 10.1186/s12889-021-11543-9] [Medline: 34362345]
75. Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. Health Aff (Millwood) 2001;20(6):64-78. [doi: 10.1377/hlthaff.20.6.64] [Medline: 11816692]
76. Venerandi A, Mellen H, Romice O, Porta S. Walkability indices—the state of the art and future directions: a systematic review. Sustainability 2024 Aug 6;16(16):6730. [doi: 10.3390/su16166730]
77. Kane NJ, Wang X, Gerkovich MM, et al. The Envirome Web Service: patient context at the point of care. J Biomed Inform 2021 Jul;119:103817. [doi: 10.1016/j.jbi.2021.103817] [Medline: 34020026]

XSL•FO
RenderX

78. Cordova-Ramos EG, Jain C, Torrice V, et al. Implementing social risk screening and referral to resources in the NICU. Pediatrics 2023 Apr 1;151(4):e2022058975. [doi: 10.1542/peds.2022-058975] [Medline: 36919445]
79. Fortin K, Vasan A, Wilson-Hall CL, Brooks E, Rubin D, Scribano PV. Using quality improvement and technology to improve social supports for hospitalized children. Hosp Pediatr 2021 Oct;11(10):1120-1129. [doi: 10.1542/hpeds.2020-005800] [Medline: 34475224]
80. Stark K, Mathur M, Fok C, et al. Evaluation of a clinic-based, electronic social determinants of health screening and intervention in primary care pediatrics. Acad Pediatr 2024 Mar;24(2):302-308. [doi: 10.1016/j.acap.2023.12.010] [Medline: 38160752]
81. Glasser NJ, Lindau ST, Wroblewski K, et al. Effect of a social care intervention on health care experiences of caregivers of hospitalized children: a randomized clinical trial. JAMA Pediatr 2023 Dec 1;177(12):1266-1275. [doi: 10.1001/jamapediatrics.2023.4596] [Medline: 37902777]
82. Buitron de la Vega P, Losi S, Sprague Martinez L, et al. Implementing an EHR-based screening and referral system to address social determinants of health in primary care. Med Care 2019 Jun;57 Suppl 6 Suppl 2(Suppl 2):S133-S139. [doi: 10.1097/MLR.0000000000001029] [Medline: 31095052]
83. Johnston DC, Mathews WD, Maus A, Gustafson DH. Using smartphones to improve treatment retention among impoverished substance-using appalachian women: a naturalistic study. Subst Abuse 2019;13:1178221819861377. [doi: 10.1177/1178221819861377] [Medline: 31312084]
84. Russell LE, Cohen AJ, Chrzas S, et al. Implementing a social needs screening and referral program among veterans: Assessing Circumstances & Offering Resources for Needs (ACORN). J Gen Intern Med 2023 Oct;38(13):2906-2913. [doi: 10.1007/s11606-023-08181-9] [Medline: 37165261]
85. Percac-Lima S, Grant RW, Green AR, et al. A culturally tailored navigator program for colorectal cancer screening in a community health center: a randomized, controlled trial. J Gen Intern Med 2009 Feb;24(2):211-217. [doi: 10.1007/s11606-008-0864-x] [Medline: 19067085]
86. Wallace AS, Luther B, Guo JW, Wang CY, Sisler S, Wong B. Implementing a social determinants screening and referral infrastructure during routine emergency department visits, Utah, 2017-2018. Prev Chronic Dis 2020 Jun 18;17:E45. [doi: 10.5888/pcd17.190339] [Medline: 32553071]
87. Tanumihardjo JP, Kuther S, Wan W, et al. New frontiers in diabetes care: quality improvement study of a population health team in rural critical access hospitals. J Gen Intern Med 2023 Mar;38(Suppl 1):56-64. [doi: 10.1007/s11606-022-07928-0] [Medline: 36864269]
88. Roth SE, Gronowski B, Jones KG, et al. Evaluation of an integrated intervention to address clinical care and social needs among patients with type 2 diabetes. J Gen Intern Med 2023 Mar;38(Suppl 1):38-44. [doi: 10.1007/s11606-022-07920-8] [Medline: 36864267]
89. Bechtel N, Jones A, Kue J, Ford JL. Evaluation of the core 5 social determinants of health screening tool. Public Health Nurs 2022 Mar;39(2):438-445. [doi: 10.1111/phn.12983] [Medline: 34628675]
90. Taber KA, Williams JN, Huang W, et al. Use of an integrated care management program to uncover and address social determinants of health for individuals with lupus. ACR Open Rheumatol 2021 May;3(5):305-311. [doi: 10.1002/acr2.11236] [Medline: 33779065]
91. Gupta D, Self S, Thomas D, Supra J, Rudisill C. Understanding the role of a technology and EMR-based social determinants of health screening tool and community-based resource connections in health care resource utilization. Med Care 2023 Jul 1;61(7):423-430. [doi: 10.1097/MLR.0000000000001800] [Medline: 36729786]
92. Britto MT, Vockell ALB, Munafo JK, et al. Improving outcomes for underserved adolescents with asthma. Pediatrics 2014 Feb;133(2):e418-e427. [doi: 10.1542/peds.2013-0684] [Medline: 24470645]
93. Loo S, Anderson E, Lin JG, et al. Evaluating a social risk screening and referral program in an urban safety-net hospital emergency department. J Am Coll Emerg Physicians Open 2023 Feb;4(1):e12883. [doi: 10.1002/emp2.12883] [Medline: 36704207]
94. Lynch S, Street NW, Sharpe L, Kellish A. Social determinants of health screening: primary care PRAPARE tool implementation. J Nurse Pract 2024 Apr;20(4):104955. [doi: 10.1016/j.nurpra.2024.104955]
95. Palakshappa D, Benefield AJ, Furgurson KF, et al. Feasibility of mobile technology to identify and address patients' unmet social needs in a primary care clinic. Popul Health Manag 2021 Jun;24(3):385-392. [doi: 10.1089/pop.2020.0059] [Medline: 32924796]
96. Wright QC, Wiencek CA, Williams IC. Utilizing a tailored approach to standardized screening of health-related social needs. J Nurs Care Qual 2024;39(2):114-120. [doi: 10.1097/NCQ.0000000000000745] [Medline: 37729002]
97. Wu L, Chang C, Lo K, et al. Telephone-based social health screening by pharmacists in the nonadherent Medicare population. J Manag Care Spec Pharm 2023 Nov;29(11):1184-1192. [doi: 10.18553/jmcp.2023.29.11.1184] [Medline: 37889865]
98. Percac-Lima S, Ashburner JM, Rigotti NA, et al. Patient navigation for lung cancer screening among current smokers in community health centers a randomized controlled trial. Cancer Med 2018 Mar;7(3):894-902. [doi: 10.1002/cam4.1297] [Medline: 29464877]
99. University of Wisconsin–Madison - Center for Health Enhancement Systems Studies. URL: https://center.chess.wisc.edu/ [accessed 2025-04-12]

XSL•FO
RenderX

100. Hosseini A, Emami H, Sadat Y, Paydar S. Integrated personal health record (PHR) security: requirements and mechanisms. BMC Med Inform Decis Mak 2023 Jul 10;23(1):116. [doi: 10.1186/s12911-023-02225-0] [Medline: 37430242]
101. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e22890. [doi: 10.2196/22890] [Medline: 34806995]
102. Fukuoka Y, Gay C, Haskell W, Arai S, Vittinghoff E. Identifying factors associated with dropout during prerandomization run-in period from an mHealth physical activity education study: the mPED trial. JMIR Mhealth Uhealth 2015 Apr 13;3(2):e34. [doi: 10.2196/mhealth.3928] [Medline: 25872754]
103. Owen JE, Jaworski BK, Kuhn E, Makin-Byrd KN, Ramsey KM, Hoffman JE. mHealth in the wild: using novel data to examine the reach, use, and impact of PTSD coach. JMIR Ment Health 2015;2(1):e7. [doi: 10.2196/mental.3935] [Medline: 26543913]
104. Oikonomidi T, Ravaud P, LeBeau J, Tran VT. A systematic scoping review of just-in-time, adaptive interventions finds limited automation and incomplete reporting. J Clin Epidemiol 2023 Feb;154:108-116. [doi: 10.1016/j.jclinepi.2022.12.006] [Medline: 36521653]
105. Kollapally NM, Chen Y, Xu J, Geller J. An ontology for the social determinants of health domain. Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 6-8, 2022. [doi: 10.1109/BIBM55620.2022.9995544]
106. Dang Y, Li F, Hu X, et al. Systematic design and data-driven evaluation of social determinants of health ontology (SDoHO). J Am Med Inform Assoc 2023 Aug 18;30(9):1465-1473. [doi: 10.1093/jamia/ocad096] [Medline: 37301740]
107. Brennan PF, Downs S, Casper G. Project HealthDesign: rethinking the power and potential of personal health records. J Biomed Inform 2010 Oct;43(5 Suppl):S3-S5. [doi: 10.1016/j.jbi.2010.09.001] [Medline: 20937482]
108. Thomas Craig KJ, Fusco N, Gunnarsdottir T, Chamberland L, Snowdon JL, Kassler WJ. Leveraging data and digital health technologies to assess and impact social determinants of health (SDoH): a state-of-the-art literature review. Online J Public Health Inform 2021;13(3):E14. [doi: 10.5210/ojphi.v13i3.11081] [Medline: 35082976]
109. Estiri H, Patel CJ, Murphy SN. Informatics can help providers incorporate context into care. JAMIA Open 2018 Jul;1(1):3-6. [doi: 10.1093/jamiaopen/ooy025] [Medline: 31984312]
110. United States, National Committee on Vital and Health Statistics. Information for health; a strategy for building the national health information infrastructure. : United States Department of Health and Human Services; 2001 URL: https://stacks.cdc.gov/view/cdc/26030 [accessed 2026-02-06]
111. Wetta RE, Severin RD, Gruhler H. An evidence-based strategy to achieve equivalency and interoperability for social-behavioral determinants of health assessment, storage, exchange, and use. Health Informatics J 2020 Jun;26(2):1477-1488. [doi: 10.1177/1460458219882265] [Medline: 31659920]
112. Deslippe AL, Soanes A, Bouchaud CC, et al. Barriers and facilitators to diet, physical activity and lifestyle behavior intervention adherence: a qualitative systematic review of the literature. Int J Behav Nutr Phys Act 2023 Feb 14;20(1):14. [doi: 10.1186/s12966-023-01424-2] [Medline: 36782207]
113. Burgess E, Hassmén P, Welvaert M, Pumpa KL. Behavioural treatment strategies improve adherence to lifestyle intervention programmes in adults with obesity: a systematic review and meta-analysis. Clin Obes 2017 Apr;7(2):105-114. [doi: 10.1111/cob.12180] [Medline: 28199047]
114. Freeman KJ, Grega ML, Friedman SM, et al. Lifestyle medicine reimbursement: a proposal for policy priorities informed by a cross-sectional survey of lifestyle medicine practitioners. Int J Environ Res Public Health 2021 Nov 5;18(21):11632. [doi: 10.3390/ijerph182111632] [Medline: 34770148]
115. Middleton KR, Anton SD, Perri MG. Long-term adherence to health behavior change. Am J Lifestyle Med 2013;7(6):395-404. [doi: 10.1177/1559827613488867] [Medline: 27547170]
116. Kim J, Kwan MP. Assessment of sociodemographic disparities in environmental exposure might be erroneous due to neighborhood effect averaging: implications for environmental inequality research. Environ Res 2021 Apr;195:110519. [doi: 10.1016/j.envres.2020.110519] [Medline: 33253702]
117. Kwan MP. How GIS can help address the uncertain geographic context problem in social science research. Ann GIS 2012 Dec;18(4):245-255. [doi: 10.1080/19475683.2012.727867]
118. Dark SJ, Bram D. The modifiable areal unit problem (MAUP) in physical geography. Progress in Physical Geography: Earth and Environment 2007 Oct;31(5):471-479. [doi: 10.1177/0309133307083294]
119. Andresen MA. Modifiable areal unit problem. In: Barnes JC, Forde DR, editors. The Encyclopedia of Research Methods in Criminology and Criminal Justice: John Wiley & Sons, Inc; 2021:854-855. [doi: 10.1002/9781119111931.ch164]
120. Kwan MP. The neighborhood effect averaging problem (NEAP): an elusive confounder of the neighborhood effect. Int J Environ Res Public Health 2018 Aug 27;15(9):1841. [doi: 10.3390/ijerph15091841] [Medline: 30150510]
121. Hinckson E, Cerin E, Mavoa S, et al. Associations of the perceived and objective neighborhood environment with physical activity and sedentary time in New Zealand adolescents. Int J Behav Nutr Phys Act 2017 Oct 25;14(1):145. [doi: 10.1186/s12966-017-0597-5] [Medline: 29065897]

XSL•FO
RenderX

122. Nyunt MSZ, Shuvo FK, Eng JY, et al. Objective and subjective measures of neighborhood environment (NE): relationships with transportation physical activity among older persons. Int J Behav Nutr Phys Act 2015 Sep 15;12(1):108. [doi: 10.1186/s12966-015-0276-3] [Medline: 26373320]
123. Jáuregui A, Salvo D, Lamadrid-Figueroa H, Hernández B, Rivera-Dommarco JA, Pratt M. Perceived and objective measures of neighborhood environment for physical activity among Mexican adults, 2011. Prev Chronic Dis 2016 Jun 9;13:E76. [doi: 10.5888/pcd13.160009] [Medline: 27281391]
124. Zhao L, Shen Z, Zhang Y, Sheng F. Study on the impact of the objective characteristics and subjective perception of the built environment on residents' physical activities in Fuzhou, China. Sustainability 2019 Dec 31;12(1):329. [doi: 10.3390/su12010329]
125. Regulation - EU - 2025/327 - EN - EUR-lex. European Union EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2025/327/oj/eng [accessed 2025-11-28]
126. 中共中央 国务院印发《"健康中国2030"规划纲要》[Article in Chinese]. The State Council of the People's Republic of China. URL: https://www.gov.cn/zhengce/2016-10/25/content_5124174.htm [accessed 2025-11-28]
127. Regulation - 2016/679 - EN - gdpr - EUR-Lex. European Union EUR-Lex. URL: http://data.europa.eu/eli/reg/2016/679/oj [accessed 2025-11-24]
128. Sittig DF, Wright A, Coiera E, et al. Current challenges in health information technology-related patient safety. Health Informatics J 2020 Mar;26(1):181-189. [doi: 10.1177/1460458218814893] [Medline: 30537881]
129. Critical appraisal tools: JBI's critical appraisal tools assist in assessing the trustworthiness, relevance and results of published papers. JBI (Joanna Briggs Institute). URL: https://jbi.global/critical-appraisal-tools?utm_source=chatgpt.com [accessed 2025-10-17]
130. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]

Abbreviations

A-CHESS: Addiction Comprehensive Health Enhancement Support System
EHR: electronic health record
EMR: electronic medical record
GRC: geo-referenced context
HIS: health information system
NCD: noncommunicable disease
NGRC: neighborhood geo-referenced context
PHR: personal health record
SDoH: social determinants of health

Edited by A Benis; submitted 11.May.2025; peer-reviewed by A Yan, A Azizi; accepted 29.Dec.2025; published 25.Feb.2026.

Please cite as:
Yang X, Dijst M, Daams JG, Suhrcke M, Yang H, Cornet R
Using Health Information Systems to Support Behavioral Interventions in Local Contexts: Scoping Review
JMIR Med Inform 2026;14:e77296
URL: https://medinform.jmir.org/2026/1/e77296
doi:10.2196/77296

© Xi Yang, Martin Dijst, Joost G Daams, Marc Suhrcke, Haoran Yang, Ronald Cornet. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Core Elements, Development, and Implementation Strategies of the Nursing Minimum Data Set: Scoping Review

Pascal Müller[1], MSc; Annabell Gangnus[1], Dr Rer Med; Katrin Gayen[1], Dip; Uta Kirchner-Heklau[1], MSc; Patrick Jahn[1], Prof Dr; Sebastian Hofstetter[1,2], PhD

[1]Health Service Research Working Group, Acute Care, Department of Internal Medicine, Faculty of Medicine, University Medicine Halle (Saale), Martin-Luther-University Halle-Wittenberg, Magdeburger Straße 12, Halle (Saale), Germany

[2]Dorothea-Erxleben-Learning-Centre, Faculty of Medicine, University Medicine Halle (Saale), Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Corresponding Author:
Pascal Müller, MSc
Health Service Research Working Group, Acute Care, Department of Internal Medicine, Faculty of Medicine, University Medicine Halle (Saale), Martin-Luther-University Halle-Wittenberg, Magdeburger Straße 12, Halle (Saale), Germany

Abstract

Background: Nursing care systems face significant challenges due to demographic changes, a workforce shortage, and rising demand for care services. Digital assistive technologies offer potential to address these challenges, but systematic and standardized nursing data are essential to evaluate both innovations and broader care processes. The Nursing Minimum Data Set (NMDS) provides a foundational framework for capturing structured information on nursing care, yet there is no international consensus on its core content, development, and practical use.

Objective: This scoping review aims to map current international literature regarding (1) the core content elements of NMDS, (2) methodological approaches used in NMDS development, and (3) implementation and use of NMDS in different nursing settings.

Methods: Following the JBI (Joanna Briggs Institute) methodology and Arksey and O'Malley framework, a systematic search was conducted on July 2, 2025, in the MEDLINE (via PubMed) and CINAHL (via EBSCO) databases using the term "nursing minimum data set." Inclusion was restricted to studies in English or German focusing on the content, development, or implementation of the NMDS. The research team reviewed studies in an independent and double-blinded fashion for eligibility based on predefined criteria, with discrepancies resolved by consensus. Eligible studies were narratively summarized, with extraction structured into categories based on the review's research questions.

Results: From 1908 initially identified articles, 26 (1.4%) studies met the inclusion criteria. Considerable heterogeneity was found in the structure and scope of the NMDS, with datasets comprising 16 to 145 items. Despite variation, 4 central domains consistently emerged: patient demographics, medical care information, nursing care elements, and institutional or organizational data. NMDS development typically followed a participatory, multistage approach involving literature analysis, stakeholder consensus building, and validation through pretesting and real-world application. Implementation and use of the NMDS serve multiple functions, including documenting nursing care processes, supporting workload measurement and resource planning, quality assurance, benchmarking, and demonstrating nursing's contribution to patient outcomes. However, successful implementation depends on technical, legal, organizational, and educational strategies. Core challenges include a lack of standardized terminology, inconsistent legal frameworks, and varying levels of staff training and acceptance.

Conclusions: The NMDS provides a robust basis for standardized nursing documentation, quality assurance, and health system planning, but international variability and ongoing challenges in harmonization, integration, and acceptability persist. Advancing the NMDS requires collaborative efforts for interoperability, investment in digital infrastructure, and targeted education. Further research should focus on comparative effectiveness, cross-context validation, and strategies to reduce documentation burden while maximizing data utility.

(JMIR Med Inform 2026;14:e84281) doi:10.2196/84281

KEYWORDS

Nursing Minimum Data Set; standardized nursing data; nursing informatics; nursing documentation; core nursing data elements; health information systems; digital health

XSL•FO
RenderX

Introduction

The nursing care system in Germany is confronted with significant challenges, including demographic shifts, a shortage of nursing personnel, and increasing demand for care services [1]. From 2014 to 2021, expenditures on long-term care insurance increased more than 2-fold, rising from €25.5 billion (US $30 billion) to €53.9 billion (US $63.4 billion) [2]. To ensure the sustainability of care provision, there is a need for innovative digital solutions. Digital assistive technologies have been shown to have considerable potential in promoting participation and autonomy, as well as reducing care dependency [3]. However, research focusing on the etiology and progression of care needs, as well as the ramifications of digital innovations, has been disproportionately underrepresented to date [4].

Robust and standardized nursing data are essential to evaluate these innovations and their impact on care. The Nursing Minimum Data Set (NMDS) is defined as "a minimum set of items of information with uniform definitions and categories concerning the specific dimension of nursing which meets the information needs of multiple data users in the health care system" [5]. The NMDS provides a structured basis for capturing nursing care across populations, settings, and regions. It facilitates the monitoring of trends in care delivery and resource use, serves as a foundational element for research by establishing connections with broader health information systems, and contributes to the formulation of health policy, thereby ensuring the visibility of nursing contributions in health care planning and decision-making [6]. With a focus on influencing factors and opportunities to prevent or reduce care dependency, the NMDS can also serve as a framework for assessing the impact of digital innovations.

Despite its potential, there is a lack of clarity regarding which elements should be included in a core nursing dataset, how such elements are developed, and what barriers or success factors influence its implementation in practice. The objective of this scoping review is to map the current state of the literature by addressing the following research questions:

1. Which content elements comprise a nursing minimum dataset?
2. Which methodological approaches have been used in its development?
3. What is the implementation and use of such a dataset in different nursing settings?

Methods

This scoping review was conducted in accordance with the JBI (Joanna Briggs Institute) methodology [7] and guided by the framework of Arksey and O'Malley [8], which structures the process into the following phases: identifying relevant studies, study selection, data presentation, and synthesis of results. To enhance rigor and transparency, the reporting process adhered to the PRISMA-ScR checklist (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Checklist 1) [9].

Search Strategy

A systematic search was conducted on July 2, 2025, in the MEDLINE (via PubMed) and CINAHL (via EBSCO) databases using the search term "nursing minimum data set," applied identically in both databases. "Nursing minimum data set" is an internationally established and unambiguous term for the concept of interest. Preliminary searches using broader terms such as "nursing data" or "nursing data set" yielded numerous irrelevant results, mostly referring to general medical datasets including nursing-related variables but not specifically designed for nursing care. Searches in other databases (eg, Google Scholar and Europe PMC) produced even more irrelevant hits.

Using the precise term "nursing minimum data set" maximized specificity and relevance. PubMed's Automatic Term Mapping further captured synonyms and MeSH (Medical Subject Headings) terms. This focused approach aligns with Freguia et al [10] but extends their work by also analyzing the development and implementation of the NMDS. The MEDLINE and CINAHL databases were selected due to their comprehensive coverage of peer-reviewed nursing and health services research. No date restrictions were applied, and reference lists of the included articles were manually screened to minimize the risk of missing eligible studies.

The detailed inclusion and exclusion criteria are enumerated in Textbox 1. Exclusion criteria were based on our aim to map the conceptual, methodological, and implementation aspects of the NMDS rather than secondary applications or specialized derivatives. Studies focusing solely on specific clinical conditions, measurement tools, isolated settings (eg, palliative care), or context (eg, quality management) were excluded. Publications on other datasets (eg, the Minimum Data Set Resident Assessment Instrument [11]) or nursing classifications were also excluded. Language was restricted to English and German to ensure reliable screening and data extraction. Although this may introduce language bias, the majority of conceptual NMDS literature is published in these languages.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Primary and secondary studies with abstracts available
- Language: German or English
- Description of content, development, implementation, and use in practical applications of the Nursing Minimum Data Set

Exclusion criteria

- Secondary analyses of the Nursing Minimum Data Set
- Reference to interventions, specific clinical pictures, or medications
- Reference to quality management, personnel indicators, or specific settings
- Comparison of measurement instruments
- In cases of duplicate publication, only the most comprehensive or earliest version was included to avoid double-counting

Study Selection and Data Extraction

The study selection followed the JBI methodology for transparency and reproducibility. Duplicate removal and screening were conducted using Rayyan (Rayyan Systems, Inc) [12], a web-based collaborative platform enabling independent and blinded review. Duplicates were first removed automatically using the software's matching algorithms based on title, author, DOI, journal, and year and then manually checked to resolve partial matches or inconsistent metadata.

Titles and abstracts of remaining records were independently screened by 2 reviewers (PM and AG) using Rayyan's blinded mode. Records identified as potentially eligible by at least 1 reviewer proceeded to full-text review. Discrepancies were resolved through discussion, with a third reviewer (KG) consulted if consensus could not be reached. The process was independent of study design and focused on studies addressing the conceptualization, development, or implementation of the NMDS. The number of records at each stage is presented in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Reference lists of the included studies were manually screened to identify further eligible studies.

To systematically extract and present relevant information, included studies were narratively summarized in a data-charting form. The studies were categorized and analyzed according to the following classifications: "overview" (eg, definition and elements); "development" (eg, methods and processes); and "implementation" (eg, evaluation, availability, reliability, and validity). The primary categories were derived deductively from the review questions. Subcategories were subject to refinement and expansion in an iterative manner as novel themes and patterns emerged from the included studies. This framework enabled the synthesis of structured and flexible approaches, capturing the breadth of the research field.

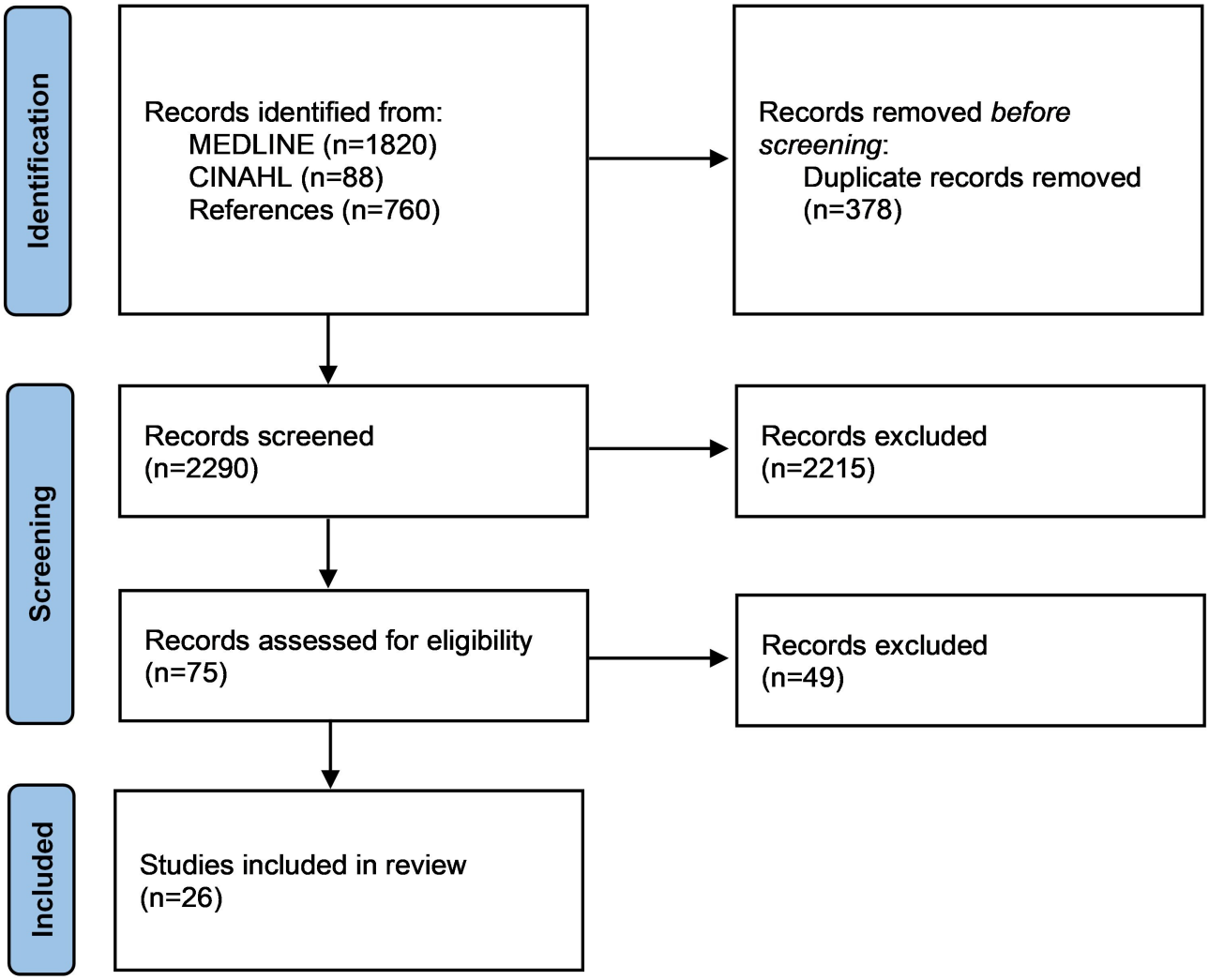
Results

Study Selection

The database search conducted in July 2025 yielded 1908 results (PubMed: n=1820, 95.4%; CINAHL: n=88, 4.6%). After removing duplicates, 1869 (N=1908, 98%) records remained for title and abstract screening. Of these, 1823 (N=1869, 98%) were excluded due to unavailability of full texts or because they focused on data infrastructure, storage, or evaluation frameworks unrelated to NMDS content, development, or implementation. A total of 36 (N=1869, 2%) articles underwent double-blind full-text review. After applying all eligibility criteria, 23 (N=1869, 1.2%) articles were included. Screening reference lists added 3 more studies, resulting in a total of 26 included studies. No conflicts occurred among independent reviewers. The selection process is presented in the PRISMA flowchart [13] (Figure 1). A comprehensive list of the included studies and the data extraction table are provided in Multimedia Appendix 1 [6,10,14-37].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart illustrating the study selection process.



Main Characteristics of the Included Studies

The 26 studies originated from 9 countries. Most were from the Netherlands (n=6, 23.1%), the United States (n=4, 15.4%), Austria (n=4, 15.4%), Belgium (n=4, 15.4%), and Ireland (n=3, 11.5%). Additional studies were conducted in Finland (n=2, 7.7%), and 1 (3.8%) each from Italy, Germany, and Thailand. The majority of reported NMDS initiatives have their origins in the United States [10,16,19,25,29], Belgium [10,19,25,29,32-34,36], Canada [10,19,25,29], Thailand [10,25,27,29,37], Switzerland [25,29], and the Netherlands [18,20-23,25,29]. Additional origins including Ireland [14,26,29], Germany [17,29], Finland [24,25,29,35], Australia [19,25,29], Sweden [25], Brazil [25], Japan [25], Austria [28,30,31], and Scotland [19] were mentioned in several studies, which also referenced a European [10,19,25] and an international NMDS [19,25]. Most of the studies focused on universal nursing settings. However, there have also been investigations of specific settings such as hospitals, long-term care facilities, nursing homes, home health care agencies, and public health contexts.

The articles encompassed a period from 1988 to 2022. The predominant language of publication was English (n=24, 92.3%), with 2 (7.7%) articles published in German. The study designs exhibited heterogeneity, encompassing theoretical papers (n=9, 34.6%), mixed methods studies (n=7, 26.9%), quantitative studies (n=9, 34.6%), and a single qualitative study. The predominant focus was the development of the NMDS [14,17,21,27,28,30,33-36], followed by evaluation [16,20,22,31] and implementation [24,26,32]. A subset of these studies was also partially cited in the included systematic review [29] and umbrella review [10].

Data Elements

The datasets have been designed to capture comprehensive and comparable information on nursing practice, patient status, and care context. This facilitates the monitoring, evaluation, and enhancement of nursing care delivery at both the individual and system levels. A review of the existing literature revealed a considerable variation in the number of elements comprising an NMDS, ranging from 16 to 145 items per dataset. Despite this variation, 4 key categories consistently emerged: patient demographics, medical care information, core nursing care elements, and institutional or organizational data (Textbox 2).

The patient's demographic information is typically documented on an initial admission to the health care facility and subsequently serves as a foundational reference point for the patient's characteristics. The medical care items in question prioritize crucial medical information. While certain elements are documented on a single occasion during the patient's hospital

XSL•FO
RenderX

stay, details concerning medical procedures and medications are frequently recorded throughout the patient's stay. Core nursing care elements are typically collected continuously throughout the duration of hospitalization. The exception to this is nursing outcomes, which are generally assessed at discharge. Risk assessments and nursing diagnoses, on the other hand, are updated as the patient's status changes. Institutional or organizational data link patient- and care-level information with structural and organizational attributes, thereby supporting contextual analysis of care processes.

Textbox 2. The datasets are designed to capture information on nursing practice, patient status, and care context.

Patient demographics (3-8 items)

- Personal identification (pseudonymized)
- Date of birth and age
- Gender
- Race and ethnicity
- Place of residence
- Address and phone number

Medical care information (4-7 items)

- *International Classification of Diseases* codes
- Medical history (patient and family)
- Drug allergies or other allergies
- Procedures
- Laboratory tests
- Medication
- Dead or alive codes and resuscitation status
- Admission or encounter date
- Discharge information (date and summary)
- Length and progression of stay
- Health status (stability, severity, complications, and critical events)

Nursing care elements

- Nursing diagnosis (17-48 items)
- Nursing interventions (8-170 items)
- Nursing outcome (4-40 items)
- Intensity of nursing care (3-4 items)
- Nursing process (10 items)
- Client status (activities of daily living)
- Discharge plan
- Care planning (goals of the shift and goals of nursing care)
- Medical instruments' use before discharge

Institutional or organizational data

- Health care setting (4-5 items): type of institution, main point of service, type of nursing delivery system, and other disposition point
- Resource allocation (4 items): number and qualification of available nurses, number of nursing hours available, number of beds, and nurse-to-patient ratios
- Institutional or provider identifiers: service or agency number, health record number, nurse identifier, and physician identifier
- Expected payer of the bills or health insurance

XSL•FO
RenderX

NMDS Development

The development of the NMDS generally follows a three-stage process: (1) selection and definition of relevant data elements based on literature analysis, document analysis, and interdisciplinary consensus; (2) validation and consensus building through expert rounds, pretesting, and reliability assessments; and (3) application and testing the feasibility in real-world settings.

Selection and Definition of Data Elements

Potential items were chosen for clinical relevance and feasibility through literature reviews, analysis of clinical records (n=15-45 per study), and collaborative expert panels or Delphi processes. Participants included field nurses, nurse managers, nursing scientists, informatics specialists, and policy representatives (n=22-287 per study). Semistructured interviews and questionnaires were central, supported by focus groups and multiple rounds of expert panels (up to 11 with n=5-88 participants). The goal was to ensure that the items reflected actual practice, captured the diversity of nursing activities, and could be coded in a feasible manner. Critical decisions included choosing terminology and coding schemes, unit of data collection (eg, patient day or episode of care), and timing of data entry.

Refinement and Validation

The refinement of items and structures was achieved through consensus-building methods, such as repeated Delphi rounds and focus groups, along with pretests and reliability assessments, including Cohen κ and intraclass correlation coefficient. These methods often incorporated interdisciplinary and practice-based feedback, facilitating the continuous improvement and refinement of the research instruments. In the preliminary phase, the expert participants engaged in deliberations concerning the relevance, frequency, and practical feasibility of each item. In addition, they addressed the conceptualization, frequency, and relevance of the items, along with the data that were routinely collected. Lists of nursing diagnoses, interventions, and outcomes were developed and discussed for clarity and practicality. Subsequent rounds entailed the refinement and evaluation of the wording and definitions for their usefulness and feasibility, as well as the systematic testing of the definitions and coding rules. The resulting data collection instruments provided multiple response options for each item, thereby generating a broad range of analyzable variables. Each data element was carefully defined and accompanied by specific operational instructions, including scoring schemes that frequently used 5-point Likert scales to evaluate interventions and diagnoses.

The documentation of nursing care is typically implemented through the use of standardized classification systems. However, a universal system has yet to be established across all NMDS implementations. Instead, a variety of established frameworks are used, including the International Classification of Nursing Practice [38], the North American Nursing Diagnosis Association-I classification [39], the Omaha System [40], and the Home Health Care Classification [41], among others. The NMDS has been designed to either specify a preferred system or permit cross-mapping between local and international standards. The ability to exchange data and information seamlessly across different systems and between different countries is paramount for the comparability of data at both the national and international levels.

Reliability and validity testing constituted a fundamental component of the second stage. Pretests were conducted using samples ranging from 4 to 214 participants, often comparing the extracted information from patient records (n=20 to nearly n=300,000) by pairs of raters (research staff and clinical nurses). The coding agreement was subsequently quantified. The validity of the content was determined through its alignment with existing literature and user evaluations.

Feasibility Testing and Application

The final development stage focused on evaluating the usability and practical implementation of the NMDS. This included structured pilot applications in real nursing settings, such as hospitals, nursing homes, ambulatory care, and home health agencies. Typically, a longitudinal design with repeated measurements over a defined observation period (eg, 7 d or 1 admission episode) was selected. Consequently, the NMDS demonstrated its efficacy in the domains of care process description, workload measurement, resource planning support, benchmarking, and quality improvement across diverse nursing settings.

One such function is to systematically describe the diversity of nursing care and patient populations through frequency analyses of nursing diagnoses, interventions, and outcomes. The application of advanced analytical approaches, such as relative to an identified distribution analysis, has been demonstrated to facilitate the illustration of variations in nursing practice and complexity across a variety of settings, age groups, and medical diagnoses. In addition, the NMDS is instrumental in evaluating the intensity of nursing care. By establishing a correlation between the implementation of each nursing intervention and the temporal and experiential parameters of the nursing personnel involved, health care institutions can facilitate a quantitative assessment of the workload associated with each patient or unit of care. Classification systems such as the nursing diagnosis index or patient dependency models (eg, the San Joaquin classification) are used to establish a correlation between patient needs and staffing requirements, thereby facilitating evidence-based staffing decisions and resource allocation.

A further central benefit of the NMDS is its ability to demonstrate nursing's contribution to patient outcomes. For instance, path analysis can be used to examine the relationship between nursing interventions and patient progress, as illustrated by improvements in functional indices (eg, the Barthel index) during hospitalization. In the context of quality management and patient safety, outcome indicators facilitate the identification of care problems and the evaluation of improvement strategies through the integration of validated assessment instruments. The longitudinal use of these tools adds further value by enabling the tracking of care trends and patient outcomes over extended periods. These analyses are valuable for clinical managers and policymakers in the context of epidemiological

XSL•FO
RenderX

analysis, benchmarking institutions, highlighting areas of best practice, and supporting the dissemination of evidence-based interventions. This facilitates the formulation of strategic decisions regarding care processes and resource distribution.

A comprehensive implementation strategy must encompass not only technical solutions, such as increasing automation and electronic data collection, but also legal, organizational, and educational strategies. For instance, the implementation of the NMDS in specific contexts was terminated due to the absence of a legal foundation or inadequate institutional endorsement. The quality and utility of collected data are contingent on nurses' familiarity with the NMDS content and vocabulary. Educational initiatives are required to enhance nurses' comprehension of the significance of structured data collection and its pertinence for clinical decision-making, quality assurance, and the development of interoperable hospital information systems. To reduce the workload associated with NMDS documentation, it may be necessary to minimize the scope of collected variables, with a push toward electronic and automated data collection.

Discussion

Principal Findings

This scoping review offers a comprehensive mapping of the extant literature on the NMDS with regard to content elements, methodological development processes, and implementation and use across nursing settings. Synthesizing evidence from diverse international contexts and research designs, this review addresses the critical knowledge gaps identified in the introduction and provides an informed basis for further research and policy development in the field of standardized nursing data.

A salient finding of this review is the substantial heterogeneity observed in the structure and length of reported NMDSs, which vary from 16 to 145 items. Nevertheless, 4 core domains consistently emerged: patient demographics, medical care information, nursing care elements, and institutional or organizational data. This convergence on key domains, despite variation in the level of detail and specific items, underscores both the complexity of capturing nursing care and the international consensus on the essential pillars of nursing-relevant information. The comprehensive integration of nursing diagnoses, interventions, and outcomes signifies a worldwide transition toward outcome-oriented, evidence-based nursing practice. However, the absence of a universally adopted classification system for these categories—variously using systems such as the North American Nursing Diagnosis Association-I classification, the International Classification of Nursing Practice, or local taxonomies—continues to challenge the comparability of nursing data both within and across countries. This diversity underscores the necessity for enhanced harmonization to facilitate broader benchmarking, interoperability, and secondary data use.

The development of the NMDS is characterized by rigorous, iterative, and participatory processes. Most of the studies used a 3-stage approach: first, systematic identification and selection of content elements (frequently drawing from the literature, expert consultations, and clinical record analyses); second, consensus building and validation (frequently relying on Delphi processes, focus groups, and pretesting); and third, pilot implementation in real-world settings. The consistent engagement of field nurses, managers, scientific experts, and policy stakeholders supports both the clinical relevance and the practical feasibility of the resulting datasets. However, methodological diversity—particularly in the degree of stakeholder inclusion, the criteria for item selection, and the approach to validation—limits the extent to which the results can be compared and generalized. A fundamental finding is that consensus and interdisciplinary involvement significantly enhance acceptance and operationalization. However, disparities in contextual adaptation and infrastructure availability necessitate context-sensitive approaches, superseding the notion of a universally optimal solution.

The review demonstrates a wide range of applications of the NMDS, encompassing care process documentation, workload measurement, resource planning, quality assurance, benchmarking, and outcome evaluation. Analytical techniques such as frequency analysis, relative to an identified distribution analysis, and path analysis facilitate the sophisticated use of NMDS data. This includes the identification of care trends and variations, the mapping of nursing workload, and the demonstration of the impact of nursing care on patient outcomes. NMDS data have demonstrated their relevance for supporting clinical management, health policy, and epidemiological surveillance in multiple health care systems.

However, the implementation of NMDS is not without challenges. Achieving success in this endeavor necessitates the implementation of a multifaceted approach, encompassing not only technical solutions such as electronic and automated data collection but also legal, organizational, and particularly educational strategies. In certain settings, the discontinuation of NMDS initiatives was attributed to a lack of legal foundation, inadequate institutional support, or suboptimal integration into health information systems. The collection of consistent and high-quality data is contingent upon nurses' familiarity with the terminology and documentation standards. Consequently, training and change management are imperative. Furthermore, the balancing of data comprehensiveness with feasibility, as well as the minimization of documentation burden, emerges as a recurring theme in the pursuit of long-term sustainability.

Strengths and Limitations

A central strength of this review is its comprehensive and systematic approach, which adheres to established methodological frameworks (JBI and PRISMA-ScR) and uses a transparent, replicable search and data extraction process. The program's emphasis on international literature and a diverse array of study types contributes to the generalizability and relevance of the findings. The use of independent, blinded screening and iterative data charting enhances the rigor and credibility of the synthesis.

Several limitations of this review should be acknowledged. First, only 2 databases and a single search term were used. Preliminary scoping checks indicated that broader terms or additional databases yielded mostly irrelevant results, suggesting

XSL•FO
RenderX

a low likelihood of missing studies specifically addressing the NMDS. As NMDS initiatives are sometimes embedded in national health IT programs, relevant reports may appear in gray literature or regional journals not indexed in the MEDLINE or CINAHL databases. Nevertheless, the potential remains that studies using alternative terminology or published outside the MEDLINE and CINAHL databases were not captured. This structural publication bias should be considered when interpreting the findings. Second, the exclusion criteria may limit the mapping breadth by underrepresenting NMDS adaptations in non–English-speaking and non–German-speaking countries or highly specialized settings. This potential exclusion could introduce publication bias and limit the capture of ongoing or context-specific innovations. Third, as is typical for scoping reviews, no formal quality appraisal of included studies was conducted, which may have affected the weight assigned to individual findings. Fourth, the heterogeneity of study designs, settings, and health care systems complicates a direct comparison and synthesis of findings. To conclude, while this review identifies fundamental content and process elements, as well as prevalent barriers and success factors, it does not provide a meta-analytic estimate of the impact of the NMDS on patient or system outcomes.

Conclusions

This scoping review highlights the heterogeneity and evolving consensus on the structure, development, and implementation of the NMDS. Despite international variations, 4 core domains consistently emerged: patient demographics, medical care information, nursing care elements, and institutional or organizational data. The NMDS can enhance the visibility, quality, and effectiveness of nursing care across health care systems, provided that challenges related to standardization, interoperability, and acceptance are addressed.

To translate these findings into practice, policymakers and health care managers should prioritize the standardization of nursing documentation through appropriate legal frameworks, investment in digital infrastructure, and targeted educational initiatives for nursing staff. Practical steps include establishing interoperable digital nursing records; providing structured training on NMDS use; and implementing pilot programs to assess the feasibility, usability, and data utility. Researchers are encouraged to evaluate the NMDS across diverse settings, conduct comparative effectiveness studies, and explore strategies to reduce documentation burden while maximizing actionable insights for care management, resource planning, and policy decisions. Overall, the review underscores that advancing the NMDS requires interdisciplinary collaboration, context-sensitive implementation, and ongoing evaluation to ensure that standardized nursing data meaningfully inform clinical practice and health system planning.

Funding

This research was funded by the Federal Ministry of Research, Technology, and Space (grant 01ZZ2401A). The study has been funded for a 5-year period, starting in November 2024. The study was funded for €4,559,301.88 (US $5,369,253.83). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. The authors acknowledge the financial support of the Open Access Publication Fund of the Martin-Luther University Halle-Wittenberg.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data extraction sheet.
[XLSX File, 82 KB - medinform_v14i1e84281_app1.xlsx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 119 KB - medinform_v14i1e84281_app2.pdf]

References

1. Behrendt S, Tsiasioti C, Schwinger A. Pflegeprävalenzen und inanspruchnahme in der räumlichen verortung. In: Schwinger A, Kuhlmey A, Greß S, Klauber J, Jacobs K, Behrendt S, editors. Pflege-Report 2024: Springer; 2024:3-25. [doi: 10.1007/978-3-662-70189-8_1]
2. Rothgang H, Müller R. Wirkungen der Pflegereformen und Zukunftstrends: Barmer; 2021.
3. Assistive technology. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/assistive-technology [accessed 2025-08-08]

XSL•FO
RenderX

4. Blüher S, Stein T, Schilling R, Grittner U, Kuhlmey A. Vermeidung von pflegebedürftigkeit – herausforderungen für forschung und praxis. In: Jacobs K, Kuhlmey A, Greß S, Klauber J, Schwinger A, editors. Pflege-Report 2021: Springer; 2021:91-102. [doi: 10.1007/978-3-662-63107-2_6]
5. Werley HH, Lang NM. Identification of the Nursing Minimum Data Set: Springer; 1988.
6. Werley HH, Devine EC, Zorn CR, Ryan P, Westra BL. The Nursing Minimum Data Set: abstraction tool for standardized, comparable, essential data. Am J Public Health 1991 Apr;81(4):421-426. [doi: 10.2105/ajph.81.4.421] [Medline: 2003618]
7. von Elm E, Schreiber G, Haupt CC. Methodische anleitung für scoping reviews (JBI-methodologie). Z Evid Fortbild Qual Gesundhwes 2019 Jun;143:1-7. [doi: 10.1016/j.zefq.2019.05.004] [Medline: 31296451]
8. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
9. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
10. Freguia F, Danielis M, Moreale R, Palese A. Nursing minimum data sets: findings from an umbrella review. Health Informatics J 2022;28(2):14604582221099826. [doi: 10.1177/14604582221099826] [Medline: 35634983]
11. Long-term care facility resident assessment instrument 3.0 user's manual. Centers for Medicare & Medicaid Services. 2025. URL: https://www.cms.gov/files/document/draftmds-30-rai-manual-v1201october2025-1.pdf [accessed 2025-08-25]
12. Hirt J, Nordhausen T. Digitale anwendungen für die studienauswahl im rahmen von systematischen evidenzsynthesen. Pflege 2019 Oct;32(5):277-278. [doi: 10.1024/1012-5302/a000690] [Medline: 31538877]
13. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
14. Butler M, Treacy M, Scott A, et al. Towards a nursing minimum data set for Ireland: making Irish nursing visible. J Adv Nurs 2006 Aug;55(3):364-375. [doi: 10.1111/j.1365-2648.2006.03909.x] [Medline: 16866831]
15. Coenen A, Schoneman D. The nursing minimum data set: use in the quality process. J Nurs Care Qual 1995 Oct;10(1):9-15. [doi: 10.1097/00001786-199510010-00004] [Medline: 7579551]
16. Devine EC, Werley HH. Test of the nursing minimum data set: availability of data and reliability. Res Nurs Health 1988 Apr;11(2):97-104. [doi: 10.1002/nur.4770110205] [Medline: 3363179]
17. Eberl I, Bartholomeyczik S. The Belgian Nursing Minimum Data Set II (B-NMDS II) and its transfer to German hospitals: results of the first investigation phase, the translation and adaption process of the instrument. Pflege 2010 Oct;23(5):309-319. [doi: 10.1024/1012-5302/a000064] [Medline: 20886449]
18. Goossen WT. Exploiting the Nursing Minimum Data Set for the Netherlands. Stud Health Technol Inform 2001;84(Pt 2):1334-1338. [Medline: 11604944]
19. Goossen WT, Epping PJ, Feuth T, Dassen TW, Hasman A, van den Heuvel WJ. A comparison of nursing minimal data sets. J Am Med Inform Assoc 1998;5(2):152-163. [doi: 10.1136/jamia.1998.0050152] [Medline: 9524348]
20. Goossen WT, Epping PJ, Feuth T, van den Heuvel WJ, Hasman A, Dassen TW. Using the Nursing Minimum Data Set for the Netherlands (NMDSN) to illustrate differences in patient populations and variations in nursing activities. Int J Nurs Stud 2001 Jun;38(3):243-257. [doi: 10.1016/s0020-7489(00)00075-4] [Medline: 11245862]
21. Goossen WT, Epping PJ, Van den Heuvel WJ, Feuth T, Frederiks CM, Hasman A. Development of the Nursing Minimum Data Set for the Netherlands (NMDSN): identification of categories and items. J Adv Nurs 2000 Mar;31(3):536-547. [doi: 10.1046/j.1365-2648.2000.01308.x] [Medline: 10718872]
22. Goossen W, Dassen T, Dijkstra A, Hasman A, Tiesinga L, van den Heuvel W. Validity and reliability of the Nursing Minimum Data Set for the Netherlands (NMDSN). Scand J Caring Sci 2003 Mar;17(1):19-29. [doi: 10.1046/j.1471-6712.2003.00112.x] [Medline: 12581291]
23. Goossen WTF. Statistical analysis of the Nursing Minimum Data Set for The Netherlands. Int J Med Inform 2002 Dec 18;68(1-3):205-218. [doi: 10.1016/s1386-5056(02)00077-1] [Medline: 12467803]
24. Kaakinen P, Torppa K. Implementation of a structured nursing documentation in a special care unit. Stud Health Technol Inform 2009;146:367-369. [Medline: 19592867]
25. Mac Neela P, Scott PA, Treacy MP, Hyde A. Nursing minimum data sets: a conceptual analysis and review. Nurs Inq 2006 Mar;13(1):44-51. [doi: 10.1111/j.1440-1800.2006.00300.x] [Medline: 16494666]
26. Morris R, Matthews A, Scott AP. Validity, reliability and utility of the Irish Nursing Minimum Data Set for General Nursing in investigating the effectiveness of nursing interventions in a general nursing setting: a repeated measures design. Int J Nurs Stud 2014 Apr;51(4):562-571. [doi: 10.1016/j.ijnurstu.2013.07.011] [Medline: 23992666]
27. Phuphaibul R. Nursing data set development in Thailand. Stud Health Technol Inform 2006;122:991. [Medline: 17102504]
28. Ranegger R, Hackl WO, Ammenwerth E. A proposal for an Austrian Nursing Minimum Data Set (NMDS): a Delphi study. Appl Clin Inform 2014;5(2):538-547. [doi: 10.4338/ACI-2014-04-RA-0027] [Medline: 25024767]
29. Ranegger R, Ammenwerth E. Nursing Minimum Data Sets (NMDS) - a literature review relating to objective and data elements. Pflege 2014 Dec;27(6):405-425. [doi: 10.1024/1012-5302/a000393] [Medline: 25416487]
30. Ranegger R, Hackl WO, Ammenwerth E. Development of the Austrian Nursing Minimum Data Set (NMDS-AT): the third Delphi round, a quantitative online survey. Stud Health Technol Inform 2015;212:73-80. [Medline: 26063260]

XSL•FO
RenderX

31. Ranegger R, Hackl WO, Ammenwerth E. Implementation of the Austrian Nursing Minimum Data Set (NMDS-AT): a feasibility study. BMC Med Inform Decis Mak 2015 Sep 17;15:75. [doi: 10.1186/s12911-015-0198-7] [Medline: 26384111]
32. Sermeus W, Delesie L, Van den Heede K, Diya L, Lesaffre E. Measuring the intensity of nursing care: making use of the Belgian Nursing Minimum Data Set. Int J Nurs Stud 2008 Jul;45(7):1011-1021. [doi: 10.1016/j.ijnurstu.2007.05.006] [Medline: 17588580]
33. Sermeus W, Van den Heede K, Michiels D, et al. A nation-wide project for the revision of the Belgian Nursing Minimum Dataset: from concept to implementation. Stud Health Technol Inform 2004;110:21-26. [Medline: 15853247]
34. Sermeus W, van den Heede K, Michiels D, et al. Revising the Belgian Nursing Minimum Dataset: from concept to implementation. Int J Med Inform 2005 Dec;74(11-12):946-951. [doi: 10.1016/j.ijmedinf.2005.07.005] [Medline: 16153887]
35. Turtiainen AM, Kinnunen J, Sermeus W, Nyberg T. The cross-cultural adaptation of the Belgium Nursing Minimum Data Set to Finnish nursing. J Nurs Manag 2000 Sep;8(5):281-290. [doi: 10.1046/j.1365-2834.2000.00194.x]
36. Van den Heede K, Michiels D, Thonon O, Sermeus W. Using nursing interventions classification as a framework to revise the Belgian nursing minimum data set. Int J Nurs Terminol Classif 2009;20(3):122-131. [doi: 10.1111/j.1744-618X.2009.01124.x] [Medline: 19659842]
37. Volrathongchai K, Delaney CW, Phuphaibul R. Nursing Minimum Data Set development and implementation in Thailand. J Adv Nurs 2003 Sep;43(6):588-594. [doi: 10.1046/j.1365-2648.2003.02757.x] [Medline: 12950564]
38. Hinz M, Dörre F, König P. In: Tackenberg P, editor. ICNP: Internationale Klassifikation für die Pflegepraxis: Huber; 2003.
39. Herdman TH, Kamitsuru S, Lopes C. NANDA-I-Pflegediagnosen: Definitionen und Klassifikation 2024-2026: Thieme RECOM; 2025.
40. Martin KS, Norris J. The Omaha System: a model for describing practice. Holist Nurs Pract 1996 Oct;11(1):75-83. [doi: 10.1097/00004650-199610000-00011] [Medline: 8945176]
41. Saba VK. Nursing classifications: Home Health Care Classification System (HHCC): an overview. Online J Issues Nurs 2002;7(3):9. [Medline: 12410634]

Abbreviations

JBI: Joanna Briggs Institute
MeSH: Medical Subject Headings
NMDS: Nursing Minimum Data Set
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Benis; submitted 17.Sep.2025; peer-reviewed by AG Kirkova-Bogdanova, MA Zoubi; revised version received 17.Nov.2025; accepted 01.Jan.2026; published 27.Feb.2026.

Please cite as:
Müller P, Gangnus A, Gayen K, Kirchner-Heklau U, Jahn P, Hofstetter S
Core Elements, Development, and Implementation Strategies of the Nursing Minimum Data Set: Scoping Review
JMIR Med Inform 2026;14:e84281
URL: https://medinform.jmir.org/2026/1/e84281
doi:10.2196/84281

© Pascal Müller, Annabell Gangnus, Katrin Gayen, Uta Kirchner-Heklau, Patrick Jahn, Sebastian Hofstetter. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Technological Strategies for the Patient Experience in Emergency Departments: Scoping Review

Andrés Santiago Santafé[1]; Juan Manuel Aranda[1], PhD; Wilmer Jair Beltrán[1]; William Javier Guerrero[1], PhD; Maryory Guevara Lozano[2], MSc; Ingrid Xiomara Bustos[3], RN

[1]Faculty of Engineering, Universidad de La Sabana, Campus del Puente del Común, Km 7 Autopista Norte de Bogotá, Chía, Cundinamarca, Colombia

[2]Faculty of Nursing and Rehabilitation, Universidad de La Sabana, Chía, Cundinamarca, Colombia

[3]Clínica Universidad de La Sabana, Chía, Cundinamarca, Colombia

Corresponding Author:
Juan Manuel Aranda, PhD
Faculty of Engineering, Universidad de La Sabana, Campus del Puente del Común, Km 7 Autopista Norte de Bogotá, Chía, Cundinamarca, Colombia

Abstract

Background: Technology has improved patient care in hospitals, enhancing the overall patient experience. However, digitalization raises questions on effectively integrating technological strategies to ensure assertive communication of information during emergency department (ED) journeys. Keeping patients well-informed boosts their service perception and satisfaction, a factor often neglected by institutions in EDs.

Objective: This paper analyzes relevant studies on technological strategies designed for EDs aimed at improving patient experience, focusing on communication and information access. We analyze the technologies, outcomes, impacts, and challenges of the strategies.

Methods: A scoping review was conducted using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines and CADIMA tool. Searches were performed in Scopus, PubMed, IEEE Xplore, and CINAHL databases. Articles published from January 2018 to December 2024 were included. Quality appraisal was performed using the Crowe Critical Appraisal Tool version 1.4. Three reviewers independently examined the title and abstract for eligibility based on the inclusion and exclusion criteria.

Results: Sixteen eligible studies were included. Four technological strategy categories were identified: artificial intelligence−based, simulation-based, infrastructure and hardware technologies, and interfaces and information systems. Mobile and web applications were the main technologies adopted in the studies.

Conclusions: Technological strategies hold significant potential to enhance patient experiences in EDs by providing real-time updates on medical status and care progress. However, their effectiveness depends on usability, literacy, and system design. Existing literature highlights the impact and challenges of deploying and using these strategies in EDs. However, no studies have systematically evaluated long-term outcomes or cost-effectiveness across diverse ED settings.

(JMIR Med Inform 2026;14:e79782) doi:10.2196/79782

KEYWORDS

scoping review; PRISMA-ScR framework; emergency department; patient journey; communication technology; patient experience

Introduction

The emergency department (ED) constitutes the gateway to health care for many people, being the place where treatment begins for a wide spectrum of diseases, some of which are life-threatening and require immediate attention [1]. The typical patient journey in the ED involves triage classification, medical evaluation, necessary examinations, medications, or procedures, reevaluation, and finally discharge or observation [2]. However, challenges such as overcrowding, uncomfortable facilities, and poor communication continue to hinder timely and efficient ED patient care, negatively impacting both patient experience and institutional performance [3]. In fact, effective and assertive communication of information about the ED patient journey has been found to be associated with patient satisfaction; thus, studies indicated that the lack of information on the progress and delays has a greater effect on patient satisfaction than perceived waiting times while visiting the ED [4]. In addition, it leads to feelings of powerlessness for patients and their family members [5]. Patients who feel well informed tend to have a better perception of the service [6], leading to higher satisfaction with the care and reduced anxiety [7], which in turn contributes to greater institutional efficiency [8].

XSL•FO
RenderX

With recent technological innovations, the potential to address communication challenges in EDs has increased significantly. Technologies that enhance communication and improve information accessibility have shown promise in improving the patient experience in the ED [9]. However, due to the rapid pace of technological advancement and the growing use of digital solutions in health care, there is an urgent need to review existing evidence to guide future implementation efforts.

This scoping review aims to analyze relevant studies and describe the evidence on technological strategies implemented in EDs worldwide that enhance patients' experiences during their journey in the ED. A technological strategy refers to the approach and implementation of technologies and supporting infrastructure within ED workflows to improve patient experience throughout their journey. Regarding the ED patient experience, the review focuses on two dimensions: communication and access to information. Communication refers to the exchange of timely, accurate, and comprehensible information between health care providers and patients or family members regarding medical status and progress updates during the ED visit. Access to information refers to patients' ability to obtain and understand relevant details about their journey in the ED. The following research questions (RQ) were addressed:

- What technological strategies exist to improve patient experience in EDs, considering dimensions such as communication and information accessibility?
- What infrastructure and technologies support technological strategies?

The scoping review offers a comprehensive and interdisciplinary perspective, integrating insights from clinical, technological, and human-centered design domains. The primary innovation of this study lies in its integrative and multidimensional approach to evaluating technological interventions in EDs, specifically through the lens of patient communication and information accessibility—dimensions often underrepresented in prior reviews. Unlike earlier works that focus predominantly on clinical outcomes or operational efficiency, this review foregrounds the patient experience and satisfaction as central evaluative criteria. It also identifies a critical knowledge gap: the lack of standardized frameworks for evaluating the long-term effectiveness, cost-efficiency, and scalability of these technologies across diverse health care settings. Additionally, the review highlights the insufficient attention given to the adaptability of these tools for vulnerable populations, such as those with low digital literacy or disabilities, and the need for participatory design approaches to ensure equitable and sustainable implementation.

Methods

Databases and Search Keywords

To analyze the current and future landscape of technological strategies in EDs, a scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [10]. Studies were retrieved from Scopus, PubMed, IEEE Xplore, and CINAHL in January 2025. The inclusion of CINAHL and PubMed ensures coverage of nursing and clinical perspectives that are highly relevant to patient care and workflow improvements, complementing the technical focus of IEEE Xplore and the multidisciplinary scope of Scopus. The keywords used in the search queries included technology, strategy, ED, communication, and access to information. The complete search strategy was as follows: ("Technolog*") AND ("Strateg*" OR "Approach*" OR "Solution*" OR "Framework*" OR "System*" OR "Model*") AND ("Emergency Department*" OR "Emergency Service, Hospital" OR "Emergency Room*" OR "Emergency Medical Service*") AND ("Communication" OR "Access to Information"). The final database search queries for the period January 2018 to December 2024 are summarized in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

To ensure the quality of the articles, only those that met the inclusion and exclusion criteria presented in Textbox 1 were considered in the review process. These criteria were chosen to focus on studies that specifically examine technological strategies designed for EDs and their impact on patient experience, particularly in the dimensions of communication and information accessibility. Three reviewers independently applied the criteria to refine the list of eligible articles. In cases of uncertainty, the decision was made collectively after discussion.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Documents published between 2018 and 2024 (including both years)
- Documents published in peer-reviewed journals or conference proceedings
- Full-text articles available in English
- The research setting is at least a hospital or an emergency department (ED), regardless of the age or sex of the patients who have used the ED
- Technological strategies specifically designed for EDs aimed at improving the patient's experience during their ED journey, focusing on communication and access to information
- Studies reporting on at least one patient-centered outcome (satisfaction, experience, or well-being)

Exclusion criteria

- Technological strategies outside the health care sector or those presented as telemedicine, remote medicine, or prehospital solutions
- Patients who have not used hospital or ED services
- Articles that do not present a descriptive evaluation, case study, or empirical evidence regarding the design, implementation, or impact of technological interventions on patient experiences in EDs
- Protocol-only studies or incomplete data reports

Study Selection

A 3-step filtering process was implemented to refine the selection of studies relevant to our RQ:

- In the first filter, *identification*, duplicates were removed. This process ensured the uniqueness of each record for accuracy and to avoid redundancy.
- The second filter involved *screening* by title and abstract, which significantly narrowed down the selection, based on our inclusion and exclusion criteria. This step focuses on identifying papers that are most relevant to ED's technological advancements.
- The third filter involved a full-text review to assess the *eligibility* of studies that addressed our RQ, culminating in the final selection of studies included in the review.

CADIMA was used for both the first and second filtering stages. CADIMA is an open-access online tool developed by the Julius Kühn-Institut during the EU-funded Global Response Against Child Exploitation project to support the conduct of systematic reviews [11].

Data Charting

The data charting process for the studies included in this review was structured into three rigorous rounds to explore the categories of technological strategies in EDs comprehensively. In the first round, we extracted fundamental attributes from each study, such as the authors, publication year, journal, and keywords. This process allowed us to understand the distribution of research over time and identify key categories within the field. A standardized data charting form was jointly developed in Microsoft Excel by three reviewers and piloted on a subset of studies to ensure consistency and clarity before full data extraction.

In the second round, we delved deeper into technological strategies applied in the studies. The charting form included variables such as the affected dimension (communication or access to information), technology and infrastructure used, impact on patient experience, measurement tools for usability or satisfaction, preferred information by patients, implementation details (approach, settings, challenges), and implementation outcomes.

Finally, in the third round, each paper was independently reviewed by three reviewers to ensure consistency and reduce bias in the analytical process. Any discrepancies were discussed and resolved by consensus. The charting process was iterative, and adjustments were made to the Excel form when new variables or categories emerged.

Quality Appraisal

Each selected paper underwent a quality appraisal conducted independently by two reviewers using the Crowe Critical Appraisal Tool version 1.4 (CCAT) [12]. This step was undertaken to assess the methodological rigor and credibility of the included evidence, supporting a more reliable interpretation of the findings in relation to our research questions. The CCAT is a widely used instrument recognized for its high reliability in evaluating research papers [13]. This instrument is organized into 8 categories comprising 22 items. Each category is scored on a 6-point scale ranging from 0 (lowest) to 5 (highest). The overall score is calculated as a percentage, and we arbitrarily considered acceptable a score above 60%. Prior to the appraisal, both reviewers were familiarized with the CCAT manual to ensure consistent interpretation of criteria. In cases of discrepancies between the reviewers' assessments, consensus was reached through discussion of the arguments and evidence.

Synthesis of Results

We grouped the studies by technological strategy category and summarized their associated challenges, as well as related patient experience outcomes and impacts. The strategy categories were defined using a content-based grouping approach: we reviewed all included studies, examined the technologies and their stated purposes, and clustered similar technologies based on their primary function and orientation toward improving ED patient

XSL•FO
RenderX

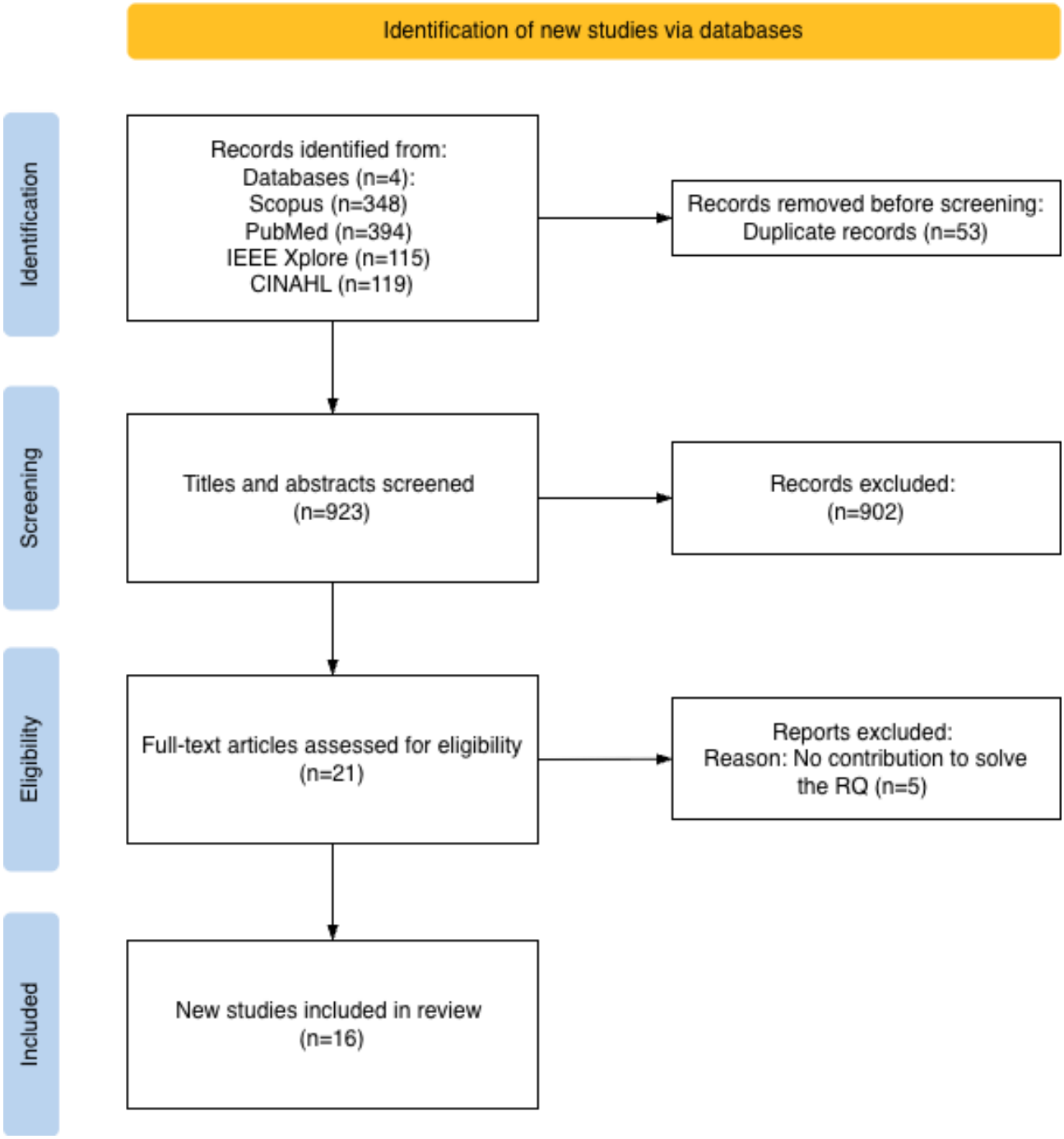
experience. The synthesis mapped the relationship between each strategy and its impact on communication and access to information, highlighting key outcomes, impacts, and challenges. Technologies and IT infrastructures were tabulated by strategy category, while patients' preferred information and satisfaction improvements were synthesized narratively.

Results

Flow Diagram and Number of Studies

The scoping review process is detailed in the flow diagram shown in Figure 1, presenting the stepwise filtering and the number of studies included at each step. In the *identification* step, 976 papers were identified and selected. Papers were imported into the CADIMA, where duplicates were removed, resulting in 923 papers. These papers were *screened*, applying the defined inclusion and exclusion criteria, where 21 papers were classified as eligible. Finally, after *full-text* reading, 16 papers were selected for further analysis; 5 articles did not contribute to the RQ.

Figure 1. Flow diagram of the scoping review process. RQ: research question.



XSL•FO
RenderX

General Description and Characteristics of Studies

The set of 16 studies, published between 2018 and 2024 in journals such as *Journal of Medical Internet Research* and *Applied Clinical Informatics*, examines a range of technological strategies designed to improve patient experience in EDs. These studies address key aspects, including the design and implementation of solutions, the technological and infrastructural components that support them, patients' preferences for information delivery, system usability, and the assessment of their impact on patient satisfaction and overall experience. Methodologically, the articles include qualitative evaluations, observational pilot studies, and mixed methods research, reflecting a diversity of approaches (see Multimedia Appendix 2).

The reviewed literature reveals an interdisciplinary orientation, with collaborations among computer scientists, clinicians, designers, and data analysts. Notably, studies originate from diverse regions, including the United States [14-19], the United Kingdom [20], Italy [21], Israel [22,23], Denmark [5], Australia [24], Thailand [25], Malaysia [26], South Korea [27], and Greece [28]. Finally, the quality appraisal of the 16 studies revealed that 87.5% (n=14) were of high quality and 12.5% (n=2) of medium quality (see Multimedia Appendix 2).

Key Findings on Technological Strategies

Technological Strategies

Based on revised studies and following the process described in the Synthesis of Results subsection, we categorized the technological strategies that help to enhance patient journey experiences in EDs into four primary categories, as outlined below.

1. Interfaces and information systems: This category encompasses tools and information systems designed to provide access to comprehensive patient medical information and ED stage updates, enhancing patients' understanding of their journeys through the ED and facilitating communication with medical staff.
2. Infrastructure and hardware technologies: This category includes strategies focused on the underlying IT infrastructure necessary for ED operations. It encompasses advanced computing systems, integration devices, and hardware that are critical for patient care and tracking.
3. Artificial intelligence (AI)–based strategy: This category includes strategies that utilize machine learning (ML) and deep learning models for predictive analytics and pattern recognition in EDs.
4. Simulation-based strategy: This category encompasses simulation strategies, which are essential for evaluating the performance of health care operations, analyzing patient flow, and modeling complex interactions within ED settings.

Figure 2 illustrates the distribution of these strategy categories by dimension within the reviewed literature. Interfaces and information systems strategies were the most frequently cited, appearing in 14 studies, followed by infrastructure and hardware technologies and AI-based strategies, each appearing in 3 studies, and the simulation-based category, appearing in 2 studies. All categories addressed both dimensions, except the simulation-based category. These findings indicate a clear preference for integrated solutions and reveal a gap in studies dedicated exclusively to communication strategies.

XSL•FO
RenderX

Figure 2. Distribution of articles by technological strategy category and dimension focus.

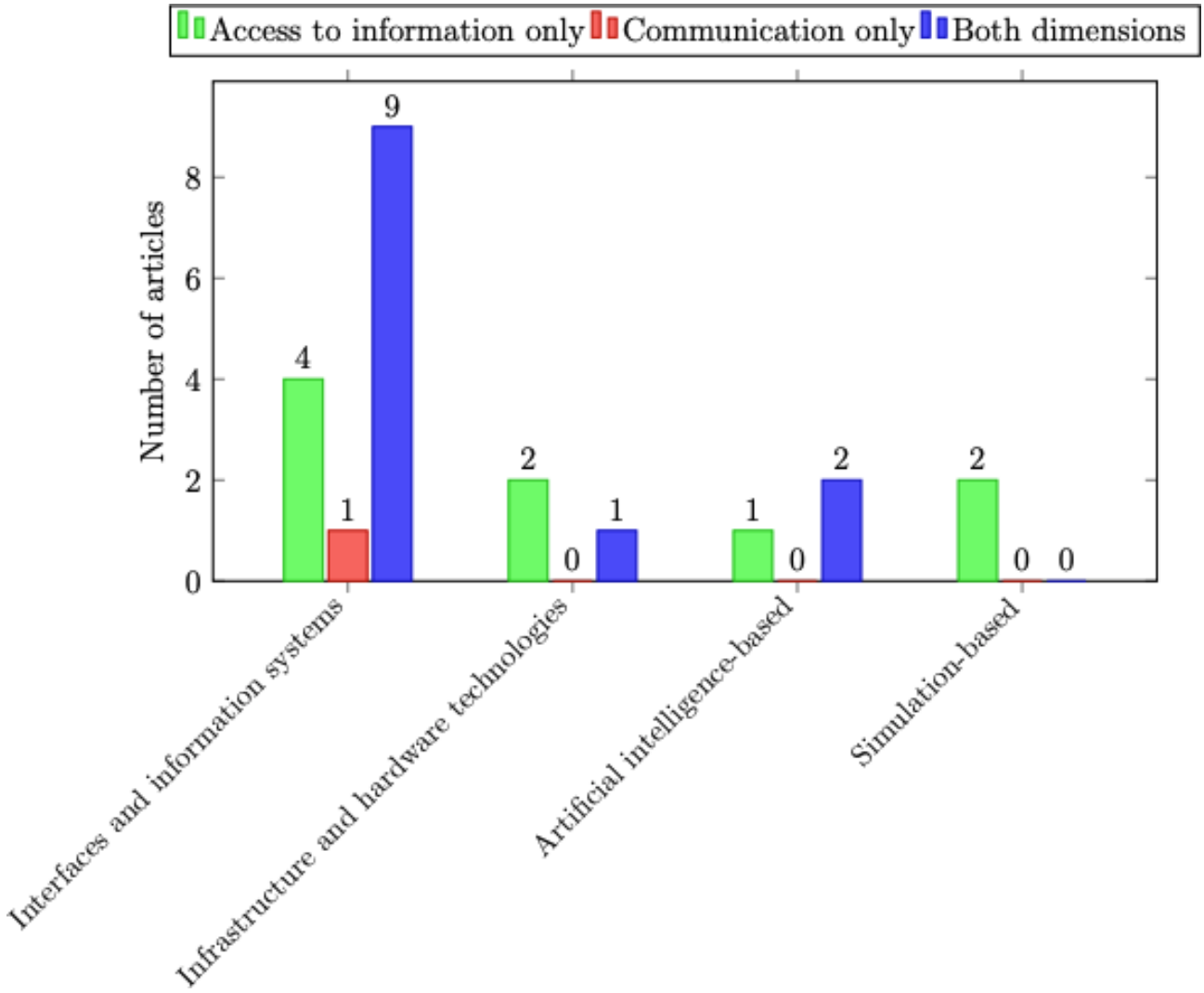


Table 1 provides a comprehensive overview of the key technological strategies implemented in EDs to enhance patient experience. For each strategy, the table details the specific dimension(s) addressed (communication and access to information), the implementation approach used and the target population (patients, family members, or health care staff), the measurement methods used to assess outcomes, the observed outcomes and their impact on patient experience and satisfaction, and the implementation challenges encountered. The results reveal a strong emphasis on systems that integrate both communication and access to information, particularly through interfaces and information systems such as digital whiteboards [16], clinical logic platforms [5], and discharge tools [17,24]. These solutions consistently aim to enhance patient and family engagement, improve satisfaction, and streamline information flow during care transitions [22].

XSL•FO
RenderX

Table . Key technological strategies for patient experience in EDs.

References	Dimension	Strategy category	Description	Measurement methods	Outcomes and impact	Challenges
Westphal et al [22]	Comm[a] and ATI[b]	AI, Int & IS[c]	• MyED: Web application providing real-time information about ED[d] journeys, procedures and expected waiting times • Target population: patients, family	Deployment survey, interviews, usage analytics	Improved patient understanding of the ED journey; positive reactions to comprehensive health information; enhanced transparency	Delays in synchronizing data from hospital databases; limited engagement period due to short ED visits
Marshall et al [16]	Comm	Int & IS	• Digital whiteboards: E-paper displays showing real-time orienting and clinical information in ED rooms • Target population: patients, Family	Satisfaction survey (Likert), custom quantitative perception survey	96% of patients preferred rooms with whiteboards; higher overall ED satisfaction; more likely to recommend hospital to family/friends	Not integrated with EHR[e]; updates not truly real-time; potential privacy concerns
Østervang et al [5]	Comm and ATI	Int & IS	• Cetrea Clinical Logistics for patients: hospital information system displaying time estimates, staff names, and educational videos • Target population: patients, family	SUS[f] surveys (Likert), qualitative semistructured interviews	Increased sense of control reduced uncertainty; improved communication with health care professionals; positive impact on family members; usability rated close to excellent (83.6 SUS)	Difficult for older patients; privacy concerns; IT support dependency; content limitations
Steel et al [17]	ATI	Int & IS	• MyEDCare: Mobile app for discharge instructions and information comprehension • Target population: patients	ED Consumer Assessment of Health Care Providers and Systems patient satisfaction scores, usage analytics	Reduced 72-hour, 9-day, and 30-day unscheduled return ED visits; positive impact on patients' perceptions of ED nursing; improved patient safety	Patient difficulties understanding complex information; technical issues; software bugs
Doyle et al [24]	Comm and ATI	Int & IS	• PETS[g]/REDCap[h] digital discharge system: discharge system: digital tools for pediatric patient registry and discharge information. • Target population: patients, family	SUS scores (Likert), confidence surveys, semistructured interviews	Parents felt more confident with digital discharge information; preferred digital engagement for clinical research; high usability ratings (94 SUS)	Parent concerns about understanding instructions; technical issues

XSL•FO
RenderX

References	Dimension	Strategy category	Description	Measurement methods	Outcomes and impact	Challenges
Jaafar et al [26]	Comm and ATI	I&H[i]	• Patient Tracking System (TrackMe): Mobile app with wireless wristbands for real-time updates on patient location • Target population: staff, family	Satisfaction surveys (Likert)	Enhanced family engagement; decreased anxiety through real-time updates about patient status and location; high user satisfaction (mean >4.0)	Resistance to adoption among staff; training needs; connectivity and infrastructure requirements

[a]Comm: communication.

[b]ATI: access to information.

[c]Int & IS: interfaces and information systems.

[d]ED: emergency department.

[e]EHR: electronic health record.

[f]SUS: System Usability Scale.

[g]PETS: Parent Engagement Through Technology Solutions.

[h]REDCap: Research Electronic Data Capture.

[i]I&H: infrastructure and hardware.

Common measurement methods include satisfaction surveys, usability scales, and structured interviews, reflecting a focus on patient-centered evaluation (see the *Measuring Usability, Patient Experience, and Satisfaction Improvement* subsection for more information). Reported outcomes and impact generally include improved understanding of care processes [22], reduced anxiety [5,26], and better communication with health care professionals [5], although challenges persist, such as technical limitations [17], integration with existing systems [16], and resistance to adoption [26]. Notably, infrastructure-based tools such as patient tracking systems [26] complement these efforts by supporting real-time updates, but they face barriers related to cost and implementation complexity. Overall, the evidence suggests a trend toward multidimensional strategies that prioritize transparency and communication while highlighting gaps in scalability and interoperability. In addition, the literature provides limited discussion on how these strategies can adapt to dynamic changes in ED protocols [22], ensure equity among diverse patient populations—including individuals with disabilities or lower digital literacy [5,18,24]—and effectively support digitally underserved or vulnerable groups [16,17].

Technologies and IT Infrastructures

Table 2 provides an overview of the main technologies and IT infrastructures used in ED settings, grouped by strategy category.

XSL•FO
RenderX

Table . Overview of technologies and IT infrastructures used in EDs, grouped by strategy category.

Strategy category and technology/infrastructure	Description and outcome	References
Interfaces and information systems		
Mobile app	Provides a patient-friendly health information and communication interface through a smartphone app, designed to deliver real-time and comprehensive medical data—such as health status, triage classification, vital signs, discharge instructions, and location within the ED[a] journey—while also enabling communication with patients' next of kin and supporting guidance for effective illness management, thereby improving health care services and enhancing patient satisfaction.	[17,20,26-28]
Web dashboard	Delivers a user-friendly, graphical web interface that provides real-time updates on patients' medical status—including triage classification, clinical conditions, and vital signs—along with their ED treatment progression, open clinical note access, care pathways, and discharge instructions. This interface enhances transparency and patient engagement during the ED stay.	[5,14,18,19,22-25,28]
Digital whiteboard	Provides an electronic paper interface via a digital whiteboard that communicates relevant real-time clinical information to ED patients, including status updates, orientation details, care team assignments, safety alerts, patient orders, and disposition information. This interface enhances patient understanding of their journey through the ED.	[16]
Patient-centered information system	Provides an information system accessible across various devices and interfaces (web and desktop applications), designed to deliver real-time updates and detailed information—including procedures, treatments, associated waiting times, nurse and physician responsible for care, and current and upcoming stages—directly to patients and their family members. This enhances patients' understanding of their individual journeys through the ED and supports informed decision-making and communication.	[5,22]
Infrastructure and hardware technologies		
Internet of Medical Things	Includes IoT[b] devices (eg, wearables) for real-time monitoring of ED patients, providing vital insights into their health status, vital signs, patient stratification for special services, and physical location within the ED. This information supports clinical decision-making and enhances communication by delivering timely and essential updates to both medical staff and patients' families about the health care process.	[26-28]
Computing infrastructure	Covers advanced computing infrastructures, such as cloud computing, that support real-time data storage, processing, and analysis in EDs.	[26,28]
Artificial intelligence–based		
ML[c] algorithms	Encompasses ML algorithms like random forest for predictive analytics in EDs, providing insights into patient treatment, stratification, and predicting waiting times based on workload and patient characteristics.	[14,15,22]
Simulation-based		

XSL•FO
RenderX

Strategy category and technology/infrastructure	Description and outcome	References
Discrete-event simulation	Provides a simulation-based model in R to estimate individual treatment times from aggregated ED data, aiming to assess delays, improve waiting time estimates, and enhance understanding of patient flow.	[21]
Petri net model	Comprises a simulation-based model in Python using Colored Petri Nets to model and analyze ED processes and workflows, supporting improvements in patient flow across treatment paths and enhancing decision-making efficiency.	[15]

[a]ED: emergency department.

[b]IoT: Internet of Things.

[c]ML: machine learning.

Interfaces and information systems emerge as the most significant strategy category for directly improving patient experience through enhanced communication and transparency (14 studies). Within this category, *mobile* and *web dashboard applications* are the most widely adopted technologies (5 and 9 studies, respectively). For example, the Parent Engagement Through Technology Solutions (PETS) [24] and MyEDCare [17] mobile apps support patients and caregivers during discharge by improving comprehension and reinforcing key information about the ED journey. Complementing these mobile tools, the MyED system [22] and the Cetrea Clinical Logistics (CCL) dashboard for patients [5] were designed as web dashboard applications to enhance patients' understanding of their care trajectory by displaying completed, ongoing, and anticipated procedures, as well as identifying the clinicians responsible for their care. Additionally, digital whiteboards, leveraging E-paper technology, deliver real-time notifications on delays and post-discharge instructions, functioning as integral components of the IT infrastructure [16].

Building on these interface-focused solutions, *infrastructure and hardware technologies* provide the foundation for real-time monitoring and data exchange. This category integrates *Internet of Medical Things* (*IoMT*) components with robust *computing infrastructure*. IoMT technologies enable real-time sensing and location tracking through dedicated hardware: beacon devices worn as patient bracelets and beacon readers (gateways) enable continuous wireless monitoring of patient location and status [26]; wearable biosensors capture physiological signals such as heart rate and oxygen saturation, supported by real-time location devices, Wi-Fi hotspots, and Bluetooth-to–Wi-Fi gateways for data transmission [28]; and near-field communication (NFC) tags paired with mobile electronic medical record (EMR) devices (eg, smartphones) allow location-based access to patient information [27]. Computing infrastructure complements these technologies by offering backend systems for secure data storage, processing, and decision support, including cloud-based databases and local servers operating within private networks [26,28]. Examples of strategies include TrackMe [26], which combines beacon-enabled bracelets with a mobile app and a 3-tier architecture for real-time updates; the NFC-integrated Mobile EMR System [27], which accelerates access to patient records; and IntelTriage [28], which uses wearable biosensors and LAN-based communication to monitor vital signs and geo-location, supported by a decision-support subsystem for dynamic prioritization.

Building on these physical infrastructures, *AI-based* strategies leverage computational techniques to enhance decision-making and operational efficiency. Predictive models such as *machine learning algorithms* [14,15] and random forest [22] approaches analyze EMR data—including length of stay, procedures, and clinical notes—to generate actionable insights. For example, the AI-enabled non-English language preference prioritization strategy [14] computes complexity scores for patients with a non-English language preference, while the AI/ML model presented in Gehlot et al [15] provides deeper insights into treatment details and patient experiences, thereby supporting more accurate patient routing based on acuity. Similarly, the MyED strategy [22] combines random forest models with operations research techniques, such as process discovery and queue mining, to predict waiting times for specific procedures and deliver real-time updates to patients through a responsive mobile interface.

Finally, *simulation-based* approaches complement predictive analytics by modeling complex ED workflows to optimize patient flow. Techniques such as *discrete-event simulation* and *Petri Net models* are used to estimate waiting times, identify bottlenecks, and validate algorithms for time-to-treatment predictions. These models support more reliable and equitable resource allocation. Reported strategies use these models to simulate care pathways and assess workflow constraints, enabling validation of predictive algorithms for waiting time and treatment prioritization [21]. A notable example is presented in Gehlot et al [15], where a Colored Petri Net–based simulation is integrated with Python-based AI/ML components to evaluate and refine triage processes, aiming to improve clinical outcomes and enhance patient experience.

Further details on usability, patient experience, and satisfaction impacts are provided in Table 1 and discussed in the subsection *Measuring Usability, Patient Experience, and Satisfaction Improvement*.

Patient's Preferred Information

The selected studies displayed a wide variety of topics that patients preferred on their journey through the ED. Specifically, patients preferred detailed information on their medical status

XSL•FO
RenderX

and progress updates [18,26], including triage status [15,28], treatment, and discharge instructions [17,19,24]. Additionally, patients valued comprehensive medical information related to their vital signs [18,25,28], laboratory results [27] and medical orders [16,18], nursing care plans [15,25,28], and the nurse and physician responsible for care [5,16]. Patients also appreciated personalized medical information and updates on procedures, associated waiting times [5,21], as well as current and upcoming stages of their ED stay [16,22]. They also favored precise tracking of their location in the ED [26,27], safety, and disposition information [16]. These preferences highlight a significant trend toward personalized health care information, underscoring the need for technological strategies to focus on delivering precise, individualized health data and highlighting the importance of accuracy and clarity in the patient communication tools of the ED.

Measuring Usability, Patient Experience, and Satisfaction Improvement

Usability evaluations are essential in health care technology to identify design flaws early and prevent ineffective implementations [24]. Among reviewed studies, the System Usability Scale (SUS) was the most frequently used instrument [5,18,24,27]. SUS is a 10-item Likert questionnaire that generates a score from 0 to 100, interpreted using a grading scheme (eg, >90: “A” [excellent]; <60: “F” [poor]). To illustrate, the PETS mobile discharge tool scored 94/100 (A grade—Excellent), indicating superior design and strong patient acceptance [24]. Similarly, the CCL patient information system achieved near-excellent usability with a mean SUS score of 83.6/100 (B+ grade—Acceptable) [5]. In contrast, NFC-integrated mobile EMR and REDCap eConsent interfaces scored 71.9/100 and 78/100 (B grade—acceptable), reflecting moderate usability and less direct impact on patient experience [24,27]. The electronic health record display for ED patient health status and care plan progress received a mean SUS score of 69.6 out of 100 (C grade—marginally acceptable) [18], suggesting limited effectiveness in improving patient experience.

While SUS provides a standardized measure, complementary approaches are often necessary to capture user perceptions more comprehensively. Likert-scale satisfaction surveys (typically 5- or 10-point) are commonly used to assess design quality, content relevance, and overall user-friendliness [16,25,26]. For example, the web application for emergency trauma patients, using a 10-point Likert scale, reported a mean usability score of 4.09 (SD 0.74) [25]. Similarly, the TrackMe patient-tracking system, using a 5-point Likert scale, reported a mean usability score of 4.0 (SD 1.2) and a mean satisfaction score of 4.4 (SD 1.0), indicating a generally positive user experience during ED waiting periods [26]. These findings reinforce that usability and satisfaction are interrelated but distinct constructs, underscoring the need for multidimensional assessment.

Patient experience and satisfaction are closely linked to usability, which strongly influences technology acceptance in clinical settings. In ED environments, these areas are typically assessed using mixed methods that combine SUS scores, Likert-scale measures, and validated instruments such as Hospital Consumer Assessment of Health Care Providers and Systems (CAHPS) and ED-CAHPS [16,17], complemented by semistructured interviews for contextual insights into patient perceptions [5,22,24]. For example, semistructured interviews in the MyED study showed that patients who used the solution demonstrated significantly better understanding of their ED journey compared with nonusers ($F_{8,299}$=2.519; P=.01) [22]. Using satisfaction surveys, the digital whiteboards intervention improved patient-reported communication about delays (1.96 vs 2.48; P=.03) and clarity of post-discharge instructions (2.41 vs 2.83; P=.02), and users were more likely to recommend the facility (9.66 vs 9.22; P=.05) [16]. Operational benefits were also reported in the MyEDCare study: patients receiving digital discharge instructions had significantly fewer unscheduled ED returns at 72 hours and 30 days compared with those receiving standard paper-based discharge (P=.003 and P<.001) and higher ED-CAHPS scores in nursing communication domains (eg, “Nurses explain in a way you understand”: 86.5 vs 81.9; P=.03) [17].

Collectively, these findings underscore that usability is not merely a design metric but a critical driver of patient experience, satisfaction, and even clinical outcomes in ED technology implementations. In addition, the evidence shows that *Interfaces and Information Systems*—particularly *mobile and web applications*—provide the greatest added value for improving patient experience, primarily by enhancing communication, transparency, and comprehension.

Discussion

Principal Findings

The integration of advanced digital technologies in EDs represents a rapidly evolving but still unstructured field. Among the 16 studies included in this review, 9 reported interventions that achieved measurable improvements in communication, patient satisfaction, or reductions in service fragmentation [5,16-18,22,24-27]. Despite the promising results (see subsection *Measuring Usability, Patient Experience, and Satisfaction Improvement*), the absence of standardized frameworks, definitions, and evaluation models limits comparability and scalability across settings [29]. Developing a shared conceptual and operational framework is essential to ensure consistency in research and guide technology implementation within ED systems.

Patient-facing technologies such as mobile apps, real-time dashboards, and web-based platforms have enabled patients to access individualized health information, often in real time, enhancing their sense of involvement in emergency care [22]. Some tools demonstrated significant benefits—for example, MyEDCare improved patient comprehension and experience (86.5 vs 81.9; P=.03) [17], and digital whiteboards reduced perceived uncertainty (2.41 vs 2.83; P=.02) [16]. However, implementation challenges were common. The MyED system faced difficulties synchronizing live data from hospital databases, and short ED visits limited patient engagement [22]. Usability issues were also reported among older adults, who requested simplified functions [5]. These findings underscore the need for participatory design approaches that incorporate patient perspectives prior to implementation [30]. Ultimately,

XSL•FO
RenderX

usability remains a critical determinant of successful adoption and effective use [31].

Technological impact extends beyond patient interaction to clinical workflows. Real-time data systems and mobile-based discharge tools require adjustments in routines, new training protocols, and additional coordination. While these tools can enhance communication and data accuracy, they often introduce extra responsibilities, increasing cognitive and operational burden on health care professionals [24]. This tension—between intended benefits such as efficiency and the reality of added workload—has been documented in multiple studies [32,33]. Understanding nursing staff perceptions and acceptance is therefore essential for successful adoption [34]. Policymakers and ED managers must consider these trade-offs when planning technological strategies to avoid overburdening staff while pursuing efficiency goals [16].

Additionally, the integration of wearable devices (eg, TrackMe, IntelTriage) and patient-facing platforms (eg, CCL, MyED) is reshaping care dynamics by fostering patient autonomy [5] and involvement in decision-making [20,23]. Although these innovations hold promise for improved outcomes and satisfaction [16,17,23], they raise concerns regarding workflow implications (eg, TrackMe increased nurses' workload [26]; fears of delays with MyED [22]), digital literacy barriers (eg, limited access to Open Notes among non-English speakers [23]; usability issues for older users of CCL [5]), and communication norms (eg, risks of patients viewing information before staff [22]; potential reduction in nurse/physician time) [5]). These factors must be addressed to ensure equitable and safe implementation.

Finally, the economic cost required for the investment, implementation, and maintenance of advanced technologies in EDs must be carefully considered. These expenses may include infrastructure upgrades, software development, technical support, and training for health care personnel [19]. Given the potential financial implications, future studies should explore the cost-benefit relationship of these technological strategies, assessing their feasibility, sustainability, and overall value in diverse health care settings [17].

Strengths and Limitations

A key strength of this study is its multidimensional approach, which provides a holistic view by examining various facets of ED care, including communication and information accessibility. This comprehensive perspective was achieved by focusing on emerging technologies, such as AI and IoT, which represent the forefront of operational efficiency and diagnostic accuracy in EDs. The recent literature (2018 - 2024) grounds the study's findings in current technological trends, ensuring relevance. However, limitations include a focus on English articles, which may have excluded relevant studies in other languages, potentially restricting generalization. Furthermore, since many technologies analyzed are relatively new (eg, IoMT, AI), few studies directly address their long-term impacts. Variability in environments where these technologies are implemented can also affect reproducibility, especially in resource-limited settings. The recent temporal range may not capture emerging trends or long-term effects, highlighting the need for ongoing research to understand their sustained impact.

Conclusions

This scoping review synthesized current technological strategies designed to enhance patient experiences in EDs, primarily through improved communication and access to information. The limited number of studies identified (n=16, published between 2018 and 2024) underscores the nascent and fragmented nature of this research area. Existing evidence suggests that patient-facing tools—such as mobile apps, real-time dashboards, and web-based platforms—can improve information transparency and engagement, while clinical workflow solutions, including digital whiteboards and mobile discharge systems, support communication accuracy and care coordination. Emerging innovations like IoT-wearable monitoring devices further demonstrate potential for promoting patient autonomy and continuous observation. However, these benefits are tempered by persistent implementation challenges. Usability barriers among older adults, technical integration issues, and increased workload for health care professionals remain significant obstacles. Equity concerns related to digital literacy and language accessibility further complicate adoption, raising questions about inclusiveness and fairness in technology-driven care. Importantly, none of the reviewed studies conducted systematic evaluations of long-term outcomes or cost-effectiveness across diverse ED contexts, limiting the evidence base for sustainable integration. Future research should move beyond descriptive assessments toward rigorous, standardized evaluation frameworks that capture clinical, economic, and experiential outcomes. Participatory and inclusive design approaches are essential to address usability and equity gaps, while comprehensive cost-effectiveness analyses will inform scalability and resource allocation. Advancing this agenda is critical to ensure that technological innovations in EDs are not only effective but also equitable, sustainable, and adaptable across varied health care settings.

Acknowledgments

The authors would like to acknowledge all the institutions that supported this work: Clínica Universidad de La Sabana, CAPSAB–Control y Automatización de Procesos Universidad de La Sabana and Sistemas Logísticos–Facultad de Ingeniería, and Cuidado de Enfermería Unisabana-Facultad de Enfermería y Rehabilitación, Universidad de La Sabana (project ING-299-2022). The authors also thank Rosario Lozano and Santiago Lozano for their valuable assistance during the screening and data extraction processes. During the preparation of this work, the authors used ChatGPT (OpenAI 2025) and Copilot (Microsoft 2025) to improve the readability of the paper. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

XSL•FO
RenderX

Funding

This research is supported by Universidad de La Sabana, grant agreement ING-299-2022—"Estrategias Tecnológicas Para Aumentar La Eficiencia En La Ruta De Atención Del Usuario En Servicios De Urgencias De La Clínica Universidad De La Sabana, 2023-2024"—call for the promotion of the academic career and categorization of researchers–2022.

Data Availability

All data generated or analyzed during this study are included in this manuscript and multimedia appendices.

Authors' Contributions

Writing – original draft: ASS, JMA
Methodology: ASS, JMA, WJB, WJG, MGL, IXB
Conceptualization: ASS, JMA, WJB, WJG, MGL, IXB
Investigation: ASS, JMA, WJB
Visualization: ASS, JMA, MGL, IXB
Data curation: ASS, JMA
Writing – review & editing: JMA, WJB, WJG
Funding acquisition: JMA, WJG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database search queries and results.
[DOCX File, 15 KB - medinform_v14i1e79782_app1.docx]

Multimedia Appendix 2
Crowe Critical Appraisal results.
[XLSX File, 15 KB - medinform_v14i1e79782_app2.xlsx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 110 KB - medinform_v14i1e79782_app3.docx]

References

1. Tsai WC, Liu CF, Lin HJ, et al. Design and implementation of a comprehensive AI dashboard for real-time prediction of adverse prognosis of ED patients. Healthcare (Basel) 2022 Aug 9;10(8):1498. [doi: 10.3390/healthcare10081498] [Medline: 36011155]
2. Mora-Meza D, Espejo-Díaz JA, Guerrero WJ. A simulation approach to analyze the operational response plans in an emergency department under the COVID-19 pandemic. In: Montoya-Torres JR, Guerrero WJ, Cortés-Murcia DL, editors. Operations Research and Analytics in Latin America: Proceedings of ASOCIO/IISE Region 16 Joint Conference 2022: Springer; 2023:149-158. [doi: 10.1007/978-3-031-28870-8_12]
3. Sonis JD, Aaronson EL, Lee RY, Philpotts LL, White BA. Emergency department patient experience: a systematic review of the literature. J Patient Exp 2018 Jun;5(2):101-106. [doi: 10.1177/2374373517731359] [Medline: 29978025]
4. Yoo J, Soh JY, Lee WH, Chang DK, Lee SU, Cha WC. Experience of emergency department patients with using the Talking Pole device: prospective interventional descriptive study. JMIR mHealth uHealth 2018 Nov 22;6(11):e191. [doi: 10.2196/mhealth.9676] [Medline: 30467105]
5. Østervang C, Jensen CM, Coyne E, Dieperink KB, Lassen A. Usability and evaluation of a health information system in the emergency department: mixed methods study. JMIR Hum Factors 2024 Feb 21;11:e48445. [doi: 10.2196/48445] [Medline: 38381502]
6. Sonis JD, Rogg J, Yun B, Raja AS, White BA. Improving the patient experience: ten high-yield interventions. : American Academy of Emergency Medicine News; 2016 URL: https://apps.aaem.org/UserFiles/file/MarApr16CommonSense_OMC.pdf [accessed 2025-06-26]
7. Bissonette S, Chartrand J, Bailey L, Lalonde M, Tyerman J. Interventions to improve nurse-family communication in the emergency department: a scoping review. J Clin Nurs 2024 Jul;33(7):2525-2543. [doi: 10.1111/jocn.17068] [Medline: 38476035]
8. Sangal RB, Orloski CJ, Shofer FS, Mills AM. Improving emergency department patient experience through implementation of an informational pamphlet. J Patient Exp 2020 Apr;7(2):225-231. [doi: 10.1177/2374373519826246] [Medline: 32851144]

XSL•FO
RenderX

9. Goldfine CE, Knapp A, Goodman GR, et al. Media and technology usage and attitudes in emergency department patients. Front Digit Health 2022;4:894683. [doi: 10.3389/fdgth.2022.894683] [Medline: 36386045]
10. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
11. Kohl C, McIntosh EJ, Unger S, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid 2018 Dec;7(1):8. [doi: 10.1186/s13750-018-0115-5]
12. Crowe M. Crowe Critical Appraisal Tool (CCAT) user guide. : Conchra; 2013 Nov 19 URL: https://conchra.com.au/wp-content/uploads/2015/12/CCAT-user-guide-v1.4.pdf [accessed 2025-11-21]
13. Crowe M, Sheppard L, Campbell A. Comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: a randomised trial. Int J Evid Based Healthc 2011 Dec;9(4):444-449. [doi: 10.1111/j.1744-1609.2011.00237.x] [Medline: 22093394]
14. Strechen I, Wilson P, Eltalhi T, et al. Harnessing health information technology to promote equitable care for patients with limited English proficiency and complex care needs. Trials 2024 Jul 4;25(1):450. [doi: 10.1186/s13063-024-08254-y] [Medline: 38961501]
15. Gehlot V, Rokowski P, Sloane EB, Wickramasinghe N. Taxonomy, tools, and a framework for combining simulation models with AI/ML models. Presented at: 2022 Annual Modeling and Simulation Conference (ANNSIM); Jul 18-20, 2022. [doi: 10.23919/ANNSIM55834.2022.9859494]
16. Marshall ADA, Hasdianda MA, Miyawaki S, et al. A pilot of digital whiteboards for improving patient satisfaction in the emergency department: nonrandomized controlled trial. JMIR Form Res 2023 Mar 21;7:e44725. [doi: 10.2196/44725] [Medline: 36943360]
17. Steel PAD, Bodnar D, Bonito M, et al. MyEDCare: Evaluation of a smartphone-based emergency department discharge process. Appl Clin Inform 2021 Mar;12(2):362-371. [doi: 10.1055/s-0041-1729165] [Medline: 33910262]
18. Wang X, Kim TC, Hegde S, et al. Design and evaluation of an integrated, patient-focused electronic health record display for emergency medicine. Appl Clin Inform 2019 Aug;10(4):693-706. [doi: 10.1055/s-0039-1695800] [Medline: 31533171]
19. Gesell SB, Golden SL, Limkakeng AT Jr, et al. Implementation of the HEART pathway: using the consolidated framework for implementation research. Crit Pathw Cardiol 2018 Dec;17(4):191-200. [doi: 10.1097/HPC.0000000000000154] [Medline: 30418249]
20. Rathbone AL, Neill S, Prime S, Thomas N, Everett K. Supporting parents with acutely ill children: environment scan and user evaluation of mobile applications (the SuPa kids project). Int J Med Inform 2024 Sep;189:105504. [doi: 10.1016/j.ijmedinf.2024.105504] [Medline: 38833841]
21. Lusa L, Bukovšek D. Providing patients visiting emergency departments with useful information using public real time data: a case study based on Italian data. J Eval Clin Pract 2019 Jun;25(3):404-411. [doi: 10.1111/jep.13059] [Medline: 30407698]
22. Westphal M, Yom-Tov GB, Parush A, et al. A patient-centered information system (MyED) for emergency care journeys: design, development, and initial adoption. JMIR Form Res 2020 Feb 25;4(2):e16410. [doi: 10.2196/16410] [Medline: 32130144]
23. Sangal RB, Powers E, Rothenberg C, et al. Disparities in accessing and reading open notes in the emergency department upon implementation of the 21st Century CURES Act. Ann Emerg Med 2021 Nov;78(5):593-598. [doi: 10.1016/j.annemergmed.2021.06.014] [Medline: 34353651]
24. Doyle S, Pavlos R, Carlson SJ, et al. Efficacy of digital health tools for a pediatric patient registry: semistructured interviews and interface usability testing with parents and clinicians. JMIR Form Res 2022 Jan 17;6(1):e29889. [doi: 10.2196/29889] [Medline: 35037889]
25. Pearkao C, Potisopha W, Wonggom P, Jumpamool A, Apiratwarakul K, Lenghong K. Outcomes of emergency trauma patients after the implementation of web application operating systems. Asian Nurs Res (Korean Soc Nurs Sci) 2023 Aug;17(3):174-179. [doi: 10.1016/j.anr.2023.06.003] [Medline: 37355036]
26. Jaafar R, Lam MC, Abdullah MN, et al. Enhancing healthcare service delivery via patient tracking system using mobile apps technology: a case study. J Eng Sci Technol 2022;17(5):3526-3538 [FREE Full text]
27. Jung KY, Kim T, Jung J, et al. The effectiveness of near-field communication integrated with a mobile electronic medical record system: emergency department simulation study. JMIR mhealth uHealth 2018 Sep 21;6(9):e11187. [doi: 10.2196/11187] [Medline: 30249577]
28. Mavropodi R, Zouka M, Bamidis P, et al. The architecture of an intelligent incident response system in health environments. Presented at: 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM); Sep 20-22, 2019. [doi: 10.1109/SEEDA-CECNSM.2019.8908329]
29. Hampiholi N. Elevating emergency healthcare - technological advancements and challenges in smart ambulance systems and advanced monitoring and diagnostic tools. Int J Comput Trends Technol 2024;72(1):1-7. [doi: 10.14445/22312803/IJCTT-V72I1P101]

XSL•FO
RenderX

30. Østervang C, Lassen A, Schmidt T, Coyne E, Dieperink KB, Jensen CM. Development of a health information system to promote emergency care pathways: a participatory design study. Digit Health 2022;8:20552076221145856. [doi: 10.1177/20552076221145856] [Medline: 36601282]
31. Rochat J, Ehrler F, Siebert JN, Ricci A, Garretas Ruiz V, Lovis C. Usability testing of a patient-centered mobile health app for supporting and guiding the pediatric emergency department patient journey: mixed methods study. JMIR Pediatr Parent 2022 Mar 15;5(1):e25540. [doi: 10.2196/25540] [Medline: 35289754]
32. Alowad A, Samaranayake P, Ahsan K, Alidrisi H, Karim A. Enhancing patient flow in emergency department (ED) using lean strategies–an integrated voice of customer and voice of process perspective. Bus Process Manag J 2021 Jan 25;27(1):75-105. [doi: 10.1108/BPMJ-11-2019-0457]
33. Al Harbi S, Aljohani B, Elmasry L, et al. Streamlining patient flow and enhancing operational efficiency through case management implementation. BMJ Open Qual 2024 Feb 28;13(1):e002484. [doi: 10.1136/bmjoq-2023-002484] [Medline: 38423585]
34. Aranda JM, Moreno-Fergusson ME, Guerrero WJ, et al. Technology acceptance and use among nursing staff in Latin American hospitals: a mixed methods study. Int J Nurs Stud Adv 2025;8:100290. [doi: 10.1016/j.ijnsa.2024.100290] [Medline: 39872435]

Abbreviations

AI: artificial intelligence
CAHPS: Consumer Assessment of Health Care Providers and Systems
CCAT: Crowe Critical Appraisal Tool
CCL: Cetrea Clinical Logistics
ED: emergency department
EMR: electronic medical record
IoMT: Internet of Medical Things
ML: machine learning
NFC: near-field communication
PETS: Parent Engagement through Technology Solutions
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SUS: System Usability Scale

Edited by A Benis; submitted 27.Jun.2025; peer-reviewed by E Bai, R Clay-williams; revised version received 27.Nov.2025; accepted 01.Jan.2026; published 09.Mar.2026.

Please cite as:
Santafé AS, Aranda JM, Beltrán WJ, Guerrero WJ, Guevara Lozano M, Bustos IX
Technological Strategies for the Patient Experience in Emergency Departments: Scoping Review
JMIR Med Inform 2026;14:e79782
URL: https://medinform.jmir.org/2026/1/e79782
doi:10.2196/79782

© Andrés Santiago Santafé, Juan Manuel Aranda, Wilmer Jair Beltrán, William Javier Guerrero, Maryory Guevara Lozano, Ingrid Xiomara Bustos. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review

Dongliang Li[1], MSc; Syaheerah Lebai Lutfi[2], PhD

[1]Artificial Intelligence & Software Engineering, School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia

[2]Medical Informatics Department, College of Medicine and Health Sciences, Sultan Qaboos University, Al Seeb, Oman

Corresponding Author:
Syaheerah Lebai Lutfi, PhD
Medical Informatics Department
College of Medicine and Health Sciences
Sultan Qaboos University
PO Box 35
Al Khoudh
Al Seeb, 123
Oman
Phone: 60 134526001
Email: s.lutfi@squ.edu.om

Abstract

Background: Large language models (LLMs), such as GPT-3.5 and GPT-4 (OpenAI), have been transforming virtual patient systems in medical education by providing scalable and cost-effective alternatives to standardized patients. However, systematic evaluations of their performance, particularly for multimorbidity scenarios involving multiple coexisting diseases, are still limited.

Objective: This systematic review aimed to evaluate LLM-based virtual patient systems for medical history-taking, addressing four research questions: (1) simulated patient types and disease scope, (2) performance-enhancing techniques, (3) experimental designs and evaluation metrics, and (4) dataset characteristics and availability.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020, 9 databases were searched (January 1, 2020, to August 18, 2025). Nontransformer LLMs and non–history-taking tasks were excluded. Multidimensional quality and bias assessments were conducted.

Results: A total of 39 studies were included, screened by one computer science researcher under supervision. LLM-based virtual patient systems mainly simulated internal medicine and mental health disorders, with many addressing distinct single disease types but few covering multimorbidity or rare conditions. Techniques like role-based prompts, few-shot learning, multiagent frameworks, knowledge graph (KG) integration (top-k accuracy 16.02%), and fine-tuning enhanced dialogue and diagnostic accuracy. Multimodal inputs (eg, speech and imaging) improved immersion and realism. Evaluations, typically involving 10-50 students and 3-10 experts, demonstrated strong performance (top-k accuracy: 0.45-0.98, hallucination rate: 0.31%–5%, System Usability Scale [SUS] ≥80). However, small samples, inconsistent metrics, and limited controls restricted generalizability. Common datasets such as MIMIC-III (Medical Information Mart for Intensive Care-III) exhibited intensive care unit (ICU) bias and lacked diversity, affecting reproducibility and external validity.

Conclusions: Included studies showed moderate risk of bias, inconsistent metrics, small cohorts, and limited dataset transparency. LLM-based virtual patient systems excel in simulating multiple disease types but lack multimorbidity patient representation. KGs improve top-k accuracy and support structured disease representation and reasoning. Future research should prioritize hybrid KG-chain-of-thought architectures integrated with open-source KGs (eg, UMLS [Unified Medical Language System] and SNOMED-CT [Systematized Nomenclature of Medicine - Clinical Terms]), parameter-efficient fine-tuning, dialogue compression, multimodal LLMs, standardized metrics, larger cohorts, and open-access multimodal datasets to further enhance realism, diagnostic accuracy, fairness, and educational utility.

(JMIR Med Inform 2026;14:e79039) doi:10.2196/79039

XSL•FO
RenderX

KEYWORDS

large language model; virtual patient; medical education; history-taking; simulated patients

Introduction

Since 2020, large language models (LLMs) such as GPT-3.5 (OpenAI) [1] and GPT-4 (OpenAI) [2] have significantly enhanced virtual patient systems in medical education. Unlike traditional methods relying on resource-intensive standardized patients or high-fidelity simulators [3], LLMs provide scalable, low-risk, and cost-effective solutions by simulating realistic patient interactions across a wide range of clinical scenarios, including internal medicine, mental health disorders, and surgical and orthopedic cases [4]. This capability addresses key challenges in medical education, such as limited exposure to diverse clinical cases and the high costs of traditional simulation approaches.

Early virtual patient systems, often based on models like BERT (Bidirectional Encoder Representations from Transformers), struggled to generate natural dialogues and adapt to complex clinical scenarios, limiting their effectiveness in medical training [5,6]. In contrast, modern LLMs, leveraging prompt-based techniques for role and scenario customization, demonstrate improved contextual understanding, enabling clinically relevant responses. However, these systems still face challenges with hallucination, defined as the generation of factually incorrect or contextually irrelevant content, which may compromise the accuracy of medical history-taking in virtual patient simulations [7,8].

To ensure effective clinical training, continuous improvement and validation of LLMs are essential to mitigate hallucination rate and ensure the reliability of generated information [7]. Recent studies have incorporated techniques such as Supervised Fine-Tuning (SFT) [9] and Retrieval-Augmented Generation (RAG) [10] to enhance contextual adaptability and diagnostic accuracy, as measured by metrics like top-k accuracy and GTPA@k. However, the specific disease types simulated by LLMs, such as such as neurological and rheumatological or rare multiple disease types and potential gaps in simulation capabilities remain underexplored partly due to variations in experimental design and datasets.

Despite progress, systematic comparative evaluations of these techniques in virtual patient systems are lacking. Existing literature reviews often broadly discuss LLMs in medical education without focusing on virtual patient history-taking, limiting insights into domain-specific challenges [11]. Thus, a systematic literature review is critical to consolidate fragmented research, identify challenges, and guide future work.

Previous systematic reviews have explored LLMs in medical education but lack specificity. For instance, Lucas et al [11] reviewed LLMs' implications for teaching effectiveness, ethics, and reliability but did not focus on virtual patient history-taking. García-Torres et al [12] used a hybrid human-LLM methodology to evaluate virtual patients' impact on clinical reasoning but provided limited technical details on prompt design, knowledge graph (KG) integration, or fine-tuning. Similarly, Fatima et al [13] conducted a cross-disciplinary review of ChatGPT in research, clinical practice, education, and patient interaction but did not systematically analyze virtual patient history-taking. Recent empirical studies on LLM-powered virtual patients with automated feedback have prioritized educational outcomes over methodological comparisons [8].

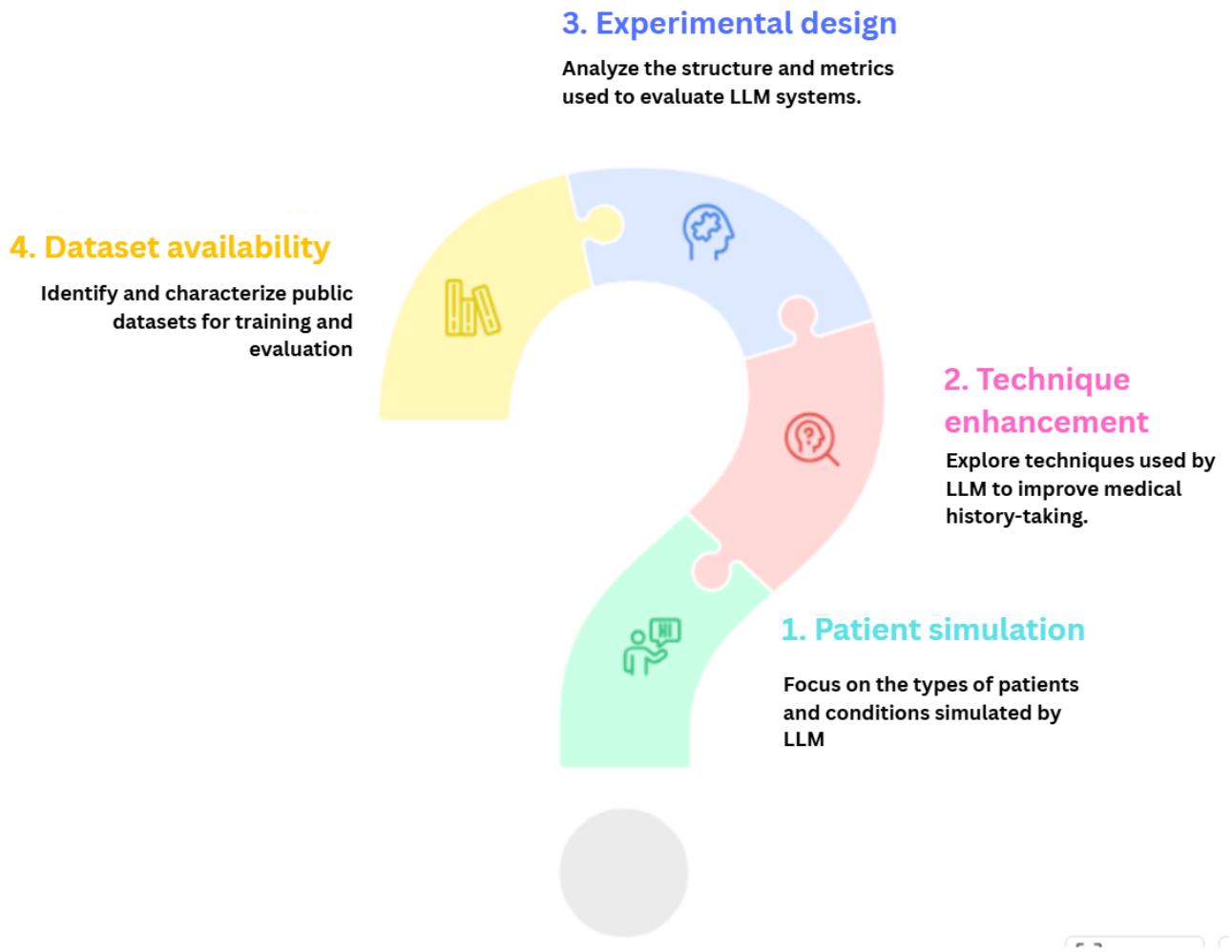
In contrast, this review specifically examines LLM-based virtual patient systems for clinical history-taking. It systematically analyzes prompt engineering, external knowledge integration, model fine-tuning, and evaluation strategies, synthesizing their implications for medical education and effective clinical training. By adopting this methodological focus, this work addresses gaps in prior reviews, which were either too broad or lacked technical depth.

Given this research gap, a systematic literature review is essential to consolidate fragmented research and guide future studies. This paper addresses four core research questions (RQs), as outlined in Figure 1:

- RQ1: what types of patients, conditions, or diseases, such as internal medicine or rare and multiple disease types, are simulated in LLM-based virtual patient systems?
- RQ2: what techniques do LLMs use to enhance medical history-taking capabilities in clinical interviews?
- RQ3: how are experimental designs structured to evaluate LLM-based virtual patient systems, and what evaluation metrics, such as top-k accuracy or System Usability Scale (SUS), are used?
- RQ4: what public datasets are available, and what are their characteristics for training, simulating, and evaluating medical history–taking in virtual patient systems?

XSL•FO
RenderX

Figure 1. Overview of 4 research questions on large language model–based virtual patient systems, covering simulation types, enhancement techniques, evaluation strategies, and datasets.



This systematic literature review addresses these questions to guide researchers in understanding the potential, limitations, and ethical challenges of LLMs in virtual patient systems. It is critical for advancing effective clinical training tools in medical education. The paper is structured as follows, the first section introduces the background of virtual patient systems and LLM applications; the second section describes the literature search methodology and evaluation criteria; the third section presents results, analyzing effectiveness and challenges; the fourth section discusses findings in the context of existing literature; and the fifth section concludes with key findings and recommendations for future research.

Methods

Overview

This study conducted a systematic literature review adhering to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. The research topic and target population were clearly defined, a comprehensive search strategy was formulated, and key search terms were identified. Search results were exported as CSV files via Mendeley (Mendeley Ltd) and Zotero (Corporation for Digital Scholarship) and manually screened. Abstracts were initially evaluated, followed by a comprehensive full-text assessment to confirm relevance and quality.

Eligibility Criteria

Population

The population includes medical students and physicians engaged in medical history–taking. Artificial intelligence (AI)–based virtual patients refer to LLM-based systems simulating patients for training in history-taking. Only transformer-based LLMs (eg, GPT series) are included, excluding older models like BERT (Google AI) or pre-Transformer architectures. Studies focusing on anesthesia, emergency procedures, or intraoperative scenarios are excluded, as these are nontraditional history-taking contexts. Preoperative history-taking for surgical and orthopedic patients is included. This ensures a focus on transformer-based LLMs for history-taking, emphasizing educational and clinical skill development while excluding irrelevant clinical settings. Only studies where LLMs serve as virtual patients for active history-taking and patient communication are included, excluding those focused solely on medical interviews without history-taking.

XSL•FO
RenderX

Intervention

The intervention involves transformer-based LLM technologies applied to medical history-taking, including prompt design, model fine-tuning, KG integration, and other LLM-related techniques. These aim to enhance diagnostic accuracy, interaction quality, and overall performance in virtual patient systems.

Outcomes

Outcomes include performance metrics such as top-k accuracy (eg, top-1 accuracy), empathy scores, readability, system stability (eg, response consistency), user experience (eg, SUS, Chatbot Usability Questionnaire [CUQ]), and other relevant indicators, such as κ (Cohen κ) and *P* value. These metrics evaluate the effectiveness of transformer-based LLM virtual patient systems in medical history–taking.

Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were established to ensure relevance and quality, as shown in Textbox 1. Inclusion criteria prioritize studies with technical depth, validated outcomes, recent publication, and a focus on Transformer-based LLM virtual patients for history-taking, excluding nontransformer models.

Textbox 1. Inclusion and exclusion criteria for transformer-based large language model (LLM) virtual patient studies.

Inclusion criteria:

- IC1: the study population included medical students or physicians involved in medical history–taking or communication training.
- IC2: the intervention involved Transformer-based LLMs (eg, GPT series) used as virtual patients for history-taking or communication training in traditional consultation settings.
- IC3: the study was published between January 1, 2020, and August 18, 2025, including peer-reviewed articles and preprints.
- IC4: the study reported measurable outcomes related to diagnostic performance, communication effectiveness, empathy, readability, or user experience.
- IC5: preoperative history-taking scenarios (eg, surgical or orthopedic patients) were included if they involved conventional patient–clinician consultation processes.

Exclusion Criteria:

- EC1: based on earlier or nongenerative transformer models (eg, BERT [Bidirectional Encoder Representations from Transformers] and GPT-2) rather than modern LLMs.
- EC2: focused solely on medical interviews without explicit medical history–taking.
- EC3: published before January 2020.
- EC4: duplicate titles or redundant publications.
- EC5: review or commentary papers.
- EC6: non-English language studies.
- EC7: focused on anesthesia, emergency procedures, or intraoperative contexts (except preoperative consultations).

Information Sources

A systematic literature search was conducted across 9 authoritative databases PubMed, Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SpringerLink, ERIC, arXiv, and ACL Anthology, covering studies published between January 2020 and August 18, 2025, in medicine, AI, education, and virtual technologies. A secondary search used a snowballing strategy, identifying additional sources from 15 initial articles and relevant publications, including theses from ProQuest. Including peer-reviewed articles and grey literature (eg, preprints from arXiv and theses) ensured comprehensive coverage of the rapidly evolving fields of AI and medical education.

Search Strategy

Two search strategies were used for this systematic review. The first query combined virtual patient terms (eg, virtual patient, simulated patient, AI patient, conversational patient, chatbot patient, intelligent virtual agent, and dialogue agent) with LLM-related keywords (eg, LLM, ChatGPT, GPT-4, GPT, transformer model, generative AI, AI-powered tutor, and natural language generation) and was applied to structured databases including PubMed, Scopus, Web of Science, IEEE Xplore, and ACM Digital Library. The second, simplified query included only virtual patient terms and was applied to broader repositories, namely SpringerLink, ERIC, arXiv, and ACL Anthology, to ensure comprehensive coverage, as strict LLM keywords reduced results in these databases (see Multimedia Appendix 1 for more details).

Selection Process and Data Collection

The literature screening and data collection process followed PRISMA 2020 guidelines [14] (checklist provided in Multimedia Appendix 2). Titles and abstracts were manually screened for relevance to LLM-based virtual patient history-taking, with GPT-3.5 Turbo used for auxiliary verification. Full-text evaluations were conducted for studies meeting preliminary inclusion criteria. A multidimensional quality assessment form recorded key study characteristics and outcomes, as detailed in Multimedia Appendices 3 and 4. The ChatPDF tool was used to cross-check content, and final inclusion decisions were made with the supervising researcher's input.

XSL•FO
RenderX

Quality and Risk of Bias Assessment

A reviewer with a computer science background, under the supervision of an experienced research advisor, conducted the quality assessment. A customized multidimensional assessment framework was developed to evaluate the technical quality of the included studies, as described in Multimedia Appendices 3 and 4. Conventional appraisal tools such as the Joanna Briggs Institute Critical Appraisal Checklist (2020) were adapted because they were not fully suitable for LLM-based virtual patient research. The framework incorporated 6 evaluation dimensions methodology clarity, dataset transparency, completeness of system evaluation, innovation or integration level, reproducibility and openness, and the presence of control or baseline comparisons. Each dimension was rated on a 3-point scale ranging from 0 to 2, with 2 indicating the highest quality. The total score for each study, therefore, ranged from 0 to 12. Based on the overall score, studies were categorized into 3 quality levels, high (9-12 points), medium (5-8 points), and low (0-4 points). This classification ensured a consistent and transparent interpretation of the technical quality across studies. The assessment emphasized methodological rigor, technical implementation, system architecture, model training strategies, multimodal integration, and evaluation methods to identify potential sources of bias.

Risk of bias was also assessed across 5 domains—selection or reporting bias, implementation bias, evaluation bias, data bias, and reporting completeness bias—to capture variations in study design, data transparency, technical implementation, and reporting quality. This combined approach ensured a systematic evaluation of both methodological soundness and potential bias in LLM-based virtual patient studies.

Based on the aforementioned screening and evaluation methods, the research results will be presented in detail in the next section.

Results

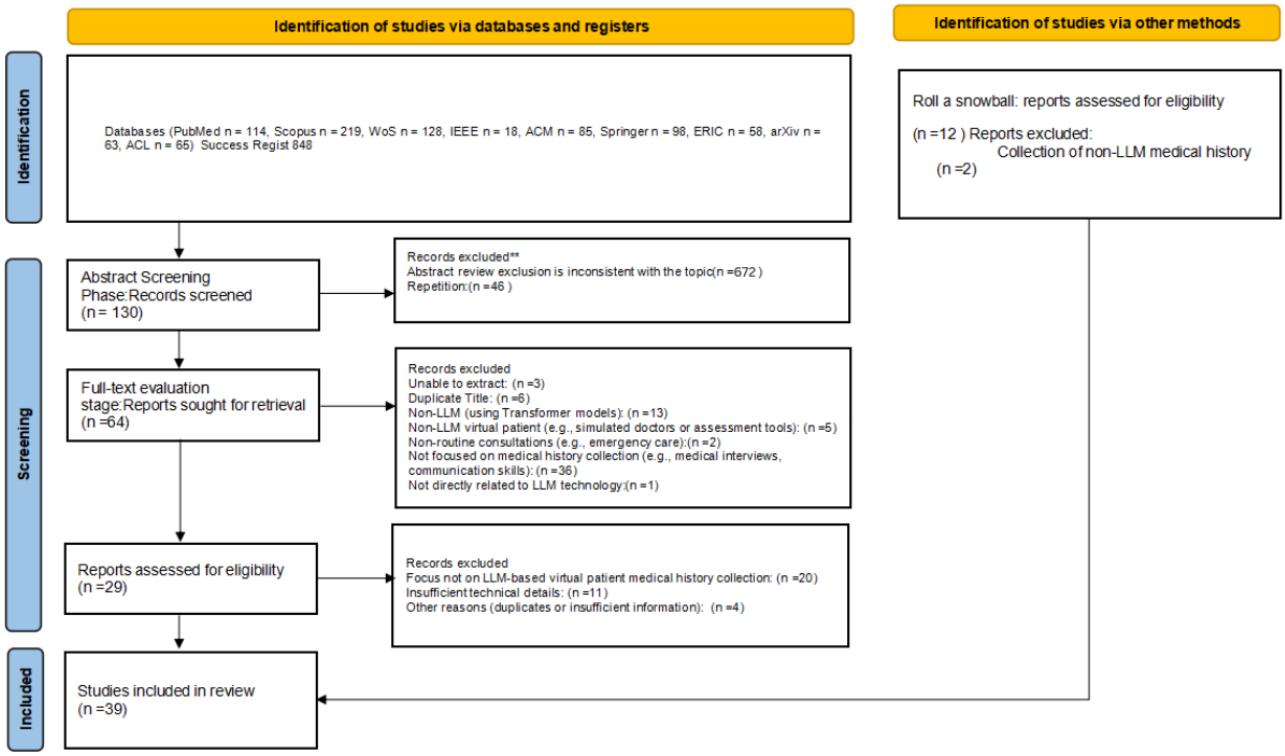
Overview

This section reports the results of literature selection, study characteristics, and findings related to the 4 RQs, avoiding interpretive discussion.

Study Selection Results

During identification (see Figure 2 for the 2020 flow diagram for systematic reviews, including database and register searches), 848 records were retrieved from 10 databases. During abstract screening, 672 records were excluded, and after removing 46 duplicates, 130 unique full-text articles were assessed. Of these, 66 were excluded for reasons including lack of extractable data (n=3), duplicate content (n=6), non-LLM models (n=13), focus not on LLM applications (n=5), nonroutine clinical settings (n=2), no history-taking focus (n=36), or nonrelevance to LLM technologies (n=1), leaving 64 eligible articles. During multidimensional evaluation, 20 studies were excluded for limited relevance, 11 for insufficient technical detail, and 4 for other reasons, yielding 29 articles from database searches. An additional 12 articles were identified through snowballing, of which 2 were excluded as irrelevant, resulting in 10 included snowballed studies. The final synthesis comprised a total of 39 studies [8,10,15-51] (6 high-quality, 33 moderate-quality) (see Multimedia Appendix 5).

Figure 2. PRISMA flow diagram.



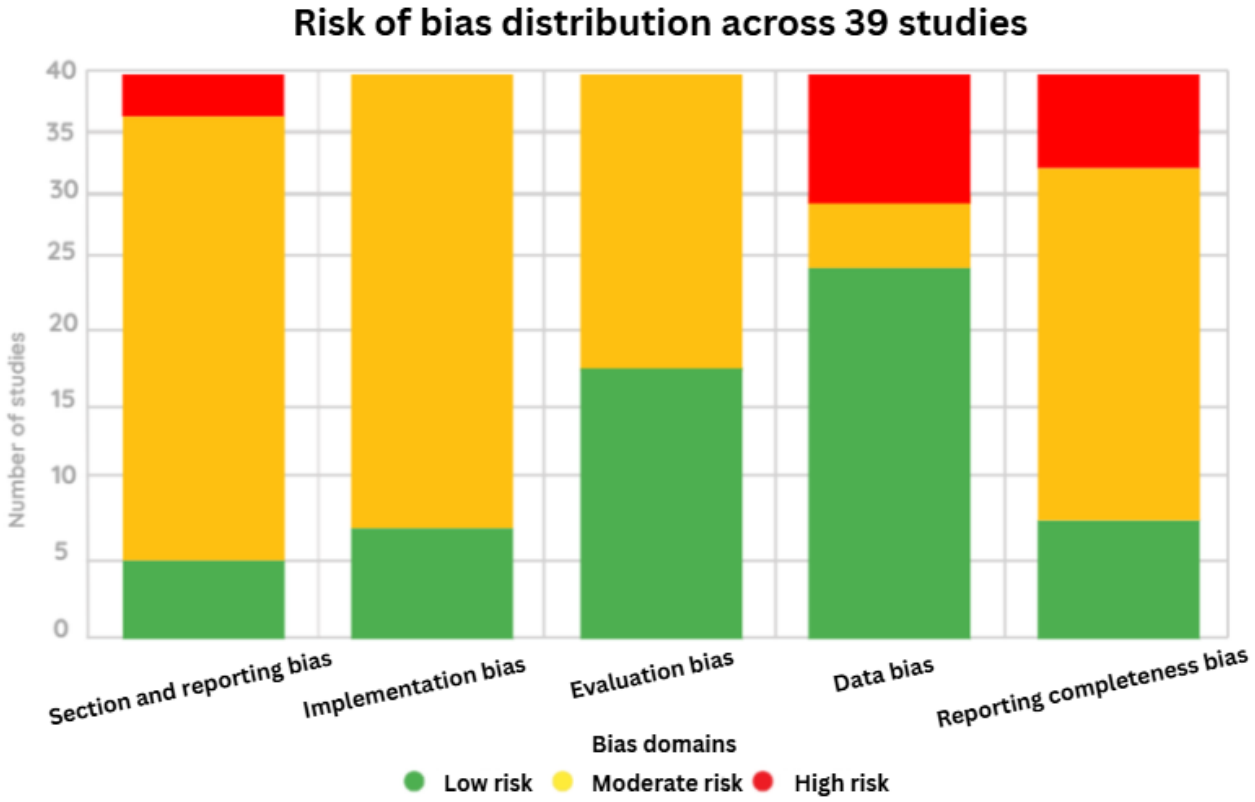
XSL•FO
RenderX

Quality and Risk of Bias Assessment Results

Most studies achieved the highest rating (2 points) for methodological clarity, indicating well-defined research procedures and consistent experimental designs. Scores for dataset transparency and system evaluation completeness varied considerably, with approximately 45%-55% of studies receiving moderate ratings (1 point) due to insufficient details regarding data sources or evaluation frameworks. Innovation and integration levels were generally high, with about 80% of studies receiving high ratings (2 points), reflecting notable progress in multimodal integration and technological creativity. Reproducibility and openness received moderate ratings (1 point) in approximately 60% of studies, as some provided replication details whereas others lacked information on model configurations or training strategies. Control or baseline comparison obtained the lowest scores (0-1 point), with only 25%-30% of studies incorporating explicit comparative or controlled analyses. Overall, 6/39 (15%) studies were rated as "high quality," and 33/39 (85%) as "moderate quality," suggesting strong methodological design and innovation but highlighting the need for improved dataset transparency, evaluation completeness, and reproducibility (see Multimedia Appendix 5 for details).

Risk of bias was assessed across five domains—selection and reporting bias, implementation bias, evaluation bias, data bias, and reporting completeness bias—as illustrated in Figure 3.

Figure 3. Overview of risk of bias distribution across studies.



Selection and reporting bias was evident in disease coverage. Approximately 95% (37/39) of studies simulated specific disease types (eg, internal medicine or mental health disorders), but only 50% (20/39) addressed multiple disease stages, thereby limiting generalizability to multimorbidity contexts.

Implementation bias was observed in the limited reporting of model development details. While 80% (31/39) of studies described model types, only 35% (14/39) provided information on training procedures, fine-tuning processes, prompt engineering, reinforcement learning with human feedback (RLHF), domain-specific adaptation, or multimodal integration. This resulted in high technical heterogeneity.

Evaluation bias was reflected in the uneven assessment of educational and user experience outcomes. Approximately 70% (27/39) of studies evaluated educational outcomes, and 65% (25/39) assessed user experience using instruments such as the SUS or the Chatbot Usability Questionnaire (CUQ). However, only 30% (12/39) of studies conducted comparative or controlled evaluations, which limited the interpretability of effectiveness findings.

Data bias stemmed from moderate dataset transparency. Fewer than half of the studies (18/39, 45%) explicitly identified data sources or quality control procedures, constraining reproducibility.

Reporting completeness bias was identified in the limited documentation of personalization mechanisms and quality control for generated outputs, reported in only 25%-30% (10-12/39) of studies, suggesting incomplete methodological reporting.

In summary, this multidimensional quality and bias assessment systematically examined the technical features, methodological design, and potential biases of LLM-based virtual patient research. The findings highlight the need to improve dataset

XSL•FO
RenderX

transparency, enhance the comprehensiveness of system evaluation, and include controlled comparative studies to strengthen reproducibility and scientific validity. Details of the screening results and full bias assessment are provided in Multimedia Appendix 6.

Finding for RQ1: What Types of Patients Are Simulated in LLM-Based Virtual Patient Systems?

This section presents a descriptive statistical analysis and classification of patient types simulated by LLMs in virtual patient history-taking systems, based on a systematic literature review. The classification uses disease categories from Table 1, grouping similar disease types (eg, internal medicine and mental health disorders). For each category, the number of studies, specific simulation scenarios, disease complexity (eg, low: single symptom; medium: multisystem involvement; high: rare or complex interactions), number of simulated cases (if specified), and disease stage (eg, acute and chronic) are summarized.

Table 1. Classification of patient types simulated in virtual patient systems.

Disease category	Simulation scenarios	Complexity	Number of cases	Stage	References
Internal medicine	Chest/abdominal pain, diabetes, COPD[a], COVID-19, hypertension, multisymptom	Medium-high	3-500+	Acute-chronic	12 studies [15,20-30]
Mental health disorders	Depression, PTSD[b], ADHD[c], TRD[d], CBT[e] models	High	1-106	Acute-chronic	8 studies [18,31-35,37]
Surgical/orthopedic	Plastic surgery, hand surgery, joint pain	Medium	3-10	Acute–chronic	4 studies [16,20,24,38]
Neurological/rheumatological	Stroke, meningitis, concussion, polymyositis, brain hemorrhage	High	1-4	Acute	6 studies [20,21,36,39-41]
Ophthalmological	Eye conditions, pain with redness/photophobia	Medium	1-24	Acute	3 studies [17,20,42]
Dermatological	Ear cyst, telogen effluvium, skin conditions	Medium	1-394	Acute-chronic	4 studies [19,25,43,44]
Rare/multiple disease types[f]	Rare diseases, unspecified conditions, broad patient scenarios	High	8-5230	Diverse	10 studies [8,10,25,45-51]

[a]COPD: chronic obstructive pulmonary disease.

[b]PTSD: posttraumatic stress disorder.

[c]ADHD: attention deficit hyperactivity disorder.

[d]TRD: treatment-resistant depression.

[e]CBT: cognitive behavioral therapy.

[f]Only [25] involves rare disease simulation, using the RareBench dataset with 421 rare disease cases. Other references in the "rare/multiple disease types" category focus on multiple disease types or unspecified/broad patient scenarios.

Summary of Research Focus and Coverage

A systematic descriptive analysis of the literature indicates that research primarily focuses on internal medicine, particularly gastrointestinal (eg, abdominal pain, heartburn, hematemesis [20-23,27,28]), respiratory (eg, cough, chronic obstructive pulmonary disease, COVID [20,21,23,29,30]), cardiovascular (eg, chest pain, hypertension [20-22,30]), metabolic and endocrine (eg, diabetes [10,23,24]), and fatigue-related conditions (eg, chronic fatigue [15,26]). These conditions are well-suited for effective clinical training, with LLMs effectively simulating both acute and chronic management scenarios, though vague symptoms like fatigue remain underexplored.

Mental health disorders are a significant focus, including depression and related disorders (eg, suicidal ideation and treatment-resistant depression) [15,32-35], posttraumatic stress disorder (posttraumatic stress disorder; eg, combat trauma) [18,31], attention-deficit/hyperactivity disorder [31], and other mental health conditions (eg, cognitive impairment, cognitive behavioral therapy [CBT] models) [32,37]. These simulations emphasize emotional and psychological complexity, suitable for empathy and CBT training, though behavioral disorders like attention-deficit/hyperactivity disorder are less explored.

Rare and multiple disease types are well represented, focusing on heterogeneous case collections and electronic health record (EHR)–driven simulations [8,46-51], as well as rare disease modeling (eg, 421 rare diseases [10,25,45]). These studies highlight the scalability and diversity of LLMs in simulating complex and varied clinical conditions.

Less studied areas include neurological and rheumatological diseases (eg, stroke, meningitis, and polymyositis [20,21,36,39-41]), dermatological diseases (eg, ear cyst, telogen effluvium, and skin conditions [19,25,43,44]), surgical and orthopedic diseases (eg, joint pain, plastic surgery, and hand surgery [16,20,24,38]), and ophthalmological diseases (eg, eye pain, photophobia, and multiple eye conditions [17,20,42]). These areas, due to specialized or visual simulation requirements, have received less attention, revealing significant research gaps.

XSL•FO
RenderX

Overall, research focuses on internal medicine and mental health disorders due to their prevalence and clinical relevance. Rare and multiple disease types demonstrate LLMs' scalability and generalization potential. In contrast, specialized domains such as surgical and orthopedic, neurological and rheumatological, dermatological, and ophthalmological diseases remain underexplored, presenting opportunities for further innovation in simulation-based medical education.

Finding for RQ2: What Techniques do LLMs Use to Enhance Medical History–Taking Capabilities in Clinical Interviews?

To address RQ2, LLMs use advanced techniques to enable realistic, dynamic, and accurate virtual patient simulations for medical history-taking, providing medical students and professionals with effective clinical training platforms that closely resemble real-world scenarios. The following sections categorize core LLM technologies for history-taking into prompt engineering, KGs and structured data, model fine-tuning and training, and speech interaction, as summarized in Table 2.

Table 2. Technical metrics, evaluation scores, and key strengths and limitations of LLM virtual patient techniques.

Category and technique (source)[a]	Top-k accuracy (%)	Hallucination rate (%)	IRR[b] (%)	AS[c] (%)	Strengths	Limitations
Prompt design						
Role-based prompts [10,23]	81.4	4.97	2.08	28	Realistic role-play; multiagent;	Long prompts inconsistent; complex cases limited
Few-shot [25]	Not used: 25%, used: 52%				Simple, effective;	Limited depth
Multiagent prompt frameworks [21,24]	—[d]	—	—	—	Consistent dialogue	Higher computation
KG[e]						
Entity-relation triples [17]	—	—	—	—	Structured triples	Small-scale realism
Entity layering [10]	97.42	—	—	—	High diagnostic accuracy and interpretability	Limited semantic layers
KG+history and few-shot [10]	97.85	—	—	—	Role continuity	Context overflow
KG+multiagent prompts [10]	97.85	—	—	—	Modular roles	High cost
Fine-tuning						
SFT[f] [17]	—	(Before fine-tuning 3.71%, after fine-tuning 0.31%)	4.79	87.00	Reduces hallucination rate; better reasoning	Limited turn validation
LoRA[g] [17]	—	—	—	—	Efficient tuning	Limited scalability evidence
Speech interaction						
TTS[h] [28,51]	—	—	—	—	Realistic voice; multimodal input	Recognition errors possible

[a]All techniques listed are LLM-based Transformer architectures.

[b]IRR: information-related response rate.

[c]AS: Anthropomorphism Score.

[d]Not available.

[e]KG: knowledge graph.

[f]SFT: supervised fine-tuning.

[g]LoRA: low-rank adaptation.

[h]TTS: text-to-speech.

Prompt Design for Realistic Patient Simulation

1. Role-based prompt design: detailed prompts embed patient demographics, personality traits (eg, Big Five), and clinical symptoms to simulate authentic patient behavior. For instance, Borg et al [41] developed virtual patient cases with approximately 2000 tokens per prompt, including detailed medical histories and contextual information, occupying a significant portion of the LLM's 4096-token context window. Implementations include structured prompts with role context, medical history, and behavioral constraints [16-20,22,24-29,31,33-41,43-46,51]. For

XSL•FO
RenderX

example, Bodonhelyi et al [37] used Satir model roles (eg, "accuser"') to achieve up to 95% dialogue consistency. This ensures consistent role-playing but requires instructor oversight to maintain medical accuracy [36].

2. Few-shot and dialogue history–based prompting: prompts incorporate dialogue history and few-shot examples to maintain conversational coherence and mimic gradual information disclosure. MEDDxAgent uses structured examples to guide diagnostic reasoning, improving top-k accuracy from 25% to 52% in complex cases [25]. Limitations include token constraints, which may restrict handling of intricate scenarios.
3. Multiagent prompt frameworks: multiple agents (eg, patient agent, doctor agent, and behavior controller) collaborate to generate realistic and personalized clinical dialogues, reducing reliance on single prompts and minimizing hallucination rate. For example, EvoPatient, developed by Du et al [18], uses a patient–doctor dual-agent system for natural dialogue via unsupervised learning. LLM-based generative agents integrate memory flows, retrievers, and cognitive mechanisms to enhance dialogue realism and training effectiveness [24]. Additionally, AI self-play agents such as AMIE simulate diagnostic conversations through internal and external self-play loops, applying multirole strategies within a single model to improve diversity and dialogue adaptation [21].

Knowledge Graphs and Structured Data

1. Dynamic KG retrieval: MedDiT uses KG agents to retrieve relevant subgraphs via SPARQL queries, linearizing them into natural language prompts to reduce token load and context loss, achieving significant token reduction [47].
2. Hierarchical KG modeling: AI Patient integrates layered KG with a Reasoning RAG multiagent framework, boosting accuracy from 68.94% to 94.15%. For the difficult "family and social history" category, accuracy improved from 13.33% to 85.56%, showing the value of structured KG reasoning. Compared to role-based prompts, the KG-based method performs better in multiclass tasks. With entity layering [10], accuracy rose by 16.02% (81.4% to 97.42%) in challenging categories such as allergy and social history.

Model Fine-Tuning and Training

1. Instruction fine-tuning and self-play: AMIE uses real and simulated medical dialogues with self-play loops (inner and outer) to optimize diagnostic performance, outperforming primary care physicians (PCPs) [19,45].
2. SFT and LoRA (Low-Rank Adaptation): Liu et al [17] fine-tuned Qwen2.5-72B-Instruct on MedDialog, reducing the hallucination rate from 3.71% to 0.31% [17]. At the same time, in terms of information-related response rate and Anthropomorphism Score (AS).
3. Chain-of-thought (CoT) and RLAIF+MoM: CureFun uses CoT and RAG, while Kumar et al's [26] RLAIF+MoM framework structures ambiguous symptoms, achieving up to 95% output clarity [23].

Speech Interaction

1. TTS and STT integration: Takata et al [28] use Google application programming interface (API) and Unity for synchronized speech [51] and emotional facial expressions, outperforming traditional platforms in interaction realism. Thesen et al [51] leverage Whisper-3 for immersive speech processing, achieving high student satisfaction.
2. Multimodal enhancements: AMIE integrates image uploads (eg, skin photos) to support history-taking, outperforming PCPs in diagnostic accuracy [19]. Ryu et al [35] enhances realism with patient image uploads.

In conclusion, LLM technologies achieve consistent and realistic virtual patient dialogues through multilevel prompt design, few-shot examples, and multiagent mechanisms. The use of KGs and structured data enhances information retrieval and medical history modeling, improving classification and recognition accuracy. Model fine-tuning techniques (SFT, LoRA, self-play, CoT, RLAIF+MoM) effectively reduce hallucination rate, optimize reasoning, and enhance diagnostic performance. Furthermore, speech interaction and multimodal integration improve interaction immersion and provide diagnostic support.

Findings for RQ3: How Are Experimental Designs Structured to Evaluate LLM-Based Virtual Patient Systems, and What Evaluation Metrics Are Used?

To address RQ3, this review synthesized 32 studies listed in Table 3 to examine the experimental design, evaluation methods, and metrics of LLM-based virtual patient systems, focusing on medical education, high diagnostic accuracy, and effective clinical training. Of the 39 identified studies (see Multimedia Appendix 5), 7 were excluded due to the absence of formal evaluations (eg, Takata et al [28] described planned behavioral experiments without data collection; Rodrigo et al [46] outlined testing plans but reported no results; Geer [31] focused on design without providing quantitative outcomes; Li et al [47] emphasized system architecture but lacked performance metrics; Staples et al [34] relied solely on qualitative feedback; Kumar et al [26] discussed workflow but lacked quantified results; Lee et al [15] provided expert Likert scores without statistical analysis). Due to high heterogeneity in evaluation metrics (eg, top-k accuracy, hallucination rate, CUQ, inconsistent scales and dimensions, lack of statistical information such as SDs or confidence intervals in many studies, and large variations in sample sizes that could bias pooled analyses), a meta-analysis was not performed. Instead, a structured narrative synthesis was adopted, supplemented with tabular summaries in Tables 2 and 3, which present the technical approaches and corresponding evaluation metrics.

XSL•FO
RenderX

Table 3. Merged quantitative and qualitative evaluation details for virtual patient systems.

Authors	Participants (N)	Identity	Comparisons	Analysis method	Results	Evaluation type
Du et al [18]	None	AI[a]	Evolved vs Nonevolved	Stats[b]	Relevance 0.7589; Faithfulness 0.8786	Q[c]
Holderried et al [8]	106	Med students	GPT-4 vs Human	κ [d]	Top-k accuracy >99%;κ=0.832	Q
Tu et al [21]	20	PCPs[e]	AMIE vs PCPs	Stats	AMIE better on 28/32 expert metrics	Q
Yamamoto et al [20]	145	Med students	AI vs Traditional	t and MWU[f]; Likert[g]	OSCE: 28.1 vs 27.1 (P^h=.01)	M[i]
Cook et al [50]	3	Physicians	GPT-4.0 vs 3.5	1–6 scale, multivar.	GPT-4.0 higher authenticity, feedback	Q
Haider et al [16]	None	AI	GPT-4o vs Claude vs Gemini	Stats	Nonsignificant differences; high baseline	Q
Brugge et al [39]	21	Med students	Feedback vs Control	Wilcoxon, ICC[j]=0.924	CRI-HTI: 3.60 vs 3.02 (P^k≈.05)	Q
Holderried et al [22]	28	Med students	None	Stats; Spearman	Script Q 60.3%; Answers 94.4%	Q
Leypold et al [38]	3	Hand surgeons	None	Likert (1–5)	Mean=4.6 (Understand 5.0; History 4.2)	Q
Borg et al [40]	15	Med students	Robot vs VP	*t* test; Thematic	Authenticity 4.5 vs 3.9 (P^l=.04)	M
Radel et al [36]	40	Med students	Feedback vs None	*t* test; Likert	Improved scores with feedback (P^m<.05)	M
Luo et al [42]	184	Med students	LLMDP vs Traditional	*t* test; Pearson	78.13 vs 67.08 (P^n<.001)	M
Thesen et al [51]	94	Med students	None	Likert; *t* test	Comfort 61% to 76% (P^o<.001)	M
Laverde et al [24]	86	Med students	Agent vs Others	CUQ[p]	CUQ: 86.25/100	Q
Benfatah et al [29]	12	Nursing students	None	Pearson; Likert	Total score 19.42; Clarity r=0.701	M
Geer [31]	None	AI	None	Descr[q]	High similarity; no quant data	QL[r]
Borg et al [41]	15	Med students	Robot vs VIC	Wilcoxon; Text	Authenticity 4.47 vs 3.93 (P^s≈.03)	M
Ng et al [49]	100	Med students	Hybrid vs Baselines	Acc[t]; failure and confusion	Top-k accuracy 98.7%; failure 2.0%	Q
Wang et al [32]	N/S	Experts	Expert vs GPT-4	Stats; Subjective	Experts rated higher, GPT-4 underestimated	QL
Zheng et al [48]	N/S	Experts	None	Weighted F_1-score; Fuzzy labels	High professionalism and ethics	M
Rose et al [25]	None	AI	GPT-4o vs Llama3.1	Stats	GTPA@1 0.96; RareBench 0.45	Q
Liu et al [17]	None	AI	Proposed vs Baselines	Stats	Hallucination Rate 0.31%; Anthropomorphism 0.87	Q
Chen et al [33]	25	Patients and Psychiatrists	Prompt D1–D4	Stats	Fluency 3.28; Empathy 3.43; Dx 55.56%	Q
Liao et al [43]	N/S	Students, laypeople	GPT-4 vs others	Stats	Dx 53.33%; Coverage rate:15.36%-33.89%%	Q
Johri et al [44]	None	AI	GPT-4 vs 3.5	Stats	MCQ 0.919; FRQ 0.684	Q

XSL•FO
RenderX

Authors	Participants (N)	Identity	Comparisons	Analysis method	Results	Evaluation type
Bodonhelyi et al [37]	N/S	Psych experts	Accuser vs Rationalizer	Likert; Emotion; Stats	Realism 3.8 vs 3.7	M
Rashidian et al [30]	2	Clinicians	AI vs Doctors	κ; Likert	Symptom top-k accuracy 97.7%; κ=0.74	M
Tu et al [45]	20	PCPs	AMIE vs PCPs	Stats	AMIE better on 28/32 expert metrics	Q
Saab et al [19]	43	Patients, specialists	AMIE vs PCPs	Stats; *P* values	*Top*-1 accuracy: 0.65 vs 0.53 (P^{u}<.001)	Q
Li et al [23]	8	Med experts	Auto vs Manual	Spearman; Pearson	ρ=0.81; *r*=0.85 (P^{v}<.05)	Q

[a]AI: artificial intelligence.

[b]Stats: statistical analysis.

[c]Q: Quantitative.

[d]κ: Cohen κ.

[e]PCP: primary care physician.

[f]t and MWU: *t* test/Mann-Whitney *U* test.

[g]Likert: Likert scale.

[h]*P*: Objective Structured Clinical Examination performance comparison between AI-trained and traditionally trained students showed a difference of 28.1 vs 27.1, respectively (Mann–Whitney *U* test, *P*=.01).

[i]M: Mixed.

[j]ICC: intraclass correlation coefficient.

[k]*P*: Feedback group scored 3.60 on CRI-HTI vs 3.02 in control group (ANOVA, ICC=0.924, *P*=.049).

[l]*P*: Robot vs virtual patient authenticity ratings were 4.5 vs 3.9 (*t* test; *P*=.04).

[m]*P*: Students receiving feedback showed higher Likert-scale scores compared to control (*t* test; *P*=.04).

[n]*P*: LLMDP-trained students scored 78.13 (SD 8.35) on history acquisition; traditional group scored 67.08 (SD 7.21), with a mean difference of 11.05 points (*P*<.001).

[o]*P*: Comfort scores improved from 61% to 76% after intervention (*t* test; *P*<.001).

[p]CUQ: Chatbot Usability Questionnaire.

[q]Descr: Descriptive.

[r]QL: Qualitative.

[s]*P*: Robot vs VIC comparison yielded authenticity ratings of 4.47 vs 3.93 (Wilcoxon test; *P*=.035).

[t]Acc: Accuracy.

[u]*P*: Original article did not provide comparative statistical values such as means or test statistics, only reported model accuracy with significance levels.

[v]*P*: Among the 8 evaluated cases, exact *P* values for comparisons 1 and 2 were *P*<.001 and *P*=.04, respectively; the remaining 6 ranged from *P*=.003 to *P*=.011. As per reporting guidelines, *P*=.000 was converted to *P*<.001.

Commonalities and Specificities in Experimental Design

Commonalities

Most studies involved medical or health professional students, typically 10-50 participants, with some including residents or practicing physicians as evaluators (usually 3-5 experts) [10,19,36]. A few studies used AI agents for large-scale automated testing [18,25]. Tasks primarily covered core clinical competencies, including history-taking accuracy [10,36], diagnostic reasoning [45], role-playing [37], and multiturn dialogues [41].

Evaluation paradigms were categorized into 3 types quantitative (18 studies) using randomized controlled trials (RCTs) or comparative experiments with statistical tests (eg, ANOVA, *t* tests, and correlation) [39,45]; qualitative (1 study) using expert interviews, thematic analysis, or questionnaires [32]; and mixed (13 studies) integrating objective metrics (eg, top-k accuracy, and F_1-score) with subjective scales (eg, Likert, CUQ) [36,40]. Comparative baselines included model comparisons (GPT-4 and 4o, Claude, and Gemini) [16], AI versus human physicians [19,45], with or without feedback or platform comparisons [40,41], and variations in prompt strategies, multimodality, or RAG and KG [49]. Most studies applied blinding and reported rater consistency (eg, κ) [8,39]. Details of each study's evaluation type are provided in Table 3 (Evaluation type column).

Specificities

Some studies implemented AI self-play for scalability [18], or AI doctor-AI patient automated evaluation [16]. Social robot-LLM hybrids were used to enhance realism [40,41]. Comparative baselines varied, including direct AI versus human physician comparisons [19,45], modular and hybrid architecture comparisons (eg, RASA or KG or LLM) [49], and multimodel comparisons (GPT-4.5 and 4o, Claude 3.7, and Gemini 2.5)

XSL•FO
RenderX

showing nonsignificant differences, indicating a high performance baseline [16].

Evaluation Metrics: Commonalities and Specificities

Evaluation metrics in LLM-based virtual patient systems generally fall into 5 categories: clinical accuracy and knowledge, communication and interaction quality, robustness and stability, training efficacy and feedback quality, and system performance. Clinical accuracy and knowledge were assessed using metrics such as top-k accuracy [19], GTPA@k [25], information coverage [43], hallucination rate [17,19], and fidelity and relevance [18], evaluating diagnostic and information-gathering capabilities. Communication and interaction quality were measured through readability (Flesch and Flesch–Kincaid) [10], CUQ [22], and Anthropomorphism Score [17,37], reflecting language clarity and interaction naturalness. Robustness and stability were evaluated via paraphrase robustness [10], leak resistance [18], and rater consistency metrics, including κ [8] and intraclass correlation coefficient (ICC) [39]. Training efficacy and feedback quality focused on learning outcomes and user experience, assessed through Objective Structured Clinical Examination (OSCE) score improvements [39], and usability scales (such as CUQ). System performance metrics, including latency, failure rate [49], and confusion or clarification rates [49], captured efficiency and reliability. Specificities included unique metrics such as span-level F_1-score [10] for knowledge extraction, cosine similarity-based fidelity and relevance [18], weighted F_1-score or fuzzy labels for professionalism and ethics [39], and system-level confusion, clarification, or failure rates. Automated evaluation metrics such as GTPA@k [25] provided standardized measures for diagnosis, while expert ratings revealed realism biases, including underestimation by GPT-4. Usability was further examined through instruments such as the CUQ (score of 77/100), 7-dimensional Likert scales combined with ANOVA for accuracy, realism, and empathy, and system latency and confusion reporting (0.5%). These metrics collectively enable comprehensive assessment of both the technical performance and experiential quality of LLM-based virtual patient systems. The calculation formula for the indicators is detailed in Multimedia Appendix 7.

Evaluation Results

The synthesis of evaluation results from the 32 included studies indicates that LLM-based virtual patient systems demonstrate high diagnostic accuracy across key metrics. Quantitative results showed top-k accuracy ranging from 0.45 to 0.98 (based on 34 models and 300 patient cases); Saab et al [19] reported that AMIE achieved a Top-1 Accuracy of 0.65, compared to 0.53 for PCPs, with $P<.001$. Information coverage averaged 33.89% to 94.4% (eg, Liao et al [43] reported 33.89%, based on 150 cases; Holderried et al [22] reported 94.4%, based on 502 questions explicitly covered in case scripts, with a total of 826 answers), while hallucination rate remained low, ranging from 0.31% to 5% (eg, Liu et al [17] conducted experiments with up to 5 dialogue rounds, achieving a hallucination rate of 0.31%).

In terms of communication quality, CUQ scores ranging from 77 to 86.25 (eg, Laverde et al [24] reported 86.25). Training effectiveness was evident in OSCE performance (eg, Yamamoto et al [20] reported posttraining 28.1 vs pretraining 27.1, $P=.01$; Luo et al [42] reported that medical students trained using the LLM-based digital patient system (LLMDP) achieved a medical history acquisition score of 78.13 (SD 8.35), while the control group trained with traditional real patients scored 67.08 (SD 7.21). The difference between groups was 11.05 points, with $P<.001$, indicating a highly statistically significant difference). Robustness metrics such as κ ranged from 0.74 to 0.832 (eg, Holderried et al [8] reported 0.832; Rashidian et al [30] reported 0.74).

Mixed methods studies reported improvements in authenticity (eg, Borg et al [40] reported postuse 4.5 vs preuse 3.9; $P=.04$) and comfort (Thesen et al [51] reported that among 69 participants, average preuse comfort was 61%, increasing to 76% postuse, indicating higher student-reported comfort after training; $P<.001$). Automated baseline systems, such as EvoPatient, achieved a relevance of 0.7589 and a faithfulness of 0.8786 (Du et al [18]), while multimodel comparisons showed no significant differences among top LLMs (Haider et al [16]).

These results highlight the effectiveness of LLM-based systems in simulating realistic interactions, although specific metrics—such as the low failure rate of 2.0% reported by Ng et al [49] (based on 200 dialogue trials) and the high ICC of 0.924 reported by Brugge et al [39]—indicate that further standardization is needed.

The reviewed studies exhibit consistent commonalities in participant composition, task objectives, and evaluation paradigms, typically involving small to medium student samples and multisource evaluations with experts and AI, focusing on history-taking and diagnostic reasoning. Specificities are evident in comparative settings (AI vs humans, multimodel comparisons) and statistical methods such as bootstrap, false discovery rate (FDR), and ICC. Across studies, 5 key evaluation aspects were assessed using objective metrics. Clinical accuracy and knowledge were measured using top-k accuracy [19,21], GTPA@k [25], information coverage [22,43], and span-level F_1-score [18], capturing high diagnostic accuracy and knowledge acquisition. Authenticity was evaluated using hallucination rate [17] and fidelity metrics [18]. Interaction experience was assessed via CUQ [24] and Anthropomorphism Score [17], reflecting communication quality and user perception. Robustness and consistency were quantified using κ [30] and ICC [39]. System performance was captured through latency, failure rate, and confusion rate [21,49]. These categories collectively enable a comprehensive assessment of both technical and experiential quality, highlighting the systems' effectiveness while indicating areas for further standardization and evaluation refinement.

Finding for RQ4: What Public Datasets Are Available, and What Are Their Characteristics for Training, Simulating, and Evaluating Medical History-Taking in Virtual Patient Systems?

A variety of datasets support the development and evaluation of virtual patient systems, ranging from real-world EHRs to structured synthetic clinical scenarios. Table 4 summarizes key datasets used in recent studies, which are either publicly

XSL•FO
RenderX

available or accessible through formal application processes. These datasets provide diverse clinical data resources for effective clinical training and evaluation.

MIMIC-III (Medical Information Mart for Intensive Care-III) is a comprehensive publicly available medical dataset, containing detailed EHR data from over 40,000 intensive care unit (ICU) patients [10]. It covers multiple disease types, including internal medicine and neurological and rheumatological conditions, serving as a critical resource for virtual patient system development. However, its ICU-focused nature, primarily reflecting acute or severe diseases, limits its applicability to noncritical care scenarios, such as outpatient consultations, mental health disorders, or chronic disease management [10].

DDxPlus provides clinical dialogue data for respiratory diseases [25], suitable for training virtual patient systems in specific domains within internal medicine. The iCraft-MD dataset focuses on dermatological cases [25], and RareBench contains region-specific subsets of rare and multiple disease types [25]. While valuable for specialized applications, their scope is limited for building generalized, multitask virtual patient systems.

The medical-NLP corpus provides a broad range of clinical dialogues and records [18], while CCKS 2019 offers a Chinese medical KG dataset, enabling multilingual history-taking simulations [17,56]. The Open-i dataset includes multimodal chest X-ray images and textual descriptions, supporting dynamic image generation alongside history-taking [47]. MedQA provides multiple-choice and long-form medical question-answering data, useful for fine-tuning LLMs for diagnostic dialogues [45]. These datasets enhance multimodal and question-answering capabilities, though their coverage is narrower than MIMIC-III.

In summary, current datasets primarily focus on critical care or specific medical domains, with limited publicly available, diverse resources. These datasets have limitations in generalizability, multitask applicability, and multilingual support for virtual patient systems. Combining MIMIC-III (broad coverage) [10], DDxPlus and iCraft-MD (specialized domains) [25], and CCKS 2019 (multilingual support) [17,56] provides partial coverage. However, further development of diverse, emotionally annotated, and non-English datasets is needed to enhance system generalizability and conversational fidelity.

Table 4. Datasets used in virtual patient systems.

Dataset Name	Index	Description	Access requirements
MIMIC-III[a]	[10]	EHR[b] database from Beth Israel Deaconess Medical Center, over 40,000 ICU[c] patients (2001-2012), including vital signs, medications, laboratory results, diagnostic codes.	Access via PhysioNet; requires CITI training and DUA [52].
DDxPlus	[25]	Synthetic dataset for respiratory diseases, with clinical dialogues and diagnoses.	CC-BY[d]; publicly available, allows commercial use with attribution [53].
iCraft-MD	[25]	Synthetic dermatology dataset from public medical question banks and expert cases.	MIT License; publicly available [54].
RareBench	[25]	Rare disease dataset with regional subsets (Europe, Canada, and China).	Apache-2.0; publicly available [55].
medical-nlp	[18]	Medical NLP[e] corpus with clinical dialogues and records.	GNU General Public License v3.0; publicly available on GitHub [18].
CCKS 2019	[17]	Chinese medical knowledge graph dataset with entity recognition, relation extraction, QA[f] tasks.	Research use only; cite source [56].
Open-i	[47]	Multimodal dataset with 3314 chest x-ray images and textual descriptions.	Open Data Commons Open Database License; publicly available via NIH Open-i project [57].
MedQA	[45]	Multiple-choice and long-form medical QA dataset for diagnostics.	Publicly available via GitHub [45].

[a]MIMIC-III: Medical Information Mart for Intensive Care-III.

[b]EHR: electronic health record.

[c]ICU: intensive care unit.

[d]CC-BY: Creative Commons Attribution.

[e]NLP: natural language processing.

[f]QA: question-answering.

Discussion

Principal Findings

Current LLM-based virtual patient systems exhibit significant limitations in disease coverage, complex case simulation, multimorbidity representation, specialty applicability, multimodal capabilities, and standardization of evaluation metrics, indicating a need for systematic optimization to enhance clinical fidelity, educational adaptability, and interaction quality.

XSL•FO
RenderX

Limitations and Future Directions in Virtual Patient Type Simulation Research

The systematic review reveals several limitations in current virtual patient systems regarding disease coverage, case complexity modeling, and support across medical specialties [20,32]. First, research is heavily concentrated on internal medicine (eg, gastrointestinal, respiratory, and metabolic disorders) and mental health disorders, accounting for over half of the studies [15,32]. This focus reflects the strengths of LLMs in language-driven tasks and the availability of relevant data, but it highlights significant gaps in surgical and orthopedic, ophthalmological, and dermatological domains. Scenarios requiring procedural operations, image recognition, or multimodal interactions are underrepresented, with current systems lacking effective modeling mechanisms [16,43]. This finding aligns with RQ1, indicating that most simulations focus on specific disease categories, leaving certain specialty areas insufficiently covered.

Although some studies involve multiple disease types, current virtual patient systems mainly use single-disease trajectories, with each virtual patient representing only one disease [21,25]. multimorbidity simulations—patients with 2 or more coexisting chronic or acute diseases—are limited. Clinical multimorbidity requires complex decisions, including drug interactions, overlapping symptoms, and conflicting management priorities. Dataset analysis for RQ4 shows most datasets focus on a single primary disease, lacking comprehensive multimorbidity cases. Without simulating this complexity, virtual patient systems have limited use in advanced medical education and clinical reasoning. Future research should emphasize role construction and dialogue design for multimorbidity scenarios to support integrated management across multiple diseases.

Furthermore, current systems exhibit instability in simulating vague or atypical symptoms (eg, chronic fatigue, low mood, attention deficits), demonstrating weak extraction of unstructured chief complaints and incoherent reasoning processes, which limits support for comprehensive clinical assessments [10,31]. Incorporating KGs and CoT reasoning mechanisms may improve knowledge organization and causal chain construction, enhancing models' reasoning and response capabilities in complex clinical scenarios [58,59]. Open-source KGs, such as UMLS (Unified Medical Language System) [60], SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) [61] can be integrated to provide structured, high-coverage medical knowledge. Leveraging these KGs allows LLMs to reference verified entities and relationships during dialogue generation, reducing hallucinations and enhancing reasoning in multimorbidity or complex symptom scenarios.

Additionally, coverage of specialty domains remains limited, particularly in surgical and orthopedic, ophthalmological, and dermatological contexts requiring procedural or visual recognition skills. These fields show constraints in case complexity, multimodal interaction support, and procedural training, limiting educational adaptability and clinical fidelity [43]. This observation is consistent with findings from RQ1 and RQ4, indicating that certain specialty cases and complex patient scenarios remain underrepresented.

In summary, current virtual patient systems require systematic optimization in disease coverage balance, case complexity, multimorbidity simulation, specialty applicability, multimodal capabilities, and interaction depth. These improvements are expected to enhance clinical fidelity, educational adaptability, and natural human-computer interaction, providing a foundation for advancing intelligent medical education and clinical decision support systems.

Technical Challenges and Solutions of LLM-Based Virtual Patient Simulation

In virtual patient simulation, prompts guide LLMs to generate dialogues conforming to specific role characteristics. However, overly long prompts lead to information loss, affecting dialogue quality and coherence. Specifically, GPT-style models based on autoregressive decoder architectures process prompts sequentially from left to right, prioritizing information at the beginning (primacy effect) and end (recency effect), while middle information may be ignored or inadequately processed, impacting content completeness and accuracy [62]. Additionally, few-shot learning maintains role settings using limited historical dialogues, but balancing information completeness and avoiding overload in multiturn complex dialogues remains challenging. To address this, core information should be prioritized at the prompt's beginning and end, with emotional content placed in the middle to ensure accurate and consistent information transmission.

Compared to single-prompt methods, multiagent frameworks decompose dialogue generation into modular components, with agents handling specific functions such as patient history or simulated emotions, or leveraging adversarial training to enhance dialogue quality [18,21,24]. However, multiagent systems increase computational costs, as each agent independently calls APIs, leading to higher financial burdens. Additionally, fine-tuning using older APIs may cause "memory loss," affecting dialogue context continuity. Dialogue saving and restoration mechanisms are necessary to ensure information consistency. To reduce costs and improve efficiency, dialogue compression and summarization mechanisms can simplify context input, and noncritical tasks can be assigned to lower-cost APIs to balance performance and expense.

Research on fine-tuning GPT-style models and integrating KGs remains limited. Liu et al [17] proposed a SFT strategy using synthetic medical records and manually annotated dialogues, significantly enhancing the realism and anthropomorphism of patient history collection dialogues on the Qwen-72B model (Alibaba Cloud), reducing Hallucination Rate from 3.71% to 0.31%. However, their training process did not explicitly embed disease names into the Transformer model, limiting specific disease knowledge mastery. Prior studies indicate that incorporating disease names as labels or input features improves semantic understanding and generalization, particularly for simulating patient histories for specific diseases [58,59]. Combining CoT methods with fine-tuning to guide step-by-step reasoning for complex medical knowledge is underexplored and presents a potential direction for medical dialogue systems.

KGs systematically represent structured medical knowledge, but there is a lack of studies comparing disease embeddings

XSL•FO
RenderX

trained by LLMs with KG embeddings to evaluate differences in knowledge representation and reasoning capabilities. KGs' clear entity and relation structures complement LLMs' limitations in sparse knowledge or verification, while LLM-trained embeddings as a baseline reveal constraints in semantic understanding and improvements from KG enhancement. Future research should establish a unified evaluation framework integrating supervised fine-tuning, LoRA, and CoT reasoning to explore the complementary roles of KG and LLM embeddings, advancing medical dialogue models in logical reasoning and knowledge accuracy.

Evaluation Design Suggestions and Summary of Metrics

Evaluation methods and metrics in LLM-based virtual patient systems are diverse, reflecting the multidimensional nature of performance and educational outcomes. However, the lack of standardized frameworks hinders cross-study comparison and generalizability. Diagnostic accuracy is often measured by top-k accuracy (eg, 65% [19]) and GTPA@k [25], though these coarse metrics may not capture system capabilities. Information Coverage and Hallucination Rate assess retrieval fidelity but usually rely on manual verification. Interaction quality metrics—such as semantic similarity (0.7589 [18]) and user scores (eg, CUQ=77 [22])—highlight interactivity but are limited by subjectivity and small samples. External studies also report high usability, such as SUS=88.1[31], a 10-item scale measuring ease of use, confidence, and learnability, reinforcing the systems' educational value despite inconsistent evaluation standards.

Moreover, system performance assessment remains insufficient. Some studies report response delays affecting conversation naturalness, but lack quantitative measures or systematic evaluations. The absence of standardized key performance indicators exacerbates framework fragmentation, hindering effective implementation and broader application in medical education and effective clinical training [41].

To address fragmented metrics, a unified and scientifically grounded evaluation framework is necessary. Key performance indicators with recommended thresholds are proposed to guide system design and assessment:

- Top-1 accuracy (≥0.80): AMIE achieved a top-1 accuracy of 0.65, outperforming primary care physicians (0.53) [19], while *GTPA*@1 reached 0.96 [25]. A threshold of 0.80 ensures reliable diagnostic performance.
- Hallucination rate (≤0.05): GPT-4o demonstrated a Hallucination Rate of 0.31% [17], with rates below 5% reported in [19], supporting medical safety standards.
- Information coverage (≥0.50): coverage of critical history items was 33.89% [43], indicating room for improvement. A 50% threshold ensures adequate information capture.
- Empathy and Anthropomorphism Score (≥0.75, standardized 0-1 scale): GPT-4o scored 0.87 [17], and [37] reported approximately 0.76, indicating desired human-like interaction and empathy.
- Usability (SUS≥80, CUQ≥75): SUS of 88.1 [59] and CUQ of 77 [22] meet standards for satisfactory usability.
- Robustness (leak resistance≥0.90): a value of 0.9412 was reported in [18], indicating compliance with privacy and ethical requirements.
- Rater consistency (κ and ICC≥0.80): high interrater reliability was observed, with ICC=0.924 [39] and κ=0.832 [8].

Challenges of Data Diversity: Limitations in Corpus Coverage and Adaptability

Existing datasets for training and evaluating LLM-based virtual patient systems are diverse, encompassing real-world EHRs, synthetic clinical scenarios, and multimodal or multilingual resources [10,17,18,25,45,47,56]. Table 4 summarizes key datasets, which are either publicly available or accessible through formal application processes. Analysis reveals several limitations impacting system development.

Mainstream datasets like MIMIC-III primarily reflect intensive care scenarios, containing records from ICU patients with acute or critical conditions, often within internal medicine or neurological and rheumatological categories [10]. This bias limits generalizability to nonacute settings, such as outpatient consultations, mental health disorders, or chronic disease management. Specialized datasets, including DDxPlus (internal medicine), iCraft-MD (dermatological), and RareBench (rare and multiple disease types), provide tailored resources [25]. While valuable for specific domains, their narrow coverage limits suitability for general-purpose or multitask system training.

Linguistic and cultural diversity is limited, as most corpora are English-based and originate from Western health care systems [17,18,45,47,56]. The lack of datasets in other languages, such as Chinese [17,56], and integration with local KGs or region-specific disease contexts constrains performance in multilingual and cross-cultural environments. Modality limitations are evident; most datasets provide textual information and lack multimodal inputs like images, speech, or physiological signals [47], restricting interaction realism and diagnostic reasoning.

Data accessibility and format heterogeneity affect usability. Access requirements, annotation styles, field definitions, and dialogue formats vary [10,17,18,25,45,47,56], hindering integration and comprehensive training. Standardization of data formats and unified interfaces is necessary to reduce development costs and support broader adoption. Additionally, task alignment poses challenges, as datasets like MedQA [45] are structured for multiple-choice or question-answering tasks, requiring extensive adaptation for dialogue generation.

Combining broadly covering datasets like MIMIC-III [10] with domain-specific (DDxPlus, iCraft-MD, RareBench) [25] and multilingual or multimodal datasets (CCKS 2019 [17,56], Open-i [47], medical-nlp [18], MedQA [45]) can partially address gaps. However, developing diverse, emotionally annotated, non-English, and multimodal datasets is essential to enhance generalizability, robustness, and interaction fidelity.

XSL•FO
RenderX

Conclusion

This systematic review, conducted per PRISMA 2020 guidelines, evaluated studies (January 2020-August 18, 2025) on LLM-based virtual patient systems for medical history collection, sourced from 9 databases PubMed, Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Springer, ERIC, arXiv, and ACL Anthology. Following rigorous screening, deduplication, and quality appraisal, 6 high-quality and 33 moderate-quality studies were included, addressing 4 research questions, simulated patient types, performance-enhancing technologies, experimental designs, and evaluation metrics.

Key findings include (1) systems primarily simulate internal medicine and mental health disorders (acute and chronic), with limited coverage of rare and multiple disease types, multimorbidity, and specialties like surgical and orthopedic, neurological and rheumatological, dermatological, and ophthalmological, restricting applicability in complex clinical reasoning and education [15,16,32,43]. (2) Technologies such as role-based prompts, few-shot learning, multiagent frameworks, KGs, and fine-tuning (eg, SFT, LoRA, CoT, RLAIF+MoM) enhance dialogue coherence, retrieval accuracy (+16.02% with KGs) [10], and high diagnostic accuracy, while multimodal integration (eg, speech) improves immersion [18,25]. (3) Evaluations involved medical students and practitioners, using mixed methods (top-k accuracy, F_1-score, SUS, CUQ, and expert ratings) with comparisons across AI models, physicians, or prompt variations; small sample sizes (10-50 students and 3-10 experts) and inconsistent metrics limit generalizability [19,40]. (4) Systems demonstrated high diagnostic accuracy: top-k accuracy 0.45-0.98, information coverage 33.89%-94.4%, Hallucination Rate 0.31%-5%, and high usability (SUS≥80), often outperforming junior physicians [18,43]. Dataset limitations (eg, MIMIC-III ICU bias, restricted access, low multilingual and multimodal diversity) hinder cross-study comparability [10,17,25].

The key discussion points are summarized: (1) Disease coverage is imbalanced, favoring internal medicine and mental health disorders over surgical and orthopedic, dermatological, ophthalmological, and multimorbidity scenarios, limiting effective clinical training [16,43]. Future systems should prioritize multimorbidity and diverse patient populations (cultural and linguistic) to enhance realism [17,56]. (2) Prompt design endures information loss in long prompts; placing critical information at prompt ends and using dialogue compression or multiagent frameworks can mitigate this [18,60]. KG-LLM integration and fine-tuning improve performance, with potential for further gains via hybrid KG-CoT approaches [58,59]. (3) Fragmented evaluation frameworks, inconsistent metrics, and small participant pools reduce reliability. A standardized framework with thresholds (eg, top-1 accuracy≥0.80, hallucination rate≤0.05, SUS≥80, CUQ≥75,κ and ICC≥0.80) and larger samples (50-100 students, 5-10 experts) is needed [19,39]. (4) Dataset biases (eg, ICU focus), format heterogeneity, and privacy restrictions limit inclusivity. Open-access, ethically compliant, multimodal, and multilingual datasets are essential for equitable systems [10,45,47].

Future research should focus on large-scale longitudinal studies, standardized evaluation metrics, diverse open-access datasets (eg, UMLS [60] and SNOMED-CT [61]), and advanced integration of KGs, multimodal training, and optimized prompts to enhance the realism, high diagnostic accuracy, and fairness of LLM-based virtual patient systems in medical education [17,18,59].

Acknowledgments

We thank Sultan Qaboos University for partially funding this project. The corresponding author is also affiliated with Universiti Sains Malaysia (USM). Additional supporting materials related to this work are provided in Multimedia Appendix 8.

Funding

The authors thank Sultan Qaboos University (Oman) for the funding of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Database Search Strategies.
[DOCX File , 13 KB - medinform_v14i1e79039_app1.docx]

Multimedia Appendix 2
PRISMA Checklist.
[DOCX File , 26 KB - medinform_v14i1e79039_app2.docx]

Multimedia Appendix 3
Joanna-Briggs Institute Quality Assessment Questionnaire.
[DOCX File , 13 KB - medinform_v14i1e79039_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Multidimensional assessment table.
[DOCX File , 15 KB - medinform_v14i1e79039_app4.docx]

Multimedia Appendix 5
Quality Assessment Result.
[DOCX File , 27 KB - medinform_v14i1e79039_app5.docx]

Multimedia Appendix 6
Risk of Bias Assessment for 39 Included Studies.
[DOCX File , 26 KB - medinform_v14i1e79039_app6.docx]

Multimedia Appendix 7
Calculation Formulas for Evaluation Metrics.
[DOCX File , 23 KB - medinform_v14i1e79039_app7.docx]

Multimedia Appendix 8
Other file screening process, source files, etc.
[ZIP File (Zip Archive), 1137 KB - medinform_v14i1e79039_app8.zip]

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 December 06; Red Hook, NY, United States p. 1877-1901.
2. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. ArXiv. Preprint posted online on March 4, 2024.
3. Lim AS, Lee SWH. Is technology enhanced learning cost-effective to improve skills?: the Monash objective structured clinical examination virtual experience. Simul Healthc 2022;17(2):131-135. [doi: 10.1097/SIH.0000000000000526] [Medline: 33273417]
4. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach 2025;47(1):40-42. [doi: 10.1080/0142159X.2024.2376879] [Medline: 38992981]
5. Song W, Hou X, Li S, Chen C, Gao D, Wang X, et al. An intelligent virtual standard patient for medical students training based on oral knowledge graph. IEEE Trans Multimedia 2023;25:6132-6145. [doi: 10.1109/tmm.2022.3205456]
6. Babu A, Boddu SB. BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. Explor Res Clin Soc Pharm 2024;13:100419 [FREE Full text] [doi: 10.1016/j.rcsop.2024.100419] [Medline: 38495953]
7. Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: current trends, challenges, and future directions in large language model-enabled medical chatbots. BioMedInformatics 2024;4(1):837-852. [doi: 10.3390/biomedinformatics4010047]
8. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
9. Yi Y, Kim KJ. The feasibility of using generative artificial intelligence for history taking in virtual patients. BMC Res Notes 2025;18(1):80 [FREE Full text] [doi: 10.1186/s13104-025-07157-8] [Medline: 39994780]
10. Yu H, Zhou J, Li L, Chen S, Gallifant J, Shi A, et al. AIPatient: simulating patients with EHRs and LLM powered agentic workflow. ArXiv. Preprint posted online on July 29, 2025. [doi: 10.48550/arXiv.2409.18924]
11. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
12. García-Torres D, Vicente Ripoll MA, Fernández Peris C, Mira Solves JJ. Enhancing clinical reasoning with virtual patients: a hybrid systematic review combining human reviewers and ChatGPT. Healthcare (Basel) 2024;12(22):2241 [FREE Full text] [doi: 10.3390/healthcare12222241] [Medline: 39595439]
13. Fatima A, Shafique M, Alam K, Fadlalla Ahmed TK, Mustafa M. ChatGPT in medicine: a cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore) 2024;103(32):e39250 [FREE Full text] [doi: 10.1097/MD.0000000000039250] [Medline: 39121303]
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]

XSL•FO
RenderX

15. Lee WY, Han SY, Kim JH, Lee BW, Han Y, Lee S. Gen-SynDi: leveraging knowledge-guided generative AI for dual education of syndrome differentiation and disease diagnosis. Applied Sciences 2025;15(9):4862. [doi: 10.3390/app15094862]
16. Haider SA, Prabha S, Gomez-Cabello CA, Borna S, Genovese A, Trabilsy M, et al. Synthetic patient-physician conversations simulated by large language models: a multi-dimensional evaluation. Sensors (Basel) 2025;25(14):4305 [FREE Full text] [doi: 10.3390/s25144305] [Medline: 40732431]
17. Liu Z, Tu Q, Ye W, Xiao Y, Zhang Z, Cui H. Exploring the inquiry-diagnosis relationship with advanced patient simulators. ArXiv. Preprint posted online on March 11, 2025.
18. Du Z, Zheng L, Hu R, Xu Y, Li X, Sun Y, et al. Llms can simulate standardized patients via agent coevolution. ArXiv. Preprint posted online on June 7, 2025.
19. Saab K, Freyberg J, Park C, Strother T, Cheng Y, Weng WH, et al. Advancing conversational diagnostic AI with multimodal reasoning. ArXiv. Preprint posted online on May 6, 2025. [doi: 10.48550/arXiv.2505.04653]
20. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024;10:e58753 [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
21. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
22. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A Generative Pretrained Transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
23. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. ArXiv. Preprint posted online on April 25, 2024.
24. Laverde N, Grévisse C, Jaramillo S, Manrique R. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. Comput Struct Biotechnol J 2025;27:2481-2491. [doi: 10.1016/j.csbj.2025.05.025] [Medline: 40547455]
25. Rose D, Hung C, Lepri M, Alqassem I, Gashteovski K, Lawrence C. Meddxagent: a unified modular agent framework for explainable automatic differential diagnosis. ArXiv. Preprint posted online on February 26, 2025. [doi: 10.48550/arXiv.2502.19175]
26. Kumar R, Gattani D, Singh K. Enhancing medical history collection using llms. 2024 Presented at: ACSW '24: Proceedings of the 2024 Australasian Computer Science Week; 2024 May 13; NSW, Sydney, Australia p. 140-143.
27. Choi D, Jung Y, Kim J, Oh N, Oh H, Lee S, et al. LLM-Based clinical history taking system: a persona-driven approach. Stud Health Technol Inform 2025;329:1866-1867. [doi: 10.3233/SHTI251254] [Medline: 40776271]
28. Takata T, Yamada R, René A. Development of a virtual patient model for Kampo medical interview: new approach for enhancing empathy and understanding of Kampo medicine pathological concepts. 2024 Presented at: 2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS); 2024 November 01; Himeji, Japan p. 1-5.
29. Benfatah M, Marfak A, Saad E, Hilali A, Nejjari C, Youlyouz-Marfak I. Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: a study on nursing students' experience. Teach Learn Nurs 2024;19(3):e486-e493. [doi: 10.1016/j.teln.2024.02.005]
30. Rashidian S, Li N, Amar J, Lee J, Pugh S, Yang E, et al. AI agents for conversational patient triage: preliminary simulation-based evaluation with real-world EHR data. ArXiv. Preprint posted online on June 4, 2025.
31. Kenny PG, Parsons TD. Virtual standardized llm-ai patients for clinical practice. Annual Review of Cybertherapy And Telemedicine 2024;22:177-182.
32. Wang R, Milani S, Chiu J, Zhi J, Eack S, Labrum T, et al. PATIENT-Ψ: using large language models to simulate patients for training mental health professionals. ArXiv. Preprint posted online on October 3, 2024.
33. Chen S, Wu M, Zhu K, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. ArXiv. Preprint posted online on May 23, 2023. [doi: 10.48550/arXiv.2305.13614]
34. Staples P, Clarke P, Leininger C, Principato C, Shafiei M, Goodwin G. Building virtual patients for training mental health professionals. PsyArXiv 2025.
35. Ryu JW, Kwon CY, Park JS, Lim SR, Jeon HL, Kim HJ. Development and application of a chatgpt-based simulation training platform for korean medicine. Journal of Oriental Neuropsychiatry 2024;35(4):413-427.
36. Rädel-Ablass K, Schliz K, Schlick C, Meindl B, Pahr-Hosbach S, Schwendemann H, et al. Teaching opportunities for anamnesis interviews through AI based teaching role plays: a survey with online learning students from health study programs. BMC Med Educ 2025;25(1):259 [FREE Full text] [doi: 10.1186/s12909-025-06756-0] [Medline: 39966894]
37. Bodonhelyi A, Stegemann-Philipps C, Sonanini A, Herschbach L, Szép M, Herrmann-Werner A. Modeling challenging patient interactions: LLMs for medical communication training. Preprint posted online on April 8, 2025. [doi: 10.48550/arXiv.2503.22250]
38. Leypold T, Schäfer B, Boos AM, Beier JP. Artificial intelligence-powered hand surgery consultation: GPT-4 as an assistant in a hand surgery outpatient clinic. J Hand Surg Am 2024;49(11):1078-1088 [FREE Full text] [doi: 10.1016/j.jhsa.2024.06.002] [Medline: 39066762]

XSL•FO
RenderX

39. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ 2024;24(1):1391 [FREE Full text] [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
40. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res 2025;27:e63312 [FREE Full text] [doi: 10.2196/63312] [Medline: 40053778]
41. Borg A, Parodis I, Skantze G. Creating virtual patients using robots and large language models: a preliminary study with medical students. 2024 Presented at: HRI '24: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; 2024 March 11 - 15; Boulder CO USA p. 273-277. [doi: 10.1145/3610978.3640592]
42. Luo M, Bi S, Pang J, Liu L, Tsui C, Lai Y, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med 2025;8(1):502 [FREE Full text] [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]
43. Liao Y, Meng Y, Wang Y, Liu H, Wang Y, Wang Y. Automatic interactive evaluation for large language models with state aware patient simulator. ArXiv. Preprint posted online on July 21, 2024. [doi: 10.2139/ssrn.4890649]
44. Johri S, Jeong J, Tran BA, et al. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. 2024 Presented at: AAAI 2024 Spring Symposium on Clinical Foundation Models; March 25-27, 2024; Stanford University, Stanford, California.
45. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024;642(3):442-450. [doi: 10.1056/aioa2300138]
46. Rodrigo MMT, Castaneda S, Hernandez RB, Alaan JAMV, Caoile PSP, Chidrome KD. Project caladrius: The design of a virtual patient for philippine medical education. 2025 Presented at: 7th International Conference, AIS 2025, Held as Part of the 27th HCI International Conference, HCII; 2025 June 22–27; Gothenburg, Sweden p. 232-244. [doi: 10.1007/978-3-031-92970-0_17]
47. Li Y, Zeng C, Zhang J, Zhou J, Zou L. Meddit: A knowledge-controlled diffusion transformer framework for dynamic medical image generation in virtual simulated patient. ArXiv. Preprint posted online on August 22, 2024. [doi: 10.24963/ijcai.2024/1267]
48. Zheng W, Turner L, Kropczynski J, Ozer M, Nguyen T, Halse S. Llm-as-a-fuzzy-judge: Fine-tuning large language models as a clinical evaluation judge with fuzzy logic. ArXiv. Preprint posted online on June 12, 2025. [doi: 10.48550/arXiv.2506.11221]
49. Ng HW, Koh A, Foong A, Ong J. Real-time hybrid language model for virtual patient conversations. 2023 Presented at: International Conference on Artificial Intelligence in Education; 2023 3 July; Tokyo, Japan p. 780-785. [doi: 10.1007/978-3-031-36272-9_71]
50. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res 2025;27:e68486 [FREE Full text] [doi: 10.2196/68486] [Medline: 39854611]
51. Thesen T, O'Brien WN, Stone S, Pinto-Powell R. Generative AI as the first patient: practice, feedback, and confidence. Med Sci Educ 2025:1-6. [doi: 10.1007/s40670-025-02473-x]
52. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
53. Fansi TA, Goel R, Wen Z, Martel J, Ghosn J. DDXPlus: a new dataset for automatic medical diagnosis. ArXiv. Preprint posted online on October 13, 2022.
54. Li J, Lai Y, Li W, Ren J, Zhang M, Kang X. Agent hospital: a simulacrum of hospital with evolvable medical agents. ArXiv. Preprint posted online on January 17, 2025. [doi: 10.48550/arXiv.2405.02957]
55. Chen X, Mao X, Guo Q, Wang L, Zhang S, Chen T. RareBench: can LLMs serve as rare diseases specialists? 2024 Presented at: KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024 August 25 - 29; Barcelona Spain p. 4850-4861. [doi: 10.1145/3637528.3671576]
56. Han X, Wang Z, Zhang J, Wen Q, Li W, Tang B. Overview of the CCKS 2019 knowledge graph evaluation trackntity, relation, event and QA. ArXiv. Preprint posted online on March 9, 2020. [doi: 10.48550/arXiv.2003.03875]
57. Demner-Fushman D, Kohli M, Rosenman M, Shooshan S, Rodriguez L, Antani S, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23(2):304-310 [FREE Full text] [doi: 10.1093/jamia/ocv080] [Medline: 26133894]
58. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. ArXiv. Preprint posted online on November 29, 2020. [doi: 10.5860/choice.189890]
59. Steenstra I, Nouraei F, Bickmore T. Scaffolding empathy: training counselors with simulated patients and utterance-level performance visualizations. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025 1 May; Yokohama Japan p. 1-22. [doi: 10.1145/3706598.3714014]
60. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

XSL•FO
RenderX

61. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc 2005;12(4):486-494 [FREE Full text] [doi: 10.1197/jamia.M1767] [Medline: 15802483]
62. Liu N, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F. Lost in the middle: how language models use long contexts. ArXiv. Preprint posted online on November 20, 2023. [doi: 10.1162/tacl_a_00638]

Abbreviations

AI: artificial intelligence
API: application programming interface
AS: Anthropomorphism Score
BERT: Bidirectional Encoder Representations from Transformers
CoT: chain-of-thought
CUQ: Chatbot Usability Questionnaire
EHR: electronic health record
FDR: false discovery rate
ICC: intraclass correlation coefficient
ICU: intensive care unit
KG: knowledge graph
LLM: large language model
LLMDP: LLM-based digital patient system
LoRA: Low-Rank Adaptation
MIMIC-III: Medical Information Mart for Intensive Care-III
OSCE: Objective Structured Clinical Examination
PCP: primary care physician
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG: retrieval-augmented generation
RCT: randomized controlled trial
RLHF: reinforcement learning with human feedback
RQ: research question
SFT: supervised fine-tuning
SNOMED-CT: Systematized Nomenclature of Medicine - Clinical Terms
SUS: System Usability Scale
UMLS: Unified Medical Language System

Edited by A Benis; submitted 14.Jun.2025; peer-reviewed by W Wang, E Chang; comments to author 11.Aug.2025; accepted 23.Oct.2025; published 02.Jan.2026.

Please cite as:
Li D, Lebai Lutfi S
Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review
JMIR Med Inform 2026;14:e79039
URL: https://medinform.jmir.org/2026/1/e79039
doi:10.2196/79039
PMID:

©Dongliang Li, Syaheerah Lebai Lutfi. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Causal Discovery in Observational Medical Research: Scoping Review

Zuting Liu[1,2], MPH; Tian Luo[1,2], MS; Hailin Ma[1,2], MBBS; Jiali Mo[1,2], MPH; Xia Yang[1,2], MS; Zhenglong Huang[1,2], MPH; Jingkun Li[3*], PhD; Jie Kuang[1,2*], PhD

[1]Department of Epidemiology, School of Public Health, Jiangxi Medical College, Nanchang University, Nanchang, China

[2]Jiangxi Provincial Key Laboratory of Disease Prevention and Public Health, Nanchang University, Nanchang, China

[3]Department of Biostatistics, School of Public Health, Harbin Medical University, Harbin, China

[*]these authors contributed equally

Corresponding Author:
Jie Kuang, PhD
Department of Epidemiology
School of Public Health, Jiangxi Medical College
Nanchang University
No. 1299 Xuefu Avenue
Nanchang, 330019
China
Phone: 86 15807008572
Email: kuangjie@ncu.edu.cn

Abstract

Background: Observational data are fundamental to medical research but present formidable challenges for causal inference. Machine learning–based causal discovery algorithms have emerged as a promising solution to identify causal structures directly from such data. However, the current literature is skewed toward theoretical and methodological innovations, with a critical gap in systematic assessments of performance in medical research settings and a lack of practical guidance for clinicians and researchers on selecting and applying these algorithms in specific medical contexts.

Objective: This study aimed to systematically map and synthesize the application of causal discovery methods in observational medical research, detailing the methodologies used, their application domains, the robustness of the findings, and the practical challenges encountered.

Methods: Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, we conducted a systematic search of Scopus, Web of Science, PubMed, MEDLINE, Embase, and CINAHL from inception to May 2025. We included studies that applied any causal discovery algorithms within a medical research context, encompassing both analyses of real-world observational data and method-validation studies using synthetic or benchmark datasets with a clear medical focus. Purely methodological papers and studies based solely on experimental data were excluded. Data were extracted and synthesized using a descriptive analysis focused on study characteristics, algorithm types, application domains, reported numerical results, and implementation challenges.

Results: Out of 2296 identified publications, 72 (3.1%) met the inclusion criteria. Our synthesis revealed three key themes. The first theme was methodological landscape, where constraint-based algorithms were the most prevalent (38/72, 52.8%), with the fast causal inference (10/72, 13.9%) and Peter-Clark algorithms (9/72, 12.5%) being most common. Score-based (19/72, 26.4%) and hybrid (14/72, 19.4%) methods also represented significant and growing segments (methods were not mutually exclusive). The second theme was application domains and findings, where the majority of studies (54/72, 75%) were in clinical research, with a strong focus on mental health (19/72, 26.4%; eg, identifying symptom networks in schizophrenia and posttraumatic stress disorder) and chronic diseases (19/72, 26.4%; eg, elucidating progression pathways in Alzheimer and diabetes). Etiological research was the primary objective (28/72, 38.9%). Public health applications (18/72, 25%) frequently assessed the causal impacts of behavioral interventions. The third theme was implementation challenges and innovations, where common challenges included pervasive unmeasured confounding, limited sample sizes (noted in more than 20% of studies), and reliance on unvalidated causal assumptions. Emerging innovations focused on longitudinal data frameworks and the integration of multimodal data sources to strengthen causal claims.

XSL•FO
RenderX

Conclusions: This review underscores the growing application of causal discovery algorithms in medical research while also highlighting challenges such as the lack of standardized validation frameworks and persistent confounding. Future efforts must focus on developing evaluation standards and fostering interdisciplinary collaboration to translate these powerful computational techniques into reliable tools for medical research and practice.

(JMIR Med Inform 2026;14:e82499) doi:10.2196/82499

KEYWORDS

causal discovery; medical research; observational data; algorithm; scoping review

Introduction

The advent of big data has driven significant advances in machine learning (ML), yet current methodologies often focus on identifying correlations rather than uncovering underlying causal relationships [1]. This limitation is particularly critical in medicine, where understanding causality is essential for effective clinical decision-making. Contemporary causal research addresses 2 interrelated challenges: causal discovery (identifying causal structures and directions between variables) and causal inference (quantifying effects within predefined causal frameworks) [2-5]. Causal discovery is concerned with deducing the causal structure and directionality among variables from data, with the primary objective of elucidating direct and indirect causal pathways by constructing causal graphs. Typically, it relies on statistical principles such as conditional independence tests to discern potential causal structures from correlational patterns. In contrast, causal inference focuses on quantifying the magnitude of the effect of specific interventions on an outcome variable, presuming a known or assumed causal structure. Therefore, inferring causal relationships, that is, causal discovery, from observational medical data is critical for advancing precision and personalized medicine. A central class of computational methods developed for causal discovery, such as the Peter-Clark (PC) algorithm and fast causal inference (FCI), aims to learn a causal graph that represents the underlying data-generating mechanism. The growing scale and accessibility of diverse medical data resources, encompassing real-world records, such as surgical, therapeutic, and diagnostic information, as well as genetic, lifestyle, and environmental data, have become invaluable for such research [6]. However, these data are often affected by confounding, selection bias, and missing data, posing substantial challenges to deriving reliable causal conclusions. Although randomized controlled trials (RCTs) remain the gold standard for medical decision-making [7], the limits of RCTs and the growing abundance of observational data underscore an urgent need for more robust causal discovery methodologies.

Observational datasets, particularly those derived from real-world settings, offer significant advantages for medical research, including large sample sizes, accessibility, and the ability to reflect real-world clinical scenarios. Previous studies have demonstrated their potential to elucidate disease etiology [8-10], optimize treatment strategies [11,12], and refine prognostic models [13,14]. However, the specific application of causal discovery methods to harness these datasets remains particularly challenging and fragmented. Key barriers include unmeasured confounding, complex data structures, and a lack of consensus on methodological best practices. As a result, there is an insufficient synthesis of how these methods can be robustly applied, validated, and translated into clinical insights. This scoping review aims to fill this gap by systematically analyzing causal discovery methods applied to observational medical studies. Specifically, we critically assess: (1) The evolving methodological landscape and algorithm suitability across clinical contexts; (2) High-impact application domains and translational outcomes; (3) Persistent implementation challenges, such as unmeasured confounding and sample limitations. By mapping these dimensions, we provide a foundation for advancing causal discovery in digital health and precision medicine.

Methods

Study Design

This scoping review adhered to Arksey and O'Malley's framework for scoping review [15] and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [16] (Multimedia Appendix 1). Our primary research question was: "What are the methodologies, clinical applications, and implementation challenges of causal discovery in observational medical research?" The aim was to synthesize evidence to guide future methodological development and clinical translation. Given the rapidly evolving nature of this field, no temporal restrictions were applied.

Search Strategy

To comprehensively review studies of causal discovery algorithms in medical research and its subtypes, 2 researchers (XY and ZH) performed a systematic literature search of Scopus, Web of Science, PubMed, MEDLINE, Embase, and CINAHL in May 2025 and was restricted to studies published in English, using the following search strategies: "causal discovery/causal structure discovery/causal structure learning/causal graph discovery/learning causal models." The search was focused on biomedical databases (PubMed, Embase, etc) as this review aims to map applications in medical research. While computer science–focused databases, such as IEEE Xplore or ACM Digital Library, include methodological developments, they primarily emphasize algorithmic innovation; seminal methodological papers with medical applications are typically also indexed within the selected biomedical sources. The search strategy was designed in consultation with an information specialist. The entire search strategy for the 6 databases is displayed in Multimedia Appendix 2.

XSL•FO
RenderX

Eligibility Criteria

Two researchers (HM and JM) independently screened studies based on predefined inclusion and exclusion criteria by examining their titles, abstracts, and full texts. Studies met the following inclusion criteria: (1) application focus: empirical implementation of causal discovery algorithms in medical or health contexts. (2) Data and methodology: the study applied a causal discovery approach within the observational research paradigm. This included analyses of real-world observational medical data as well as method-validation studies that used synthetic or benchmark datasets to address a medical research question or simulate a medical data environment. (3) Output: the study generated an explicit causal graph or causal structure. (4) Language: English-language publications. The exclusion criteria were: (1) The study purely focused on methodological or theoretical papers without medical applications. (2) Experimental (RCT) or basic science studies. (3) Duplicate publications, brief abstracts, or conference summaries lacking full methodological and results sections. (4) Studies where causal discovery was not the primary methodology. Discrepancies in study eligibility were resolved through consensus between 2 independent reviewers (JK and JL).

Data Extraction and Synthesis

Screening was performed using Rayyan [17], with dual-reviewer verification at all stages. A standardized electronic form was used to extract the following data items: (1) study characteristics: year of publication, country or region of corresponding author affiliation, study design, and sample size; (2) causal discovery methods: specific algorithms, methodological category, and the primary software or package used for implementation; (3) data properties: data type, dimensionality, and temporal features; (4) application domain and specific clinical or public health objectives; and (5) validation approaches and key reported limitations.

Data synthesis was conducted in 2 complementary ways to align with the scoping review's descriptive aims. First, a descriptive quantitative analysis summarized structured characteristics (eg, causal discovery algorithms, software tools, and application domains) using frequencies and percentages. Second, for textual data on reported methodological considerations and challenges, an inductive content summary was performed. Two reviewers (ZL and TL) independently identified recurrent issues and patterns, which were then discussed, consolidated, and organized into overarching descriptive themes (eg, methodological landscape and challenges). This process ensured the results reflected both the quantitative distribution and qualitative substance of the included literature. While this scoping review did not undertake a formal critical appraisal of individual studies using standardized quality assessment or risk-of-bias tools, we systematically extracted study-reported methodological limitations and implementation challenges as a prespecified data item during full-text screening. These author-reported limitations were subsequently subjected to inductive qualitative synthesis, alongside other extracted textual data, to identify recurrent patterns and to develop the descriptive themes presented in the "Challenges" section. This approach was intended to map commonly reported methodological issues in the literature rather than to appraise or grade the quality of individual studies.

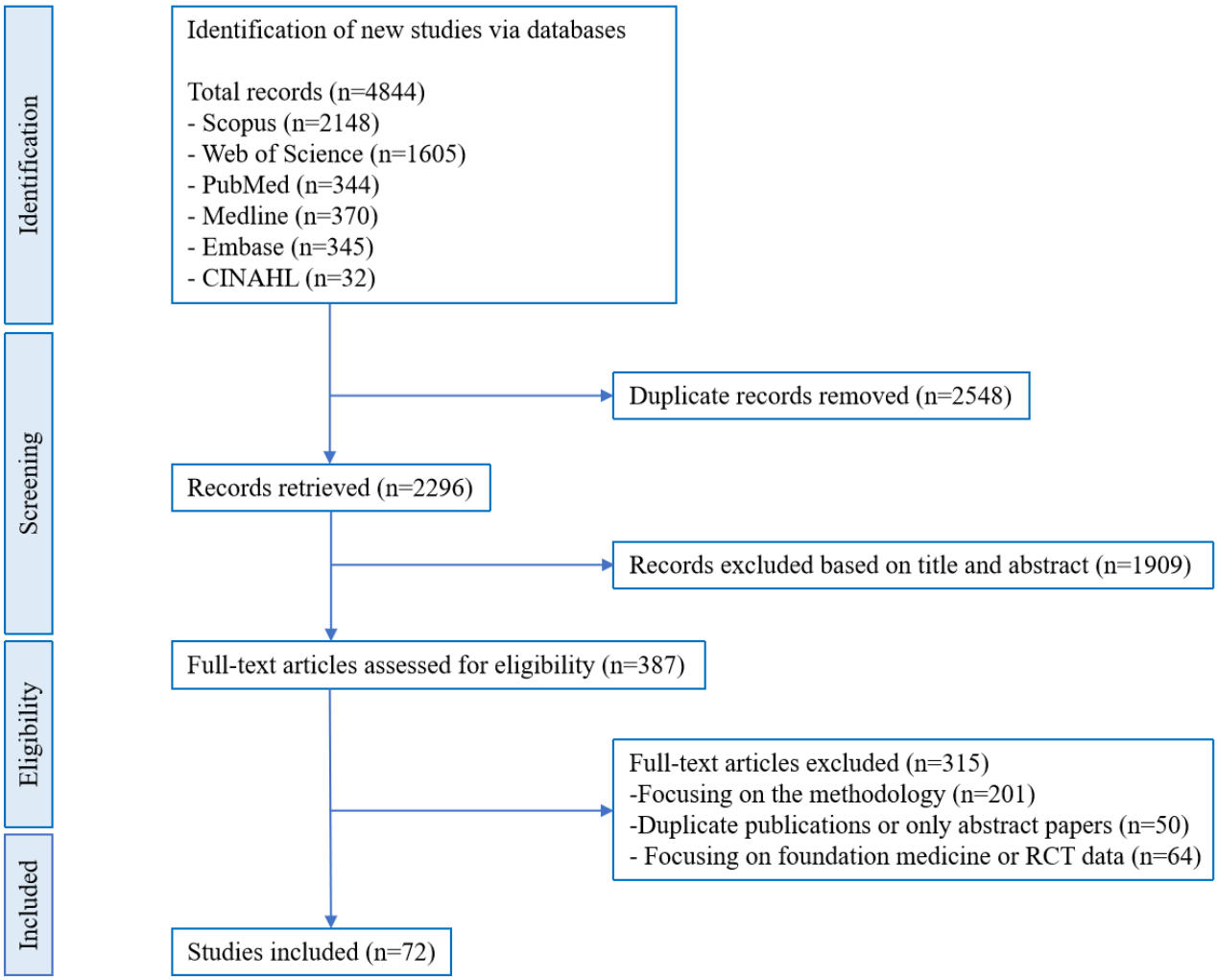
Results

Article Selection

After the removal of duplicates, 2296 unique records were identified for screening. Of these, 72 studies [8-14,18-82] (3.1%) met the inclusion criteria and were included in the final synthesis (Figure 1). After eliminating 2548 duplicates (52.6%), title or abstract screening excluded 1909 records (39.4%). A full-text review of 387 articles led to the exclusion of 315 (81.4%) due to reasons such as a methodological focus (201/315, 63.8%), duplicate publications or abstracts (50/315, 15.9%), and basic research or experimental designs (64/315, 20.3%).

XSL•FO
RenderX

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flowchart for the study selection process. RCT: randomized controlled trial.



Characteristics of the Articles

The characteristics of the included studies (N=72) [8-14,18-82] are summarized in Table 1, covering publication year, country or region, and document type. The number of annual publications progressively increased from 2017 to 2025, with the years 2023-2025 accounting for the majority (46/72, 63.9%). Publications peaked in 2023 and 2024. The majority of studies were peer-reviewed journal articles (64/72, 88.9%), with the remaining studies being conference proceedings. The studies originated from 18 distinct countries or regions, with the United States contributing the largest proportion (25/72, 34.7%), followed by China (8/72, 11.1%), the Netherlands (6/72, 8.3%), and Japan (4/72, 5.6%), accounting for a combined total of 59.7% of publications. Three primary themes emerged from the analysis: (1) methodological landscape, that is, dominant causal discovery approaches and typologies in medical research; (2) application domains, that is, multidimensional implementation scenarios across observational studies; and (3) challenges, that is, methodological limitations and barriers to clinical translation.

XSL•FO
RenderX

Table 1. Characteristics of the studies used in the review, including the year, type, and country of publication (N=72).

Features	Studies, n (%)	References
Year of publication		
2025	14 (19.4)	[8-10,13,18-26,82]
2024	16 (22.2)	[11,14,27-40]
2023	16 (22.2)	[41-56]
2022	8 (11.1)	[57-64]
2021	5 (6.9)	[12,65-68]
2020	7 (9.7)	[69-75]
2019	2 (2.8)	[76,77]
2018	1 (1.4)	[78]
2017	3 (4.2)	[79-81]
Type of publication		
Journal article	64 (88.9)	[8-14,18-31,33,34,37,38,41-48,50,51,53-63,65-82]
Conference article	8 (11.1)	[32,35,36,39,40,49,52,64]
Country of publication		
United States	25 (34.7)	[18,20,22,23,26,28,36,38,43,46,47,51,56-58,60,61,63,65-67,70,73,74,79]
China	8 (11.1)	[19,29,33,42,52,69,77,78]
Netherlands	6 (8.3)	[9,27,59,71,80,81]
Japan	4 (5.6)	[39,40,49,75]
Other countries	29 (40.3)	[8,10-14,21,24,25,30-32,34,35,37,41,44,45,48,50,53-55,62,64,68,72,76,82]

Themes

Theme 1: Methodological Landscape

Causal discovery methodologies are broadly divided into several major categories: constraint-based, score-based, continuous optimization, functional causal models, and hybrid methods. It should be noted that these categories are not mutually exclusive; a single study could use methods from more than one category. Thus, the percentages reported below are calculated based on the total number of included studies (N=72) [8-14,18-82] and may sum to more than 100%. Among the 72 studies analyzed, constraint-based methods constituted the predominant approach (38/72, 52.8%) [8-10,12,14,18,24,25,27,29-32,34-38,43,45,48,52,53,55,57-59,63,64,68,70,71,73,74,77,79-81], followed by score-based (19/72, 26.4%) [8,11,18-20,23,24,26,28,32,33,35,39,42,61,67,73,74,82], hybrid (14/72, 19.4%) [8,13,22,23,33,41,46,47,51,56,60,63,65,69], continuous optimization (9/72, 12.5%) [29,33,36,40,44,49,50,54,62], and functional causal models (5/72, 6.9%) [21,32,66,75,78]. At the algorithmic level, FCI (10/72, 13.9%) [10,18,29,34,38,43,58,63,73,74] and PC (9/72, 12.5%) [12,18,25,27,29,35,36,53,70] algorithms were the most implemented constraint-based methods, whereas greedy fast causal inference (GFCI) (10/72, 13.9%) [13,22,23,41,46,47,51,60,63,65] and Non-combinatorial Optimization via Trace Exponential and Augmented Lagrangian for Structure learning (6/72, 8.3%) [36,40,49,50,54,62] demonstrated notable dominance within the hybrid and continuous optimization categories, respectively. Additionally, time-series causal discovery (eg, Peter-Clark momentary conditional independence) was actively explored in 11 studies [32,44,48,55,64,66,68,72,76,77,82]. These findings highlight a clear trajectory toward the adoption of more integrated, scalable, and nonparametric causal discovery frameworks. The reviewed studies used methodologies from 2 distinct conceptual frameworks. The majority used structural causal discovery algorithms (eg, PC, FCI, and NOTEARS) to learn causal graphs or networks. An analysis of implementation tools revealed the software landscape for causal discovery (Table 2). Of the 72 included studies, 45 (62.5%) [8-10,12,13,18-22,24,26-28,30,32,36-39,41,43,44,46-48,50,51,54,55,57,60-63,65,68-71,74,75,77,79] explicitly specified the software or package used. Among these, R was the most prevalent environment, used in 22 studies [8-10,12,18,20,22,24,26-28,30,37,38,48,55,57,60,61,68,74,82] (22/45, 48.9% of reporting studies, or 22/72, 30.6% of all studies), with the *pcalg* package being the most frequently cited. The standalone Tetrad suite was used in 10 studies [13,18,41,43,46,47,51,63,65,74] (10/45, 22.2% of reporting studies, or 10/72, 13.9% overall). Out of 45 studies, Python and MATLAB were implemented in 7 [9,21,32,36,50,54,62] (15.5%) and 8 [32,44,69-71,75,77,79] (17.7%) of the software-reporting studies, respectively. A notable finding regarding transparency was that 27 studies [11,14,23,25,29,31,33-35,40,42,45,49,52,53,56,58,59,64,66,67,72,73,76,78,80,81] (37.5% of the total) did not report any software details, posing a significant barrier to reproducibility. For details, refer to Table 2.

XSL•FO
RenderX

Table 2. Characteristics of the studies used in the review, including author, causal discovery method, method type, software, disease or area, and type of data.

Authors	Causal discovery method	Method type	Software	Disease or area	Type of data	Validation methods
Gururaghavendran and Murray [18]	PC[a], FCI[b], FGES[c], GRaSP[d]	Constraint-based; Score-based	R (*pcalg*), Tetrad	Cardiovascular mortality	Cohort	Internal stability assessment; Expert review
Petrungaro et al [8]	PC-stable[e], Inter-IAMB[f], HC[g], Tabu[h], MMHC[i], H2PC[j]	Constraint-based; Score-based; Hybrid	R (*bnlearn*)	Sepsis	Cross-sectional	None reported
Li et al [19]	GES[k]	Score-based	Python (DoWhy, Causal Discovery Toolbox)	Alzheimer disease	Cohort	None reported
Korvink et al [20]	GES	Score-based	R (*pcalg*, *dagitty*)	Public health	Cross-sectional	None reported
Penson et al [9]	BCCD[l]	Constraint-based	R (*lavaan*, *RU-causal*)	Chronic fatigue	Cross-sectional	None reported
Noh et al [21]	LiNGAM[m]	Functional-based	Python (DoWhy)	Diabetes	Cross-sectional	Statistical verification
Tseng et al [22]	GFCI[n]	Hybrid	R (*lavaan*)	Autism	Cohort	None reported
Yu et al [10]	FCI and MGM[o]	Constraint-based	R (*rCausalMGM*)	Dementia	Cohort	None reported
Bronstein et al [23]	GFCI, GraSP, GraSP-FCI[p]	Score-based; Hybrid	Not specified	Clinical high risk for psychosis	Cohort	Internal stability assessment
Giuliani et al [13]	GFCI	Hybrid	Tetrad	Schizophrenia	Cross-sectional	None reported
Colineaux et al [24]	Hill climbing, Inter-IAMB, ARACNE[q]	Constraint-based; Score-based	R (*bnlearn*)	Health care use	Cohort	Expert review
Ribeiro-Dantas et al [14]	iMIIC[r]	Constraint-based	Not specified	Breast cancer	Cohort	Expert review; Internal stability analysis
Verveen et al [27]	PC	Constraint-based	R (*pcalg*)	COVID-19	Cohort	None reported
Ren et al [28]	GGES[s]	Score-based	R (*pcalg*)	Esophageal cancer	Cohort	None reported
Guo et al [29]	PC, FCI, and GAE[t]	Constraint-based; Continuous optimization-based	Not specified	Nonsuicidal self-injury	Cross-sectional	None reported
Foraita et al [30]	Modified PC[u]	Constraint-based	R (*pcalg*, *tPC*, *micd*)	Childhood obesity	Cohort	None reported
Ribeiro et al [31]	AnchorFCI[v]	Constraint-based	Not specified	Cardiometabolic diseases	Cross-sectional	Internal stability analysis
Sabzevar et al [32]	RDCM[w], MVGC[x], PCMCI[y], Var-LINGAM[z]	Constraint-based; Score-based; Functional-based	MATLAB, Python	Schizophrenia	Cross-sectional (fMRI time series)	None reported
Zhang et al [33]	MMHC, H2PC, CBAMN[aa]	Score-based; Continuous optimization-based; Hybrid	Not specified	Acute kidney injury	Cohort	Internal stability analysis; Statistical verification; Independent dataset validation
Bernasconi et al [11]	SEM[ab]	Score-based	Not specified	Breast cancer	Cohort	Expert review
Fay et al [25]	PC	Constraint-based	Not specified	Aortic Morphology	Cross-sectional	None reported

XSL•FO
RenderX

Authors	Causal discovery method	Method type	Software	Disease or area	Type of data	Validation methods
Gąsior et al [41]	GFCI	Hybrid	Tetrad	Congenital heart disease	Cross-sectional	None reported
Gao et al [42]	RL[ac]-based causal discovery (IIE[ad] strength)	Score-based	Not specified	Type 2 diabetes	Cross-sectional	None reported
Lee et al [43]	FCI	Constraint-based	Tetrad	Cardiac surgery (CABG[ae] and AVR[af])	Cohort	None reported
Miley et al [51]	GFCI	Hybrid	Tetrad	Schizophrenia	Cohort	Independent dataset validation
Thomas et al [44]	DYNOTEARS[ag]	Continuous optimization-based	MATLAB	Sleep disorders	Cross-sectional	None reported
Vowels et al [45]	SAM[ah]	Constraint-based	Not specified	Mental health (depression and anxiety)	Cross-sectional and cohort	None reported
Bao et al [57]	Modified PC, MMPC, IAMB[ai], GS[aj]	Constraint-based	R (*bnlearn*)	HIV	Cross-sectional	Statistical verification; Internal stability assessment
Singh et al [58]	FCI	Constraint-based	Not specified	In-hospital mortality	Cross-sectional	None reported
Pierce et al [46]	GFCI	Hybrid	Tetrad	PTSD[ak]	Longitudinal time-series	None reported
Langwerden et al [59]	BCCD	Constraint-based	Not specified	Mental disorders	Cross-sectional	None reported
Bronstein et al [60]	GFCI	Hybrid	R (*rcausal*)	Eating disorders	Cross-sectional	None reported
Ganopoulou et al [12]	PC	Constraint-based	R (*MXM*)	Coronary chronic total occlusions	Cross-sectional	None reported
Rawls et al [65]	GFCI	Hybrid	Tetrad	Alcohol use disorder	Cross-sectional	None reported
Li et al [66]	ANM[al]	Functional-based	Not specified	COVID-19	Longitudinal time-series and cross-sectional	None reported
Wang et al [69]	PKCL[am]	Hybrid	MATLAB	T2DM[an] and osteoporosis	Cross-sectional	Expert review
Saxe et al [70]	GLL-PC[ao] and TPC[ap]	Constraint-based	MATLAB	PTSD	Cohort	None reported
Lee et al [47]	GFCI	Hybrid	Tetrad	Nursing home hospitalization	Cohort	Expert review
Andersen et al [48]	TPC	Constraint-based	R (*causalDisco*)	Mental health (stress and depression)	Cross-sectional and cohort	Independent dataset validation
Shen et al [67]	FGES, Modified FGES (temporal constraints)	Score-based	Not specified	T2DM	Cohort	Independent dataset validation
Petersen et al [68]	TPC	Constraint-based	R (*causalDisco*)	Depression	Cohort	Internal stability assessment
Schoenmacker et al [71]	BCCD	Constraint-based	MATLAB	ADHD[aq]	Cohort	None reported
Ahangaran et al [72]	PCDSD[ar]	Probabilistic graphical model	Not specified	Amyotrophic lateral sclerosis	Longitudinal time-series	None reported
Wang et al [73]	FCI[b] and LSTM[as]	Constraint-based; Score-based	Not specified	Knee osteoarthritis	Cohort	None reported
Chen et al [78]	McDSL[at]	Functional-based	Not specified	Acute kidney injury	Cohort	None reported

XSL•FO
RenderX

Authors	Causal discovery method	Method type	Software	Disease or area	Type of data	Validation methods
Galatzer-Levy et al [79]	SVM[au], LGMM[av], and HITON-PC[aw]	Constraint-based	MATLAB	PTSD	Cohort	None reported
Sokolova et al [80]	BCCD	Constraint-based	Not specified	ADHD and ASD[ax]	Cross-sectional	None reported
van Dijk et al [81]	BCCD	Constraint-based	Not specified	The Wechsler Adult Intelligence Scale-IV	Cross-sectional	None reported
Li et al [49]	NOTEARS[ay]	Continuous optimization-based	Not specified	General health status	Longitudinal time-series	None reported
Shen et al [74]	FCI, FGES, and SEM	Constraint-based; Score-based	Tetrad, R (*lavaan*)	Alzheimer disease	Cohort	Expert review
Mouches et al [62]	NOTEARS	Continuous optimization-based	Python (CausalNex)	Brain aging and CVD[az]	Cross-sectional	None reported
Shen et al [61]	FGES	Score-based	R (*rcausal*)	White matter hyperintensities	Cohort	None reported
Zhang et al [63]	FCI and GFCI	Constraint-based; Hybrid	Tetrad	COVID-19-associated AKI[ba]	Longitudinal time-series	None reported
Park et al [50]	NOTEARS	Continuous optimization-based	Python (CausalNex)	Quality of life and mental health	Cross-sectional	None reported
Fu et al [26]	GES	Score-based	R (package unspecified)	Progressive supranuclear palsy	Cohort	None reported
Nogueira et al [64]	ItsPC[bb]	Constraint-based	Not specified	ICU[bc] survival analysis	Cohort	None reported
Li et al [34]	FCI and PPCIT[bd]	Constraint-based	Not specified	Hospital-acquired pressure injury and spinal cord injury	Cohort	Statistical verification
Kazemi et al [35]	PC, GES, and GIES[be]	Constraint-based; Score-based	Not specified	Maternal quality of life (obesity and stress-related)	Cohort	Statistical verification
Naik et al [36]	NOTEARS and PC	Constraint-based; Continuous optimization-based	Python (CausalNex, gcastle, pgmpy)	Non–small cell lung cancer	Cross-sectional	None reported
Ribeiro Santiago et al [82]	Bayesian Structure Learning (partition MCMC[bf])	Score-based	R (*BiDAG*, *blavaan*)	Adolescent emotional disorders (depression or anxiety)	Cohort	Expert review
McCormick et al [37]	GGM[bg], PC-stable, and SEM	Constraint-based	R (*bnlearn*, *lavaan*, *EGAnet*)	Oral health	Cross-sectional	None reported
Zeng et al [38]	FCI	Constraint-based	R (*rCausalMGM*)	Childhood Sjögren's	Cohort	None reported
Li et al [52]	PC	Constraint-based	Not specified	Primary liver cancer	Cohort	None reported
Nagaraj et al [53]	PC	Constraint-based	Not specified	Stress-related disorders	Longitudinal time-series	None reported
Vigneshwaran et al [54]	NOTEARS	Continuous optimization-based	Python (CausalNex)	Brain aging and gray matter atrophy	Cross-sectional	None reported
Petersen et al [55]	TPC	Constraint-based	R (*causalDisco*)	CVD and depression	Cohort	Expert review
Zhang et al [40]	NOTEARS	Continuous optimization-based	Not specified	Psychological stress and obesity-related risks	Longitudinal time-series	Expert review
Tosaki et al [39]	Causal Bayesian network	Score-based	INGOR	Atherosclerosis, hypertension, diabetes, dyslipidemia, osteopenia, chronic kidney disease, COPD[bh], and obesity	Cohort	None reported

XSL•FO
RenderX

Authors	Causal discovery method	Method type	Software	Disease or area	Type of data	Validation methods
Fung et al [56]	Causal Bayesian network	Hybrid	Not specified	COVID-19	Cross-sectional	Statistical verification
Ahangaran et al [76]	PCDSD	Probabilistic graphical model	Not specified	Amyotrophic lateral sclerosis	Longitudinal time-series	None reported
Yang et al [77]	CD-SF^{bi} and CD-SU-SF^{bj}	Constraint-based	MATLAB	Lung cancer	Cross-sectional	None reported

XSL•FO
RenderX

Authors	Causal discovery method	Method type	Software	Disease or area	Type of data	Validation methods
Itahashi et al [75]	LiNGAM[bk]	Functional-based	MATLAB	Adolescent mental issues	Cohort	None reported

[a]PC: Peter-Clark algorithm.
[b]FCI: fast causal inference.
[c]FGES: fast greedy equivalence search.
[d]GRaSP: greedy relaxation of the sparsest permutation.
[e]PC-stable: Peter-Clark stable.
[f]Inter-IAMB: interleaved incremental association Markov blanket.
[g]HC: hill climbing.
[h]Tabu: tabu search.
[i]MMHC: max-min hill climbing.
[j]H2PC: hybrid of HPC and PC.
[k]GES: greedy equivalence search.
[l]BCCD: Bayesian Constraint-based Causal Discovery.
[m]LiNGAM: linear non-Gaussian acyclic model.
[n]GFCI: greedy fast causal inference.
[o]MGM: mixed graphical model.
[p]GraSP-FCI: greedy relaxation of the sparsest permutation fast causal inference.
[q]ARACNE: algorithm for the reconstruction of accurate cellular networks.
[r]iMIIC: interpretable multivariate information-based inductive causation.
[s]GGES: greedy generalized equivalence search.
[t]GAE: graph autoencoder.
[u]Modified PC: modified Peter-Clark algorithm.
[v]AnchorFCI: anchor fast causal inference.
[w]RDCM: regression dynamic causal modeling.
[x]MVGC: multivariate Granger causality.
[y]PCMCI: Peter-Clark momentary conditional independence.
[z]VarLINGAM: vector autoregressive linear non-Gaussian acyclic model.
[aa]CBAMN: cycle-breaking algorithm based on modified NOTEARS.
[ab]SEM: structural expectation-maximization.
[ac]RL: reinforcement learning.
[ad]IIE: inverse information entropy strength.
[ae]CABG: isolated coronary artery bypass grafting.
[af]AVR: isolated aortic valve replacement.
[ag]DYNOTEARS: dynamic noncombinatorial optimization via trace exponential and augmented lagrangian for structure learning.
[ah]SAM: structural agnostic modeling.
[ai]IAMB: incremental association Markov blanket.
[aj]GS: grow-shrink.
[ak]PTSD: posttraumatic stress disorder.
[al]ANM: additive noise model.
[am]PKCL: prior-knowledge-driven local causal structure learning.
[an]T2DM: type 2 diabetes mellitus.
[ao]GLL-PC: global-local learning Peter-Clark algorithm.
[ap]TPC: temporal Peter-Clark algorithm.
[aq]ADHD: attention-deficit/hyperactivity disorder.
[ar]PCDSD: probabilistic causal discovery in sequential datasets.
[as]LSTM: long short-term memory.
[at]McDSL: multiple-cause discovery combined with structure learning.
[au]SVM: support vector machine.
[av]LGMM: latent growth mixture modeling.
[aw]HITON-PC: HITON-PC algorithm.
[ax]ASD: autism spectrum disorder.

XSL•FO
RenderX

[ay]NOTEARS: noncombinatorial optimization via trace exponential and augmented Lagrangian for structure learning.

[az]CVD: cardiovascular disease.

[ba]AKD: acute kidney injury.

[bb]ItsPC: irregular time-series PC.

[bc]ICU: intensive care unit.

[bd]PPCIT: predictive permutation conditional independence tests.

[be]GIES: greedy interventional equivalence search.

[bf]MCMC: Markov chain Monte Carlo.

[bg]GGM: Gaussian graphical model.

[bh]COPD: chronic obstructive pulmonary disease.

[bi]CD-SF: causal discovery based on the streaming feature.

[bj]CD-SU-SF: causal discovery with symmetrical uncertainty based on the streaming feature.

[bk]LiNGAM: linear non-Gaussian acyclic model.

Theme 2: Application Domains

Examination of the 72 studies [8-14,18-82] indicated that causal discovery applications were dominated by clinical research (54/72, 75%), with public health comprising the remainder (18/72, 25%). Etiological investigation emerged as the predominant focus within clinical research, followed by predictive modeling, highlighting the method's central role in elucidating disease mechanisms and enabling clinical prognosis. This application profile demonstrated the adaptability of causal discovery methodologies across the spectrum from individual-level clinical research to population health studies.

Causal Discovery Methodological Approaches to Disease Categories

Among the 54 clinical studies [8-14,18,19,21-23,25-29,31-34,36,38,39,42-46,51-54,58-61,63-65,67,69-81], mental health represented the most extensively examined area (19/72, 26.4%) [13,22,23,27,29,32,45,46,51,53,59,60,65,70,71,75,79-81], with significant attention to schizophrenia (3/72, 4.2%) [13,32,51], posttraumatic stress disorder (3/72, 4.2%) [46,70,79], autism spectrum disorder (2/72, 2.8%) [22,80], and attention-deficit/hyperactivity disorder (2/72, 2.8%) [71,80]. Chronic disease investigations equally comprised 26.4% (19/72) of studies [9-12,14,19,21,26,28,31,36,38,42,52,69,72-74,76], spanning neurodegenerative diseases [10,19,26,72-74,76] (particularly Alzheimer disease [10,19,26,74] and amyotrophic lateral sclerosis [72,76]), malignancies [11,14,28,36,52], diabetes [21,42], and other chronic conditions [9,12,31,38,69]. Additional clinical applications included acute illnesses [8,33,78], pharmacological research [18,63], and other various specialized clinical scenarios [25,34,39,43,44,54,58,61,64,67,77], collectively demonstrating the extensive penetration of causal discovery methodologies in clinical research. For details, refer to Table 2.

Research Focus and Clinical Applications

The 54 clinical studies [8-14,18,19,21-23,25-29,31-34,36,38,39,42-46,51-54,58-61,63-65,67,69-81] showed a distinct concentration in research aims. Etiological exploration represented the predominant focus (28/72, 38.9%) [8-10,18,22,23,25,27,29,31,39,42-46,51,54,60,61,63,65,69-71,74,75,80], establishing causal discovery as a core methodology for investigating disease mechanisms. Predictive modeling constituted the second major application area (13/72, 18.1%) [19,21,26,33,34,58,64,72,73,76-79], while prognostic evaluation [13,14,28,52] and disease diagnosis [32,36,38,53] each accounted for 4 studies respectively. Collectively, these 4 domains accounted for 49 investigations (49/72, 68.1%), underscoring the method's primary use in addressing fundamental clinical questions from causation to outcome assessment.

For example, in a study by Zhang et al [63], causal discovery methods were applied to construct causal networks of COVID-related acute kidney injury using longitudinal electronic health record (EHR) data. Specifically, the FCI and GFCI algorithms were used to automatically identify the development of COVID-19–related acute kidney injury with minimal prior assumptions about pathway connectivity. The results provided a more refined understanding of how remdesivir and other risk factors contribute to acute kidney injury, identifying critical time points for potential intervention and offering valuable insights for clinical decision-making [63]. Similarly, Sokolova et al [80] demonstrated the application of the Bayesian constraint-based causal discovery algorithm to cross-sectional phenotypic data, which revealed directional relationships between attention-deficit/hyperactivity disorder and autism spectrum disorder symptom domains that traditional correlational methods failed to elucidate, generating novel hypotheses regarding the etiological mechanisms underlying their comorbidity [80].

However, the limited applications in clinical decision support development (1/72, 1.4%) [67] and measurement scale implementation (2/72, 2.8%) [59,81] suggested that the translation of causal discovery into practical clinical tools remained underdeveloped, highlighting substantial potential for future research and implementation in real-world health care settings.

Public Health Research Applications

Public health domains were addressed in 25% (18/72) of studies [20,24,30,35,37,40,41,47-50,55-57,62,66,68,82], demonstrating the expanding relevance of causal discovery beyond clinical settings. Health behavior interventions constituted the largest subgroup (8/72, 11.1%) [30,35,37,40,41,49,56,62], primarily focusing on lifestyle quantification and health promotion. Disease management decision-making and health care optimization strategies were examined in 9.7% (7/72) of studies [20,24,47,50,66,68,82], encompassing social determinants of

XSL•FO
RenderX

health, health care use patterns, and nursing home hospitalization reduction. The remaining studies addressed infectious disease transmission research [57,66] and life-course epidemiology [55]. These applications collectively demonstrated the methodological versatility of causal discovery in tackling population-level health challenges and informing public health policy.

Theme 3: Challenges

Common Observational Study Challenges

Medical observational studies using causal discovery methods consistently faced several fundamental limitations. Unmeasured confounding represented the most prevalent challenge (32/72, 44.4%) [11,12,18,20,21,25-27,29,37,38,41,43,44,52-57,59,62,63,65,66,70,71,73,75,82-84], potentially introducing spurious discoveries and biased effect estimates. Sample size limitations were reported in 29.2% (21/72) of studies [9,10,23,24,27,30,31,35,38,40,41,49,55,61,63,69,71,73,78,81,85], restricting statistical power and generalizability. Data quality issues emerged across multiple investigations, including substantial missing data and discretization-induced biases [10,12,30,33,47,49,54,77]. The absence of longitudinal designs in several studies further hindered the verification of temporal directionality in causal relationships [10,13,19,29,54,74,86].

Causal Discovery-Specific Methodological Issues

The application of causal discovery algorithms revealed distinct methodological challenges. Unvalidated algorithmic assumptions were identified in 18.1% (13/72) of studies [8,9,37,43,44,48,52,57,61,71,82,87,88], compromising validity when core assumptions such as causal sufficiency and the Markov property were violated. Limited cross-disease validation [21] and insufficient external validation [59] constrained the broader applicability of findings. Additionally, overreliance on prior knowledge introduced subjectivity into the discovery process, potentially influencing causal structure learning and interpretation [24,36,74,89,90].

Proposed Methodological Improvements

To address these challenges, future research should focus on the following. First, developing confounder-robust algorithms, as demonstrated by AnchorFCI in cardiometabolic disease research [31], and integrating domain knowledge in neurodegenerative studies [10]. Second, advancing longitudinal causal discovery frameworks for irregularly sampled temporal data in EHRs, as seen in nephrotoxicity and intensive care unit studies [63,64]. Third, enabling multimodal data integration strategies, such as coupling neuroimaging with EHRs or synthesizing wearable data with clinical assessments to explore lifestyle diseases [40,49,61]. Finally, incorporating artificial intelligence (AI) architectures like large language models, deep learning, and reinforcement learning to enhance clinical interpretability and model transferability across institutions [21,25,36,42,58,67].

Discussion

Principal Findings

This scoping review represents the first systematic mapping of causal discovery algorithms in observational medical research. It synthesizes key methodological, translational, and implementation gaps across clinical and public health domains. Despite the field's nascent stage, research momentum has accelerated markedly, reflecting growing translational potential. The scarcity of qualifying studies (72 of 2296 screened) underscores both the emergent nature of these methods and the novelty of this review. Our integration of fragmented evidence establishes a foundational framework for interdisciplinary refinement, advancing beyond traditional observational research limitations and supporting the causal discovery's role in medical settings.

Translational Gaps in Application Domains

Clinical research concentrates on etiological exploration, leveraging causal graphs to deconstruct complex diseases (eg, mental health and chronic conditions). However, a severe imbalance exists between mechanism discovery and clinical decision support development (1/72, 1.4%), revealing a critical mechanism-application dissociation. This translational gap primarily stems from 4 interconnected barriers: most methods remain at the proof-of-concept stage without large-scale clinical validation, creating a fundamental trust deficit; existing tools lack seamless integration with clinical workflows and EHR systems; real-world data quality issues and computational demands hinder practical implementation; and crucially, statistically robust findings often fail to translate into clinically actionable recommendations that clinicians can readily understand and apply at the point of care. Public health applications focus on behavior interventions (eg, lifestyle-outcome linkages) and system optimization (eg, health care use patterns), but face domain-specific confounders (eg, policy shifts). Cross-domain translation requires context-adapted methodological frameworks.

Methodological Challenges to Practical Translation

Causal discovery in observational medical research remains dominated by constraint-based (eg, PC and FCI) and score-based methods (eg, GES and FGES), valued for their reliability in exploratory analysis and high-dimensional data, as well as their inherent interpretability; their graphical outputs and testable assumptions align well with clinical reasoning. However, their limitations in identifying causal directions, handling unmeasured confounding, and computational efficiency have spurred new approaches: continuous optimization (eg, NOTEARS) enhances scalability, although often at the cost of transparency; hybrid methods (eg, GFCI) improve robustness; functional models (eg, LiNGAM) enable direction identification; and time-series approaches (eg, PCMCI) capture dynamic causality. This interpretability-scalability trade-off critically guides model selection in clinical settings, where trust and actionable insight depend on understandable outputs. This evolution marks a transition toward more robust, scalable, and temporally-aware causal inference.

XSL•FO
RenderX

Unmeasured confounding remains fundamental, distorting effect estimates and causal structures. Partial ancestral graph algorithms (eg, FCI and RFCI [91]) and negative control designs [92] offer partial solutions but lack universal robustness. Small-sample limitations undermine statistical power and generalizability, particularly in rare diseases or early-phase studies. Transfer learning [93], multimodal data fusion (eg, Cov-Pneum dataset [94]), and cross-disease validation mitigate data scarcity [95], while embedding expert knowledge (eg, in Bayesian networks) reduces model uncertainty. Beyond these challenges, translating causal discovery into practice requires concerted action on 2 fronts: establishing benchmark datasets and implementing a rigorous, explainable, multitiered validation framework. We strongly recommend the curation of medical benchmark datasets derived from gold-standard sources, such as large-scale RCTs or well-validated longitudinal cohorts, to provide a critical foundation for fairly evaluating and comparing algorithms. For instance, the study by Gururaghavendran and Murray [18] used a zero placebo effect as the benchmark to compare multiple causal discovery algorithms, demonstrating that guidance from expert insight and prior knowledge significantly enhances their performance.

Explainable artificial intelligence techniques hold significant potential for improving the causal discovery process. Hasan and Gani [96] introduced a knowledge-guided causal AI system that integrates domain-specific prior knowledge as structural constraints with observational data to refine causal graph discovery. The results show that this approach also lowers computational demands while improving reliability [96]. One study introduced a new causal discovery approach, called REX, leveraging ML models coupled with interpretability techniques, showing REX's effectiveness and robustness in accurately recovering true causal structures, along with its applicability to real-world problems [97]. In addition, moving beyond theoretical sensitivity analyses, a robust validation strategy must integrate internal validation (eg, using bootstrap to assess graph stability), external validation on independent datasets, and most importantly, rigorous evaluation of biological and clinical plausibility against established knowledge or new experimental findings. This comprehensive approach is essential for bridging the current trust gap and generating clinically actionable insights.

Future Directions

To propel the field beyond current methodological constraints and accelerate clinical translation, the following strategic advancements warrant prioritized investigation:

1. Advancing temporal causal inference: integrate time-series data (eg, EHRs) with target trial emulation frameworks to resolve temporal ambiguity, mitigate time-varying confounding, and strengthen causal directionality inference [98].
2. Enabling multimodal data integration: leverage causal ML methodologies to unify heterogeneous data sources (eg, EHRs, genomics, and behavioral data), thereby facilitating robust estimation of individualized treatment effects and bridging critical evidence gaps [10,36,99].
3. Developing AI-enhanced causal discovery: synthesize deep learning (for high-dimensional feature extraction), reinforcement learning (for adaptive decision-making), and causal representation learning to enhance the accuracy, scalability, and interpretability of causal inference in complex biomedical systems [21,25,42,100].
4. Accelerating clinical translation: it is imperative to establish reproducible benchmarking frameworks and standardized reporting guidelines. We encourage journals to adopt policies that actively promote code and data availability, thereby enhancing transparency and reproducibility. The integration of sensitivity analyses into routine methodological reporting should be considered essential for evaluating the robustness of inferred causal relationships. Furthermore, the development of cross-institutional validation protocols and standardized tools that facilitate the incorporation of clinical domain knowledge into causal discovery pipelines will be critical for fostering broader clinical adoption.
5. Enhancing interpretability: improving the interpretability of causal discovery outputs and building clinical trust requires heightened methodological transparency and consistent reporting standards. The field should move toward developing domain-specific reporting guidelines for causal discovery in medicine, which would detail the algorithms used, parameter settings, and evaluation metrics. Systematically embedding interpretability-focused techniques, such as feature importance analysis or causal pathway explanation, can help bridge inferred causal structures with established clinical evidence.

Ultimately, the real-world impact of causal discovery hinges on its ability to deliver comprehensible and actionable insights that resonate with clinical researchers and practitioners. However, whether causal discovery can genuinely drive scientific progress, generate novel insights, and inform scholarly discourse in fields can ultimately be determined only through its repeated application in practice, coupled with rigorous and transparent reporting of the results [101].

Strengths and Limitations

This review provides a timely, comprehensive synthesis of causal discovery algorithms across diverse medical domains. Nevertheless, limitations exist: First, restricting the search to English-language publications may have excluded valuable non-English evidence. Second, our review reveals a pronounced geographic concentration of research, with the majority of studies originating from the United States and China. This likely reflects disparities in access to large-scale digital health infrastructure, advanced computational resources, and targeted research funding. Consequently, the current evidence base and its associated methodological trends may not be fully representative of global research activity or directly generalizable to health care settings in other regions. Third, the search was conducted within biomedical databases and relied on a core set of terms centered on "causal discovery." While this approach efficiently captured studies self-identifying with this paradigm, it is possible that some relevant applications published under different terminology (eg, "Bayesian network structure learning") in computational science venues were not retrieved. Finally, while we mapped application breadth, a formal assessment of reporting quality or bias risk was not

XSL•FO
RenderX

conducted; future work should implement standardized appraisal tools to systematically evaluate methodological rigor.

Conclusions

Causal discovery is emerging as a pivotal approach in medical observational research, offering transformative potential for etiological exploration and clinical decision-making. However, critical barriers impede clinical translation, notably the absence of standardized validation frameworks and persistent confounding. Future progress necessitates rigorous evaluation standards and AI-causal methodology integration to enable robust, generalizable real-world decision support.

Acknowledgments

In accordance with journal policy, we confirm that no generative artificial intelligence tools were used in the writing of this manuscript or in the generation of its scientific content. Artificial intelligence–assisted tools were used solely for minor language polishing and proofreading. All conceptualization, analysis, intellectual interpretation, and authorship of the final manuscript are solely attributable to the human authors.

Data Availability

All data generated or analyzed during this study are included in this published article and Multimedia Appendices 1-2.

Funding

This study was supported by the National Natural Science Foundation of China (grant numbers 82360667 and 82160645), the Natural Science Foundation of Jiangxi Province (grant number 20212BAB206091), and the National Undergraduate Training Program for Innovation and Entrepreneurship (grant number 202510403077).

Authors' Contributions

JK and JL were involved in the concept and design of the review. ZL, TL, HM, JM, XY, and ZH made major contributions to literature search, data collection, data analysis, and data interpretation. ZL and JK wrote and edited the manuscript. All authors read and approved the final manuscript.

JK and JL are co-corresponding authors of this paper. JL can be reached at ljk545464@163.com.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 72 KB - medinform_v14i1e82499_app1.docx]

Multimedia Appendix 2
Queries used for the database search.
[DOCX File , 19 KB - medinform_v14i1e82499_app2.docx]

References

1. Hernán MA. The c-word: scientific euphemisms do not improve causal inference from observational data. Am J Public Health 2018;108(5):616-619. [doi: 10.2105/AJPH.2018.304337] [Medline: 29565659]
2. Nogueira AR, Pugnana A, Ruggieri S, Pedreschi D, Gama J. Methods and tools for causal discovery and causal inference. WIREs Data Min Knowl Discov 2022;12(2):e1449. [doi: 10.1002/widm.1449]
3. Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Appl Inform (Berl) 2016;3:3 [FREE Full text] [doi: 10.1186/s40535-016-0018-x] [Medline: 27195202]
4. Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. In: Computer Reviews. Cambridge: The MIT Press; 2017.
5. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Cambridge, MA: MIT Press; 2000.
6. Kaufman H, Rappoport N, Gilad A, Linial M. Advancing causal inference in medicine using biobank data. J Biomed Inform 2025;171:104903. [doi: 10.1016/j.jbi.2025.104903] [Medline: 40953713]
7. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016;374(22):2175-2181. [doi: 10.1056/NEJMms1604593] [Medline: 27248626]
8. Petrungaro B, Kitson NK, Constantinou AC. Investigating potential causes of sepsis with bayesian network structure learning. Appl Intell 2025;55(7):17. [doi: 10.1007/s10489-025-06405-5]

XSL•FO
RenderX

9. Penson A, Bucur IG, Walraven I, Grootenhuis MA, Maurice-Stam H, van der Heiden-van der Loo M, et al. Structural equation modeling to explore putative causal factors for chronic fatigue in childhood cancer survivors: a DCCSS LATER study. J Cancer Surviv 2025. [doi: 10.1007/s11764-024-01738-5] [Medline: 40019719]
10. Yu X, Lophatananon A, Holmes V, Muir KR, Guo H. Investigating causal networks of dementia using causal discovery and natural language processing models. NPJ Dement 2025;1(1):4. [doi: 10.1038/s44400-025-00006-2] [Medline: 40351535]
11. Bernasconi A, Zanga A, Lucas PJF, Scutari M, Trama A, Stella F. A causal network model to estimate the cardiotoxic effect of oncological treatments in young breast cancer survivors. Prog Artif Intell 2024. [doi: 10.1007/s13748-024-00348-7]
12. Ganopoulou M, Kangelidis I, Sianos G, Angelis L. Causal models for the result of percutaneous coronary intervention in coronary chronic total occlusions. Applied Sciences 2021;11(19):9258. [doi: 10.3390/app11199258]
13. Giuliani L, Sanmarchi F, Mucci A, Rucci P, Caporusso E, Bucci P, et al. Investigating the causal pathways among psychopathological variables, cognitive impairment, and real-life functioning in people with schizophrenia. Schizophrenia (Heidelb) 2025;11(1):1 [FREE Full text] [doi: 10.1038/s41537-024-00545-2] [Medline: 39753575]
14. Ribeiro-Dantas MDC, Li H, Cabeli V, Dupuis L, Simon F, Hettal L, et al. Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients. iScience 2024;27(5):109736 [FREE Full text] [doi: 10.1016/j.isci.2024.109736] [Medline: 38711452]
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
16. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
17. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
18. Gururaghavendran R, Murray EJ. Can algorithms replace expert knowledge for causal inference? A case study on novice use of causal discovery. Am J Epidemiol 2025;194(5):1399-1409. [doi: 10.1093/aje/kwae338] [Medline: 39218433]
19. Li C, Wang S, Xia Y, Shi F, Tang L, Yang Q, et al. Risk factors and predictive models in the progression from MCI to alzheimer's disease. Neuroscience 2025;565:312-319. [doi: 10.1016/j.neuroscience.2024.11.056] [Medline: 39645072]
20. Korvink M, Biondolillo M, Van Dijk JW, Banerjee A, Simenz C, Nelson D. Detection of potential causal pathways among social determinants of health: a data-informed framework. Soc Sci Med 2025;373:118025. [doi: 10.1016/j.socscimed.2025.118025] [Medline: 40164049]
21. Noh MJ, Kim YS. Diabetes prediction through linkage of causal discovery and inference model with machine learning models. Biomedicines 2025;13(1):124 [FREE Full text] [doi: 10.3390/biomedicines13010124] [Medline: 39857708]
22. Tseng A, Francis SM, Rawls E, Conelea C, Grissom NM, Kummerfeld E, et al. Integrating causal discovery and clinically-relevant insights to explore directional relationships between autistic features, sex at birth, and cognitive abilities. Psychol Med 2025;55:e89. [doi: 10.1017/S0033291725000571] [Medline: 40098476]
23. Bronstein MV, Kummerfeld E, Bearden CE, Cornblatt BA, Walker EF, Woods SW, et al. Delineating empirically plausible causal pathways to suicidality among people at clinical high risk for psychosis. J Psychopathol Clin Sci 2025;134(3):239-250. [doi: 10.1037/abn0000969] [Medline: 39913476]
24. Colineaux H, Lepage B, Chauvin P, Dimeglio C, Delpierre C, Lefèvre T. Contribution of structure learning algorithms in social epidemiology: application to real-world data. Int J Environ Res Public Health 2025;22(3):348 [FREE Full text] [doi: 10.3390/ijerph22030348] [Medline: 40238329]
25. Fay L, Hepp T, Winkelmann MT, Peters A, Heier M, Niendorf T, et al. Determinants of ascending aortic morphology: cross-sectional deep learning-based analysis on 25 073 non-contrast-enhanced NAKO MRI studies. Eur Heart J Cardiovasc Imaging 2025;26(5):895-907. [doi: 10.1093/ehjci/jeaf081] [Medline: 40052574]
26. Fu M, Sankararaman S, Vossel KA, Chang TS. Identifying common disease trajectories of progressive supranuclear palsy with electronic health records. Mov Disord Clin Pract 2025;12(10):1528-1538. [doi: 10.1002/mdc3.70119] [Medline: 40344390]
27. Verveen A, Nugroho FA, Bucur IG, Wynberg E, van Willigen HDG, Davidovich U, et al. Long-term sequelae of SARS-CoV-2 two years following infection: exploring the interplay of biological, psychological, and social factors. Psychol Med 2024;54(15):1-11. [doi: 10.1017/S0033291724002721] [Medline: 39618323]
28. Ren S, Beeche CA, Iyer K, Shi Z, Auster Q, Hawkins JM, et al. Graphical modeling of causal factors associated with the postoperative survival of esophageal cancer subjects. Med Phys 2024;51(3):1997-2006. [doi: 10.1002/mp.16656] [Medline: 37523254]
29. Guo X, Wang L, Li Z, Feng Z, Lu L, Jiang L, et al. Factors and pathways of non-suicidal self-injury in children: insights from computational causal analysis. Front Public Health 2024;12:1305746 [FREE Full text] [doi: 10.3389/fpubh.2024.1305746] [Medline: 38532971]
30. Foraita R, Witte J, Börnhorst C, Gwozdz W, Pala V, Lissner L, et al. A longitudinal causal graph analysis investigating modifiable risk factors and obesity in a European cohort of children and adolescents. Sci Rep 2024;14(1):6822 [FREE Full text] [doi: 10.1038/s41598-024-56721-y] [Medline: 38514750]

XSL•FO
RenderX

31. Ribeiro AH, Crnkovic M, Pereira JL, Fisberg RM, Sarti FM, Rogero MM, et al. AnchorFCI: harnessing genetic anchors for enhanced causal discovery of cardiometabolic disease pathways. Front Genet 2024;15:1436947 [FREE Full text] [doi: 10.3389/fgene.2024.1436947] [Medline: 39717478]
32. Sabzevar S, Masoudnia S, Araabi BN, Nazemzadeh M, Tavakoli H. Enhanced diagnosis of schizophrenia using integrated node embedding and causal discovery on imbalanced fMRI data. 2024 Presented at: Proceedings of the 10th International Conference on Signal Processing and Intelligent Systems (ICSPIS) 2024; December 25-26, 2024; Shahrood, Iran p. 234-239. [doi: 10.1109/icspis65223.2024.10931071]
33. Zhang M, Zhang X, Dai M, Wu L, Liu K, Wang H, et al. Development and validation of a multi-causal investigation and discovery framework for knowledge harmonization (MINDMerge): A case study with acute kidney injury risk factor discovery using electronic medical records. Int J Med Inform 2024;191:105588. [doi: 10.1016/j.ijmedinf.2024.105588] [Medline: 39128399]
34. Li Y, Scheel-Sailer A, Riener R, Paez-Granados D. Mixed-variable graphical modeling framework towards risk prediction of hospital-acquired pressure injury in spinal cord injury individuals. Sci Rep 2024;14(1):25067 [FREE Full text] [doi: 10.1038/s41598-024-75691-9] [Medline: 39443567]
35. Kazemi K, Ryhta I, Azimi I, Niela-Vilen H, Axelin A, Rahmani AM, et al. Impact of physical activity on quality of life during pregnancy: a causal ML approach. Annu Int Conf IEEE Eng Med Biol Soc 2024;2024:1-6. [doi: 10.1109/EMBC53108.2024.10782584] [Medline: 40040140]
36. Naik N, Khandelwal A, Joshi M, Atre M, Wright H, Kannan K. Applying large language models for causal structure learning in non small cell lung cancer. 2024 Presented at: Proceedings of the IEEE 12th International Conference on Healthcare Informatics (ICHI); June 3-6, 2024; Orlando, FL p. 688-693. [doi: 10.1109/ichi61247.2024.00110]
37. McCormick KM, Ribeiro Santiago PH, Jamieson L. The impact of COVID-19 on the oral health self-care practices of Australian adults. J Public Health (Berl.) 2024 Jun 05;34(3):525-534. [doi: 10.1007/s10389-024-02287-y]
38. Zeng W, Thatayatikom A, Winn N, Lovelace TC, Bhattacharyya I, Schrepfer T, et al. The Florida scoring system for stratifying children with suspected Sjögren's disease: a cross-sectional machine learning study. Lancet Rheumatol 2024;6(5):e279-e290. [doi: 10.1016/S2665-9913(24)00059-6] [Medline: 38658114]
39. Tosaki T, Uchino E, Harada Y, Sakuragi M, Koyanagi Y, Okajima S. Subgrouping causal networks of disease onset in large-scale health and medical data using supercomputer Fugaku. 2024 Presented at: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6, 2024; Lisbon, Portugal p. 4384-4391. [doi: 10.1109/bibm62325.2024.10822105]
40. Zhang J, Cong R, Deng O, Li Y, Lam K, Jin Q. Analyzing lifestyle and behavior with causal discovery in health data from wearable devices and self-assessments. 2024 Presented at: Proceedings of the IEEE International Conference on E-Health Networking, Application and Services, HealthCom 2024; November 18-20, 2024; Nara. [doi: 10.1109/healthcom60970.2024.10880758]
41. Gąsior JS, Młyńczak M, Williams CA, Popłonyk A, Kowalska D, Giezek P, et al. The discovery of a data-driven causal diagram of sport participation in children and adolescents with heart disease: a pilot study. Front Cardiovasc Med 2023;10:1247122 [FREE Full text] [doi: 10.3389/fcvm.2023.1247122] [Medline: 38075979]
42. Gao XE, Hu JG, Chen B, Wang YM, Zhou SB. Causal discovery approach with reinforcement learning for risk factors of type II diabetes mellitus. BMC Bioinformatics 2023;24(1):296 [FREE Full text] [doi: 10.1186/s12859-023-05405-x] [Medline: 37480046]
43. Lee JJR, Srinivasan R, Ong CS, Alejo D, Schena S, Shpitser I, et al. Causal determinants of postoperative length of stay in cardiac surgery using causal graphical learning. J Thorac Cardiovasc Surg 2023;166(5):e446-e462 [FREE Full text] [doi: 10.1016/j.jtcvs.2022.08.012] [Medline: 36154975]
44. Thomas A, Niranjan M, Legg J. Causal analysis of physiological sleep data using granger causality and score-based structure learning. Sensors (Basel) 2023;23(23):9455 [FREE Full text] [doi: 10.3390/s23239455] [Medline: 38067827]
45. Vowels LM, Vowels MJ, Carnelley KB, Millings A, Gibson-Miller J. Toward a causal link between attachment styles and mental health during the COVID-19 pandemic. Br J Clin Psychol 2023;62(3):605-620 [FREE Full text] [doi: 10.1111/bjc.12428] [Medline: 37300241]
46. Pierce B, Kirsh T, Ferguson AR, Neylan TC, Ma S, Kummerfeld E, et al. Causal discovery replicates symptomatic and functional interrelations of posttraumatic stress across five patient populations. Front Psychiatry 2022;13:1018111 [FREE Full text] [doi: 10.3389/fpsyt.2022.1018111] [Medline: 36793783]
47. Lee K, Kummerfeld E, Robinson E, Anderson L, Rantz M. Data-driven analytics to discover APRN's impact on nursing home hospitalization: causal discovery analysis. J Am Med Dir Assoc 2023;24(11):1746-1754. [doi: 10.1016/j.jamda.2023.05.003] [Medline: 37302798]
48. Andersen TO, Sejling C, Jensen AK, Drews HJ, Ritz B, Varga TV, et al. Nighttime smartphone use, sleep quality, and mental health: investigating a complex relationship. Sleep 2023;46(12):zsad256. [doi: 10.1093/sleep/zsad256] [Medline: 37758231]
49. Li YX, Deng O, Ogihara A, Nishimura S, Jin Q. Causal discovery of health features from wearable device and traditional Chinese medicine diagnosis data. 2023 Presented at: Proceedings of the 25th International Conference on Human-Computer Interaction, HCI International 2023; July 23, 2023; Copenhagen, Denmark p. 556-569. [doi: 10.1007/978-3-031-48041-6_37]

XSL•FO
RenderX

50. Park CH, Kwon J, Lee JT, Ahn S. Impact of criterion versus norm-referenced assessment on the quality of life in Korean medical students. J Korean Med Sci 2023;38(17):e133 [FREE Full text] [doi: 10.3346/jkms.2023.38.e133] [Medline: 37128877]
51. Miley K, Meyer-Kalos P, Ma S, Bond DJ, Kummerfeld E, Vinogradov S. Causal pathways to social and occupational functioning in the first episode of schizophrenia: uncovering unmet treatment needs. Psychol Med 2023;53(5):2041-2049 [FREE Full text] [doi: 10.1017/S0033291721003780] [Medline: 37310333]
52. Li X, Qian XL, Liang LT, Kong LJ, Dong QL, Chen JJ. Causally-aware intraoperative imputation for overall survival time prediction. 2023 Presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 17-24, 2023; Vancouver, BC p. 15681-15690. [doi: 10.1109/cvpr52729.2023.01505]
53. Nagaraj S, Goodday S, Hartvigsen T, Boch A, Garg K, Gowda S, et al. Dissecting the heterogeneity of "in the wild" stress from multimodal sensor data. NPJ Digit Med 2023 Dec 20;6(1):237. [doi: 10.1038/s41746-023-00975-9] [Medline: 38123810]
54. Vigneshwaran V, Wilms M, Forkert ND. The causal link between cardiometabolic risk factors and gray matter atrophy: an exploratory study. Heliyon 2023;9(11):e21567 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e21567] [Medline: 38027770]
55. Petersen AH, Ekstrøm CT, Spirtes P, Osler M. Constructing causal life-course models: comparative study of data-driven and theory-driven approaches. Am J Epidemiol 2023;192(11):1917-1927 [FREE Full text] [doi: 10.1093/aje/kwad144] [Medline: 37344193]
56. Fung H, Sgaier SK, Huang VS. Discovery of interconnected causal drivers of COVID-19 vaccination intentions in the US using a causal Bayesian network. Sci Rep 2023 May 16;13(1):6988. [doi: 10.1038/s41598-023-33745-4] [Medline: 37193707]
57. Bao L, Li C, Li R, Yang S. Causal structural learning on MPHIA individual dataset. J Am Stat Assoc 2022;117(540):1642-1655 [FREE Full text] [doi: 10.1080/01621459.2022.2077209] [Medline: 36620488]
58. Singh H, Mhasawade V, Chunara R. Generalizability challenges of mortality risk prediction models: a retrospective analysis on a multi-center database. PLOS Digit Health 2022;1(4):e0000023 [FREE Full text] [doi: 10.1371/journal.pdig.0000023] [Medline: 36812510]
59. Langwerden RJ, Van der Heijden PT, Claassen T, Derksen JJL, Egger JIM. The structure of dimensions of psychopathology in normative and clinical samples: applying causal discovery to MMPI-2-RF scales to investigate clustering of psychopathology spectra and -factors. Front Psychiatry 2022;13:1026900 [FREE Full text] [doi: 10.3389/fpsyt.2022.1026900] [Medline: 36440421]
60. Bronstein MV, Everaert J, Kummerfeld E, Haynos AF, Vinogradov S. Biased and inflexible interpretations of ambiguous social situations: associations with eating disorder symptoms and socioemotional functioning. Int J Eat Disord 2022;55(4):518-529 [FREE Full text] [doi: 10.1002/eat.23688] [Medline: 35132668]
61. Shen X, Raghavan S, Przybelski SA, Lesnick TG, Ma S, Reid RI, et al. Causal structure discovery identifies risk factors and early brain markers related to evolution of white matter hyperintensities. Neuroimage Clin 2022;35:103077 [FREE Full text] [doi: 10.1016/j.nicl.2022.103077] [Medline: 35696810]
62. Mouches P, Wilms M, Bannister JJ, Aulakh A, Langner S, Forkert ND. An exploratory causal analysis of the relationships between the brain age gap and cardiovascular risk factors. Front Aging Neurosci 2022;14:941864 [FREE Full text] [doi: 10.3389/fnagi.2022.941864] [Medline: 36072481]
63. Zhang J, Kummerfield E, Hultman G, Drawz PE, Adam TJ, Simon G, et al. Application of causal discovery algorithms in studying the nephrotoxicity of remdesivir using longitudinal data from the EHR. AMIA Annu Symp Proc 2022;2022:1227-1236 [FREE Full text] [Medline: 37128413]
64. Nogueira AR, Ferreira CA, Gama J. Temporal nodes causal discovery for in intensive care unit survival analysis. 2022 Presented at: Proceedings of the EPIA Conference on Artificial Intelligence; August 31, 2022; Lisbon, Portugal p. 587-598. [doi: 10.1007/978-3-031-16474-3_48]
65. Rawls E, Kummerfeld E, Zilverstand A. An integrated multimodal model of alcohol use disorder generated by data-driven causal discovery analysis. Commun Biol 2021;4(1):435 [FREE Full text] [doi: 10.1038/s42003-021-01955-z] [Medline: 33790384]
66. Li Z, Xu T, Zhang K, Deng HW, Boerwinkle E, Xiong M. Causal analysis of health interventions and environments for influencing the spread of COVID-19 in the United States of America. Front Appl Math Stat 2021;6:611805. [doi: 10.3389/fams.2020.611805]
67. Shen X, Ma S, Vemuri P, Castro MR, Caraballo PJ, Simon GJ. A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus. Sci Rep 2021;11(1):21025 [FREE Full text] [doi: 10.1038/s41598-021-99990-7] [Medline: 34697394]
68. Petersen AH, Osler M, Ekstrøm CT. Data-driven model building for life-course epidemiology. Am J Epidemiol 2021;190(9):1898-1907. [doi: 10.1093/aje/kwab087] [Medline: 33778840]
69. Wang W, Hu G, Yuan B, Ye S, Chen C, Cui Y, et al. Prior-knowledge-driven local causal structure learning and its application on causal discovery between type 2 diabetes and bone mineral density. IEEE Access 2020;8:108798-108810. [doi: 10.1109/access.2020.2994936]

XSL•FO
RenderX

70. Saxe GN, Ma S, Morales LJ, Galatzer-Levy IR, Aliferis C, Marmar CR. Computational causal discovery for post-traumatic stress in police officers. Transl Psychiatry 2020;10(1):233 [FREE Full text] [doi: 10.1038/s41398-020-00910-6] [Medline: 32778671]
71. Schoenmacker GH, Groenman AP, Sokolova E, Oosterlaan J, Rommelse N, Roeyers H, et al. Role of conduct problems in the relation between attention-deficit hyperactivity disorder, substance use, and gaming. Eur Neuropsychopharmacol 2020;30:102-113. [doi: 10.1016/j.euroneuro.2018.06.003] [Medline: 30292416]
72. Ahangaran M, Jahed-Motlagh MR, Minaei-Bidgoli B. A novel method for predicting the progression rate of ALS disease based on automatic generation of probabilistic causal chains. Artif Intell Med 2020;107:101879. [doi: 10.1016/j.artmed.2020.101879] [Medline: 32828438]
73. Wang Y, You L, Chyr J, Lan L, Zhao W, Zhou Y, et al. Causal discovery in radiographic markers of knee osteoarthritis and prediction for knee osteoarthritis severity with attention-long short-term memory. Front Public Health 2020;8:604654 [FREE Full text] [doi: 10.3389/fpubh.2020.604654] [Medline: 33409263]
74. Shen X, Ma S, Vemuri P, Simon G, Alzheimer's Disease Neuroimaging Initiative. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. Sci Rep 2020;10(1):2975 [FREE Full text] [doi: 10.1038/s41598-020-59669-x] [Medline: 32076020]
75. Itahashi T, Okada N, Ando S, Yamasaki S, Koshiyama D, Morita K, et al. Functional connectomes linking child-parent relationships with psychological problems in adolescence. Neuroimage 2020;219:117013 [FREE Full text] [doi: 10.1016/j.neuroimage.2020.117013] [Medline: 32504815]
76. Ahangaran M, Jahed-Motlagh MR, Minaei-Bidgoli B. Causal discovery from sequential data in ALS disease based on entropy criteria. J Biomed Inform 2019;89:41-55 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.004] [Medline: 30339928]
77. Yang J, Li N, Fang S, Yu K, Chen Y. Semantic features prediction for pulmonary nodule diagnosis based on online streaming feature selection. IEEE Access 2019;7:61121-61135. [doi: 10.1109/access.2019.2903682]
78. Chen W, Hu Y, Zhang X, Wu L, Liu K, He J, et al. Causal risk factor discovery for severe acute kidney injury using electronic health records. BMC Med Inform Decis Mak 2018;18(Suppl 1):13 [FREE Full text] [doi: 10.1186/s12911-018-0597-7] [Medline: 29589567]
79. Galatzer-Levy IR, Ma S, Statnikov A, Yehuda R, Shalev AY. Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. Transl Psychiatry 2017;7(3):e0 [FREE Full text] [doi: 10.1038/tp.2017.38] [Medline: 28323285]
80. Sokolova E, Oerlemans AM, Rommelse NN, Groot P, Hartman CA, Glennon JC, et al. A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). J Autism Dev Disord 2017;47(6):1595-1604 [FREE Full text] [doi: 10.1007/s10803-017-3083-7] [Medline: 28255761]
81. van Dijk MJAM, Claassen T, Suwartono C, van der Veld WM, van der Heijden PT, Hendriks MPH. Evaluating WAIS-IV structure through a different psychometric lens: structural causal model discovery as an alternative to confirmatory factor analysis. Clin Neuropsychol 2017;31(6-7):1141-1154. [doi: 10.1080/13854046.2017.1352029] [Medline: 28726544]
82. Ribeiro Santiago PH, Smithers L, Townsend M, Quintero A, Sawyer A, Soares G, et al. The longitudinal network of peer problems and emotional symptoms among Australian adolescents: Bayesian structure learning of directed acyclic graphs. Dev Psychol 2025;61(8):1479-1494. [doi: 10.1037/dev0001915] [Medline: 39869684]
83. Cheek C, Zheng H, Hallstrom BR, Hughes RE. Application of a causal discovery algorithm to the analysis of arthroplasty registry data. Biomed Eng Comput Biol 2018;9:1179597218756896 [FREE Full text] [doi: 10.1177/1179597218756896] [Medline: 29511363]
84. Bilgel F, Karahasan BC. Understanding covid-19 mobility through human capital: a unified causal framework. Comput Econ 2023:1-41 [FREE Full text] [doi: 10.1007/s10614-023-10359-6] [Medline: 36844967]
85. Młyńczak M, Krysztofiak H. Discovery of causal paths in cardiorespiratory parameters: a time-independent approach in elite athletes. Front Physiol 2018;9:1455 [FREE Full text] [doi: 10.3389/fphys.2018.01455] [Medline: 30425645]
86. Constantinou A, Kitson NK, Liu Y, Chobtham K, Amirkhizi AH, Nanavati PA, et al. Open problems in causal structure learning: a case study of COVID-19 in the UK. Expert Syst Appl 2023;234:121069. [doi: 10.1016/j.eswa.2023.121069]
87. Huie JR, Vashisht R, Galivanche A, Hadjadj C, Morshed S, Butte AJ, et al. Toward a causal model of chronic back pain: challenges and opportunities. Front Comput Neurosci 2022;16:1017412 [FREE Full text] [doi: 10.3389/fncom.2022.1017412] [Medline: 36714527]
88. Anderson LM, Lim KO, Kummerfeld E, Crosby RD, Crow SJ, Engel SG, et al. Causal discovery analysis: a promising tool in advancing precision medicine for eating disorders. Int J Eat Disord 2023;56(11):2012-2021. [doi: 10.1002/eat.24040] [Medline: 37548100]
89. Tao Z, Chi M, Chen L, Ban T, Tu Q, Gao F, et al. Clinical causal analysis via iterative active structure learning. Memetic Comp 2025;17(1):13. [doi: 10.1007/s12293-025-00439-5]
90. Jin W, Ni Y, Spence AB, Rubin LH, Xu Y. Directed cyclic graphs for simultaneous discovery of time-lagged and instantaneous causality from longitudinal data using instrumental variables. J Mach Learn Res 2025;26:22. [Medline: 41394884]
91. Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann Statist 2012;40(1):294-321. [doi: 10.1214/11-aos940]

XSL•FO
RenderX

92. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21(3):383-388 [FREE Full text] [doi: 10.1097/EDE.0b013e3181d61eeb] [Medline: 20335814]
93. Taghados Z, Azimifar Z, Monsefi M, Jahromi MA. CausalCervixNet: convolutional neural networks with causal insight (CICNN) in cervical cancer cell classification-leveraging deep learning models for enhanced diagnostic accuracy. BMC Cancer 2025;25(1):607 [FREE Full text] [doi: 10.1186/s12885-025-13926-2] [Medline: 40181353]
94. Agarwal S, Arya KV, Meena YK. MultiFusionNet: multilayer multimodal fusion of deep neural networks for chest X-ray image classification. Soft Comput 2024;28(19):11535-11551. [doi: 10.1007/s00500-024-09901-x]
95. Tang X, Guo R, Mo Z, Fu W, Qian X. Causality-driven candidate identification for reliable DNA methylation biomarker discovery. Nat Commun 2025;16(1):680 [FREE Full text] [doi: 10.1038/s41467-025-56054-y] [Medline: 39814752]
96. Hasan U, Gani M. A knowledge-guided framework to enhance causal reasoning and human-AI collaboration. : IEEE; 2025 Presented at: Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); March 7-21, 2025; Washington DC, DC, USA. [doi: 10.1109/percomworkshops65533.2025.00041]
97. Renero J, Maestre R, Ochoa I. ReX: causal discovery based on machine learning and explainability techniques. Pattern Recognition 2026;172:112491. [doi: 10.1016/j.patcog.2025.112491]
98. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016;183(8):758-764 [FREE Full text] [doi: 10.1093/aje/kwv254] [Medline: 26994063]
99. Weberpals J, Feuerriegel S, van der Schaar M, Kehl KL. Opportunities for causal machine learning in precision oncology. NEJM AI 2025;2(8):AIp2500277. [doi: 10.1056/aip2500277]
100. Jiao L, Wang Y, Liu X, Li L, Liu F, Ma W, et al. Causal inference meets deep learning: a comprehensive survey. Research (Wash D C) 2024;7:0467 [FREE Full text] [doi: 10.34133/research.0467] [Medline: 39257419]
101. Petersen AH, Ekstrøm CT, Spirtes P, Osler M. Causal discovery and epidemiology: a potential for synergy. Am J Epidemiol 2024;193(10):1341-1342. [doi: 10.1093/aje/kwae101] [Medline: 38825335]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
FCI: fast causal inference
GFCI: greedy fast causal inference
ML: machine learning
PC: Peter-Clark algorithm
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial

Edited by A Coristine; submitted 15.Aug.2025; peer-reviewed by JMI Arockiasamy, C Ma; comments to author 26.Oct.2025; revised version received 16.Feb.2026; accepted 19.Feb.2026; published 13.Mar.2026.

Please cite as:
Liu Z, Luo T, Ma H, Mo J, Yang X, Huang Z, Li J, Kuang J
Causal Discovery in Observational Medical Research: Scoping Review
JMIR Med Inform 2026;14:e82499
URL: https://medinform.jmir.org/2026/1/e82499
doi:10.2196/82499
PMID:

©Zuting Liu, Tian Luo, Hailin Ma, Jiali Mo, Xia Yang, Zhenglong Huang, Jingkun Li, Jie Kuang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Ethics of AI Scribes as Epistemic Agents

Frank Ursin, Dr phil; Sabine Salloch, Prof Dr med, Dr phil

Institute for Ethics, History and Philosophy of Medicine, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, Hannover, Lower Saxony, Germany

Corresponding Author:
Frank Ursin, Dr phil
Institute for Ethics, History and Philosophy of Medicine, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, Hannover, Lower Saxony, Germany

Abstract

Artificial intelligence (AI) scribes using ambient documentation technology that capture clinician-patient dialogue and auto-generate visit notes promise to alleviate documentation burden and reduce clinician burnout. In discussing empirical evidence, highlighting research gaps, and emphasizing technology-related ethical issues beyond established AI and data ethics, we show how this promise comes along with epistemic and relational risks. We proceed in 5 steps: first, we conceptually distinguish ambient documentation from broader ambient intelligence, frame it as a "tech-fix" for documentation-related burnout, and establish the notion of AI scribes as epistemic agents rather than mere transcription tools; second, we summarize empirical evidence on AI scribes, especially with regard to their impact on physicians, highlighting risks such as cognitive deskilling, clinical deprofessionalization, and shifts in epistemic accountability; third, we analyze effects on the patient-physician relationship, focusing on relational and interpretive dimensions, including changes in communication patterns and the omission of narrative nuance; fourth, we highlight risks to patient agency and epistemic justice; and fifth, we propose a design framework for ethical deployment beyond techno-solutionism. We argue that the usefulness of AI scribes should not be justified by short-term effects, but must be assessed in the context of clinical reasoning to improve not only the working conditions of physicians, but also the quality of patient care. The paper proposes a research and design agenda to counter simple "tech-fixes" for systemic problems, envisioning AI scribes that safeguard clinical reasoning and honor patient narratives while delivering relief from documentation burdens.

(JMIR Med Inform 2026;14:e88235) doi:10.2196/88235

KEYWORDS

bioethics; artificial intelligence; electronic health records; physician-patient relationship; documentation; clinical decision making; ambient documentation; epistemic agency; cognitive deskilling; narrative medicine; patient narratives

Artificial Intelligence Scribes in Clinical Care

Clinical Documentation, Clinician Burnout, and Ethical Issues

The requirement for detailed clinical documentation imposes a heavy burden on health care professionals, manifesting both in time use and cognitive load [1]. Clinicians spend up to 50% of their shifts interacting with computers and up to 50% of that time documenting in electronic health record (EHR) systems [2], depending on the medical specialty and clinical setting [3]. In observational and survey studies, high documentation burden is strongly associated with clinician burnout, job dissatisfaction, turnover intention, and reduced time for direct patient care [4,5]. Furthermore, increased documentation demands are linked to a higher frequency of documentation errors or omissions, risks to patient safety, and lower perceived quality of care, as clinicians may rush or shortcut parts of the record [5].

Ambient documentation is envisioned to solve these problems with a combination of 2 artificial intelligence (AI)–based technologies, resulting in so-called "AI scribes" [6]. AI scribes combine ambient listening through audio recording, speech recognition, and natural language processing of the conversation during a patient visit with generative AI that extracts key clinical details from the transcript in real time, produces structured notes, drafts the documentation for the clinician to review, and then integrates it into the EHR [7-9]. Therefore, AI scribes are more than transcription tools [10,11]. In contrast to other ambient intelligence systems monitoring various practices at the patient's bedside [12], however, AI scribes are restricted to documenting and processing oral communication and should be assessed against the common practice of physician-only documentation and the alternative of human medical scribes (Table 1).

XSL•FO
RenderX

Table . Comparison of documentation options.

	Physician	Human medical scribe	AI[a] scribe
Presence in consultation	Physician and patient	Physician, patient, medical scribe (in-person or remote)	Physician, patient, AI scribe
Capture of conversation	Listens and records key points	Medical scribe listens and records	AI scribe records, transcribes, and processes
Note production	Writes or dictates note	Drafts note; physician edits	Generates summaries from transcripts; physician edits
Role in knowledge production	Sole epistemic agent	Recorder under supervision	Epistemic agent selects and rearranges information
Responsibility	Physician	Physician supervises scribe	Shared socio-technical process under physician's responsibility
Scalability	Limited by physician time	Limited by staffing and costs	Highly scalable, automated, vendor-dependent

[a]AI: artificial intelligence.

Clinicians and nurses increasingly use AI scribe products with the intention to optimize their productivity [13]. While industry meets this demand in order to capture a promising market, the evidence on AI scribe prevalence is poor and is mostly based on surveys conducted by companies, for example, stating that nearly 40% of primary care clinicians use AI daily for clinical documentation [14]. A survey of general practitioners in Australia found that 11% have installed or tried an AI scribe, and their regular use more than doubled within 4 months, from 3% to 8% in 2024 [15]. According to the weekly online polls conducted by the Royal Australian College of General Practitioners, the percentage of general practitioners currently using AI scribes increased from 22% in 2024 to 40% in 2025 [16].

However, commentators warn that "the technology is currently too unsophisticated to improve productivity" [17]. It summarizes words spoken during a consultation but cannot replace clinical reasoning, and the time saved with the technology is lost again because clinicians need to revise AI-generated summaries and add their own inputs. In addition to these significant objections, there are concerns stemming from AI and data ethics, such as issues of data protection and privacy, bias, patient consent, hallucinations, omissions, misinterpretations, and speaker attribution errors [10,18-20]. On the one hand, there are concerns that AI scribes may lack higher levels of explicability in terms of algorithmic transparency, which poses an epistemic hurdle in identifying misinterpretations of medical terminology, leading to inaccurate documentation [21,22]. On the other hand, process transparency as the lowest level of opacity, in terms of patients not being aware that an AI scribe is used at all due to lack of disclosure, is a concern [22]. Furthermore, equity issues are raised by the observation that some clinicians avoid using AI in shared offices because of privacy worries, and patients with speech impairments might be excluded [23]. Additionally, long-term effects such as "cognitive debt" from overreliance on AI are a matter of concern [24,25].

While many studies report effects such as reduction in burnout or "pajama time," there is a lack of ethics research examining the normative challenges of deploying AI scribes for physicians, patients, and the patient-physician relationship. For example, although loss of physician autonomy is an important topic [24], existing ethics literature currently focuses on institutional governance and justice rather than on professional autonomy [26]. Our approach will start from the concrete circumstances in clinical encounters and focus on health care providers and patients as relevant stakeholders, as well as their specific relationship in clinical consultations. We will map and describe ethical risks based on empirical evidence and ethical-epistemic concepts that are suitable for identifying the dimensions of these issues (Table 2).

XSL•FO
RenderX

Table . Ethical-epistemic dimensions of analysis of the current practice compared to new or intensified concerns regarding artificial intelligence scribes.

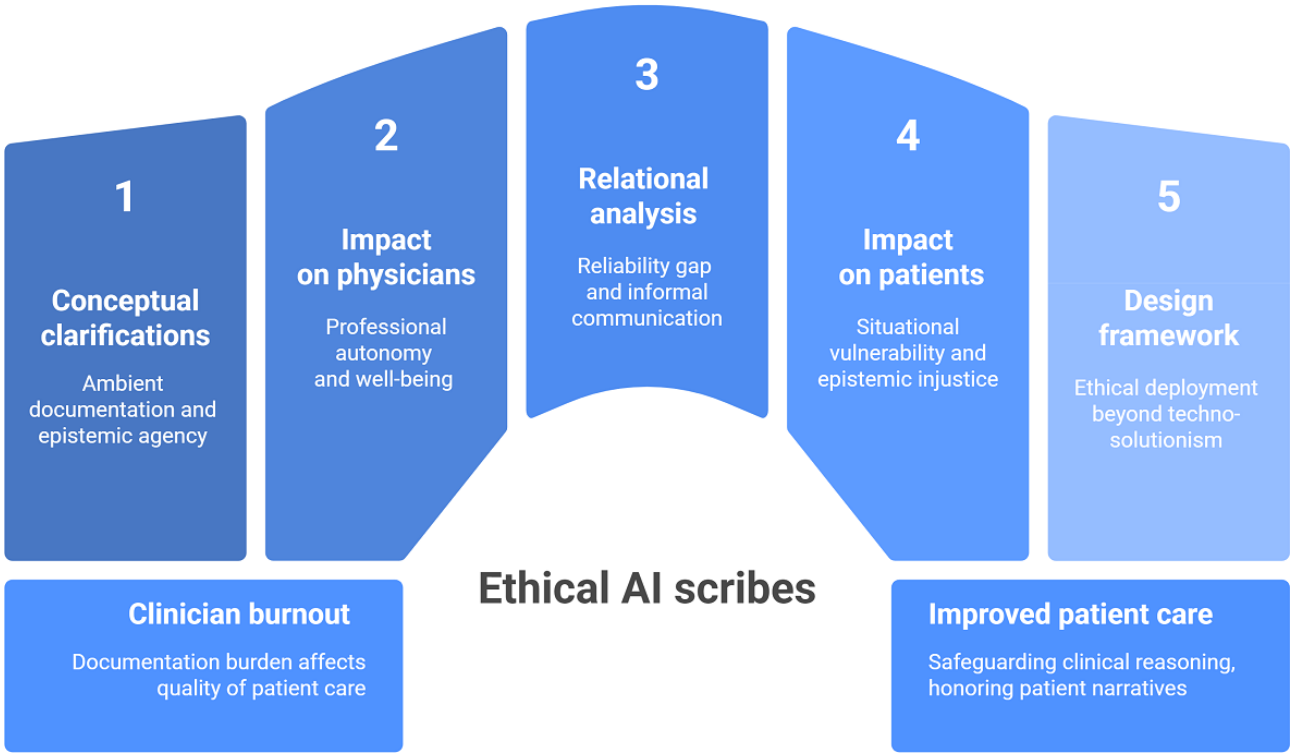
Ethical dimension	Current practice	New or intensified issues with AI[a] scribes
Professional autonomy	Documentation supports reflection and memory	Automation may shift physicians into editors
Epistemic agency	Physician authors and curates the record	AI influencing knowledge formation and reliability gaps
Narrative integrity	Physician summarizes patient narratives	AI may omit narrative nuance or filter "nonclinical" content
Epistemic justice	Risk of testimonial injustice by clinicians	Epistemic failures such as speaker misattribution or errors
Situational vulnerability	Context-specific privacy and trust issues	Specialty-specific AI modes and privacy safeguards
Political economy of data	Data stored within institutional systems	Data processed by vendors and consent to (commercial) secondary uses

[a]AI: artificial intelligence.

Some terminological clarity is required. We will conceive AI scribes as epistemic agents, that is, entities that participate in the pursuit of epistemic goals (ie, goals in knowledge production) within clinical practice and control the formation of beliefs [27,28]. Their outputs operate at a higher level in the data-information-knowledge-wisdom hierarchy [29]: rather than merely recording or transcribing speech, they structure, select, and formulate clinically relevant information, thereby shaping the record that underpins subsequent decision-making. As reasoning increasingly unfolds within hybrid systems of human-machine interaction, AI scribes become part of a socio-technical arrangement in which data are captured, filtered, accepted, or discarded through interactions between clinicians and algorithmic tools [27]. In such settings, epistemic agency should no longer be understood as being represented solely in humans or machines; rather, it emerges relationally through the ways humans and machines jointly shape what becomes recorded as clinically relevant knowledge.

At the same time, AI scribes do not qualify as full artificial agents in a strong philosophical sense. According to Dung's 5 dimensions of agency (autonomy, goal-directedness, efficacy, planning, and intentionality), current systems at least lack intentionality [30,31]. Therefore, they cannot be considered independent epistemic subjects. Nevertheless, because they perform epistemically relevant tasks, influence what becomes recorded as knowledge, and are treated as contributors to clinical reasoning, their agency is tethered to human design choices, institutional settings, and clinical use contexts [32] (p. 162). On this basis, we analyze AI scribes as epistemic agents with respect to the ethical and epistemic issues they raise for physicians, patients, and the patient-physician relationship (Table 2; Figure 1).

Figure 1. Visual abstract of steps leading to an ethical development of artificial intelligence (AI) scribes.



XSL•FO
RenderX

Impacts of AI Scribe Use on Physician Practice

Initial evidence on how the use of AI scribes affects clinicians' professional and personal well-being, as well as the quality of care, shows promising results, even if systematic research is still in its infancy. A survey with 1430 clinicians from 2 US centers, for example, found that ambient documentation technology was associated with reductions in burnout as well as improved well-being scores compared with baseline before ambient documentation technology use [6]. Olson et al [7] found that after 30 days with an ambient AI scribe, burnout among clinicians working in ambulatory clinics decreased from 51.9% to 38.8%. Pearlman et al [33] report from a retrospective cohort study in which clinicians who had used an AI scribe reported reductions in time spent in the electronic documentation systems compared with the control group who had not used the AI assistant. No significant differences between the 2 groups were observed, however, for after-hours time spent documenting, mean appointment length, or monthly number of completed office visits [33,34].

Even if prospective controlled research on the effects of AI scribes as health care interventions is missing so far, such initial empirical studies hint at positive consequences for physician well-being, often using the amount of after-hours or burnout indices as end points. Regarding effects on physicians, risks of deprofessionalization and loss of autonomy are highlighted by authors who see documentation as an integral part of clinical reasoning. According to McCoy et al [35], the process of taking notes, particularly in complex cases, is not "epiphenomenal" to clinical reasoning but constitutes an inherent part of it. They warn that text generated by a large language model (LLM) could reduce the overall quality of patient charts and draw a parallel: "EHR has shaped the way we think, practice, and record our patients' stories—and so will LLMs" [35] (p. 1564). At the same time, it would be simplistic to equate typing with reasoning. The counterargument holds that reducing clerical documentation load may free cognitive resources for higher-order tasks such as diagnostic reflection and attentive listening. From this perspective, AI scribes could support rather than undermine clinical reasoning if they relieve physicians from mechanical aspects of documentation. The ethical concern, therefore, is not the delegation of typing itself but the possible erosion of epistemically productive practices such as synthesizing information, articulating provisional assessments, and identifying knowledge gaps in case of clinicians becoming passive recipients of AI-generated notes.

In addition, a "potential loss of physician autonomy" [24] is feared to be associated with the introduction of AI scribes. According to Funer and Wiesing [36], physician autonomy as an ethical principle is fundamentally anchored in the purpose of promoting a patient's health and well-being. With the growing use of AI-based clinical decision support systems, physicians must be able to critically integrate AI outputs into their clinical reasoning to preserve professional judgment and discretion. With regard to AI scribes, the central issue is whether they retain control over how the clinical encounter is represented in the record. This requires that automatically generated transcripts and protocols remain transparent to the physician and can be revised without increasing the overall workload.

At a more foundational level, the ethical evaluation of AI scribes is intrinsically linked to the question of which purposes clinical note-taking is actually directed at. Only at first glance, AI-driven protocols can be seen as a mere documentation of what has been found, discussed, and decided in a concrete patient appointment [17]. A deeper analysis of the phenomenon needs to account for the fact that clinical notes (as all written documents) have addressees, purposes, and contexts they are serving. As indicated, AI scribes do not only (and not primarily) deliver verbatim transcripts but are designed to summarize the transcripts. Such summaries may be used not only by other health care professionals but also for insurance communication, institutional management, and statistics, and potentially by patients themselves.

Depending on the addressee and context, it is obvious that different information is needed, other parts need to be removed, and styles of language must be taken that account for the respective addressee. Due to their flexibility, LLMs might perform well to account for such different styles of reports. Ethical trade-offs, however, can occur when customizing the AI scribe for such generative functions. For example, AI-generated protocols might either serve the purposes of intraprofessional communication (and not be useful for a lay readership) or they might be appropriate for patients as readers (and thereby potentially losing medical precision). Decisions for and against certain styles for documentation, therefore, have an ethical dimension in potentially excluding certain stakeholder groups as a readership.

AI Scribes Affecting Patient-Physician Relationship

The relationship between patient and physician is a fundamental element of health care and is even more important in times of AI-driven applications [37,38]. It is critical for ensuring effective treatment, patient satisfaction, and overall well-being. At its core, the relationship builds on trust, communication, and mutual respect, enabling physicians to provide personalized care while patients engage actively in their health management, facilitated, for example, by informed consent procedures [39]. Trust in the physician facilitates openness and encourages patients to share honest and intimate information that is needed for accurate diagnosis and effective treatment planning. At first glance, AI scribes might not influence this trustful relationship, as clinical notes usually serve the aim of professional communication. On closer look, however, automated forms of documentation can alter patient-physician communication and their relationship: "When you are being recorded, you're going to be more careful about what you say—that's the Hawthorne effect" [40] (p. 6).

On the one hand, first evidence shows that AI scribes strengthen the therapeutic alliance, and patients tend to trust clinicians even if they use AI scribes [41]. A consistently reported benefit is improved clinician-patient interaction during visits: clinicians focus on the patient, maintaining eye contact and active listening instead of typing on a computer [9,40,42]. An obvious advantage of using AI scribes, therefore, lies in the opportunity to better build rapport with the patient during the clinical encounter.

On the other hand, physicians' trustworthiness can be compromised due to epistemic hurdles that prevent patients from assessing how physicians reach clinical decisions. There

XSL•FO
RenderX

is a reliability gap regarding the epistemic asymmetry that arises when patients cannot clearly determine whether to trust physicians' judgments when AI systems are involved [43]. Applied to AI scribes, this reliability gap complicates the patient-physician relationship by obscuring who accounts for and how the phrasing, selection, or omission of particular information in the medical record occurs. However, when AI scribes act as epistemic agents controlling clinical content, their output becomes part of the basis on which care is delivered. This epistemic opacity can undermine trust in clinical relationships because trust depends not only on interpersonal rapport but also on the transparency and credibility of how knowledge is formed and communicated. AI scribes reconfigure these relational dynamics.

Furthermore, there are critical issues in using AI scribes that relate to the question of which parts of the consultation are documented and which are not. Gordon Schiff [44], as a primary care practitioner, for example, emphasizes the importance of informal social conversation ("chitchat") as an integral part of medical consultations, serving for both a fuller understanding of the patient's health-related living situation and a sustainable patient-provider relationship. This "chitchat," however, is deliberately omitted in AI-generated notes, not as a bug but as a feature of the system. Vendors of AI scribes often advertise their products by claiming that they only summarize "useful" information based on the SOAP (Subjective, Objective, Assessment, and Plan) categories. Schiff [44], therefore, asks whether "our society want(s) an empathetic, highly relational care delivery system built around primary care and trusting relationships" or rather develops in the direction of an efficient, convenient, highly transactional, and dehumanizing care.

It should be noted, however, that the omission of informal conversation from the formal medical record is not unique to AI-generated documentation. Clinical notes have always been selective, legally and administratively oriented documents, and social pleasantries are rarely recorded regardless of the documentation method. The ethical concern therefore does not lie primarily in the absence of such "chitchat" in the written record but in the possibility that the logic of AI-assisted documentation may gradually reshape the interaction itself in terms of communication and behavior patterns. If clinicians and patients adapt their communication styles to what is perceived as "recordable" or "clinically relevant," informal exchanges may become less frequent even at the primary level of interaction.

Impacts of AI Scribes on Patients

In medical encounters, clinicians often chart while patients narrate their concerns. This practice can divide clinician attention, disrupt conversational flow, and impair rapport, ultimately compromising both documentation accuracy and the quality of the patient experience [45]. Introduction of AI scribes in the clinical encounter can directly affect patients' experiences in terms of affective states such as shame, anger, or despair, thus showing ethical significance. One immediate effect is simply a reduced human presence in the exam room, which can improve patient comfort and privacy in case of sensitive examinations [46]. By taking over the role of a medical scribe—the human alternative to AI scribes—AI-driven ambient documentation can eliminate the need for additional staff, for example, in dermatology and venereology, meaning that patients may no longer have a third party observing intimate exams [45,47].

However, it is an open question which clinical contexts are more ethically sensitive or expose more situational vulnerability [48,49]. In acute emergencies, patients may be distressed, disoriented, or unable to consent meaningfully; in psychiatric care, communication often concerns highly intimate, identity-related, or legally sensitive issues; in pediatrics, the triadic constellation between child, parents, and clinician complicates questions of privacy, assent, and representation. Deploying AI scribes identically across all these settings risks ignoring context-specific vulnerabilities and may undermine both trust and epistemic justice, for instance, if patients censor themselves because they feel surveilled or if sensitive relational nuances are systematically filtered out by the system. Design principles for AI scribes should vary by these contexts, considering pertinent issues of patient comfort and privacy as well as other aims, for example, related to security and interprofessional care.

Although we know that patients tend to judge physicians who use AI as less competent and have a significantly lower willingness to make an appointment with physicians who are using AI in general [50], there is little evidence on how patients perceive AI scribes, how much choice or control they wish to have, or how their own narratives are mediated by the AI. In an industry-initiated online survey, 57% of participants favored AI use if it reduces screen time and improves direct interaction with physicians, indicating that patients are willing to accept such technology when it enhances the human connection in care [51].

An advantage lies in the potentially increased attention physicians can now devote to capturing their patients' illness narratives. Kleinman [52] distinguishes between *disease* (biomedical condition), *illness* (the person's lived experience of symptoms, suffering, and changes to their identity), and *sickness* (societal meaning of being ill). By highlighting the patient's story and the gap between physician and patient perspectives (*disease* vs *illness*), Kleinman argues that health care must attend to how people make sense of their illness, not just to how bodies malfunction. Although malfunctioning bodies are also a concern from the patient's perspective, solely reducing health care to the professional perspective on this misrepresents holistic human experience. This aligns with the principles of narrative medicine, which emphasize the importance of honoring patients' stories of illness as a path to better health care and empathy [53]. Narrative medicine argues that medicine practiced with "narrative competence," that is, the ability to absorb and reflect on patient stories, leads to more humane and effective care. It is not about AI scribes recording the transcript of the patient's illness narrative correctly but rather the positive effect of enabling doctors to practice active listening. To unlock the ethos of narrative medicine, AI scribes must help physicians listen more deeply and ensure that what enters the medical record reflects not just data but the moral and experiential truth of the patient's story.

XSL•FO
RenderX

Capturing patient narratives is not just an esthetic or humanistic goal but can also counter epistemic injustice in health care. Epistemic injustice refers to the unfair treatment of someone as a knowledge holder [54]. In health care settings, patients have historically faced testimonial injustice, that is, their symptoms or accounts are dismissed or disbelieved, and hermeneutical injustice, that is, gaps in collective understanding that leave their experiences poorly understood due to, for example, lack of medical terminology. The framework of epistemic injustice helps to explain why, for instance, women or minority patients with pain conditions are often not taken seriously: pervasive stereotypes lead some clinicians to give their testimonies less credence. Similarly, patients with illnesses that lack clear biomedical tests (such as fibromyalgia or chronic fatigue) frequently report not having the language or credibility to explain the reality of their suffering, that is, a hermeneutical gap that leaves them feeling misunderstood.

Beyond these structural forms of epistemic injustice, AI scribes may also introduce more immediate technical biases. Speech recognition and language models often perform less accurately for speakers with strong accents, nonstandard dialects, speech impairments, or atypical prosody. In such cases, patients' accounts may be partially misrecognized, simplified, or omitted altogether, resulting in a technologically mediated form of testimonial injustice. These disparities are not merely technical shortcomings but carry ethical significance, as they systematically reduce the epistemic credibility of certain patient groups within the clinical record.

Integrating narrative-focused AI scribes could help address such epistemic injustices by just recording the patient's narrative to counter testimonial injustice by ensuring the patient's account is documented in their own voice, making it harder to ignore or downplay. Likewise, broader availability of patient narratives can start to fill hermeneutical gaps that are relevant to medical research: patterns in patient-described symptoms and experiences may emerge and be analyzed, expanding the understanding of conditions that patients struggle to articulate. Hence, if AI scribes are to contribute to epistemic justice, their design should prioritize the secure preservation of verbatim patient narratives. Deleting these records, as some vendors do, risks erasing precisely the voices that narrative medicine seeks to amplify. However, the ethical appeal of preserving verbatim patient narratives must be weighed against data minimization and storage-limitation requirements as well as cybersecurity risks when retaining identifiable raw audio or transcripts. Retention of raw records can also create medicolegal exposure if a discoverable verbatim transcript diverges from the finalized clinical note. In addition, AI-generated documentation introduces the risk of "hallucinated compliance," in which standardized sections of a note are inserted even though they were not discussed or performed. Such fabricated entries represent a distinct epistemic failure, as they can overwrite the patient's actual testimony and create a misleading clinical record that appears complete or compliant while lacking factual grounding.

It should be emphasized that the verbatim transcript of the patient's account follows a different structure than the SOAP framework for the purposes of physicians. From the patients' perspective, there are also practical advantages in terms of remembering and understanding medical information when they have access to the results of the AI scribe, which are then tailored to their needs. Patients often struggle to remember what has been said during a clinical encounter: up to 80% of the medical information given by physicians is forgotten immediately, and much of what is remembered is incorrect [55]. Providing patients with a transcript or summary of the conversation can be a remedy to this catastrophe, which has significant disadvantages for the therapeutic alliance and adherence.

Indeed, clinical studies have found that patients have much better recall of information when they receive audio recordings or transcripts of their visits [56], but this is not the exact application area of AI scribes (but may be a beneficial technical design aspect). Nevertheless, by automatically generating written summaries of the consultation, an AI scribe can help patients review the doctor's instructions, medication changes, and explanations at their own pace after the appointment. This memory aid may benefit the patient's autonomy, therapeutic adherence, and compliance, since patients who understand and remember their options are better equipped to participate in their care decisions [17]. However, any recommendation to increase retention of transcripts must be reconciled with data minimization principles, cybersecurity risks, and liability exposure created by discoverable verbatim records.

Ethical Design Recommendations for AI Scribes

From a critical perspective, the use of AI scribes can be conceived as a "tech-fix" [57] for systemic problems in health care systems without correcting the underlying causes in organizational structures, documentation logic, or incentive systems. A well-known feature of technical fixes is that the solution to a problem may become a problem itself [57]. However, if we are willing to accept the consequences of technological deployment, then this tech-fixing may ethically be justified through its overall net benefits. We identified ethical risks ranging from potential deterioration of clinical reasoning and physician autonomy to threats against patients' trust in physicians, reliability gaps, and the call for narrative integrity and epistemic justice. Against this background, we outline 4 key design recommendations for AI scribe developers. Implementing these recommendations could help ensure that AI scribes alleviate clinician burnout without undermining the ethical and epistemic foundations of clinical practice.

Mitigate Cognitive Deskilling and Support Clinical Reasoning

If physicians become passive editors of AI-produced notes, their reflective practice and case memory may atrophy over time, rendering documentation more extensive but less meaningful. Automation bias, that is, the tendency to uncritically accept AI outputs, further exacerbates the risk of deskilling by undermining independent verification and careful reasoning—at least a risk patients fear to witness [58]. Therefore, we recommend designing AI scribes as *cognitive aids* rather than replacements for clinical reasoning. AI scribes should require active clinician engagement in the documentation process to assert their abilities and competencies. For example, the AI

XSL•FO
RenderX

scribe could prompt physicians to confirm or refine key assessments and conclusions rather than auto-populating them. In this scenario, clinicians are not only the human in the loop who must review and edit the AI's note before finalizing. It rather preserves the physician's mental models, experiences, and clinical judgment but also serves as a check against AI errors.

Support Physician Autonomy and Epistemic Agency

Physician autonomy is a cornerstone of professional ethics, grounding the clinician's responsibility to exercise independent judgment in the care of patients. Yet, the introduction of AI scribes may erode physician autonomy, resulting in a slippery slope of deprofessionalization: delegation of clinical reasoning to AI scribes. If AI systems assume too much control over how encounters are represented, they also impair physicians' epistemic authority [31], that is, the competence to define what counts as relevant, true, and meaningful in the clinical record. In doing so, AI scribes risk threatening the epistemic agency of physicians through their own epistemic agency. Therefore, AI scribes must be designed to reinforce, not replace, professional judgment.

Safeguard Patients' Authentic Narratives

Beyond professional autonomy, AI scribes also raise epistemic concerns for patients. When AI scribes selectively capture, summarize, or "clean" patients' illness narratives, then they may inflict epistemic injustices on patients. If AI scribes effectively become the gatekeepers of what enters the medical record, they function as epistemic agents tethered to the design choices of their developers. Therefore, AI scribes should be designed for narrative preservation under data minimization constraints. Rather than default long-term retention of raw audio or transcripts, AI scribe systems should only preserve narrative data (1) if patients have given informed consent to the processing of these data. Furthermore, (2) these data should be stored within a "contestability window" during which they can be accessed for verification and as a memory aid before default deletion. Another future challenge is the navigation of situations in which options for selective pausing or revision at the patient's or physician's request should be reasonably provided.

However, preserving verbatim patient narratives also raises questions about data ownership and the political economy of clinical AI systems. Many AI scribes are developed and operated by private vendors who may gain access to large-scale datasets of intimate patient-physician conversations. Without appropriate governance, the ethical goal of narrative preservation could unintentionally enable the commercial accumulation and secondary use of highly sensitive data for model training or product development. Therefore, governance should aim at purpose limitation, institutional data stewardship, consent procedures for data protection, and explicit restrictions on secondary uses of conversational data.

Tailor AI Scribes to Clinical Settings and Situational Vulnerability

Different clinical contexts such as psychiatry, pediatrics, emergency medicine, oncology, or sexual and reproductive health expose patients to varying degrees and types of situational vulnerability [49]. Therefore, AI scribes should be designed and implemented as context-sensitive technologies, not as generic, one-size-fits-all tools. This includes enabling configurable "modes" that reflect different ethical priorities in different contexts, which must be weighed against excessive alerts or mandatory confirmations leading to workflow friction and "alert fatigue." Complementary approaches such as periodic post hoc auditing or selective review mechanisms may provide safeguards without burdening the clinical encounter. In practice, however, encounter-level configuration may be difficult to implement in enterprise-wide EHR environments with standardized licensing and infrastructure. Context sensitivity may therefore need to be realized at different levels, for example, through institutional policies, specialty-specific templates, workflow settings, or deployment guidelines rather than through fully individualized system modes. For example, there could be default data minimization and stronger consent routines in psychiatry and sexual health; emphasis on continuity-of-care documentation in chronic care; or heightened attention to parental and surrogate roles in pediatrics. In high-vulnerability settings, the default may need to be more restrictive in terms of clearer in-room signaling when recording is active. Context-sensitive deployment also has a temporal dimension: some phases of care (eg, initial crisis intervention, breaking bad news, and end-of-life conversations) may justifiably be documented more sparsely or with delayed AI support, to protect emotional safety.

Conclusions

AI scribes promise meaningful relief from documentation burden, yet their integration into clinical practice raises ethical and epistemic questions that reach far beyond mere workflow efficiency. Viewing these systems as epistemic agents makes visible how they participate in shaping clinical reasoning, mediating patient narratives, and influencing trust within the patient-physician relationship. Their deployment must therefore be guided by context-sensitive design principles that preserve professional autonomy, safeguard patients' epistemic agency, and respect the situational vulnerabilities inherent in different clinical settings. Rather than functioning as simple "tech-fixes," AI scribes should be developed and evaluated as tools that augment clinicians' cognitive work and the relational foundations of health care. If AI scribes are designed not only to reduce documentation burden, they may also contribute to better health care.

Acknowledgments

ChatGPT (GPT-5.2) has been used to check the spelling, grammar, and style of the manuscript. Based on the manuscript's abstract, Napkin has been used to draft Figure 1, which has been revised and corrected by FU.

XSL•FO
RenderX

Funding

There was no funding supporting this work.

Data Availability

Data sharing is not applicable to this paper, as no datasets were generated or analyzed during the current study. All evidence is publicly available.

Authors' Contributions

Conceptualization: FU, SS
Investigation: FU, SS
Methodology: FU, SS
Visualization: FU
Writing – original draft: FU, SS
Writing – review & editing: FU, SS
Both authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Wang Z, West CP, Vaa Stelling BE, et al. Measuring Documentation Burden in Healthcare: Agency for Healthcare Research and Quality; 2024. [doi: 10.23970/AHRQEPCTB47]
2. Cox ML, Farjat AE, Risoli TJ, et al. Documenting or operating: where is time spent in general surgery residency? J Surg Educ 2018 Nov;75(6):e97-e106. [doi: 10.1016/j.jsurg.2018.10.010] [Medline: 30522828]
3. Pinevich Y, Clark KJ, Harrison AM, Pickering BW, Herasevich V. Interaction time with electronic health records: a systematic review. Appl Clin Inform 2021 Aug;12(4):788-799. [doi: 10.1055/s-0041-1733909] [Medline: 34433218]
4. Levy DR, Withall JB, Mishuris RG, et al. Defining documentation burden (DocBurden) and excessive DocBurden for all health professionals: a scoping review. Appl Clin Inform 2024 Oct;15(5):898-913. [doi: 10.1055/a-2385-1654] [Medline: 39137903]
5. Ball CG, McBeth PB. The impact of documentation burden on patient care and surgeon satisfaction. Can J Surg 2021 Aug;64(4):E457-E458. [doi: 10.1503/cjs.013921]
6. You JG, Dbouk RH, Landman A, et al. Ambient documentation technology in clinician experience of documentation burden and burnout. JAMA Netw Open 2025 Aug 1;8(8):e2528056. [doi: 10.1001/jamanetworkopen.2025.28056] [Medline: 40839265]
7. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025 Oct 1;8(10):e2534976. [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
8. Galloway JL, Munroe D, Vohra-Khullar PD, et al. Impact of an artificial intelligence-based solution on clinicians' clinical documentation experience: initial findings using ambient listening technology. J Gen Intern Med 2024 Oct;39(13):2625-2627. [doi: 10.1007/s11606-024-08924-2] [Medline: 38980463]
9. Shah SJ, Crowell T, Jeong Y, et al. Physician perspectives on ambient AI scribes. JAMA Netw Open 2025 Mar 3;8(3):e251904. [doi: 10.1001/jamanetworkopen.2025.1904] [Medline: 40126477]
10. Martinez-Martin N, Luo Z, Kaushal A, et al. Ethical issues in using ambient intelligence in health-care settings. Lancet Digit Health 2021 Feb;3(2):e115-e123. [doi: 10.1016/S2589-7500(20)30275-2] [Medline: 33358138]
11. Cook DJ, Augusto JC, Jakkula VR. Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob Comput 2009 Aug;5(4):277-298. [doi: 10.1016/j.pmcj.2009.04.001]
12. Gerke S, Yeung S, Cohen IG. Ethical and legal aspects of ambient intelligence in hospitals. JAMA 2020 Feb 18;323(7):601-602. [doi: 10.1001/jama.2019.21699] [Medline: 31977033]
13. Topaz M. Invisible scribes: can nurses trust ambient AI for clinical documentation? J Contin Educ Nurs 2025 Sep;56(9):358-359. [doi: 10.3928/00220124-20250814-03] [Medline: 40857680]
14. Rajaee L. Nearly 40% of primary care clinicians now use AI daily for clinical documentation, Elation survey finds. Elation. 2025. URL: https://www.elationhealth.com/resources/elation-health-ehr/state-of-ai-blog?utm_sourcedetail=BusinessWire&utm_source=PR&utm_medium=WebReferral&utm_campaign=Marketing-2025-Q3-StateofAI [accessed 2026-03-15]
15. Knibbs J. GP AI scribe use more than doubles in four months. The Medical Republic. 2025. URL: https://www.medicalrepublic.com.au/gp-ai-scribe-use-more-than-doubles-in-four-months/111429 [accessed 2026-03-15]
16. Wisbey M. How do patients feel about doctors using AI in consults? Royal Australian College of General Practitioners. 2025. URL: https://www1.racgp.org.au/newsgp/professional/how-do-patients-feel-about-doctors-using-ai-in-con [accessed 2026-03-15]

XSL•FO
RenderX

17. Whitaker P. Doctors, beware the AI scribe. The New Statesman. 2025. URL: https://www.newstatesman.com/politics/health/2025/07/doctors-beware-the-ai-scribe [accessed 2026-03-15]
18. Topaz M, Peltonen LM, Zhang Z. Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. NPJ Digit Med 2025 Sep 24;8(1):569. [doi: 10.1038/s41746-025-01895-6] [Medline: 40993221]
19. Wang H, Yang R, Alwakeel M, et al. An evaluation framework for ambient digital scribing tools in clinical applications. NPJ Digit Med 2025 Jun 13;8(1):358. [doi: 10.1038/s41746-025-01622-1] [Medline: 40514413]
20. Sezgin E, Sirrianni JW, Kranz K. Evaluation of a digital scribe: conversation summarization for emergency department consultation calls. Appl Clin Inform 2024 May 15;15(3):600-611. [doi: 10.1055/a-2327-4121] [Medline: 38749477]
21. Haberle T, Cleveland C, Snow GL, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc 2024 Apr 3;31(4):975-979. [doi: 10.1093/jamia/ocae022] [Medline: 38345343]
22. Ursin F, Lindner F, Ropinski T, Salloch S, Timmermann C. Levels of explicability for medical artificial intelligence: what do we normatively need and what can we technically reach? Ethik Med 2023 Jun;35(2):173-199. [doi: 10.1007/s00481-023-00761-x]
23. Nguyen OT, Turner K, Charles D, et al. Implementing digital scribes to reduce electronic health record documentation burden among cancer care clinicians: a mixed-methods pilot study. JCO Clin Cancer Inform 2023 Mar;7(7):e2200166. [doi: 10.1200/CCI.22.00166] [Medline: 36972488]
24. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform 2025 Aug 1;13:e80898. [doi: 10.2196/80898] [Medline: 40749188]
25. Brosnahan H. Totalitarian technics: the hidden cost of AI scribes in healthcare. Med Health Care Philos 2025 Mar;29(1):155-166. [doi: 10.1007/s11019-025-10312-4] [Medline: 41284135]
26. Herington J, Cho MK. A justice-first approach to ambient intelligence in healthcare. Am J Bioeth 2026 Feb;26(2):10-21. [doi: 10.1080/15265161.2025.2497972] [Medline: 40340546]
27. Ponti M, Kasperowski D, Gander AJ. Narratives of epistemic agency in citizen science classification projects: ideals of science and roles of citizens. AI & Soc 2024 Apr;39(2):523-540. [doi: 10.1007/s00146-022-01428-9]
28. Coeckelbergh M. AI and epistemic agency: how AI influences belief revision and its normative implications. Soc Epistemol 2026 Jan 2;40(1):59-71. [doi: 10.1080/02691728.2025.2466164]
29. Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy. J Inf Sci 2007 Apr;33(2):163-180. [doi: 10.1177/0165551506070706]
30. Dung L. Understanding artificial agency. Philos Q 2025 Mar 24;75(2):450-472. [doi: 10.1093/pq/pqae010]
31. Hauswald R. Artificial epistemic authorities. Soc Epistemol 2025 Nov 2;39(6):716-725. [doi: 10.1080/02691728.2025.2449602]
32. Rubeis G. Ethics of Medical AI: Springer; 2024. URL: https://link.springer.com/book/10.1007/978-3-031-55744-6 [accessed 2026-04-11]
33. Pearlman K, Wan W, Shah S, Laiteerapong N. Use of an AI scribe and electronic health record efficiency. JAMA Netw Open 2025 Oct 1;8(10):e2537000. [doi: 10.1001/jamanetworkopen.2025.37000] [Medline: 41071550]
34. Rotenstein L, Melnick ER, Iannaccone C, et al. Virtual scribes and physician time spent on electronic health records. JAMA Netw Open 2024 May 1;7(5):e2413140. [doi: 10.1001/jamanetworkopen.2024.13140] [Medline: 38787556]
35. McCoy LG, Manrai AK, Rodman A. Large language models and the degradation of the medical record. N Engl J Med 2024 Oct 31;391(17):1561-1564. [doi: 10.1056/NEJMp2405999] [Medline: 39465898]
36. Funer F, Wiesing U. Physician's autonomy in the face of AI support: walking the ethical tightrope. Front Med (Lausanne) 2024;11:1324963. [doi: 10.3389/fmed.2024.1324963] [Medline: 38606162]
37. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. JAMA 1992;267(16):2221-2226. [doi: 10.1001/jama.1992.03480160079038] [Medline: 1556799]
38. Zohny H, Allen JW, Wilkinson D, Savulescu J. Which AI doctor would you like to see? Emulating healthcare provider-patient communication models with GPT-4: proof-of-concept and ethical exploration. J Med Ethics 2025 Mar 3:jme-2024-110256. [doi: 10.1136/jme-2024-110256] [Medline: 40032513]
39. Ludewigs S, Narchi J, Kiefer L, Winkler EC. Ethics of the fiduciary relationship between patient and physician: the case of informed consent. J Med Ethics 2024 Jan;51(1):59-66. [doi: 10.1136/jme-2022-108539]
40. Lawrence K, Kuram VS, Levine DL, et al. Informed consent for ambient documentation using generative AI in ambulatory care. JAMA Netw Open 2025 Jul 1;8(7):e2522400. [doi: 10.1001/jamanetworkopen.2025.22400] [Medline: 40694347]
41. Evans K, Papinniemi A, Ploderer B, et al. Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: perspectives of clinicians and patients. Musculoskelet Sci Pract 2025 Aug;78:103333. [doi: 10.1016/j.msksp.2025.103333] [Medline: 40328004]
42. Stults CD, Deng S, Martinez MC, et al. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Netw Open 2025 May 1;8(5):e258614. [doi: 10.1001/jamanetworkopen.2025.8614] [Medline: 40314951]
43. Buhr E, Onder O, Rudra P, Ursin F. Trust and artificial intelligence in the doctor-patient relationship: epistemological preconditions and reliability gaps. Ethics Inf Technol 2025 Dec;27(4):60. [doi: 10.1007/s10676-025-09867-w]
44. Schiff GD. AI-driven clinical documentation - driving out the chitchat? N Engl J Med 2025 May 15;392(19):1877-1879. [doi: 10.1056/NEJMp2416064] [Medline: 40353584]

XSL•FO
RenderX

45. Itauma O, Itauma I. AI scribes: boosting physician efficiency in clinical documentation. Int J Bioinform Biosci 2024 Mar 10;14(1):09-18. [doi: 10.5121/ijbb.2024.14102]
46. Chambers S, Gangal A, Stoff B. Ethics of ambient documentation. J Am Acad Dermatol 2026 Jan;94(1):e3-e4. [doi: 10.1016/j.jaad.2025.03.007] [Medline: 40081665]
47. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribes in dermatology: implications for practice. J Am Acad Dermatol 2022 Apr;86(4):968-969. [doi: 10.1016/j.jaad.2021.11.011] [Medline: 34793929]
48. Mackenzie C, Rogers W, Dodds S. Introduction: what is vulnerability, and why does it matter for moral theory. In: Vulnerability: New Essays in Ethics and Feminist Philosophy: Oxford University Press; 2014:1-29. [doi: 10.1093/acprof:oso/9780199316649.003.0001]
49. Luna F. Elucidating the concept of vulnerability: layers not labels. Int J Fem Approaches Bioeth 2009 Mar;2(1):121-139. [doi: 10.3138/ijfab.2.1.121]
50. Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. JAMA Netw Open 2025 Jul 1;8(7):e2521643. [doi: 10.1001/jamanetworkopen.2025.21643] [Medline: 40674054]
51. Babia P. Patients want doctors, not data entry: ModMed finds nearly 60% support use of AI if it means more face time. Businesswire. 2025. URL: https://www.businesswire.com/news/home/20250624882329/en/Patients-Want-Doctors-Not-Data-Entry-ModMed-Finds-Nearly-60-Support-Use-of-AI-if-it-Means-More-Face-Time [accessed 2026-03-15]
52. Kleinman A. The Illness Narratives: Suffering, Healing, and the Human Condition: Basic Books; 1988.
53. Charon R. The patient-physician relationship. Narrative medicine: a model for empathy, reflection, profession, and trust. JAMA 2001 Oct 17;286(15):1897-1902. [doi: 10.1001/jama.286.15.1897] [Medline: 11597295]
54. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing: Oxford Academic; 2007. URL: https://academic.oup.com/book/32817 [accessed 2026-04-11]
55. Kessels RPC. Patients' memory for medical information. J R Soc Med 2003 May 1;96(5):219-222. [doi: 10.1258/jrsm.96.5.219]
56. Pitkethly M, Macgillivray S, Ryan R. Recordings or summaries of consultations for people with cancer. Cochrane Database Syst Rev 2008 Jul 16(3):CD001539. [doi: 10.1002/14651858.CD001539.pub2] [Medline: 18646075]
57. Sætra HS, Selinger E. Technological remedies for social problems: defining and demarcating techno-fixes and techno-solutionism. Sci Eng Ethics 2024 Dec 2;30(6):60. [doi: 10.1007/s11948-024-00524-x] [Medline: 39621170]
58. Kuppanda PM, Janda M, Soyer HP, Caffery LJ. What are patients' perceptions and attitudes regarding the use of artificial intelligence in skin cancer screening and diagnosis? Narrative review. J Invest Dermatol 2025 Aug;145(8):1858-1865. [doi: 10.1016/j.jid.2025.01.013] [Medline: 40019459]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
LLM: large language model
SOAP: Subjective, Objective, Assessment, and Plan

Edited by A Coristine; submitted 21.Nov.2025; peer-reviewed by J Aarts, M Iscoe; revised version received 18.Mar.2026; accepted 18.Mar.2026; published 30.Apr.2026.

Please cite as:
Ursin F, Salloch S
The Ethics of AI Scribes as Epistemic Agents
JMIR Med Inform 2026;14:e88235
URL: https://medinform.jmir.org/2026/1/e88235
doi:10.2196/88235

© Frank Ursin, Sabine Salloch. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing Clinical Decision (Support) Systems Combining the Scientific and Regulatory Perspective: European Insights on Challenges, Requirements, and Practical Guidance

Sanne E W Vrijlandt[1], MD; Sanne W C M Janssen[2], MD; Annemarie M C van Rossum[2], Prof Dr; Niek B Achten[3], PhD; Rianne Oostenbrink[1], PhD

[1]Department of Pediatrics, Division of General Pediatrics, Erasmus MC-Sophia Children's Hospital, University Medical Center, Dr. Molewaterplein 40, Rotterdam, The Netherlands

[2]Department of Pediatrics, Division of Pediatric Infectious Diseases and Immunology, Erasmus MC-Sophia Children's Hospital, University Medical Center, Rotterdam, The Netherlands

[3]Department of Pediatrics, Erasmus MC-Sophia Children's Hospital, University Medical Center, Rotterdam, The Netherlands

Corresponding Author:
Rianne Oostenbrink, PhD
Department of Pediatrics, Division of General Pediatrics, Erasmus MC-Sophia Children's Hospital, University Medical Center, Dr. Molewaterplein 40, Rotterdam, The Netherlands

Abstract

Clinical decision-making is a critical process where physicians balance risks and benefits. Clinical Decision Support Systems (CDSSs) are increasingly used to help in this process. The regulatory landscape for CDSSs is evolving significantly, with the new European Medical Device Regulation (MDR) now requiring, CE certification for certain CDSSs. This shift poses challenges for health care providers to develop CDSSs in an effective and useful manner while adhering to regulations. This viewpoint comments on diverse challenges and provides solutions to develop a reliable, well integrated and practical tool for clinical use. Using three tools (the Early Onset Sepsis Calculator, Feverkidstool, and Neonatal Procalcitonin Intervention Study algorithm) as examples, we explore the development of CDSSs across four core characteristics: scientific basis, technical aspects, safety, and sustainability. These characteristics recur across the main development processes; scientific development, regulatory assessment, and implementation in routine practice. Successful integration of CDSSs into clinical practice requires a comprehensive understanding of the interconnections between these processes. For example, decisions on algorithm validation and platform selection in the scientific process influence choices for technical safety during the regulatory process. Developers should consider both regulation requirements and clinical needs, to create CDSSs that are not only compliant but also adaptable to the rapidly changing healthcare landscape. We outline a developer's checklist, for practical guidance, but also appeal for structural support, including national protocols and dedicated hospital roles, to help developers implement CDSSs successfully.

(JMIR Med Inform 2026;14:e72809) doi:10.2196/72809

KEYWORDS

clinical decision support systems; regulatory requirements; pediatrics; infection diseases; diagnostic reasoning

Juggling Risks and Benefits Using Tools—A Doctor's Evolving Job

Clinical decision-making involves a delicate balancing act between risks and benefits. It requires the synthesis of a physician's clinical intuition with objective variables. For instance, deciding whether to start or stop antibiotic treatment in children at risk for severe bacterial infections illustrates this process. Distinguishing between early-stage bacterial illness and viral or physiological causes in neonates and febrile children is particularly challenging, and wrong decisions can have fatal consequences. To aid in making accurate decisions, clinicians may use clinical decision support systems (CDSSs). As these tools become more integral to clinical practice, the regulatory framework governing them is evolving, with significant implications for their development and implementation.

Until 2017, European regulations allowed CDSSs to be used without Conformité Européenne (CE) certification under the Medical Device Directive. However, the introduction of the Medical Device Regulation (MDR) (Textbox 1) expanded the definition of medical devices to include CDSSs [1]. The MDR's implementation has made CE certification more stringent, linking it directly to a CDSS's risk classification (Figure 1) [2]. For instance, any CDSS classified from Class II onwards now requires approval from a notified body. Proper classification is the responsibility of the developer. And using only CE-certified CDSSs is the responsibility of the user. Currently, a transition period allows for the legal use of both CE-certified and noncertified CDSSs until 2028. After this period, hospitals and

XSL•FO
RenderX

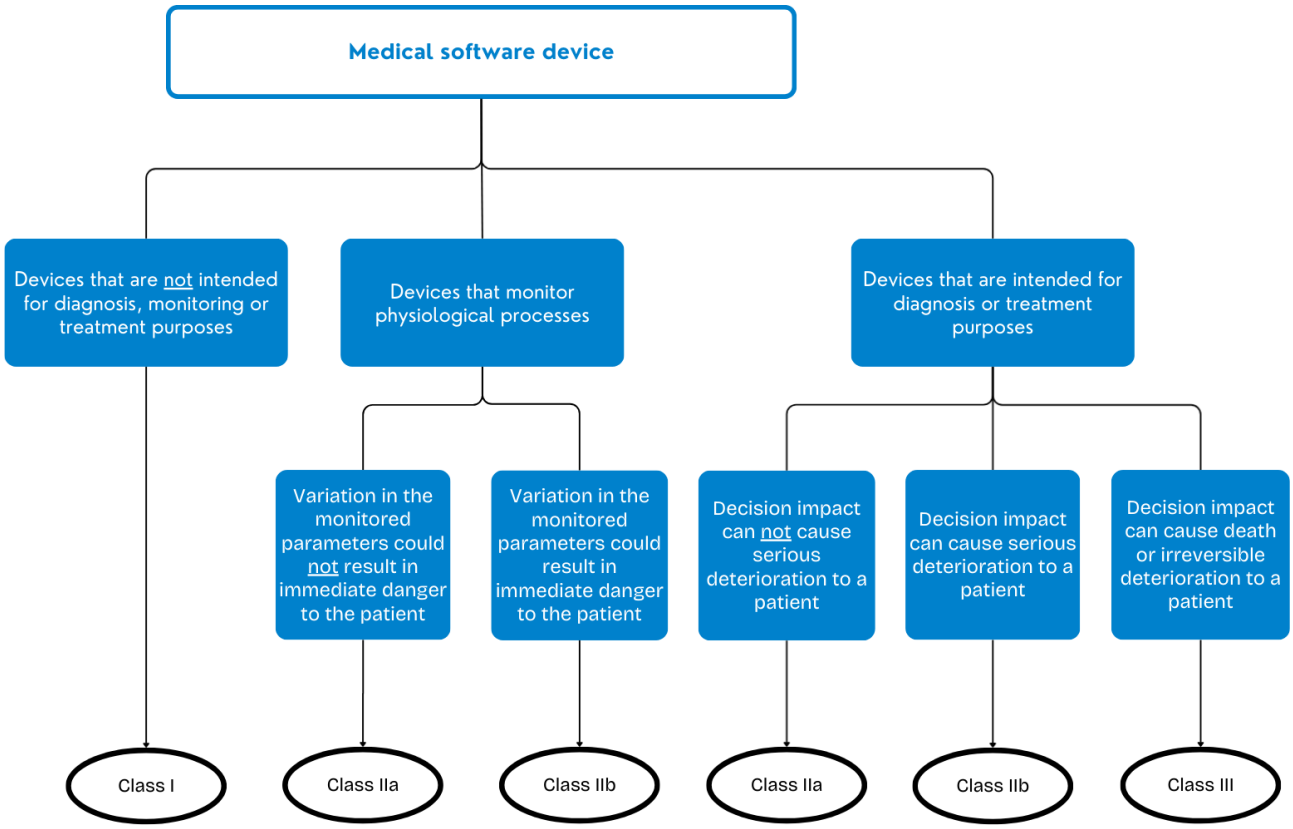
healthcare professionals will bear the responsibility for verifying the certification status of CDSSs used in practice. To aid this verification process, certified devices will be listed in the European Database on Medical Devices (EUDAMED), creating a centralized repository for medical devices used within the European Union.

Textbox 1. Highlights from the Medical Device Regulation (MDR) regulatory process

Key changes introduced by the MDR for Clinical Decision Support Systems (CDSSs):

- Risk classification: To ensure safe usage, a 4-scale classification system was added to address appropriate safety measures according to their risk. Additionally, Rule 11 of the MDR regulation states that "software intended tools to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa". As a result, CDSSs are at least classified as class IIa.
- Notified bodies: With more devices falling into higher risk classifications, including CDSSs, more devices need to be revised by notified bodies. Also, stricter requirements account for the approval of the impartial notified bodies.
- Post marketing surveillance: It is mandatory for manufacturers to actively monitor and report on the performance and safety of a device (article 10). This obligation also requires data collection post-implementation ensuring ongoing effectiveness and safety of CDSSs.
- Transparency: The MDR mandates the registration of CDSSs in the European Database on Medical Devices (EUDAMED). This is mandatory to improve insight in which tools are CE-certified by different stakeholders (regulatory authorities, manufacturers, and practitioners).

Figure 1. Medical software device classification.



Successful adoption of CDSSs in routine practice depends on overcoming regulatory and (scientific) developmental hurdles [3]. This viewpoint addresses these obstacles and identifies key improvements needed in the development of CDSSs. While the focus is on the European regulatory environment as regulatory pathways differ across regions, for example, in the United States, where the Food and Drug Administration regulates similar tools under a different framework, some principals may be relevant internationally. Our aim is to support clinical researchers and developers who are at the early stages of CDSS development by providing early insight into the scientific and regulatory considerations that will shape the trajectory of their tool development. This viewpoint synthesizes existing requirements and real-world experiences to clarify key obstacles and areas where alignment is needed. By understanding these challenges early, developers may better anticipate what is required to progress toward successful evaluation and clinical implementation.

Learning From the Past: Exploring CDSSs With Diverse Characteristics

In this viewpoint we explore three CDSSs designed to assist in the diagnosis and management of pediatric bacterial infections: the Early Onset Sepsis (EOS) Calculator [4], the Feverkidstool (FKT) [5], and the Neonatal Procalcitonin Intervention Study (NeoPInS) algorithm [6]. We described variability in the development and functionality of the CDSSs using previously

XSL•FO
RenderX

defined characteristics, including the scientific foundation, the technical design, safety considerations, and sustainability (Table 1) [7].

XSL•FO
RenderX

Table . Key characteristics of three pediatric clinical decision support systems.

Characteristics of CDSS	Feverkidstool	EOS[a] calculator	NeoPInS[b]	Regulated by the MDR[c] [1]
Usage	Helps pediatric ED[d] physicians assess the risk of serious bacterial infection and bacterial pneumonia in febrile children (>1 mo and <18 y) presenting to the emergency department	Helps doctors and nurses estimate the risk of early-onset neonatal sepsis to guide antibiotic use and clinical management in newborns born at ≥34 weeks	Helps doctors guide the management of suspected early onset neonatal sepsis by supporting decisions on safely discontinuing g antibiotics in neonates born at ≥34 weeks of gestational age	
Scientific basis				
Scientific evidence (highest support)	Stepped-wedge cluster-randomized controlled trial	Meta-analysis and parallel cluster randomized trial	Individually randomized controlled trial	No
Current evidence/validation	(international) external validation	(international) external validation	(international) external validation	No
Reproducibility	In the application a link to the algorithm is provided	In the application a link to the algorithm is provided	No direct link to the algorithm is provided	No
Technical aspect				
CDSS[e] requirement for algorithm	Tool is required for algorithm	Tool is required for algorithm	A flowchart with reference values is sufficient	No
Platform	Web-based application; accessible through smartphone, tablet, and computer, and integration in the EHR[f]	Webbased application accessible through smart-phone, tablet, and computer, open-source Shiny application, and integration in the EHR	A smart-phone application Print out the flowchart and use it on paper	Yes
Safety				
Accuracy and reliability	Technical assessment	Technical assessment	Superior and noninferior intervention study	Yes
CE-certification	In 2024 under the Medical Device Regulation as a class IIa product	The application in 2021 under the MDD as a class Ia product. The original web-application is not CE-certified[g] by itself but covered by the EHR certification	In 2021 under the Medical Device Directive as a class Ia product	Yes
CDSS's final outcome	Risk score with advice	Risk score and prescriptive guidance	Prescriptive guidance	No
Privacy	Data is stored anonymously according to the in ISO13485 and NEN7510-1 mentioned manners	Usage is anonymous, and data is not stored (apart from general storage by using a webpage). If embedded in the EHR, privacy is guaranteed by hospital security systems	Data within the software is stored by the application supplier according to the applicable laws and regulations	Yes
Sustainability				
Funding sources for development and implementation	National funding supported by ZonMW[h] for development and implementation	Private funding by Kaiser Permanente for development and implementation. EHR implementation is provided by EHR manufacturers.	Private funding by The Thrasher Foundation, The NutsOhra Foundation and the Sophia Foundation for Scientific research for development and implementation	No
Safety monitored during usage	Clinical evaluation report and a post marketing surveillance plan	Clinical evaluation RCT[i]	Clinical evaluation RCT and superior and noninferior intervention study	Yes

[a]EOS: early onset sepsis.

[b] NeoPInS: Neonatal Procalcitonin Intervention Study.

XSL•FO
RenderX

[c]MDR: Medical Device Regulation.
[d]ED: emergency department.
[e]CDSS: clinical decision support system.
[f]EHR: electronic health record.
[g]CE: Conformité Européenne.
[h]ZonMW: the Dutch National Health Council.
[i]RCT: randomized controlled trial.

Scientific Basis

Across the three CDSSs we examined, the scientific basis emerged as a central yet variably addressed component. The tools differed in the type and strength of evidence underpinning their recommendations. The EOS Calculator mainly draws from meta-analyses of observational studies [8] with a later parallel cluster-randomized trial [9], while NeoPInS and the FKT are based on an individually randomized and stepped-wedge cluster-randomized controlled trials, respectively [6,10].

All three tools included in our analysis underwent international external validation; however, validation strategies can vary considerably, and some CDSSs rely solely on internal validation [3]. Despite the centrality of the scientific foundation, the MDR does not specify requirements for the level of evidence, leaving developers to navigate this crucial aspect independently. Both clinicians and developers should be aware that the MDR procedure does not confirm a tool's scientific validity.

Transparency also varied across tools. All three publicly display their algorithms in their initial publications [4-6], and both the EOS Calculator and FKT additionally provide algorithm access via their websites, with the EOS Calculator even providing the full algorithm directly on their website. As reproducibility of a CDSS is closely linked to algorithm transparency, this variability may have implications for independent evaluation nevertheless; transparency is not regulated by the MDR.

Finally, the scope plays a significant role in the MDR classification process, as the intended use directly affects the tool's risk classification (Figure 1). While CDSSs often focus on diagnostic support, as the EOS Calculator, FKT and NeoPINS, their scope can extend beyond diagnosis as they may offer insights into prognosis, treatment, or risk assessment.

Overall, the scientific basis, while not comprehensively covered by MDR regulations, remains critical in CDSS development. Particularly, external validity remains a key consideration, requiring careful attention during development to ensure broad adoption [11]. Moreover, the strength of the scientific foundation and the chosen validation approach influence the market scan and validation processes in the regulatory process.

Technical Aspects

In our three example tools, we observed substantial variation in how the tools were technically designed and implemented. Platforms ranged from protocol- or paper-based formats to web-based applications, mobile apps, and full integration into electronic health record (EHR) systems. The EOS Calculator and FKT both rely on more complex algorithms, requiring digital platforms, whereas NeoPInS could also be delivered in a protocol or on-paper format. Ultimately, all three tools were developed into applications to support broader clinical use. These differences illustrate how algorithm complexity and intended use shape the technical form a CDSS can realistically take.

Protocol or on-paper tools, such as NeoPInS can be embedded within clinical protocols, offering continuous access at low cost. However, they face challenges in scalability and ease of use, as manual calculations are often required. The responsibility for updates and modifications lies with individual hospitals or local departments. Importantly, on-paper tools are not directly subject to the MDR certification requirements as their use does not necessitate an additional CE-mark.

Web-based applications, such as the EOS Calculator and FKT benefit from global accessibility and ease of integration in EHRs with hyperlinks. However, reliance on web-based platforms introduces vulnerabilities like downtime and security concerns. The MDR places the certification responsibility on developers, necessitating a representative for CDSSs used within the EU but developed elsewhere [1]. The complexity increases when web-based tools developed outside the EU, such as the EOS Calculator web-based version, are accessed from inside the EU.

Mobile apps like the NeoPInS (deprecated) and EOS Calculator app (currently for research purpose in the Netherlands), offer tailored user experience on devices like smartphones or tablets. Although apps provide tailored interactions, they are costly to develop and maintain, requiring agreements on intellectual property and updates.

EHR system integration, as seen with the EOS Calculator and FKT, offers user-friendliness by aligning with existing healthcare infrastructure. While cost-effective, integration across diverse EHR systems can be complex due to the lack of a universal standard. Maintenance typically falls under the purview of EHR providers, but algorithm updates may require prior agreements as they are not automatically managed by the provider.

Taken together, our comparison shows that technical design decisions, shaped by algorithm complexity, workflow integration needs, and resource constraints, directly influence reproducibility, long-term sustainability, and regulatory obligations. Because these factors determine whether a CDSS can be reliably adopted, maintained, and scaled, they must be considered early in the development process rather than treated as secondary concerns.

Safety

Our comparison shows that safety considerations are shaped not only by the underlying data flows but also by how directive

XSL•FO
RenderX

a CDSS‘s recommendation is. Although both the FKT and EOS Calculator generate risk scores, the EOS Calculator offers more prescriptive guidance compared to the FKT. The NeoPInS, without a risk score, also provides prescriptive recommendations. These functional differences influence potential clinical risk. However, although the MDR considers intended scope for classification, it does not account for recommendation type.

Compliance with privacy regulations is mandatory, as CDSSs often handle sensitive patient data. Adherence to the MDR ensures comprehensive coverage of data privacy requirements. Privacy responsibilities also differ by platform: for the EOS Calculator and FKT embedded within an EHR, hospital security systems guarantee privacy compliance, whereas for the web applications of the three tools, the developer ensures it.

Beyond privacy, safety also depends on accuracy and reliability; this involves the technical correctness of a tool’s output and its scientific validity. A reliable CDSS consistently produces accurate results, as demonstrated by the validation in our examples. Risk assessment and mitigation strategies address potential issues such as misdiagnosis or cybersecurity threats. Developers can implement security measures like dual authentication to safeguard CDSSs against cyber threats. However, user account requirements and time-consuming login processes, as seen with the NeoPInS and FKT applications, may hinder adoption and have even led to discontinuation of the NeoPInS app.

Overall, safety is paramount for CDSSs, ensuring that these tools benefit patients without causing harm. Safety is not a single requirement but an interplay of privacy protection, accuracy, reliability, and risk mitigation, all of which require continuous monitoring throughout a CDSS’s lifecycle.

Sustainability

The sustainability of a CDSS relies on its integration into daily practice, maintenance costs, and available funding. Funding sources vary, ranging from private investors to governmental organizations and research institutions. The funding party’s priorities can influence platform choice, impacting long-term sustainability. In CDSS development, project-based funding is common, posing challenges for continued financial support once projects conclude.

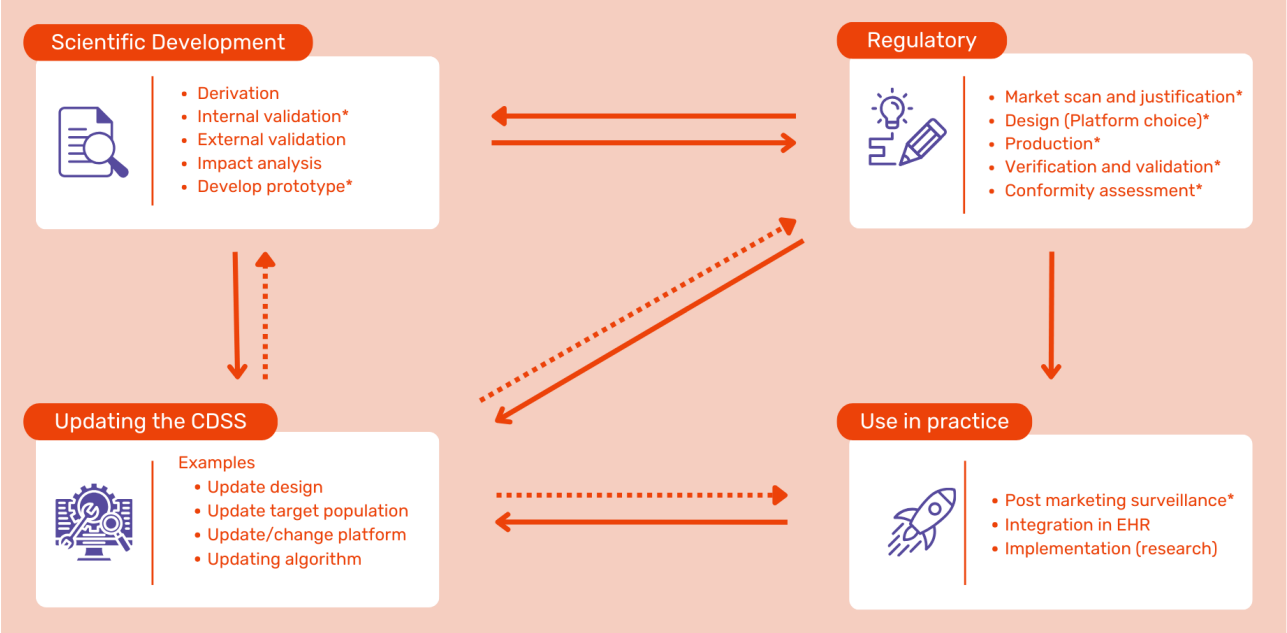
Integration into the EHR, as seen for both the FKT and EOS calculator, can enhance sustainability by facilitating routine use and reducing the need for parallel systems. At the same time, embedding a CDSS within an EHR may introduce challenges related to ownership and governance, as responsibility for the tool, its maintenance, and updates can become less transparent once it is no longer managed as a standalone application. These issues are particularly relevant in the context of MDR requirements, for example post-market surveillance, which mandates ongoing monitoring and data collection after implementation. For EHR-integrated CDSSs, it may be unclear whether such responsibilities lie with the original developer, the EHR vendor, or the health care organization, potentially complicating compliance and long-term accountability.

Leveraging the Process of Developing Scientifically Valid, Regulatory Proven, and Sustainable CDSSs for Clinical Practice

From the above, we argue that the developmental processes involved integrating CDSSs into clinical practice should not be portrayed as three consecutive phases but are dynamic and interconnected in practice (Figure 2). Also, these processes are crucial in determining how effectively a CDSS can be integrated into clinical practice, ensuring that it is both reliable and useful for healthcare professionals. This requires developers to navigate them with a strategic approach. Understanding the interplay between these processes is crucial for creating CDSSs that are scientifically sound, regulatory-compliant, practically effective, and ultimately sustainable. While the MDR addresses certain aspects of these processes (eg, internal validity, market scan and justification and post marketing surveillance) it does not comprehensively regulate every characteristic (Table 1), leaving gaps that require careful consideration by developers and clinicians. To translate the lessons learned from the three CDSSs into actionable guidance, we therefore synthesized our observations into a developer’s checklist (Textbox 2), which highlights key considerations across scientific development, technical design, safety, and sustainability.

XSL•FO
RenderX

Figure 2. Processes in the clinical decision support system development. CDSS: clinical decision support system; EHR: electronic health record. *Aspects part of the Medical Device Regulation are marked with an asterisk.



Textbox 2. Developer's checklist for a Clinical Decision Support System (CDSS).

1. Select the technical platform early and ensure it aligns with all integration and implementation requirements.
2. Engage and communicate with end-users early and throughout the development and design of the CDSS to ensure clinical workflow fits.
3. Ensure the tool's guidance matches user expectations (e.g. if it provides advice rather than full decision-making, clarify this to avoid overreliance).
4. Plan for both internal validation and external real-world testing.
5. Integrate privacy and security requirements from the earliest stage to avoid delays later.
6. Realistically estimate Conformité Européenne certification workload, expertise, and timelines.
7. Define clear ownership and a sustainable long-term maintenance plan.
8. Prepare fallback strategies to manage tool downtime safely.
9. Keep access simple (e.g. avoid unnecessary login steps).
10. Ensure full integration with existing hospital and electronic health record systems.

A critical, yet often overlooked, aspect of CDSS development is the early involvement of end-users in the scientific development process, as this can improve both tool design, effectiveness, and facilitate smoother transitions to later stages. Early strategic choices, such as selecting an appropriate technical platform, also have important downstream consequences for validity assessment and future usability. For example, although the EOS Calculator was initially developed as a web-based tool to maximize accessibility, this choice later complicated retrospective validation due to the need for manual data handling.

Validation should therefore be planned as a comprehensive, iterative process, combining internal testing with external, real-world evaluation, while accounting for the workload, expertise, and timelines required for CE certification. Beyond validation, long-term success depends on clear ownership, sustainable maintenance strategies, and predefined fallback procedures to ensure patient safety during tool downtime. Practical design considerations, such as minimizing burdensome login procedures and enabling seamless integration with existing hospital and EHR systems, further support adoption and routine use. Addressing these elements across scientific development, regulatory assessment, and clinical implementation processes enhances not only the reliability of a CDSS but also its lasting impact in clinical practice. Although all three CDSSs underwent these three key processes, their approaches differed (Table 1). As illustrated by the publication dates of manuscripts on development, the processes were timed sequentially from which important lessons were learned regarding the need for earlier and more continuous alignment.

A limitation of this viewpoint is that it focuses primarily on European MDR regulations nevertheless, similar challenges have been observed in other regions of the world. Also, currently novel AI techniques, performing the functions of CDSSs, such as dynamic risk prediction and providing targeted clinical advice, are often not yet recognized as CDSSs by clinicians. As a result, these tools are used without formal consideration of

XSL•FO
RenderX

their regulatory status, posing challenges to the current MDR regulation.

This viewpoint highlights that the complexity of CDSS development currently places a disproportionate burden on individual developers. While foresight and adaptability remain essential, developers cannot be expected to navigate scientific validation, regulatory compliance, and practical implementation alone. To support clinical scientists, our developer's checklist (Textbox 2) provides concrete guidance for successful CDSS development. Beyond these practical recommendations, structural changes are needed at the system level, for example, the creation of national protocols, standardized regulatory guidance, and dedicated support structures within hospitals. The MDR provides an important safety framework, but realizing the full potential of CDSSs requires both practical guidance for developers and systemic changes to ensure that these tools can be safely, efficiently, and effectively implemented to support clinicians and improve patient outcomes.

Acknowledgments

The manuscript text was refined with the assistance of the generative AI tool ChatGPT, version GPT-5 (OpenAI). All revisions were critically reviewed, edited, and approved by the authors.

Funding

The development of the Feverkidstool was supported by funding from the Dutch National Health Council (ZonMW), grant 170992503. Further validation of the tool was also funded by the Dutch National Health Council (ZonMW), grant 836041001. The implementation of the tool and SV were also supported by the Dutch National Health Council (ZonMW), grant 10390162210012. The Neonatal Procalcitonin Intervention Study trial received financial support from the Trasher Foundation, the NutsOhra Foundation, grant 1101-059, and the Sophia Foundation for Scientific Research, grant 681. Additionally, Thermo Fisher provided procalcitonin kits and an unrestricted grant to facilitate the organization of four investigator meetings in 2008, 2009, 2013, and 2015. SJ is supported by the Sophia Foundation for Scientific Research and Tergooi.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: NA, RO
Supervision: AW, NA, RO
Visualization: SV
Writing – Original Draft: SV, SJ
Writing – Review & Editing: SV, SJ, AW, NA, RO

Conflicts of Interest

None declared.

References

1. Regulation (EU) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. EUR-Lex (European Union). URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj [accessed 2024-11-24]
2. Keutzer L, Simonsson US. Medical device apps: an introduction to regulatory affairs for developers. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e17567. [doi: 10.2196/17567] [Medline: 32589154]
3. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res 2019;3:16. [doi: 10.1186/s41512-019-0060-y] [Medline: 31463368]
4. Puopolo KM, Draper D, Wi S, et al. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. Pediatrics 2011 Nov;128(5):e1155-e1163. [doi: 10.1542/peds.2010-3464] [Medline: 22025590]
5. Nijman RG, Vergouwe Y, Thompson M, et al. Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. BMJ 2013 Apr 2;346:f1706. [doi: 10.1136/bmj.f1706] [Medline: 23550046]
6. Stocker M, van Herk W, El Helou S, et al. Procalcitonin-guided decision making for duration of antibiotic therapy in neonates with suspected early-onset sepsis: a multicentre, randomised controlled trial (NeoPIns). Lancet 2017 Aug 26;390(10097):871-881. [doi: 10.1016/S0140-6736(17)31444-7] [Medline: 28711318]
7. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: a systematic review. PLoS One 2022;17(8):e0272846. [doi: 10.1371/journal.pone.0272846] [Medline: 35969526]

XSL•FO
RenderX

8. Achten NB, Klingenberg C, Benitz WE, et al. Association of use of the neonatal early-onset sepsis calculator with reduction in antibiotic therapy and safety: a systematic review and meta-analysis. JAMA Pediatr 2019 Nov 1;173(11):1032-1040. [doi: 10.1001/jamapediatrics.2019.2825] [Medline: 31479103]
9. van der Weijden BM, Janssen SWCM, van der Weide MC, et al. Safety and effectiveness of the early-onset sepsis calculator to reduce antibiotic exposure in at-risk newborns: a cluster-randomised controlled trial. EClinicalMedicine 2025 Sep;87:103419. [doi: 10.1016/j.eclinm.2025.103419] [Medline: 40831464]
10. van de Maat JS, Peeters D, Nieboer D, et al. Evaluation of a clinical decision rule to guide antibiotic prescription in children with suspected lower respiratory tract infection in The Netherlands: a stepped-wedge cluster randomised trial. PLoS Med 2020 Jan;17(1):e1003034. [doi: 10.1371/journal.pmed.1003034] [Medline: 32004317]
11. Abell B, Naicker S, Rodwell D, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023 Jul 26;18(1):32. [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]

Abbreviations

CDSSs: clinical decision support systems
CE: Conformité Européenne
EHR: electronic health records
EOS: Early Onset Sepsis
EUDAMED: European Database on Medical Devices
FKT: Feverkidstool
JMIR: Journal of Medical Internet Research
MDR: Medical Device Regulation
NeoPInS: Neonatal Procalcitonin Intervention Study

Edited by A Coristine; submitted 20.Feb.2025; peer-reviewed by C Duesing, G Su, SS Goswami; revised version received 29.Dec.2025; accepted 09.Jan.2026; published 09.Mar.2026.

Please cite as:
Vrijlandt SEW, Janssen SWCM, van Rossum AMC, Achten NB, Oostenbrink R
Developing Clinical Decision (Support) Systems Combining the Scientific and Regulatory Perspective: European Insights on Challenges, Requirements, and Practical Guidance
JMIR Med Inform 2026;14:e72809
URL: https://medinform.jmir.org/2026/1/e72809
doi:10.2196/72809

© Sanne E W Vrijlandt, Sanne W C M Janssen, Annemarie M C van Rossum, Niek B Achten, Rianne Oostenbrink. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telepathology Development and Stakeholder Perspectives in China: Cross-Sectional Survey

Linlin Wang[1*], MM; Fangfang Cui[1*], MM; Xianying He[1,2], MM

[1]National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, 1 Jianshe Road, Zhengzhou, Henan, China

[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Xianying He, MM
National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, 1 Jianshe Road, Zhengzhou, Henan, China

Abstract

Background: Telepathology is widely recognized for improving diagnostic access in remote areas and is strongly promoted in China to address the shortage of pathologists. However, a comprehensive national assessment of its implementation and multistakeholder perspectives remains lacking.

Objective: This study aims to conduct a comprehensive analysis of the infrastructure, practical application, and user experience of telepathology in China from 2018 to 2023 and to identify the key factors affecting its adoption and satisfaction levels.

Methods: A repeated cross-sectional survey was administered to hospitals across China (2018-2023), alongside a parallel survey of telepathology users (2023). Data were analyzed from hospitals (sample sizes ranging from 464 to 188 across years), pathologists (n=262), and patients (n=141). Regression analysis was applied to investigate factors associated with professional satisfaction and willingness to adopt the technology.

Results: The proportion of hospitals with telepathology software grew at an annual average rate of 7.60%, increasing from 29.5% (137/464) in 2018 to 42.6% (80/188) in 2023. The median number of telepathology cases per hospital increased from 51 to 200, with the fastest growth observed in the eastern region. Among pathologists, 61.1% (160/262) used telepathology weekly, 61.5% (161/262) spent 11 to 20 or 21 to 30 minutes. Pathologists reported a 95.8% (251/262) satisfaction rate, and 95% (249/262) were willing to continue participation. Convenience of system operation and service processes were associated with enhanced satisfaction and sustained engagement willingness (P<.05). Among patients, 92.2% (130/141) received diagnostic results within 24 hours, with 100% (141/141) satisfaction, and 99.3% (140/141) were willing to recommend telepathology.

Conclusions: Telepathology in China grew significantly from 2018 to 2023, with high satisfaction among both pathologists and patients. For sustained development, a multilevel approach is essential: improving system usability, implementing supportive policies (eg, funding and insurance coverage), and conducting targeted promotion of its proven benefits. This integrated strategy is key to making telepathology a sustainable solution for equitable diagnostic access nationwide.

(JMIR Med Inform 2026;14:e83514) doi:10.2196/83514

KEYWORDS

telepathology; digital slide scanners; software; service; usage; satisfaction; perceptions

Introduction

Pathological diagnosis is the gold standard for disease diagnosis, particularly for cancer. However, China's primary medical institutions face a significant shortage of pathologists, especially experienced ones. According to the 2019 China Pathology Quality Report [1], there are only 0.55 pathologists per 100 hospital beds, and 46.3% of secondary hospitals have just 1 pathologist, which is far below the demand for timely and high-quality pathology services. Data from the Pathology Quality Control Center of China [2] reveal that 82% of senior pathologists are concentrated in tertiary hospitals, leaving many primary hospitals without senior-title pathologists. In central and western regions, such as Tibet, Xinjiang, Qinghai, and Guizhou, more than 20% of institutions have no more than 1 pathologist, highlighting a significant imbalance in pathology resource distribution across China.

Telepathology, first proposed in 1987 [3], addresses resource shortages by enabling the transmission of static images, dynamic videos, and whole-slide images (WSIs). As defined in the Basic Functions Specification of Telemedicine Information System (WS/T529-2016), telepathology involves the submission of

XSL•FO
RenderX

diagnostic requests and patient data by an inviting party to an invited expert, who then provides diagnostic opinions and reports. Central to this practice is WSI technology, which converts pathological slides into high-resolution, zoomable digital images and supports the diagnosis of all pathology-dependent diseases, including tumor classification and grading, as well as the identification of specific pathogens or special inflammatory lesions. Its diagnostic reliability is supported by multiple guidelines and studies. The College of American Pathologists (CAP) released relevant validation guidelines in 2013 and 2021 [4,5], and the Royal College of Pathologists issued "Best Practice Recommendations for Implementing Digital Pathology" [6]. Studies by Montgomery et al [7] and Shinohara et al [8] confirmed the high diagnostic accuracy and effectiveness of WSIs in remote diagnosis. Furthermore, Girolami et al's [9] meta-analysis validated the diagnostic concordance between WSIs and light microscopy, reinforcing its role in delivering high-quality diagnostic services to resource-limited regions. At the regulatory and application levels, pathological diagnosis based on WSIs has been approved by the Food and Drug Administrations of the United States, the European Union, and Japan [10-12]. Since the US Food and Drug Administration first authorized a WSI system for primary diagnosis in 2017, multiple integrated digital pathology platforms, combining scanning, storage, viewing, and analysis functions, have received 510(k) clearance [13]. As of December 31, 2025, China's National Medical Products Administration (NMPA) had also approved 71 digital pathology–related medical devices [14]. In practice, WSI-based telepathology has been widely adopted in numerous countries and regions, including Europe [15], Nigeria [16], India [17], Tanzania [18], and Canada [19].

In China, developing a WSI-based telepathology system is a core strategy to address the deficiencies in grassroots pathology services [20]. Preliminary studies have demonstrated the effectiveness of the national cloud-based telepathology platform [21] and its potential to optimize resource allocation [22]. However, the existing literature lacks a comprehensive, long-term, and systematic evaluation of the overall development of telepathology in China. Key questions remain unanswered: What are the national trends and regional disparities in its adoption? As primary users, what are the adoption behaviors, perceptions, and influencing factors among pathologists? Moreover, what are patients' experiences and satisfaction levels with telepathology services? The absence of such critical information limits a thorough understanding of current implementation barriers and hinders targeted interventions.

To address these gaps, this study utilizes nationwide survey data from 2018 to 2023 with the following objectives: (1) to delineate the developmental trends and spatial distribution characteristics of telepathology in China over the past 6 years; (2) to analyze, from the perspective of pathologists, their usage behaviors, attitudes, and key influencing factors; and (3) to conduct the first systematic assessment of patient satisfaction with telepathology services. By integrating multidimensional evidence, including macro-level trends, insights from pathologists, and patient experiences, this study aims to provide an empirical foundation and decision-making support for optimizing China's telepathology infrastructure and promoting the equitable distribution of high-quality pathology resources.

Methods

Study Setting

To understand the current status of and identify obstacles to telemedicine services (telemedicine service items include teleconsultation, telepathology, teleradiology, telemonitoring, tele-outpatient services, etc), the Telemedicine Informationization Professional Committee of China (TIPC) and the National Telemedicine Center of China (NTCC) have conducted web-based surveys on telemedicine since 2018.

Survey on the Development of Telemedicine in Hospitals

This survey targeted hospitals approved by Chinese health administrative departments. The questionnaire was developed by experts with more than 5 years of telemedicine experience and refined annually based on prior findings. Following the initial design, 10 to 18 hospitals were selected for pilot testing, and feedback was incorporated to finalize the questionnaire.

Survey on the Usage and Perceptions of Telemedicine

A survey was conducted among doctors and patients participating in telemedicine services. The questionnaires were designed through needs analysis, a review of telemedicine practices, and relevant literature. After drafting, a pilot test was administered at the Henan Provincial Telemedicine Center and selected subcenters. The instruments were further optimized based on pilot feedback to form the final versions.

Implementation of the Survey

The electronic open survey was an online questionnaire on NTCC's website [23]. The survey announcement was formally issued by TIPC via an official red-headed document. Each annual theme had its own webpage, with up to 42 mandatory questions (the exact number depended on the respondent's answers) across 18 to 25 pages in Chinese. Completion required approximately 8 to 12 minutes based on pilot testing. The surveys were open from July to September 2018, October to November 2019, September to October 2020, and September to October 2023. To ensure data quality, duplicate submission prevention was implemented through IP address verification, response-time logic control, and cross-validation with official hospital records. Response metrics for each questionnaire type were as follows:

- The hospital questionnaire recorded 896 views with a 63.3% (567/896) participation rate and an 81.8% (464/567) completion rate in 2018
- The corresponding figures for 2019, 2020, and 2023 were 482 views (60.8% participation, 82.6% completion), 401 views (61.1% participation, 83.3% completion), and 364 views (63.2% participation, 81.7% completion).
- The doctor questionnaire received 3269 views, achieving a 53.1% (1736/3269) participation rate and a 77.9% (1353/1736) completion rate, including 262 pathologists among the completers.

XSL•FO
RenderX

- The patient questionnaire garnered 3528 views, with a 42.6% (1502/3528) participation rate and a 59.3% (891/1502) completion rate, of whom 141 patients had used telepathology services.

Data Cleaning and Valid Questionnaire

After the collection deadline, data were exported to Excel for cleaning. Hospital questionnaires from identical or conflicting IP addresses that could not be verified against official hospital records (eg, outpatient visits, bed counts) were excluded. Additionally, hospital questionnaires completed in less than 2 minutes and multiple doctor or patient questionnaires submitted from the same IP address within a 3-minute interval were omitted. Valid hospital questionnaires totaled 464 in 2018, 242 in 2019, 204 in 2020, and 188 in 2023.

Study Design

To explore the current status of telepathology development and application in Chinese hospitals, this study used data from TIPC and NTCC. Surveys were conducted in 2018, 2019, 2020, and 2023 across eastern, central, and western China. Due to the COVID-19 pandemic, restrictions on personnel movement during the outbreak disrupted normal work routines and medical services, resulting in the cancellation of surveys that were planned for 2021 and 2022. The eastern, central, and western regions are classified according to standards set by the National Bureau of Statistics of China (Multimedia Appendix 1). The survey coverage area is presented in Multimedia Appendix 2. Pearson chi-square test indicated no significant difference in provincial distribution among these regions across the 4 years (P=.90). Telepathology-related data were selected to assess construction progress and development trends. Based on 2023 survey responses from pathologists and patients who participated in telepathology services, application effectiveness was further analyzed.

This study was reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines (Checklist 1) [24]. The hospital questionnaire included 11 repeated survey questions on telepathology, comprising 9 single-choice questions and 2 fill-in questions. Among them, 6 questions related to telepathology infrastructure and service volume, and 5 questions related to telepathology operations and billing. The questionnaire was completed by telemedicine organizers at each hospital. The doctor questionnaire consisted of 24 questions, including basic information (7 questions), telepathology service usage and experience (9 questions), evaluation and attitudes (6 questions), and open-ended questions regarding problems and expectations (2 questions). The patient questionnaire consisted of 19 questions, including basic information (7 questions), usage (7 questions), and evaluation and attitudes (5 questions) (Multimedia Appendix 3).

Data Analysis

Data cleaning was performed to avoid logical problems and ensure data quality. Descriptive statistics were used to analyze the data. Median values and IQR are reported for quantitative data, and percentages are reported for qualitative data. Chi-square tests and Fisher exact probability tests were used to assess differences. Variables selected through univariate analysis were further analyzed using multivariate ordinal logistic regression to identify factors influencing satisfaction and usage intention [25]. All statistical analyses were conducted using SPSS (version 20.0; IBM Corp). All tests were 2-tailed, and P<.05 was considered statistically significant. Chinese maps were generated using ArcMap (version 10.8.1; Esri Corp) and R language (version 4.4.1; R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the institutional review board at the First Affiliated Hospital of Zhengzhou University (protocol 2021-KS-HNSR181). Ethics approval was obtained with the stipulation that collected data must not expose respondents' identities and only aggregated information would be addressed in the research paper. Informed consent was obtained from all participants. The survey was conducted without compensation. The questionnaire was anonymous, and the data file was stored on the NTCC intranet server. Only research team members who signed confidentiality agreements were permitted to access the data.

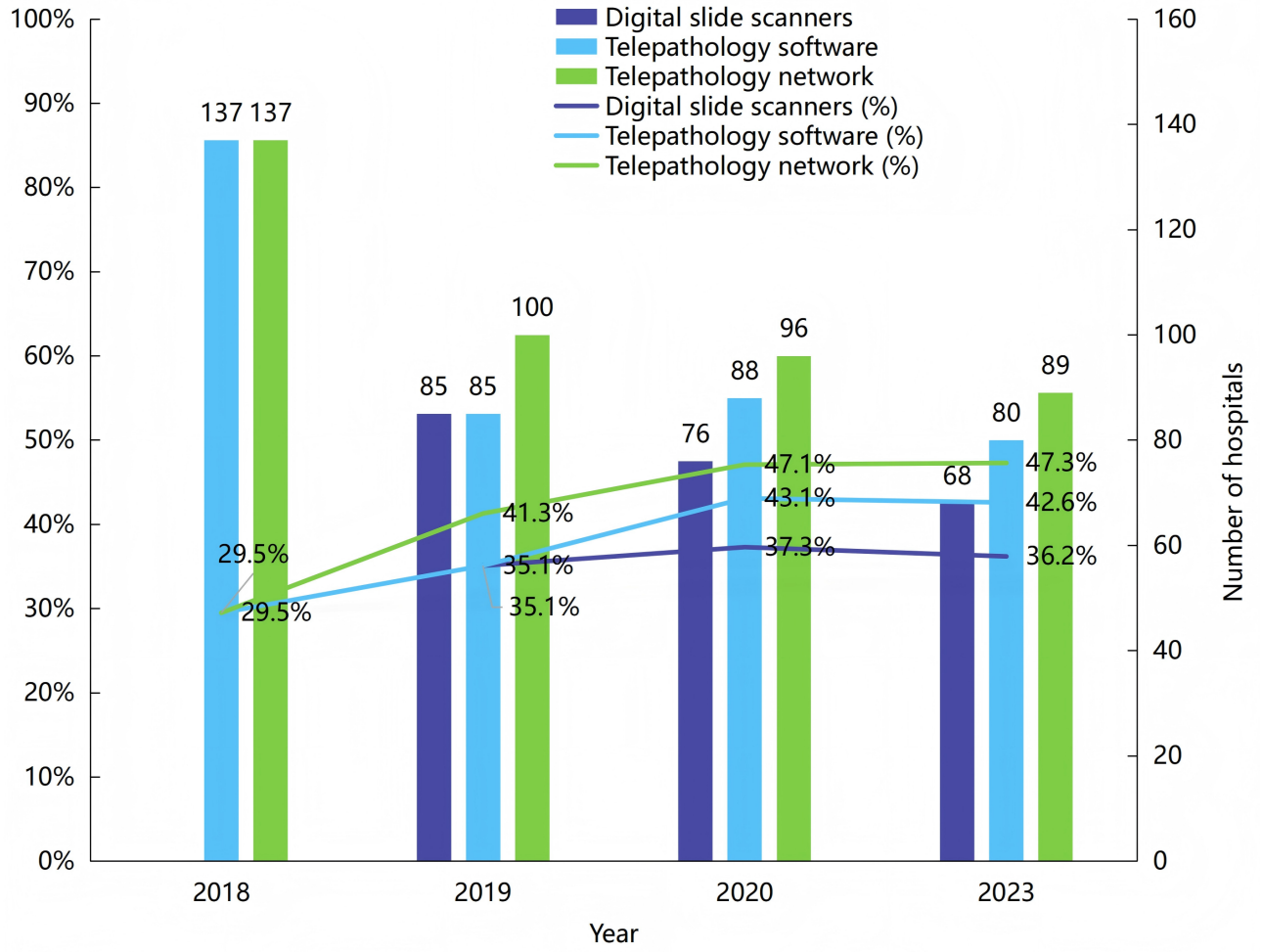
Results

The Construction of Telepathology

Digital slide scanners, software, and networks were investigated, as shown in Figure 1. The configuration status of digital slide scanners was not investigated in 2018. In 2019, 2020, and 2023, the number of hospitals equipped with scanners were 85 of 242 (35.1%), 76 of 204 (37.3%), and 68 of 188 (36.2%), respectively, with an annual average growth rate of 0.75%. From 2018 to 2023, the percentages of hospitals deploying telepathology software were 29.5% (137/464), 35.1% (85/242), 43.1% (88/204), and 42.6% (80/188), respectively, with an annual average growth rate of 7.6%. The telepathology service network comprises private networks and the internet. In 2018, 2019, 2020, and 2023, there were 137 (137/464, 29.5%), 100 (100/242, 41.3%), 96 (96/204, 47.1%), and 89 (89/188, 47.3%) hospitals connected to the telepathology service network, respectively. Private networks were consistently more utilized than the internet, with annual ratios of 68.6% (94/137) vs 31.4% (43/137), 74% (74/100) vs 26% (26/100), 67% (64/96) vs 33% (32/96), and 93% (83/89) vs 7% (6/89).

XSL•FO
RenderX

Figure 1. In 2018 - 2020 and 2023, the construction of telepathology infrastructure.



The Current Status of Telepathology Services

In 2018, 2019, 2020, and 2023, the proportions of hospitals providing telepathology services were 23.3% (108/464), 28.9% (70/242), 33.3% (68/204), and 31.4% (59/188), respectively, with an annual average growth rate of 6.1%. Among hospitals with established telepathology systems, the proportions providing telepathology services were 78.8% (108/137) in 2018, 82.3% (70/85) in 2019, 77.3% (68/88) in 2020, and 73.8% (59/80) in 2023. The median numbers of telepathology cases for these years were 51 (IQR 10 - 200), 95 (IQR 18.25 - 228.5), 100 (IQR 27.5 - 223.75), and 200 (IQR 93 - 331.5), respectively. A statistically significant increase was observed from 2018 to 2023, with an annual average growth rate of 31.4%. Table 1 details the regional and hospital-level distribution. From 2018 to 2020, the Central region led in telepathology services; however, by 2023, the Eastern region took over. Significant differences were observed in telepathology services between the Central and Western regions across the years, with annual average growth rates of 17.7% and 27.2%, respectively. Tertiary hospitals consistently reported higher median telepathology case volumes than secondary hospitals, with a statistically significant annual average growth rate of 26.2%.

XSL•FO
RenderX

Table . Distribution of telepathology services in different regions and different levels of hospitals.

Characteristics	Telepathology service volume per year, median (IQR)				*H*	*P* value
	2018	2019	2020	2023		
District						
East	36.5 (10-105)	80 (27.75-215)	31 (11.25-120.75)	359 (164.75-676.25)	4.55	.21
Center	93 (30-226)	141 (60-287)	148 (47-258)	210 (112.5-331.5)	10.11	.02
West	24 (2-180)	13 (2-123.5)	124 (60-776)	80 (36.75-146.75)	11.32	.01
Hospital grade						
Third-level	64 (5-292.25)	124 (60.5-245)	120 (25-274.5)	205 (109.75-420.25)	14.15	<.001
Second-level	50 (14.5-165)	55 (6.5-213)	80 (30-157)	100 (32.5-218)	2.15	.54
Total	51 (10-200)	95 (18.25-228.5)	100 (27.5-223.75)	200 (93-331.5)	17.61	.001

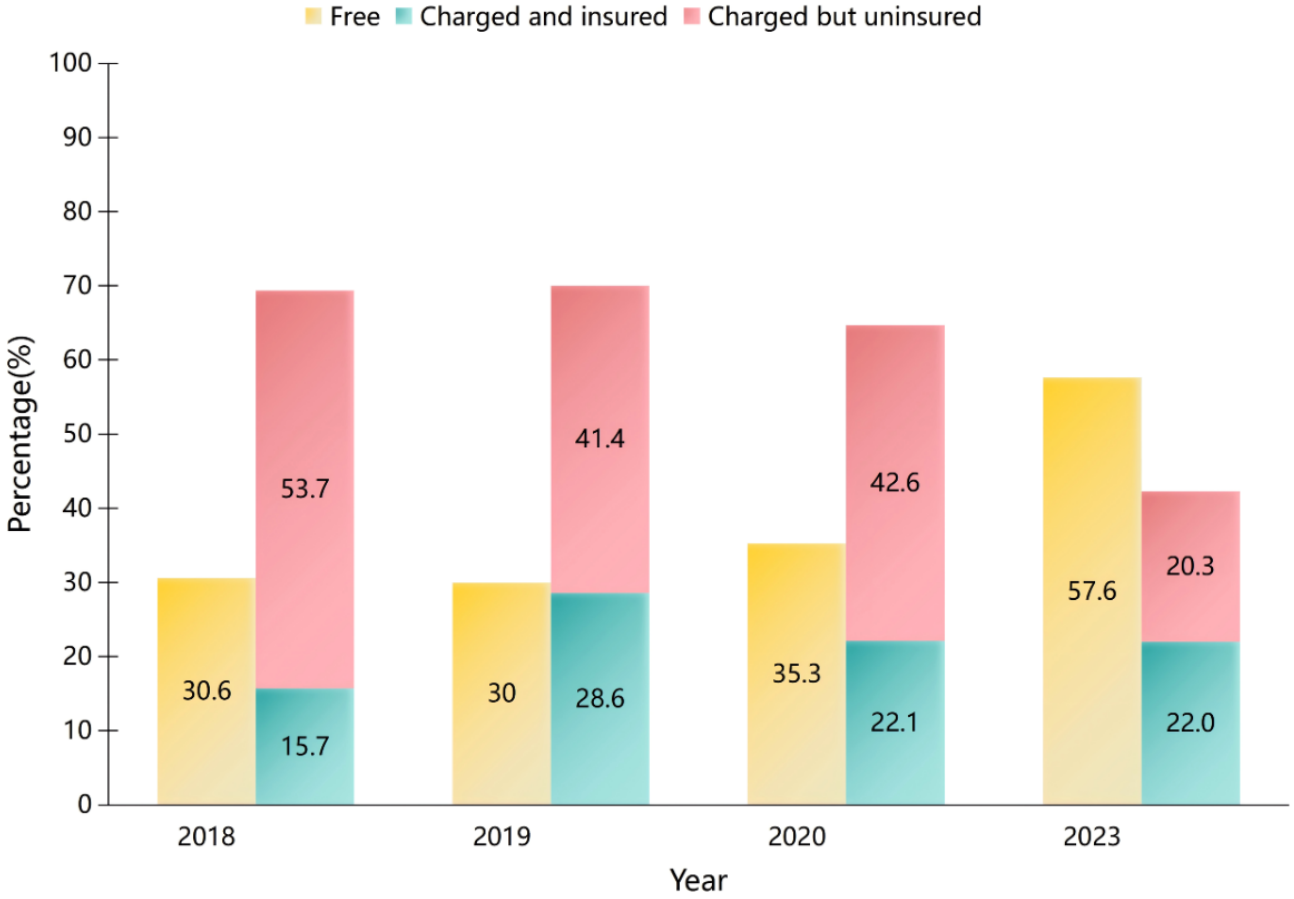
The Operation and Charging Situation of Telemedicine in Hospitals That Carry Out Telepathology Services

Among the hospitals providing telepathology services in 2018, 2019, 2020, and 2023, the percentages with established telemedicine departments were 69.4% (75/108), 77.1% (54/70), 76.5% (52/68), and 72.9% (43/59), respectively. The median numbers of telemedicine staff allocated were 3 (IQR 2 - 5) in 2018 and 2019, 4.5 (IQR 2 - 7) in 2020, and 5 (IQR 2.5 - 7) in 2023. Telemedicine operation modes included hospital self-operation, third-party operation, and mixed operation. Surveys over the 4 years showed that hospital self-operation was the primary mode, accounting for 67.59%, 50%, 67.65%, and 72.88%, respectively (Table S1 in Multimedia Appendix 4).

In 2018, 2019, 2020, and 2023, among hospitals providing telepathology services, the percentages of free services were 30.6% (33/108), 30% (21/70), 35.3% (24/68), and 57.6% (34/59), respectively. For charged services reimbursed by medical insurance, the percentages were 15.7% (17/108), 28.6% (20/70), 22.1% (15/68), and 22% (13/59), respectively (Figure 2). Fee standards were based on 3 factors: administrative department guidelines, hospital-specific standards, and telemedical service provider criteria (Table S1 in Multimedia Appendix 4).

XSL•FO
RenderX

Figure 2. Proportions of hospitals providing telepathology services (free, charged and insured, charged but uninsured) in 2018 - 2020 and 2023



Usage of Telepathology Among Pathologists

A total of 262 valid questionnaires were collected from pathologists, as shown in Table 2. The distribution of these pathologists by region was 69 from the eastern, 61 from the central, and 132 from the western regions. There were 129 (49.2%) males and 133 (50.8%) females with a mean age of 35.3 (SD 9.1, range 20-73) years. The dominant frequency was once a week, accounting for 61.1% (n=160). Combined, those who serviced once a week or 2 to 3 times a week made up 84.4% (n=221) of the total, whereas those servicing 4 to 5 times a week or more accounted for 15.6% (n=41). The average time spent on telepathology services was mainly 11 to 20 minutes (n=78, 29.8%) and 21 to 30 minutes (n=83, 31.7%).

XSL•FO
RenderX

Table . Demographic characteristics and telepathology usage among pathologists.

Characteristics	Participants, n (%)
District	
East	69 (26.3)
Center	61 (23.3)
West	132 (50.4)
Gender	
Male	129 (49.2)
Female	133 (50.8)
Age, y	
<30	76 (35.1)
30-39	106 (39.6)
40-49	60 (20.1)
≥50	20 (5.2)
Education	
Doctoral degree	13 (4.9)
Master's degree	35 (13.3)
Bachelor's degree	149 (56.8)
College diploma and below	65 (24.8)
Professional title	
Senior	23 (8.7)
Deputy senior	30 (11.4)
Intermediate	71 (27)
Primary and below	138 (52.6)
Frequency of telepathology use, time/wk	
1	160 (61.1)
2-3	61 (23.3)
4-5	23 (8.8)
6-7	3 (1.1)
≥8	15 (5.7)
Average duration of telepathology, min	
≤10	44 (16.8)
11-20	78 (29.8)
21-30	83 (31.7)
31-60	42 (16)
>60	15 (5.7)

Pathologists' Evaluation and Satisfaction With Telepathology

As shown in Table 3, 92.7% (n=243) of pathologists considered the system easy to operate, and the service process was convenient. Additionally, 95.5% (n=250) believed that telepathology could reduce patient burdens. The overall satisfaction rate with telepathology services was 95.8% (n=251), and 95% (n=249) of pathologists were willing to continue participation.

XSL•FO
RenderX

Table . The evaluation and satisfaction of telepathology.

Characteristics	Participants, n (%)
Convenience of system operation	
Very convenient	168 (64.1)
Convenient	75 (28.6)
Neutral	8 (3.1)
Inconvenient	9 (3.4)
Very inconvenient	2 (0.8)
Convenience of service process	
Very convenient	149 (56.9)
Convenient	94 (35.9)
Neutral	8 (3.1)
Inconvenient	10 (3.8)
Very inconvenient	1 (0.4)
The role of telepathology in reducing the medical burden of patients	
Very helpful	166 (63.4)
Helpful	84 (32.1)
Neutral	5 (1.9)
Unhelpful	4 (1.5)
Very unhelpful	3 (1.1)
Using intention of telepathology	
Very willing	153 (58.4)
Willing	96 (36.6)
Neutral	10 (3.8)
Unwilling	2 (0.8)
Very unwilling	1 (0.4)
Overall satisfaction	
Very satisfied	147 (56.1)
Satisfied	104 (39.7)
Neutral	6 (2.3)
Dissatisfied	4 (1.5)
Very dissatisfied	1 (0.4)

Factors Influencing Pathologists' Satisfaction and Use Intention of Telepathology

According to the univariate analyses shown in Tables 2 and 3, 6 variables were associated with overall satisfaction (P<.25), and 5 variables were associated with use intention (P<.25) (Table 4). Multivariate ordinal logistic regression analysis revealed that 3 variables were included in the models for pathologists' overall satisfaction and use intention, respectively. Both models met the proportional odds (parallel lines) assumption, with test results showing P=.83 for the satisfaction model and P=.44 for the usage intention model. Convenience of system operation and service processes significantly improved both overall satisfaction and willingness to use telepathology. Compared with an average service duration of ≤10 minutes per session, longer service times were associated with improved willingness to continue using the service. In addition, a moderate frequency of participation enhanced pathologists' overall satisfaction.

XSL•FO
RenderX

Table . Ordinal logistic regression results of the influencing factors of the attitude toward telepathology among pathologists.

Variables	Overall satisfaction		Using intention	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
Threshold (ref=very satisfied/very willing)				
Very dissatisfied/very unwilling	—[b]	<.001	—	<.001
Dissatisfied/unwilling	—	<.001	—	<.001
Neutral	—	<.001	—	<.001
Satisfied/willing	—	.13	—	.19
Education (ref=doctoral degree)				
College diploma and below	2.53 (0.52-12.39)	.25	1.92 (0.41-8.91)	.41
Bachelor's degree	1.26 (0.30-5.30)	.76	0.87 (0.21-3.53)	.85
Master's degree	0.50 (0.11-2.28)	.37	0.36 (0.08-1.57)	.17
Professional title (ref=senior)				
Primary and below	1.26 (0.40-3.99)	.70	0.54 (0.17-1.70)	.29
Intermediate	1.48 (0.46-3.81)	.51	1 (0.31-3.23)	>.99
Deputy senior	0.89 (0.25-3.21)	.86	0.67 (0.19-2.39)	.54
Frequency of telepathology use (ref=≥8 time/wk)			—	—
1 time/week	0.26(0.06-1.13)	.07		
2-3 times/week	0.21 (0.04-0.96)	.045		
4-5 times/week	0.36 (0.06-1.98)	.24		
6-7 times/week	0.15 (0.01-2.56)	.19		
Average duration of telepathology (ref=≤10 min)				
≥61 min	4.03 (0.93-17.51)	.06	6.68 (1.48-30.23)	.01
31-60 min	1.98 (0.74-5.34)	.18	1.92 (0.75-4.94)	.18
21-30 min	1.76 (0.75-4.12)	.19	2.49 (1.09-5.67)	.03
11-20 min	1.78 (0.75-4.21)	.19	2.30 (1.01-5.26)	.049
Convenience of system operation (ref=Very convenient)				
Very inconvenient	6.21 (0.23-169.52)	.28	2.76 (0.12-64.33)	.53
Inconvenient	0.76 (0.08-7.50)	.81	0.70 (0.08-5.82)	.74
Neutral	0.06 (0.01-0.50)	.01	0.02 (0.00-0.18)	<.001
Convenient	0.44 (0.20-0.94)	.03	0.49 (0.23-1.04)	.06
Convenience of service process (ref=Very convenient)				
Very inconvenient	0.05 (0.00-7.92)	.25	0.08 (0.00-9.46)	.30
Inconvenient	0.07 (0.01-0.54)	.01	0.20 (0.03-1.32)	.09
Neutral	0.08 (0.01-0.70)	.02	0.25 (0.04-1.77)	.17
Convenient	0.22 (0.10-0.46)	<.001	0.29 (0.14-0.59)	.001

[a]OR: odds ratio.

[b]Not applicable.

Patients' Perceptions of Telepathology Service Quality

A total of 141 patients who had used telepathology services participated in the survey, as shown in Table 5. The distribution across eastern, central, and western provinces was 31.9% (n=45), 32.6% (n=46), and 35.5% (n=50), respectively. There were 64 males and 77 females, with a male-to-female ratio of 0.83:1. Ages ranged from 19 to 72 years, with a mean age of 33.5 (SD 10.5) years. Most patients held a bachelor's degree (n=59, 41.8%), followed by an associate degree (n=50, 35.5%). Urban residents accounted for 61% (n=86) of patients. Employee medical insurance was the most common type of coverage

XSL•FO
RenderX

(n=81, 57.5%), followed by the New Rural Cooperative Medical Insurance (n=38, 27%). Notably, 76.6% (n=108) of patients reported an average annual family income below 30,000 yuan (US $4335), which is lower than the national per capita disposable income of 39,218 yuan (US $5667) reported in the 2024 China Statistical Yearbook.

Table . Characteristics of the patients participating in telepathology.

Characteristics	Participants, n (%)
District	
East	45 (31.9)
Center	46 (32.6)
West	50 (35.5)
Age (y)	
<30	61 (43.3)
30-39	46 (32.6)
40-49	23 (16.3)
≥50	11 (7.8)
Gender	
Male	64 (45.4)
Female	77 (54.6)
Education	
Doctoral degree	2 (1.4)
Master’s degree	6 (4.3)
Bachelor’s degree	59 (41.8)
Associate degree	50 (35.5)
High school and below	21 (14.9)
Other	3 (2.1)
Average annual family income (¥)[a]	
<5000	42 (29.8)
5000-9999	26 (18.4)
10,000-14,999	17 (12.1)
15,000-19,999	7 (5)
20,000-24,999	9 (6.4)
25,000-29,999	7 (5)
30,000-49,999	8 (5.7)
≥50,000	25 (17.7)
Location	
Urban	86 (61)
Rural	55 (39)
Insurance	
Employee medical insurance	81 (57.5)
Medical insurance for urban residents	16 (11.4)
New rural cooperative medical insurance	38 (27)
Business insurance	2 (1.4)
No medical insurance	1 (0.7)
Other	3 (2.1)

[a]Exchange rate at the time of the study: US $1=¥6.92.

XSL•FO
RenderX

A total of 99.3% (140/141) of patients reported that the doctor thoroughly explained telepathology, and 96.5% (136/141) had signed the informed consent form, as shown in Table 6. All patients received diagnostic opinions from senior experts. Within 24 hours, 92.2% (130/141) received results; within 48 hours, 98.6% (139/141) received results. Only 1 patient (1/141, 0.7%) received results within 72 hours. Overall, 98.6% (139/141) of patients were satisfied with the waiting time. All 141 patients found telepathology convenient and beneficial for their condition. Regarding service cost, 63.1% (89/141) were charged, with 52 of 89 (58.4%) considering the charges very reasonable, 36 of 89 (40.5%) reasonable, and only 1 of 89 (1.1%) unreasonable.

XSL•FO
RenderX

Table . Telepathology service quality and patient satisfaction (N=141).

Characteristics	Participants, n (%)
Did doctor introduce telepathology in detail?	
Very detailed	97 (68.8)
Detailed	43 (30.5)
Not detailed	1 (0.7)
Did you sign the informed consent form?	
Yes	136 (96.5)
No	5 (3.5)
Convenience of telepathology service	
Very convenient	96 (68.1)
Convenient	45 (31.9)
Inconvenient	0 (0)
Have you received the expert diagnostic opinion?	
Yes	141 (100)
No	0 (0)
Waiting time (h)	
≤12	101 (71.6)
12.1-24	29 (20.6)
24.1-48	9 (6.4)
48.1-72	1 (0.7)
>72	1 (0.7)
Are you satisfied with the waiting time?	
Very satisfied	88 (62.4)
Satisfied	51 (36.2)
Dissatisfied	2 (1.4)
Is telepathology helpful for diseases?	
Very helpful	100 (70.9)
Helpful	41 (29.1)
Unhelpful	0 (0)
Is there a fee for this service?	
Yes	89 (63.1)
No	52 (36.9)
Is the fee reasonable? (n=89)	
Very reasonable	52 (58.4)
Reasonable	36 (40.5)
Unreasonable	1 (1.1)
Has medical expenses been saved?	
Yes	134 (95)
No	7 (5)
Willingness to recommend to others?	
Very willing	93 (66)
Willing	47 (33.3)
Unwilling	0 (0)

XSL•FO
RenderX

Characteristics	Participants, n (%)
Very unwilling	1 (0.7)
Overall satisfaction	
Very satisfied	91 (64.5)
Satisfied	50 (35.5)
Dissatisfied	0 (0)

Patient Satisfaction With Telepathology Services

The overall satisfaction rate was 100% (141/141), with 64.5% (91/141) reporting being “very satisfied”. Among patients, 95% (134/141) believed that telepathology could save medical expenses, and 99.3% (140/141) were willing to recommend it to others. The main reasons for recommendation included convenience (91/140, 65%), access to quality medical services (22/140, 15.7%), cost saving (11/140, 7.9%), time saving (6/140, 4.3%), and necessity (1/140, 0.7%).

Discussion

Principal Results

Diagnosis consistency between WSI and light microscopy has been confirmed, and telepathology is increasingly adopted worldwide [26]. It is widely used for remote primary diagnosis, second-opinion consultations, quality assurance, education, and research [27]. In China, the government has actively promoted telepathology to address the shortage of pathologists in primary health care institutions. However, although most medical institutions in China have implemented telemedicine systems, the utilization rate has remained below 20%, as highlighted at the 2019 subforum of the Global Health Forum of Boao Forum for Asia: Innovation for Health—The Future of Internet Medicine [28]. The actual use of telepathology in Chinese hospitals remains systematically understudied. Although annual surveys by TIPC and NTCC provide cross-sectional snapshots of telemedicine development, they are largely descriptive and lack dedicated analysis of telepathology or longitudinal assessment. To address this gap, this study provides the first systematic examination of telepathology development in China through long-term trend analysis, regional comparisons, and dual perspectives from both clinicians and patients. Based on these findings, targeted recommendations are proposed to inform policy-making and support the sustainable advancement of telepathology in China.

Regarding telepathology infrastructure construction, the highest proportion is networks, while the lowest is digital slide scanners. This is because invited experts only need to provide diagnostic suggestions based on WSI uploaded by the inviting institutions, eliminating the need for slide scanners. From 2018 to 2023, the number of telepathology service cases in China grew steadily, with the median per hospital increasing from 51 to 200 cases, reflecting an average annual growth rate of 31.4%. However, surveys revealed that the proportion of hospitals with established telepathology systems exceeded the proportion of those actually providing services each year. The ratio of service implementation to system construction ranged from 73.8% (59/80) to 82.3% (70/85), a trend consistent with the reported service utilization rates of 72.3% (60/83) to 89.1% (74/83) in the China CNCTPS telepathology platform report [21]. Hitchcock [29] pointed out that the limited implementation of telepathology in developing countries primarily stems from a lack of standardized equipment, inconsistent image formats, and insufficient training for pathologists. Practices in Nigeria [16] further highlighted the significant time cost associated with image scanning and uploading (5 - 10 min per slide for scanning and 10 - 30 min for uploading to the cloud), which has become a critical bottleneck affecting service efficiency. Research by Indian scholars [17] also indicated that challenges such as the large file sizes of WSI, insufficient transmission speeds, and the need for security protection of patient data and images during internet transmission are key factors limiting the application of telepathology. In response to these challenges, various technical approaches have been proposed in the academic community. Emerging technologies such as IoT [30], blockchain-based digital identity authentication [31], and aspect-oriented programming [32] have been proven effective in reducing data transmission latency, enhancing medical information security, and improving system interoperability. These technologies offer feasible solutions for addressing the challenges of cross-institutional secure retrieval of large WSI files. Therefore, promoting the integrated application of these technologies in telepathology holds significant potential for increasing service utilization rates in the future. Moreover, policy and standardization improvements are crucial. Health authorities should promote WSI standardization, clarify cross-institutional licensing rules, and include telepathology in insurance reimbursement.

From a regional distribution perspective, the development of telepathology varies across different areas. Similarly, a national survey in Ireland [33] indicated that the advancement of telemedicine differs significantly across regions, with multiple regional networks established in the east, while development in the west predominantly consisted of isolated links. Our study reveals that the eastern region experienced the fastest growth, with an average annual growth rate of 58%, likely attributable to its more abundant medical resources, stronger economic foundation, and the advantages of policy piloting and early implementation. From 2018 to 2020, the central region reported the highest median number of telepathology diagnostic cases, which may be associated with its lower urbanization rate and relatively scarce rural healthcare resources. According to the China Statistical Yearbook, the urbanization rate in the eastern region was 68.5%, compared with 56.8% in the central region and 54.1% in the western region. Additionally, data from the China Health Statistical Yearbook indicate that from 2018 to

XSL•FO
RenderX

2020, the central region consistently had the lowest number of health technicians per thousand rural population. These factors likely contributed to a higher demand in the central region. To address regional disparities in China's telepathology services, a systematic policy intervention is essential. This includes: (1) Investment: establishing a dedicated fund for central/western regions and integrating telepathology into medical insurance; (2) Operations: implementing unified technical standards and regional diagnostic centers with performance incentives; (3) Capacity: training local pathology technicians and physicians.

The dominant frequency of telepathology usage among pathologists in China was once a week, accounting for 61.1% (160/262). In contrast, a national survey in the United Kingdom revealed that telepathology was rarely used on a weekly or monthly basis [34]. Meanwhile, surveys in South Africa showed that 11% (8/71) used telepathology weekly [35], and a study in Switzerland reported a 31% (42/134) weekly usage rate for telepathology [36]. The average time spent on telepathology services is mainly 11 to 20 and 21 to 30 minutes (161/262, 61.5%), which aligns with the reported average duration of telemedicine services in China [37]. In contrast, a study from the United Kingdom reported that the time taken for telepathology diagnosis averaged 3.9 minutes [38]. Pathologists' acceptance of telepathology demonstrates a strong global consensus. In our study, 95.8% (251/262) of participating pathologists reported overall satisfaction, with 95% (249/262) willing to continue its use. This aligns with international findings: a Saudi Arabian study noted only 1.6% (1/64) had no intention to use telepathology [39]; a UK study involving 18 pathologists found 88.89% (16/18) were likely to continue digital reporting [40]; and a Japanese study showed 70% (16/23) were interested in the system for consultations [41]. Notably, our survey revealed a higher rate of strong intention to use (153/262, 58.4%) compared to the Saudi study (23/64, 35.9%) [39]. This trend extends beyond pathology, as evidenced by a global survey of urologists from 58 countries, where 80.9% (197/244) intended to continue using telemedicine in clinical practice [42]. When analyzing the influencing factors of satisfaction and using intention among pathologists, this study demonstrated that the convenience of system operations and service process had a significant positive impact on pathologists' satisfaction and willingness to use telepathology. Additionally, longer service durations were also associated with positive impact, consistent with prior research on telemedicine services [37]. However, no significant associations were found between pathologists' satisfaction or willingness to use telepathology and factors such as gender, age, education level, or professional title. This is consistent with the findings from a survey of 64 pathologists in Saudi Arabia [39]. The manual navigation tools required for digital slides and the single-layer 2D nature of WSIs demand careful scrutiny by pathologists to ensure diagnostic accuracy. Shorter duration of telepathology may compromise diagnostic precision. Conversely, an intuitive system interface and streamlined service workflow facilitate smoother telepathology diagnosis, thereby enhancing pathologists' engagement and platform retention. The research suggests that improving pathologists' engagement in telepathology depends on optimizing both system design and workflow. To achieve this, medical institutions should: (1) enhance system usability and streamline the consultation process; and (2) reasonably schedule telepathology sessions based on workload and diagnostic quality requirements.

While research on telemedicine from an end-user perspective is extensive, studies focusing on telepathology user attitudes remain limited. This may be because telepathology often does not require direct patient involvement. Nevertheless, we surveyed 141 telepathology patients to assess their experiences and satisfaction. Among them, 92.2% (130/141) and 98.6% (139/141) received diagnostic results within 24 and 48 hours, respectively, surpassing the rates from our previous analysis of Henan Province's telepathology network (64%, 14,835/23,167 and 82.9%, 19,201/23,167) [21] and exceeding those reported by the Québec Telepathology Network in Canada [43]. All patients expressed satisfaction, aligning with the high patient satisfaction reported in other telemedicine surveys in Canada (248/272, 91%) [44], Norway [45], Argentina (116/116, 100%) [46], and Africa [47]. Higher than the survey in India, which indicated that 78% (39/50) of patients with cancer expressed interest in future telemedicine opportunities [48], 99.3% (140/141) were willing to recommend telepathology to others, citing high-quality service, time savings, and cost savings as key reasons. Based on the findings of this study and international comparative data, telepathology services have demonstrated clear and significant advantages in diagnostic turnaround time and patient satisfaction. Therefore, it is recommended that in future public education and service promotion, the core messages of "high quality (expert diagnosis), high efficiency (rapid response), convenience (localized diagnosis and treatment)" and "high satisfaction (patient endorsement)" be precisely communicated. This approach aims to enhance patient awareness and trust, thereby encouraging patients to more proactively choose telepathology services when faced with diagnostic options.

Limitations

This study analyzes the construction and application of telepathology in China from 2018 to 2023 through a nationwide survey, focusing on digital slide scanners, telepathology software, and network connectivity. It also evaluates the progress of telepathology services across hospitals of varying levels and regions. As the first detailed nationwide report on telepathology in China, this study uniquely examines perspectives from both pathologists and patients, covering usage, satisfaction, and influencing factors analysis. The findings provide a comprehensive understanding of telepathology in China and offer actionable recommendations for its enhanced application.

Although our findings offer valuable insights into telepathology in China, this study has several limitations. The sample was based on voluntary participation rather than a longitudinal cohort follow-up, and the significant shift in policy focus toward internet hospitals may have reduced engagement with this telemedicine-specific survey, contributing to the annual decline in sample size. The survey was conducted through TIPC, the national academic organization for telemedicine, and administered via the NTCC website, with participating hospitals spanning most regions. However, the online nature of the survey

XSL•FO
RenderX

may have introduced selection bias. The survey spanned from 2018 to 2023, a period marked by significant advancements in telemedicine and a shift toward internet-based healthcare, driven by national policies. Policy shifts and potential survey fatigue contributed to a decline in valid responses each year. Additionally, the surveys for 2021 and 2022 were disrupted due to the pandemic. Despite these limitations, the survey encompassed a sufficient number of medical institutions across China, and data analysis revealed a clear growth trend in telepathology development. Therefore, while these limitations exist, their impact on the study's conclusions is likely minimal. Besides, in-depth interviews with surveyed pathologists and patients are still lacking to fully understand the barriers of telepathology, so as to put forward more reasonable suggestions.

Given that this study is based on a comprehensive national survey in which the items related to telepathology were relatively general, future research will involve the design and implementation of a specialized telepathology survey. This survey will categorize questions according to service types—such as routine histopathological diagnosis and intraoperative frozen section consultation—to obtain more granular and in-depth data. This approach aims to comprehensively and thoroughly reveal the actual clinical application landscape and differentiated needs of telepathology. Additionally, we will further explore the perspectives of pathologists and patients towards telepathology through semi-structured interviews, which can help to devise corresponding strategies for the better development of telepathology in China.

Conclusion

This study provides the first systematic examination of telepathology development in China through long-term trend analysis, regional comparisons, and dual perspectives of both clinicians and patients. The findings indicate significant growth in the construction and application of telepathology in China from 2018 to 2023, with high levels of acceptance from both pathologists and patients. Factors such as the convenience of system operation, smoothness of service processes, frequency of use, and average duration per session were significantly associated with pathologists' satisfaction and willingness to use telepathology. To enhance the adoption rate of telepathology services, targeted recommendations are proposed around a synergistic "technology-institution-promotion" framework: at the technological level, integrating cloud computing and artificial intelligence to improve system usability; at the institutional level, establishing dedicated funds for central and western regions, promoting the inclusion of telepathology in medical insurance reimbursement, and refining cross-institutional licensing and performance incentive policies; at the promotional level, accurately communicating its core values of "high quality, high efficiency, convenience, and high satisfaction" to the public to encourage proactive patient choice. Through multidimensional systemic interventions, the technological advantages of telepathology can be translated into sustainable primary healthcare service capabilities.

Funding

This research was supported by grants from the National Key R&D Program of China (2022ZD0160705), the Science and Technology Research Program in Henan province (242102311012, 242102311021), the Basic Research and Appled Basic Research Program in Zhengzhou City (Grant No. 2024ZZJCYJ017), the Key Research of Higher Education in Henan province (23A330007), and the Joint Co-construction Program of Medical Science and Technology of Henan province (LHGJ20210313).

Data Availability

The datasets generated during this study are available upon individual request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Map of China Showing the Division of Eastern, Central, and Western Regions.
[PNG File, 23 KB - medinform_v14i1e83514_app1.png]

Multimedia Appendix 2
Distribution of the survey coverage areas in China from 2018 to 2023.
[PNG File, 2983 KB - medinform_v14i1e83514_app2.png]

Multimedia Appendix 3
Survey on the development of telemedicine in Chinese hospitals.
[DOCX File, 42 KB - medinform_v14i1e83514_app3.docx]

Multimedia Appendix 4
Table S1. The operation and charging situation of telemedicine in hospitals that carry out telepathology services.

XSL•FO
RenderX

[DOCX File, 14 KB - medinform_v14i1e83514_app4.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 18 KB - medinform_v14i1e83514_app5.docx]

References

1. Lu ZH, Chen J. National pathology quality report in 2019 [Article in Chinese]. Zhonghua Bing Li Xue Za Zhi 2020 Jul 8;49(7):667-669. [doi: 10.3760/cma.j.cn112151-20200331-00273] [Medline: 32610375]
2. Yefan Y, Shafei W, Yuan L, et al. Survey report on personnel, equipment, and workload in Chinese pathology departments, 2019-2020 [Article in Chinese]. Chin J Pathol 2023;52(9):985-988. [doi: 10.3760/cma.j.cn112151-20230314-00192]
3. Weinstein RS, Bloom KJ, Rozek LS. Telepathology and the networking of pathology diagnostic services. Arch Pathol Lab Med 1987 Jul;111(7):646-652. [Medline: 3606341]
4. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med 2013 Dec;137(12):1710-1722. [doi: 10.5858/arpa.2013-0093-CP] [Medline: 23634907]
5. Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. Arch Pathol Lab Med 2022 Apr 1;146(4):440-450. [doi: 10.5858/arpa.2020-0723-CP] [Medline: 34003251]
6. Cross S, Furness P, Igali L, Snead D, Treanor D. Best practice recommendations for implementing digital pathology. : The Royal College of Pathologists; 2018 Jan URL: https://www.rcpath.org/static/f465d1b3-797b-4297-b7fedc00b4d77e51/Best-practice-recommendations-for-implementing-digital-pathology.pdf [accessed 2025-11-25]
7. Montgomery ND, Liomba NG, Kampani C, et al. Accurate real-time diagnosis of lymphoproliferative disorders in Malawi through clinicopathologic teleconferences: a model for pathology services in sub-Saharan Africa. Am J Clin Pathol 2016 Oct;146(4):423-430. [doi: 10.1093/ajcp/aqw118] [Medline: 27594430]
8. Shinohara S, Bychkov A, Munkhdelger J, et al. Substantial improvement of histopathological diagnosis by whole-slide image-based remote consultation. Virchows Arch 2022 Aug;481(2):295-305. [doi: 10.1007/s00428-022-03327-2] [Medline: 35672584]
9. Girolami I, Pantanowitz L, Marletta S, et al. Diagnostic concordance between whole slide imaging and conventional light microscopy in cytopathology: a systematic review. Cancer Cytopathol 2020 Jan;128(1):17-28. [doi: 10.1002/cncy.22195] [Medline: 31600035]
10. García-Rojo M. International clinical guidelines for the adoption of digital pathology: a review of technical aspects. Pathobiology 2016;83(2-3):99-109. [doi: 10.1159/000441192] [Medline: 27100834]
11. Hanna MG, Pantanowitz L, Evans AJ. Overview of contemporary guidelines in digital pathology: what is available in 2015 and what still needs to be addressed? J Clin Pathol 2015 Jul;68(7):499-505. [doi: 10.1136/jclinpath-2015-202914] [Medline: 25979986]
12. Hanna MG, Reuter VE, Samboy J, et al. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. Arch Pathol Lab Med 2019 Dec;143(12):1545-1555. [doi: 10.5858/arpa.2018-0514-OA] [Medline: 31173528]
13. 510(k) clearances. US Food and Drug Administration (FDA). URL: https://www.fda.gov/medical-devices/device-approvals-and-clearances/510k-clearances [accessed 2026-01-04]
14. Digital pathology-related medical device [Article in Chinese]. National Medical Products Administration. URL: https://www.nmpa.gov.cn/datasearch/search-result.html [accessed 2026-01-04]
15. Kayser K. Telepathology in Europe. Anal Cell Pathol 2000;21(3-4):95-96. [doi: 10.1155/2000/757486] [Medline: 11339573]
16. Silas OA, Abdulkareem F, Novo JE, et al. Telepathology in Nigeria for global health collaboration. Ann Glob Health 2022;88(1):81. [doi: 10.5334/aogh.3673] [Medline: 36196362]
17. Feroz A, Mohammed Feroz TP, Bastian TS, Selvamani M. Telepathology: an update on applications, latest advances, and current status in Indian scenario. J Cancer Res Ther 2020;16(4):703-707. [doi: 10.4103/jcrt.JCRT_477_17] [Medline: 32930106]
18. Mremi A, Bentzer NK, Mchome B, et al. The role of telepathology in diagnosis of pre-malignant and malignant cervical lesions: implementation at a tertiary hospital in Northern Tanzania. PLoS ONE 2022;17(4):e0266649. [doi: 10.1371/journal.pone.0266649] [Medline: 35421156]
19. Pare G, Meyer J, Trudel MC, Tetu B. Impacts of a large decentralized telepathology network in Canada. Telemed J E Health 2016 Mar;22(3):246-250. [doi: 10.1089/tmj.2015.0083] [Medline: 26252758]
20. Tong S, Lili N, Zhouping T, Yanyan C. Construction of a remote pathology consultation service platform based on internet hospitals [Article in Chinese]. Chin J Hosp Admin 2024;40(3):236-240. [doi: 10.3760/cma.j.cn111325-20230815-00078]
21. He X, Wang L, Wang L, et al. Effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021 Jul 29;23(7):e23799. [doi: 10.2196/23799] [Medline: 34326037]

XSL•FO
RenderX

22. Chen J, Jiao Y, Lu C, Zhou J, Zhang Z, Zhou C. A nationwide telepathology consultation and quality control program in China: implementation and result analysis. Diagn Pathol 2014;9 Suppl 1(Suppl 1):S2. [doi: 10.1186/1746-1596-9-S1-S2] [Medline: 25565398]
23. NTCC. URL: https://www.ntcc.org.cn [accessed 2026-02-26]
24. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
25. Wang P, Hong J, Tang ZQ, et al. The acceptance of traditional Chinese medicine among patients with Parkinson's disease: a hospital survey. J Integr Med 2024 Mar;22(2):180-187. [doi: 10.1016/j.joim.2024.03.005] [Medline: 38519276]
26. Tsuchihashi Y. Expanding application of digital pathology in Japan--from education, telepathology to autodiagnosis. Diagn Pathol 2011 Mar 30;6 Suppl 1(Suppl 1):S19. [doi: 10.1186/1746-1596-6-S1-S19] [Medline: 21489189]
27. Farahani N, Pantanowitz L. Overview of telepathology. Clin Lab Med 2016 Mar;36(1):101-112. [doi: 10.1016/j.cll.2015.09.010] [Medline: 26851668]
28. Xiaohui W. The utilization rate of telemedicine is less than 20%. Related prices and medical insurance policies will be introduced before September. URL: https://www.chinatimes.net.cn/article/87356.html [accessed 2024-06-02]
29. Hitchcock CL. The future of telepathology for the developing world. Arch Pathol Lab Med 2011 Feb;135(2):211-214. [doi: 10.5858/135.2.211] [Medline: 21284440]
30. Harasees A, Al-Ahmad B, Alsobeh A, Abuhussein A. A secure IoT framework for remote health monitoring using fog computing. Presented at: 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS); Sep 24-27, 2024; Dubrovnik, Croatia p. 17-24. [doi: 10.1109/ICCNS62192.2024.10776425]
31. Hasan FA, Ashqar HI, AlSobeh A, Darwish O. Blockchain-based national digital identity framework – case of Palestine. Presented at: 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS); Sep 24-27, 2024; Dubrovnik, Croatia p. 76-83. [doi: 10.1109/ICCNS62192.2024.10776538]
32. AlSobeh AMR, Hammad R, Tamimi AKA. A modular cloud-based ontology framework for context-aware EHR services. Int J Comput Appl Technol 2019;60(4):339-350. [doi: 10.1504/IJCAT.2019.101181]
33. Maher L, Craig A, Menezes G. A national survey of telemedicine in the Republic of Ireland. J Telemed Telecare 2007;13(7):348-351. [doi: 10.1258/135763307782215361] [Medline: 17958936]
34. Dennis T, Start RD, Cross SS. The use of digital imaging, video conferencing, and telepathology in histopathology: a national survey. J Clin Pathol 2005 Mar;58(3):254-258. [doi: 10.1136/jcp.2004.022012] [Medline: 15735155]
35. Eloff E, Schoeman R. Digital pathology in South Africa: a survey and cross-sectional analysis of the current level of use and comparison with the international community. Int J Telemed Appl 2025;2025:7990652. [doi: 10.1155/ijta/7990652] [Medline: 40917478]
36. Unternaehrer J, Grobholz R, Janowczyk A, Zlobec I, Swiss Digital Pathology Consortium (SDiPath). Current opinion, status and future development of digital pathology in Switzerland. J Clin Pathol 2020 Jun;73(6):341-346. [doi: 10.1136/jclinpath-2019-206155] [Medline: 31857377]
37. Ma Q, Sun D, Tan Z, et al. Usage and perceptions of telemedicine among health care professionals in China. Int J Med Inform 2022 Oct;166:104856. [doi: 10.1016/j.ijmedinf.2022.104856] [Medline: 36037593]
38. Leong FJ, Graham AK, Schwarzmann P, McGee JO. Clinical trial of telepathology as an alternative modality in breast histopathology quality assurance. Telemed J E Health 2000;6(4):373-377. [doi: 10.1089/15305620050503834] [Medline: 11242544]
39. Alawashiz R, AlDossary SA. Pathologists' acceptance of telepathology in the Ministry of National Guard Health Affairs Hospitals in Saudi Arabia: a survey study. Digit Health 2023;9:20552076231163672. [doi: 10.1177/20552076231163672] [Medline: 36937697]
40. Browning L, Fryer E, Roskell D, et al. Role of digital pathology in diagnostic histopathology in the response to COVID-19: results from a survey of experience in a UK tertiary referral hospital. J Clin Pathol 2021 Feb;74(2):129-132. [doi: 10.1136/jclinpath-2020-206786] [Medline: 32616541]
41. Nakayama I, Matsumura T, Kamataki A, et al. Development of a teledermatopathology consultation system using virtual slides. Diagn Pathol 2012 Dec 13;7:177. [doi: 10.1186/1746-1596-7-177] [Medline: 23237667]
42. Dubin JM, Wyant WA, Balaji NC, et al. Telemedicine usage among urologists during the COVID-19 pandemic: cross-sectional study. J Med Internet Res 2020 Nov 5;22(11):e21875. [doi: 10.2196/21875] [Medline: 33031047]
43. Perron E, Louahlia S, Nadeau L, et al. Telepathology for intraoperative consultations and expert opinions: the experience of the Eastern Québec Telepathology Network. Arch Pathol Lab Med 2014 Sep;138(9):1223-1228. [doi: 10.5858/arpa.2013-0466-OA] [Medline: 25171705]
44. Strasser L, Hayawi L, Webster RJ, Venkateswaran S, Muir K. Telemedicine in pediatric neurology: a survey of patient and provider experience. J Child Neurol 2025 Feb;40(2):99-115. [doi: 10.1177/08830738241287243] [Medline: 39539164]
45. Moseng D. Teledermatology--experiences from Northern Norway [Article in Norwegian]. Tidsskr Nor Laegeforen 2000 Jun 20;120(16):1893-1895. [Medline: 10925620]
46. Fortini S, Espeche A, Caraballo R. Telemedicine and epilepsy: a patient satisfaction survey of a pediatric remote care program. Epilepsy Res 2020 Sep;165:106370. [doi: 10.1016/j.eplepsyres.2020.106370] [Medline: 32516743]

XSL•FO
RenderX

47. Kurji K, Kiage D, Rudnisky CJ, Damji KF. Improving diabetic retinopathy screening in Africa: patient satisfaction with teleophthalmology versus ophthalmologist-based screening. Middle East Afr J Ophthalmol 2013;20(1):56-60. [doi: 10.4103/0974-9233.106388] [Medline: 23580853]
48. Sharma SK, Kumar M, Datla PKV, et al. Assessment of IPD patients' knowledge and attitudes toward telemedicine at tertiary care hospital of Metro City. J Pharm Bioallied Sci 2025 Sep;17(Suppl 3):S2376-S2380. [doi: 10.4103/jpbs.jpbs_1471_24] [Medline: 41164588]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
NTCC: National Telemedicine Center of China
TIPC: Telemedicine Informationization Professional Committee of China
WSI: whole-slide image

Edited by A Rosania; submitted 04.Sep.2025; peer-reviewed by A AlSobeh, C Ling, W Zhang; accepted 01.Feb.2026; published 16.Mar.2026.

Please cite as:
Wang L, Cui F, He X
Telepathology Development and Stakeholder Perspectives in China: Cross-Sectional Survey
JMIR Med Inform 2026;14:e83514
URL: https://medinform.jmir.org/2026/1/e83514
doi:10.2196/83514

© Linlin Wang, Fangfang Cui, Xianying He. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Dynamic Personalized Optimization: An AI Functionality Framework for Digital Therapeutics

Dohyoung Rim, PhD

Rowan Corporation, 18F Yonsei Severance Bld, 10, Tongil-ro, Jung-gu, Seoul, Republic of Korea

Corresponding Author:
Dohyoung Rim, PhD
Rowan Corporation, 18F Yonsei Severance Bld, 10, Tongil-ro, Jung-gu, Seoul, Republic of Korea

Abstract

Dynamic Personalized Optimization (DPO) is introduced as a conceptual framework that defines core artificial intelligence (AI) functions required to deliver real-time, personalized, and optimized treatment in digital therapeutics (DTx). DPO continuously refines therapeutic strategies by integrating patient data, treatment content, usage feedback, and status measurements to provide real-time, personalized treatment. Using predictive AI models, DPO adapts treatment approaches based on patient responses, thereby improving therapeutic effectiveness. Furthermore, this paper explores the potential role of large language models (LLMs) in supporting DPO by processing diverse and complex data formats. By addressing current limitations in real-time personalization within DTx, DPO establishes a structured, AI-driven approach to delivering tailored digital interventions. This framework ultimately aims to enhance treatment efficacy and improve patient engagement.

(JMIR Med Inform 2026;14:e75256) doi:10.2196/75256

KEYWORDS

dynamic personalized optimization; digital therapeutics; AI functionality framework; personalized treatment; large language models

Introduction

Digital Therapeutics (DTx) are software-based interventions designed to prevent, manage, or treat medical conditions and provide direct therapeutic benefits through digital technology [1]. Such digital interventions are particularly important for neurological disorders, which require complex, personalized approaches [2]. Neurological disorders such as tinnitus, depression, and mild cognitive impairment require individualized care that DTx can offer [3,4], but the full potential of DTx is realized only with artificial intelligence (AI) integration [5].

These DTx solutions can continuously monitor patients, deliver real-time feedback, and dynamically adjust treatments to improve outcomes [6]. AI plays a crucial role in DTx by analyzing large-scale patient data to enable real-time personalization and improved treatment outcomes [3,7,8].

This study introduces a functional framework for AI in DTx, which supports real-time personalization through dynamic adaptation. Existing DTx lack a formal framework for real-time personalization. Prior work has identified persistent challenges in dynamic personalization [8,9], including the need for more adaptable models and limitations of static content in DTx. This paper addresses this gap by introducing a new framework called Dynamic Personalized Optimization (DPO).

DPO Framework

Definition and Characteristics of DPO

DPO is a conceptual framework that delineates the core AI functions necessary for DTx to achieve dynamic personalization and optimal therapy. It integrates and analyzes patient data to refine treatment strategies in real time. DPO tailors treatment plans by incorporating patient data and responses to offer individualized therapeutic approaches. A predictive AI model anticipates the patient's future status and optimizes treatment accordingly. DPO dynamically optimizes personalized therapy by integrating clinical data, feedback, and treatment content.

Building on just-in-time adaptive interventions [10], DPO formalizes relevant data types and introduces adaptive treatment content optimization through AI-based prediction of patient feedback or status. While traditional just-in-time adaptive interventions primarily focus on determining the timing of interventions, DPO differentiates itself by continuously optimizing treatment content through AI-based prediction of patient feedback, thereby extending the concept from temporal adaptation to content-level personalization.

Necessity of DPO in DTx

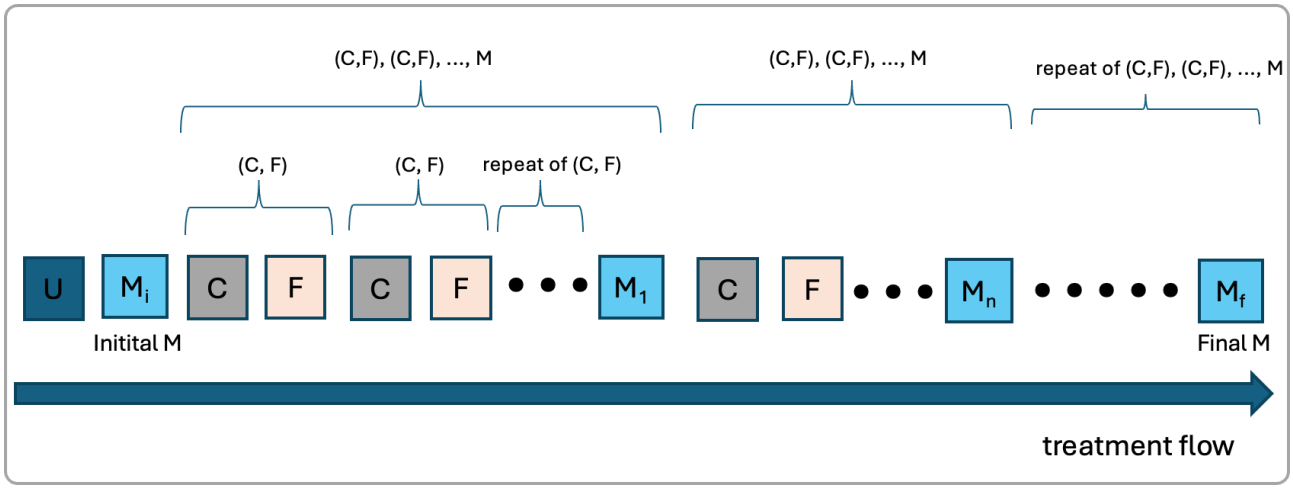
Neurological and other complex disorders require personalized, responsive treatment. DPO is essential to achieving this in DTx, as it continuously tailors therapy based on patient-specific data in real time.

XSL•FO
RenderX

The Four Data Types of DPO

DPO uses four types of data to tailor therapy. User data (U) include information about the patient, such as demographics, medical history, and lifestyle. Status measurement (M) data refer to objective health metrics, including symptom severity and test results. Treatment content data (C) encompass information related to the therapy itself, such as descriptions, types, and formats. Finally, feedback data (F) consist of feedback received after therapy, such as treatment game scores or the patient's reported feelings. AI in DTx uses these 4 types of data to create personalized treatment plans. Figure 1 illustrates how these 4 data types are sequentially processed throughout the therapeutic process.

Figure 1. The treatment process begins with a sequence of 4 data types: initial user data (U), initial patient status (Mi), and a repeated cycle of treatment content (C) and feedback (F), which leads to updated status measurements (M). This treatment-feedback sequence and patient status update continues throughout the treatment period.



Adaptive Treatment Content Optimization

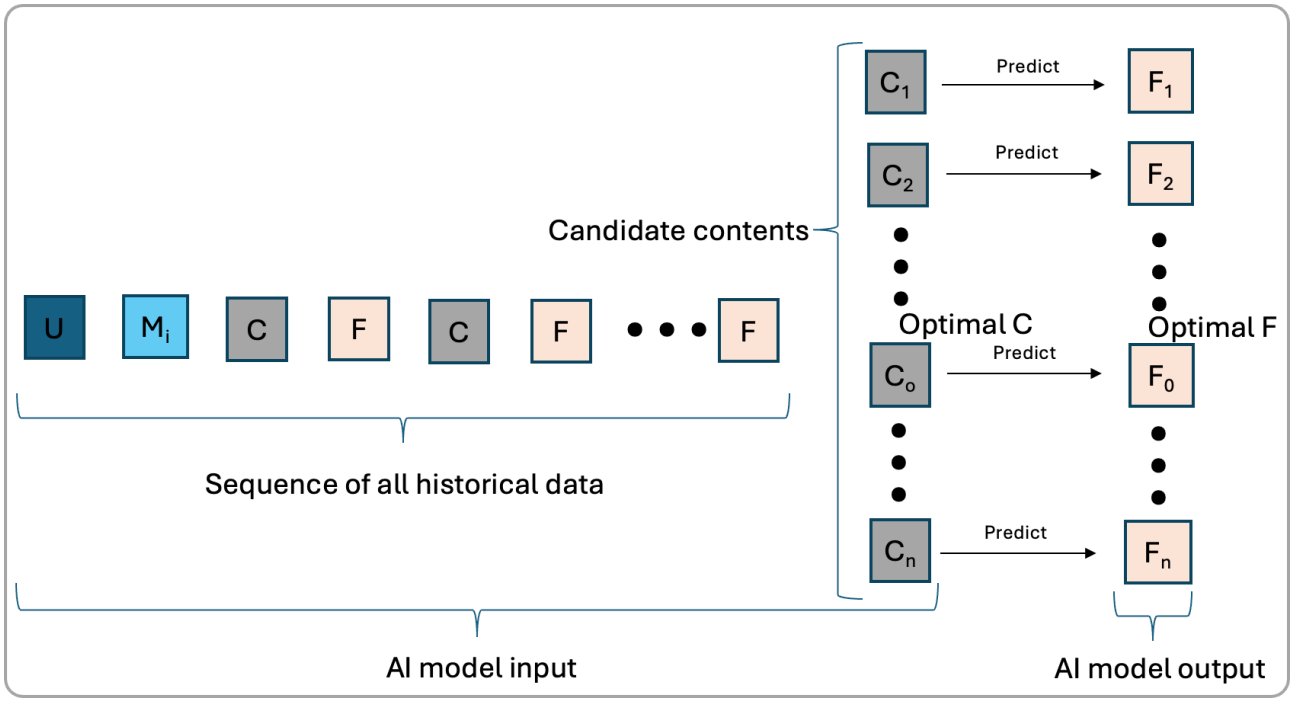
Within the DPO framework, AI is expected to continuously select the treatment content (C) that yields the greatest improvement in a patient's final status (M_f) relative to the initial status (M_i). However, identifying the "best" content in advance for AI model training is nearly impossible because it varies by individual and context and cannot be known until tried.

Predicting F

As Figure 2 illustrates, instead of directly predicting the optimal content C, this approach focuses on predicting F for a given C. In a DTx environment, real-world data collection yields sequential data (U, M, C, and F) throughout the treatment. An AI model can be trained on these historical data sequences to predict subsequent F. Although F does not directly represent treatment efficacy, it informs future treatment decisions. This approach assumes that feedback (F) is a valid proxy for treatment efficacy. Prior research shows a strong correlation between user engagement in digital interventions and improved treatment outcomes [11], though engagement alone may not fully capture clinical efficacy. Nonetheless, leveraging this correlation, the model can predict F for candidate treatment contents and select the optimal one (C_o) with the highest predicted F.

XSL•FO
RenderX

Figure 2. One optimization approach involves predicting the expected feedback (F) for each candidate content. The optimal content (C_o) is selected based on the highest predicted F. An artificial intelligence (AI) model is trained on historical data sequences to predict F for a given candidate C.

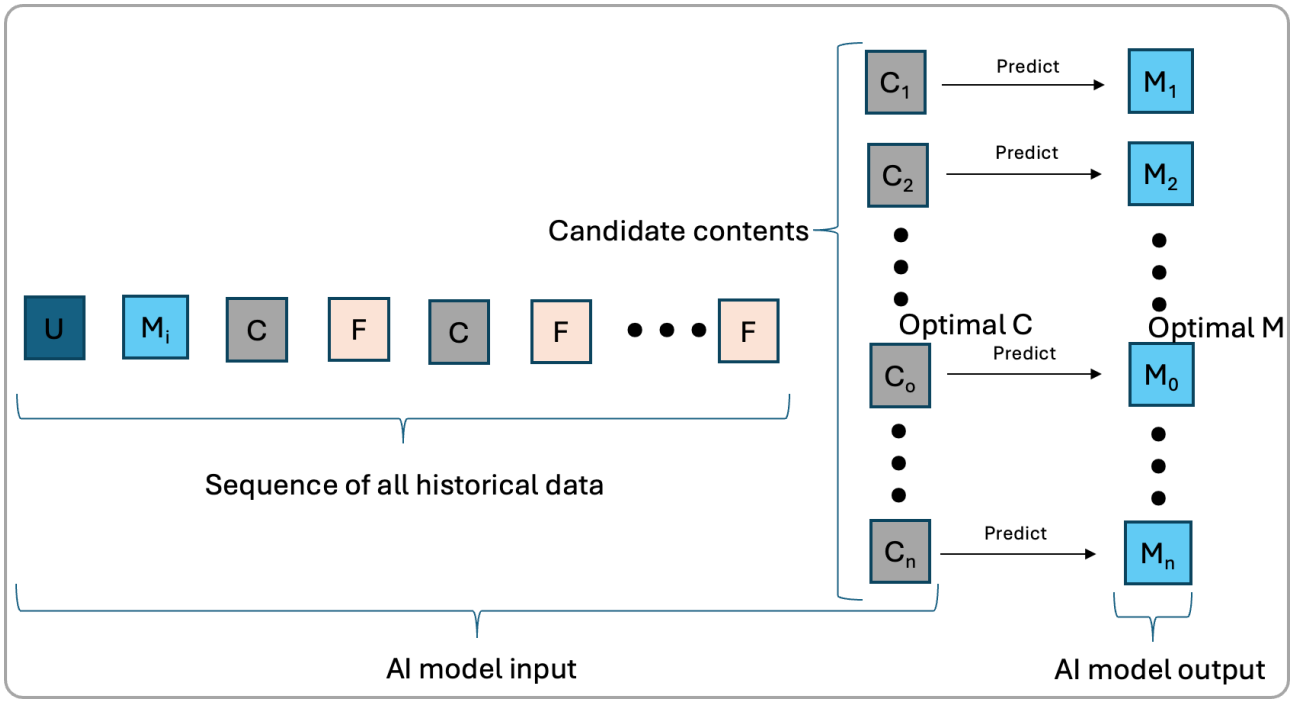


Predicting M

Figure 3 illustrates another approach: predicting the patient's M_f. In this approach, an AI model is trained with prior data sequences as input and M_f as the output. If the model learns this relationship accurately, it can predict M_f for a given input sequence and a candidate C. By predicting M for all candidate C options, the model can select the C_o that most effectively improves M.

Together, these approaches eliminate the need to explicitly label C as the correct answer. Instead, AI predicts F or M_f to dynamically determine the optimal C, ensuring that patients engage with the most effective therapeutic content. However, implementing these AI-driven strategies in practice requires managing complex and heterogeneous DTx data, which poses significant challenges.

Figure 3. An alternative approach involves predicting the final patient status (M_f) for each candidate content. The artificial intelligence (AI) model takes historical data and candidate C as input and predicts the resulting status (M_f). The optimal content (C_o) is selected to maximize improvement in M_f relative to the initial status.



XSL•FO
RenderX

In a cognitive intervention context for mild cognitive impairment (MCI) [12], DPO operates as follows. The system first collects U such as demographic and behavioral information including age, education level, daily activity, and sleep patterns. It also gathers initial measurement data (M_i), such as the Mini-Mental State Examination score [13] representing the patient's baseline cognitive state. Based on these inputs and the history of previously delivered C—for example, cognitive games that train memory, attention, or problem-solving skills—and corresponding F, such as game scores reflecting engagement and task performance, the AI model predicts either the expected F or the postsession M for each candidate content option. The predicted value then guides the system in selecting the next C_o for the patient's subsequent training session. This process of providing C and receiving F repeats over time, allowing the system to refine personalized cognitive training. Through this iterative mechanism, DPO continuously optimizes therapeutic content to enhance cognitive function and engagement in patients with MCI.

Challenges of Data in DTx and LLM Strategy

Overview

DTx data appear in diverse forms—text (eg, clinical notes), numeric measurements, images, and video—varying widely in length and quality. This heterogeneity hinders uniform processing of data by AI. Addressing this issue requires careful data preprocessing, stringent quality control, and AI models capable of integrating disparate data types.

LLMs have significant potential to overcome data processing challenges in DTx [14]. The following strategies describe how LLMs can support DPO.

Integrated Processing of Various Data Formats

LLMs can process diverse data formats (including text, numerical data, and images) thereby enabling comprehensive analysis to improve treatment outcomes [15]. They analyze unstructured text, such as medical records and patient feedback, to extract key details and interpret numerical data like vital signs and laboratory results. Additionally, LLMs use embedding techniques to represent each data modality as a numeric vector, enabling integrated analysis across different data types. This approach allows processing of disparate data types within the same mathematical space.

Processing Variable Data Sizes

LLMs can handle data inputs of widely varying lengths [16,17]—from brief notes and detailed medical records to extensive patient histories and multisource health logs. They can also process time-series data to identify patterns and trends in symptom logs and vital signs. This capability enables real-time monitoring and more adaptive care.

Discussion

The DPO framework proposed in this study has been designed as a structural model applicable to a range of DTx scenarios, offering a systematic functional definition of AI capabilities required for the personalization and adaptive coordination of digital therapeutic content. In DPO, LLMs are not used to generate new therapeutic content but rather serve as analytical tools to integrate and interpret multimodal data for personalized optimization. This paper presents a conceptual design framework and does not include empirical validation for 2 primary reasons. First, DPO is formulated as a high-level functional abstraction rather than a model-specific architecture, and its applicability may be realized through various algorithmic implementations. Therefore, the focus of this work is on the completeness and coherence of the framework design rather than the comparative performance of specific models. Second, the DPO framework is currently undergoing pilot-level integration in several ongoing DTx development pipelines within an industry setting, and early experimental outcomes related to these efforts are planned to be reported in a separate study. Future work will empirically evaluate the implementation and performance of each functional component of the DPO framework—prediction, personalization, and optimization—within actual digital therapeutic contexts using real user data. Additional studies will also explore the generalizability and scalability of DPO across diverse clinical domains, including MCI, depression, and tinnitus, to further establish the framework's practical utility and robustness.

Conclusion

The proposed DPO framework offers a structured approach to operationalizing AI in DTx, with the potential to improve both engagement and therapeutic efficacy. Integrating LLMs into DPO could greatly enhance personalized care delivery and improve health care outcomes. At the same time, ensuring data privacy, security, and safety in digital therapeutics remains essential. Moreover, the responsible use of AI-based DTx requires addressing ethical issues such as transparency, fairness, cybersecurity, and data protection, which must be implemented with accountability [18]. Recent studies on responsible AI deployment further emphasize the importance of governance, transparency, and privacy as key foundations for ethical implementation in health care [19,20]. AI must generate transparent recommendations while ensuring clinical validation and ethical oversight. Overcoming these challenges could enable a new level of personalized treatment and substantially improve patient outcomes. Leveraging advanced AI models for DPO may facilitate the integration of heterogeneous patient data in ways previously unattainable, essentially serving as a "brain" for DTx that learns and adapts to each patient. Going forward, the practical implementation and clinical validation of DPO (potentially with LLMs) are crucial next steps. These efforts must be accompanied by careful attention to ethical and regulatory considerations. If DPO can maintain patient engagement by tailoring content to what is therapeutically optimal for each individual, the therapeutic efficacy of DTx could be substantially enhanced. Such personalized adaptation prevents patient dropout or disengagement caused by monotonous or ineffective content [1], thereby further improving outcomes.

XSL•FO
RenderX

Funding

The author declares no financial support was received for this work.

Authors' Contributions

DR confirms being solely responsible for all aspects of this work, including conceptualizing and designing the study, conducting the analysis, and drafting and revising the entire manuscript.

Conflicts of Interest

None declared.

References

1. Digital therapeutics definition and core principles. : Digital Therapeutics Alliance; 2019.
2. Griñán-Ferré C, Bellver-Sanchis A, Guerrero A, Pallàs M. Advancing personalized medicine in neurodegenerative diseases: The role of epigenetics and pharmacoepigenomics in pharmacotherapy. Pharmacol Res 2024 Jul;205:107247. [doi: 10.1016/j.phrs.2024.107247] [Medline: 38834164]
3. Choi H, Shin J, Kim JH, et al. Digital therapeutics for patients with mild cognitive impairment: Design of a mobile health program. Electronics (Basel) 2024;13(19):3892. [doi: 10.3390/electronics13193892]
4. Armeni P, Polat I, De Rossi LM, Diaferia L, Meregalli S, Gatti A. Exploring the potential of digital therapeutics: An assessment of progress and promise. Digit Health 2024;10:20552076241277441. [doi: 10.1177/20552076241277441] [Medline: 39291152]
5. Darcy A, Beaudette A, Chiauzzi E, et al. Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices 2022 Apr;19(4):287-301. [doi: 10.1080/17434440.2022.2075726] [Medline: 35748029]
6. Palanica A, Docktor MJ, Lieberman M, Fossat Y. The need for artificial intelligence in digital therapeutics. Digit Biomark 2020;4(1):21-25. [doi: 10.1159/000506861] [Medline: 32399513]
7. Smith ME, Sharma D, Rivero-Arias O, et al. Digital thErapy For Improved tiNnitus carE Study (DEFINE): Protocol for a randomised controlled trial. PLoS ONE 2024;19(1):e0292562. [doi: 10.1371/journal.pone.0292562] [Medline: 38180996]
8. Niazi SK. The coming of age of AI/ML in drug discovery, development, clinical testing, and manufacturing: The FDA perspectives. Drug Des Devel Ther 2023;17:2691-2725. [doi: 10.2147/DDDT.S424991] [Medline: 37701048]
9. Hu P, Hu L, Wang F, Mei J. Editorial: Computing and artificial intelligence in digital therapeutics. Front Med 2024;10:1330686. [doi: 10.3389/fmed.2023.1330686]
10. Patel M, Butte A. Challenges in digital therapeutics: Personalization and evidence generation. Journal of Digital Health 2020;5(2):98-105. [doi: 10.1093/ehjdh/ztaa008] [Medline: 37056293]
11. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
12. Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. Arch Neurol 1999 Mar;56(3):303-308. [doi: 10.1001/archneur.56.3.303] [Medline: 10190820]
13. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
14. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: A systematic review and meta-analysis. Front Digit Health 2021;3:764079. [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
15. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: The transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 7;27(1):e59069. [doi: 10.2196/59069] [Medline: 39773666]
16. Belyaeva A, Cosentino J, Hormozdiari F, et al. Multimodal LLMs for health grounded in individual-specific data. Presented at: First International Workshop on Machine Learning for Multimodal Healthcare Data; Jul 29, 2023. [doi: 10.1007/978-3-031-47679-2_7]
17. Gpt-4 technical report. : OpenAI; 2023. [doi: 10.48550/arXiv.2303.08774]
18. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
19. Radanliev P, Santos O, Brandon-Jones A, Joinson A. Ethics and responsible AI deployment. Front Artif Intell 2024;7:1377011. [doi: 10.3389/frai.2024.1377011] [Medline: 38601110]
20. Radanliev P. AI Ethics: Integrating transparency, fairness, and privacy in AI development. Appl Artif Intell 2025 Dec 31;39(1). [doi: 10.1080/08839514.2025.2463722]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
C: treatment content data
C_o: optimal content
DPO: Dynamic Personalized Optimization
DTx: digital therapeutics
F: feedback data
LLM: large language model
M: status measurement data
MCI: mild cognitive impairment
M_f: final status
M_i: initial status
U: user data

Edited by A Benis; submitted 31.Mar.2025; peer-reviewed by B Tasci, P Radanliev, S Aammou; revised version received 18.Oct.2025; accepted 17.Dec.2025; published 18.Mar.2026.

Please cite as:
Rim D
Dynamic Personalized Optimization: An AI Functionality Framework for Digital Therapeutics
JMIR Med Inform 2026;14:e75256
URL: https://medinform.jmir.org/2026/1/e75256
doi:10.2196/75256

© Dohyoung Rim. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint on the Consequences and Mitigation of Cognitive Bias in the Radiological Interpretation of Breast Cancer Imaging Using Artificial Intelligence

Lorenzo Conti[1], MSc; Benedetta Capetti[1,2], MSc; Ottavia Battaglia[2,3], MD; Roberto Grasso[1,2], MSc; Filippo Pesapane[3], MD; Dario Monzani[1,4,5], Prof Dr; Gabriella Pravettoni[1,2], Prof Dr

[1]Applied Research Division for Cognitive and Psychological Science, European Institute of Oncology IRCCS, Via Ripamonti 435, Milan, Italy

[2]Department of Oncology and Haemato-Oncology, University of Milan, Milan, Italy

[3]Breast Imaging Division, European Institute of Oncology IRCCS, Milan, Italy

[4]Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy

[5]WeSearch Lab – Laboratory of Behavioral Observation and Research on Human Development, University of Palermo, Palermo, Italy

Corresponding Author:
Lorenzo Conti, MSc
Applied Research Division for Cognitive and Psychological Science, European Institute of Oncology IRCCS, Via Ripamonti 435, Milan, Italy

Abstract

Artificial intelligence (AI) is increasingly integrated into breast imaging workflows, offering the potential to enhance diagnostic accuracy, efficiency, and early cancer detection. Image interpretation plays a pivotal role in the breast cancer diagnostic pathway, directly influencing therapeutic decisions and patient outcomes. However, the effective implementation of AI-assisted systems relies not only on technical performance but also on radiologists' trust, acceptance, and readiness to incorporate these tools into clinical practice. In addition, system-related, perceptual, and cognitive factors may contribute to diagnostic errors, ultimately affecting overall accuracy and reliability. This paper provides a comprehensive overview of the cognitive and systemic sources of diagnostic inaccuracies in breast imaging, emphasizing the growing role of AI as both a supportive and potentially bias-modulating tool. Recent prospective studies have demonstrated the clinical safety and effectiveness of AI-assisted mammography screening, reporting improved cancer detection rates and reduced workload. Nonetheless, the integration of AI into diagnostic workflows without an appropriate knowledge of the consequences may introduce new cognitive biases, such as anchoring, automation, and confirmation bias, that influence radiologists' decision-making and counteract the intended benefits. To address these challenges, the paper outlines strategies to mitigate diagnostic errors and foster appropriate integration of AI into clinical practice. These include targeted training programs, enhanced interdisciplinary communication, and standardized interpretation workflows that promote consistent evidence-based practice. Furthermore, the adoption of explainable AI approaches is identified as a key factor in improving model transparency and interpretability, allowing radiologists to understand algorithmic reasoning and engage in a more informed, confidence-based human-AI collaboration. Ultimately, a balanced and context-sensitive integration of AI, grounded in continuous professional education and cognitive awareness, is essential for improving diagnostic accuracy while preserving radiologists' critical analytical skills.

(JMIR Med Inform 2026;14:e78955) doi:10.2196/78955

KEYWORDS

radiological assessment; artificial intelligence; AI; cognitive bias; breast cancer; imaging

Radiological Interpretation Workflow

The increasing integration of artificial intelligence (AI) into clinical medicine has created new opportunities to enhance diagnostic accuracy and therapeutic decision-making across multiple fields. AI is defined as a machine's capability to mimic intelligent human behavior, such as learning, reasoning, and problem-solving [1]. Despite its promise, AI adoption in routine clinical practice faces significant challenges, including disparities in medical imaging access, and remains in early stages across many specialties [2-4].

In oncology, AI facilitates early tumor detection and personalized treatment planning, with a prominent role in radiology, where it analyzes medical images and patient data to detect patterns and abnormalities, accelerating and improving diagnosis. However, its extensive use in clinical decision-making increases exposure to bias with potentially serious clinical repercussions [2].

XSL•FO
RenderX

Radiological interpretations critically influence therapeutic decisions and patient management outcomes [5,6]. These interpretations are based on the careful evaluation of available evidence but remain subject to limitations because of the subjective nature of image assessment, without replacing gold-standard pathological examinations [5]. In health care, AI has been introduced to support complex tasks that require reasoning and decision-making [7]. However, its application must be approached cautiously, as it does not eliminate the possibility of errors.

Given these limitations, it is essential to understand the factors contributing to diagnostic variability and potential errors in radiology.

The process of interpreting medical images is complex, requiring radiologists to evaluate imaging data meticulously to identify potential health issues. Despite their expertise, the diagnostic error rate among practicing radiologists is estimated to be between 3% and 6%, amounting to approximately 40 million errors globally each year, given annual imaging volumes [8,9]. This error rate is even higher, between 31% and 37%, in cross-sectional imaging such as computed tomography (CT) [10]. Given the growing reliance on cross-sectional imaging, it is essential to understand the causes, prevalence, and impact of diagnostic errors to minimize their effect on patient care.

Diagnostic errors have significant consequences for patient outcomes and place a substantial burden on health care systems. They can be categorized into 3 primary groups: system errors, visual-perception errors, and cognitive errors [4,11]. System-related errors include equipment failures or suboptimal image quality. Visual-perception errors occur when radiologists either fail to detect abnormalities or mistakenly classify them as normal, influenced by factors such as fixation duration [5]. Finally, cognitive errors result from incorrect interpretation of detected abnormalities and represent the largest proportion of diagnostic mistakes, reaching up to 74% in some imaging modalities [12].

This paper aims to provide an in-depth overview of the implementation of AI in medical imaging, with a specific focus on breast cancer radiology, based on a comprehensive literature search (see Multimedia Appendix 1). This focus distinguishes our work from previous reviews, such as that of Stogiannos and colleagues [4], which adopt a modality-independent perspective. In contrast, the present review concentrates on how AI is being integrated into real-world breast imaging workflows across screening, diagnostic, and interventional settings. Particular attention is given to the practical challenges of implementation and the evolving dynamics of human-AI interaction in daily clinical practice.

Adopting both a conceptual and practice-oriented lens, it explores the dual nature of AI as both a supportive adjunct and a potential source of diagnostic error, emphasizing strategies for informed and responsible AI use. Finally, it addresses cognitive biases that arise from human-AI collaboration, their implications for clinical decision-making, and possible approaches to mitigate these effects.

Through this perspective, we argue that understanding and managing these biases is crucial to ensuring the safe and effective integration of AI in breast imaging, with the ultimate goal of improving diagnostic accuracy and optimizing patient care.

The Adoption of AI in Radiological Senology

Breast cancer is one of the most diagnosed cancers worldwide, accounting for approximately 2.3 million new cases and 0.7 million deaths among women each year [13]. Early diagnosis is of pivotal importance in reducing morbidity and mortality, and, in this scenario, mammographic screening plays a crucial role, currently being the only screening test that has been shown to reduce breast cancer–related mortality [14].

Despite the proven efficacy of mammographic screening, several challenges persist, particularly in terms of workload and diagnostic consistency. Screening programs are typically addressed to women aged between 40/50 and 70/80 years, with European guidelines recommending double reading of screening mammograms to ensure high sensitivity [15]. Nevertheless, the shortage of experienced breast radiologists and the large workload render the double reading difficult to sustain in several countries [16].

A possible solution might be represented by image analysis tools based on AI, a field of computer science dedicated to the creation of systems performing tasks that usually require human intelligence, aimed, for example, at facilitating triage of screening examinations according to risk of malignancy or supporting the radiologist's decision with computer-aided detection (CAD) highlighting suspicious findings, ultimately reducing the workload and the number of interval cancers [17].

Retrospective studies [18-20] suggest that the accuracy of AI is similar to or better than that of breast radiologists and that AI could help radiologists reduce false negative screening results when used as detection support, as it has been shown that AI retrospectively classifies screening examinations as high-risk before a diagnosis of interval cancer [21-23].

A breakthrough point was reached in 2023, when Lång et al [24] published the results of their randomized, controlled mammography screening with artificial intelligence (MASAI) trial, a study investigating an AI-supported screen-reading procedure involving triage of screening examinations to single or double reading, along with detection support, compared to standard double reading.

The MASAI trial was designed as a randomized, parallel, noninferiority, single-blinded, controlled, screening accuracy study in which, after screening, mammograms were acquired, and examinations were automatically randomized to AI-supported screening (intervention group) or standard double reading without AI (control group).

The examinations randomized to the intervention group were analyzed using Transpara version 1.7.0 (ScreenPoint Medical), a deep learning system to identify and interpret mammographic regions suspicious for cancer. The AI system provided an

XSL•FO
RenderX

examination-based malignancy-risk score on a continuous scale ranging from 1 to 10: examinations were considered as low risk (risk scores 1 - 7), intermediate risk (risk scores 8 and 9), or high risk (risk score 10). Furthermore, the AI system provided CAD marks at suspicious regional findings [24].

In the intervention group, examinations with risk scores of 1 - 9 (low and intermediate risk) underwent single reading, and examinations with risk scores of 10 (high risk) underwent double reading, done by two different breast radiologists [24].

Readers first read the examination without CAD marks and then with CAD marks, if available.

In the control group, screening examinations were read with standard unblinded double reading without AI.

This clinical safety analysis testifies that a screen-reading procedure using an AI tool to triage screening examinations to single or double reading and with the use of AI as detection support in mammography screening was safe because the cancer detection rate (6.1 per 1000 participants screened) was above the prespecified lower limit for safety and was similar to that of double reading without AI (5.1 per 1000). The use of AI did not influence the rates of recalls, false positives, or consensus meetings, while the screen-reading workload was reduced by almost half. Indeed, this study found that the benefit of AI-supported screening regarding screen-reading workload reduction was considerable.

Furthermore, the results of the trial showed that AI-supported screening detected 20% more cancers (244 vs 203) compared to standard screening and increased detection of in situ cancers with AI-supported screening compared with standard screening (60 vs 38), which could be concerning in terms of overdiagnosis.

Nevertheless, improvement in one performance measure, such as cancer detection rates, does not always imply an improvement in terms of outcome, such as recurrence rates or mortality reductions. The results of future screening cycles are going to highlight whether the introduction of new strategies, such as AI-supported screening, has a positive impact on the overall effect of the intervention on outcomes and if the costs are justified [25].

The results of this prospective randomized trial corroborate the findings of several retrospective studies that have shown AI has sufficient diagnostic accuracy to make radiological reads as an independent reader of screening mammograms. Nevertheless, there is currently an absence of specific guidelines regarding the potential introduction of AI as an independent reader or, more broadly, for the application of AI tools in everyday clinical practice [26,27].

A comprehensive systematic review performed by Anderson and collaborators [28], focused on the external validation of AI technologies applied to mammographic screening in real-world clinical settings, highlighted that the majority of studies evaluating either stand-alone AI systems or AI used in conjunction with radiologists reported improved diagnostic accuracy compared to radiologist-only interpretations. Some studies included in the review indicate that combining outputs from several AI models—ensemble modeling—can enhance overall performance compared to relying on a single algorithm [29,30]. However, implementing such approaches in real-world clinical environments may be challenging, particularly when it involves integrating multiple AI system owners. Concerning prospects, they emphasized a significant gap in external validation research, which is essential to assess how well AI models perform across large and varied populations in terms of different race/ethnicities or breast cancer risk profiles, and on different imaging techniques such as digital breast tomosynthesis (DBT) versus digital mammography. Additionally, it was highlighted that there is a need to determine the specific screening scenarios, such as age groups and screening intervals, where AI applications would be most beneficial.

Similarly, Houssami and collaborators [31] reported a satisfying diagnostic accuracy for the AI models proposed by the studies involved in their scoping review, with a median area under the curve of 88.2%. What was pointed out in their work was that most of the existing AI research in breast imaging has relied on data from screen-film or digital mammography. Given that DBT is increasingly becoming a central character in breast cancer screening, future AI development should incorporate DBT imaging data, as AI tools trained solely on traditional 2D mammography may quickly become obsolete as DBT adoption grows in clinical practice [31].

Diagnostic Errors in Radiology

Due to the complex interpretational process executed by radiologists, it may be subject to diagnostic errors in up to 6% of cases [5,6]. Diagnostic errors include system errors, such as those related to the equipment used (ie, poor image quality or suboptimal technique), and visual-perceptual errors, which occur when clinically significant findings are not identified due to specific risk factors, including reader fatigue, rapid interpretation, distractions, or interruptions [5]. However, there is a third type of error, called “cognitive errors,” related to the clinicians’ judgment, when clinically significant findings are correctly identified, but an incorrect interpretation is attributed due to mental shortcuts or preconceptions [32]. Identifying these errors is crucial in the medical field, as it enables early action to prevent inaccurate interpretations.

This type of error is caused by the influence of cognitive biases, systematic and unconscious thinking errors that occur when processing and interpreting information in the environment, and that influence people’s decisions and judgments [33,34].

Over time, various classifications have been proposed to categorize cognitive bias within the field of radiology, with multiple reviews documenting the main typologies encountered during medical image interpretation. These cognitive biases have been identified as significant contributors to radiological errors and diagnostic inaccuracies [5,35]. Furthermore, with the increasing integration of AI into radiological clinical practice, it becomes essential to consider its impact within the theoretical framework. In medical AI, a bias is defined as any factor or prejudice that drives an AI algorithm to generate differential or inequitable outcomes [2]. Consequently, the implementation of AI must be carefully evaluated, as it holds the potential to both mitigate certain diagnostic errors and exacerbate existing

XSL•FO
RenderX

cognitive biases [8]. Indeed, AI has shown promise in reducing subjective prejudices, enhancing image analysis clarity, decreasing interobserver variability, and improving the speed and accuracy of radiological reporting [1]. However, AI algorithms trained on biased datasets may perpetuate or even amplify existing health disparities, leading to compromised diagnostic accuracy and inequities in patient treatment [36]. These issues introduce new challenges that health care professionals must acknowledge, particularly regarding the emergence and influence of cognitive biases in AI-augmented radiology [1].

Overview of Cognitive Bias and AI-Driven Insights

This section describes the main cognitive biases highlighted in the literature regarding radiological interpretation and explores the potential contribution and pitfalls associated with AI application in this medical context. To facilitate a clear understanding and contextualization of these errors, Table 1 summarizes each cognitive bias with a specific clinical example that realistically illustrates its occurrence in practice.

XSL•FO
RenderX

Table . Overview of cognitive bias.

Cognitive bias	Explanation	Practical example	AI[a] contribution
Influence of AI (evidence-based)			
Availability	Overemphasizes easily recalled cases over statistical prevalence	*A breast radiologist has recently dealt with several cases of aggressive triple-negative breast cancer in young patients. Shortly afterward, when performing a preventive check-up in a 30-year-old with vague symptoms and no risk factors, the radiologist overestimates the likelihood of malignancy and recommends an aggressive workup, despite weak evidence. The recent memory of similar high-impact cases is fresh and vivid (availability bias), so the radiologist gives disproportionate weight to that experience, leading to an overestimation of the probability of cancer in situations that statistically may not warrant it.*	AI shifts focus to statistical data rather than anecdotal memory [32]
Nonavailability	Ignores diagnoses not previously encountered	*A breast radiologist, during routine screening, fails to consider a rare presentation of breast cancer because they have never seen a similar case before. This is nonavailability bias, as the radiologist underestimates or overlooks the possibility of cancer because that diagnostic pattern isn't readily available in their memory or experience.*	AI trained on large datasets can identify rare cases effectively [32]
Confirmation	Selectively interprets data to confirm existing hypothesis	*A general practitioner suspects a benign fibroadenoma in a patient with a palpable lump and refers the patient to a breast radiologist for an ultrasound. The radiologist identifies a round, circumscribed mass and, believing that this finding aligns with their first impression, the radiologist selectively focuses on features that support a benign interpretation and ignores subtle signs, like nonparallel orientation and heterogeneous echostructure, that could suggest a phyllodes tumor or malignancy. This is confirmation bias: the radiologist seeks out or gives more weight to evidence that supports an existing belief or hypothesis (in this case, that the mass is benign) while not giving enough importance to contradictory evidence, which can lead to diagnostic error or delayed treatment.*	AI may amplify bias if radiologists unquestioningly accept AI suggestions [37]

XSL•FO
RenderX

Cognitive bias	Explanation	Practical example	AI[a] contribution
Anchoring	Fixates on initial information despite conflicting later evidence	*Example 1: A real-life scenario might be represented by a patient with a history of a benign breast cyst returning for follow-up imaging examinations. The breast radiologist identifies a similar-appearing mass in the same location and assumes it's the same cyst, labeling it as benign without carefully evaluating prior images or subtle changes. Nevertheless, the lesion has grown and, in the present examination, new features of malignancy have appeared (eg, irregular margins and vascularization). The radiologist fixates on the initial impression or past diagnosis (a benign cyst) and fails to adequately update (anchoring bias) their judgment in light of new evidence.* *Example 2: A breast radiologist is reviewing a mammogram assisted by an AI tool that highlights a lesion and assigns a low malignancy score. Trusting the AI's judgment, the radiologist anchors on this benign assessment and disregards their initial concern about subtle spiculations, ultimately reporting the lesion as likely benign. However, a biopsy performed later confirms that the lesion was an invasive carcinoma.*	AI may exacerbate if the radiologist overly trusts the AI's initial assessment [37]
Automation	Prefers automated AI outputs even when conflicting info exists	*During a busy screening session, a radiologist uses an AI tool that flags suspicious lesions with heat maps on mammograms. One mammographic examination shows no AI-detected abnormalities, so the radiologist quickly signs the report as negative, without further close review of the mammograms. In this way, a small spiculated mass, missed by the AI tool goes undetected.*	Risk of overreliance on AI leading to missed diagnoses [4]
Framing	Interpretation influenced by how clinical info is presented	*A surgeon sends a patient complaining of a palpable lump to a breast radiologist for an ultrasound, highlighting the patient's strong family history of breast cancer and assessing that, according to him, this finding is suspicious. The radiologist, influenced by this alarming clinical context, interprets a likely benign nodule as suspicious and recommends a biopsy, even though imaging features alone don't strongly support malignancy.*	AI could reinforce framing if trained on biased clinical data [36]

XSL•FO
RenderX

Cognitive bias	Explanation	Practical example	AI[a] contribution
Premature closure	Accepts the diagnosis too early without considering alternatives	*A breast radiologist identifies a well-circumscribed, round opacity on a mammogram and immediately labels it as a benign nodule (eg, possible expression of a fibroadenoma) based on appearance and the patient's young age, without recommending further examinations; however, the mass was a rare presentation of a malignant phyllodes tumor. This is premature closure bias: the radiologist settles too quickly on a diagnosis without fully considering alternative possibilities or completing the necessary diagnostic workup, leading to missed or delayed diagnosis, especially in atypical cases.*	AI might help by suggesting alternative diagnoses for consideration [36]
Influence of AI (theoretical/hypothesis-driven)[b]			
Satisfaction of report	Accepts previous radiological reports without independent reassessment	*A patient is referred for a diagnostic mammogram after a recent ultrasound (done at another facility), reporting a suspicious lesion in the left breast at 2 o'clock. The breast radiologist, assisted by an AI system that flags suspicious regions, identifies on the mammogram, in the same location where the lesion was reported on ultrasound, a spiculated lesion with a high-risk score for malignancy assigned by the AI tool. The radiologist "satisfied" that the main finding had been identified, both by the AI and by their own review, and stopped further searching, and missed other suspicious lesions. Indeed, on the contralateral breast, there was a small cluster of irregular microcalcifications, suspicious for an in situ lesion.*	AI could provide independent second reads to limit this bias
Attribution	Preconceptions or patient stereotypes influence diagnosis, neglecting relevant info	*A breast radiologist is interpreting a screening mammogram of a woman with a history of prior false-positive findings (eg, cluster of microcalcifications that warranted a vacuum-assisted biopsy under stereotactic guidance, with an oncologically negative result) and assumes her current finding is, even this time, a benign one. This is an attribution bias as the radiologist allows prior patient characteristics or history to inappropriately influence interpretation. Instead of evaluating the image solely on current findings, the radiologist may attribute it to past behavior or outcomes, potentially leading to a missed diagnosis.*	AI may be reduced by focusing on current image data, not patient history

XSL•FO
RenderX

Cognitive bias	Explanation	Practical example	AI[a] contribution
Satisfaction of search	Stops searching for additional abnormalities after finding one	*A breast radiologist identifies a clear spiculated mass in the left breast on a mammogram and correctly reports it as suspicious for malignancy. Satisfied with having found an obvious abnormality, the radiologist stops looking carefully and misses a second suspicious finding: a small cluster of microcalcifications in the contralateral breast.*	AI could help by prompting a full image review
Inattentional blindness	Attention fixation causes missed visible findings	*A breast radiologist is reviewing a mammogram that was previewed by a resident, who believes that there are suspicious microcalcifications on the right upper quadrant; this primes the attending physician to look for calcifications in that specific quadrant. While intensely focused on identifying calcification patterns in that area, the radiologist fails to notice an opacity in a different quadrant.*	AI may highlight all findings objectively, reducing this bias
Hindsight	Overestimates the predictability of diagnosis based on outcome knowledge	*A breast radiologist is reviewing a mammogram taken 12 months before a patient was diagnosed with breast cancer. Knowing the eventual diagnosis, the radiologist now clearly identifies the lesion: a small spiculated mass. Nevertheless, at the time of the original screening, the lesion was subtle, not clearly identifiable due to the high breast density, and not highly suspicious without the knowledge of the outcome.*	AI may provide consistent, unbiased analysis, unaffected by outcome knowledge
Regret	Overcompensation due to fear of past diagnostic miss	*A breast radiologist previously missed a case of subtle ductal carcinoma in situ (DCIS) that led to delayed diagnosis. For this reason, in subsequent cases, the radiologists begin to define benign-appearing calcifications as suspicious, overrecommending biopsies, driven by a fear of repeating the error. Emotional reactions to a past mistake influence the radiologist's judgment: this is regret bias, and the desire to avoid the regret of another missed cancer leads to overcompensation, increasing unnecessary procedures and patient anxiety.*	AI could mitigate by providing probabilistic assessments, reducing emotion-driven bias

XSL•FO
RenderX

Cognitive bias	Explanation	Practical example	AI[a] contribution
Commission	Taking an action without an objective basis for the course of action	*A breast radiologist, while analyzing a clinical mammography of an asymptomatic patient with no personal or family history of breast cancer, identifies a vague asymmetry in the outer quadrant of the left breast, only visible in one projection and basically unchanged in comparison to previous exams, thus likely representing a normal glandular area. Despite the report of this exam being concluded as negative/ BI-RADS 1, the radiologist decided to assign a BI-RADS 3 to the finding, recalling the patient for additional examinations that were all negative.*	AI could mitigate commission bias by acting as a second objective reader, limiting excessive subjective assessments of benign findings

[a]AI: artificial intelligence.

[b]For biases marked as "theoretical/hypothesis-driven," current evidence is extrapolated from the broader literature on cognitive biases, AI-assisted decision-making, and diagnostic radiology. Direct empirical evidence specific to breast cancer imaging workflows is currently limited. Language such as "may," "could," and "theoretically" indicates effects that are conceptually plausible but await clinical validation.

One prominent cognitive bias is the availability bias, where radiologists may overemphasize experiences that are easily recalled to elaborate on the diagnosis, ignoring the base rate or prevalence rate of a particular disease [33,35,38,39]. Conversely, radiologists who have never experienced a specific symptom may exhibit a nonavailability bias, resulting in the oversight of potential diagnoses associated with that symptom [5].

AI could help to limit the availability bias, shifting the focus from anecdotal evidence to statistical patterns by processing comprehensive datasets that human cognition cannot effectively synthesize [32].

Similarly, the confirmation bias is another critical cognitive error where radiologists may selectively gather and interpret clinical data to reaffirm an existing hypothesis while disregarding alternative possibilities [5,35]. Also, in this context, AI may contribute to increasing the occurrence of this cognitive bias by encouraging radiologists to confirm AI-suggested diagnoses without adequately considering alternative possibilities [37]. Conversely, when AI is trained using controlled databases, it provides objective insights that are independent of human preconceptions, thereby reducing the frequency of this bias [32]. This bias is particularly concerning as it can significantly impair diagnostic accuracy.

An opposite example is provided by the anchoring bias that manifests when radiologists tend to fixate on preliminary information despite receiving subsequent data that conflicts with their initial diagnosis [5,33,38,40]. In this case, the use of AI could accentuate the probability of making errors, leading the radiologist to be influenced by the initial results provided by it [37]. This fixation can lead to a failure in adapting their assessment to new evidence.

A further type of bias associated with the adoption of AI in radiology is the automation bias. It consists of preferring automated responses despite the presence of conflicting information, effectively deferring decision-making competence to the automated system [5,12,37]. This phenomenon occurs when radiologists, especially those with less experience, overrely on AI systems, assuming that the absence of findings meant there was no abnormality, which can lead to missed diagnoses [4].

In a related vein, the satisfaction of report bias refers to the tendency of a radiologist to perpetuate radiological error that occurred in a previous assessment [5,39]. A radiologist could tend to accept a prior report or AI finding as sufficient, which prematurely halts the search for additional abnormalities. As a result, a second radiologist may be influenced by the judgment of a first radiologist or by AI tools, perpetuating mistakes. This does not mean that comparison with previous exams [41] or the use of AI systems should be avoided, as it can improve diagnostic accuracy. To mitigate these different types of errors, radiologists should always evaluate images independently before consulting past reports or AI interpretation outcomes.

One prominent cognitive bias is the attribution bias. It occurs when the characteristics of the patients, often shaped by stereotypes, inappropriately influence the radiologist's diagnostic reasoning process, causing them to neglect other peculiar and relevant information [5,35,39]. AI has the potential to reduce this bias by focusing solely on current imaging data, without being influenced by the patient's medical history or previous clinical context, thereby helping radiologists avoid inappropriate assumptions and improve diagnostic accuracy. Another possible bias is the framing bias that highlights the tendency to interpret a diagnosis according to how the clinical scenario is presented to the radiologist, leading them to misjudge the significance of the findings [5,39,42]. A specific symptom or diagnostic imaging finding located within one anatomical-physiological system may potentially divert the radiologist's attention from the involvement of an alternative system [43]. In this case, the dataset on which the AI is trained must be optimal [36]. If the AI is trained on clinical datasets that contain inherent biases or distorted presentations, it could inadvertently reinforce framing bias by reflecting the same distortions in its analyses and recommendations.

XSL•FO
RenderX

Furthermore, the satisfaction of search bias occurs when a radiologist ceases a visual search for additional clinically significant abnormalities upon the initial pathology being identified [5,35,43,44]. This can lead to missed diagnoses of additional issues that may also be present. If the radiologist merely responds to the attending physician's question, other important findings may be overlooked if the radiologist does not perform a comprehensive assessment [43]. AI tools can counteract this phenomenon by prompting comprehensive image reviews, flagging secondary findings, and encouraging comprehensive assessment.

Similarly, some symptoms or entirely visible and clinically significant findings can be neglected because one's attention is fixed on another task or object (ie, inattentional blindness bias) [35,39]. Compelling evidence for this phenomenon was observed in a study conducted in the radiological field, where 24 radiologists were asked to detect lung nodules. In the last CT scan, a gorilla 48 times larger than the average nodule was inserted. About 83% of radiologists failed in their recognition [45,46]. This error occurred not because the lesion was subtle or invisible, but because attention was narrowly directed to the suspected finding, leading to inattentional blindness: the failure to see a visible but unexpected object when attention is engaged elsewhere. This type of bias could be reduced by AI systems that objectively highlight all anomalies detected in the image, drawing attention to findings that radiologists might overlook due to the limitations of cognitive concentration.

The phenomenon of hindsight also plays a crucial role in radiological errors, as radiologists may underestimate the complexity of reaching the initial diagnosis and overestimate the probability of a diagnosis based on prior knowledge of the outcome [5,35,43].

The radiologist's judgment during the review is influenced by hindsight bias, the tendency to see events as more predictable or obvious after they have already occurred. AI, relying on algorithms and statistical models rather than retrospective knowledge, provides consistent and unbiased analysis that is not influenced by previous results, potentially protecting against this bias.

Conversely, regret bias leads to the overreporting of certain diseases due to the negative consequences associated with missed diagnoses [5,43]. Similar to the previous bias, AI can mitigate this emotional and experience-based bias by offering probabilistic assessments based on large datasets, helping physicians strike a balance between caution and evidence-based reasoning.

Another bias to which breast radiologists are commonly exposed is the commission bias, representing the urge to do something that might be unnecessary, such as recalling a patient, ordering additional imaging, or recommending a biopsy, instead of doing nothing [43]. The commission bias might happen in different contexts, but breast imaging is very vulnerable to this bias due to the high psychological, clinical, and medicolegal pressures of this field. Nevertheless, this approach might lead to avoidable expenses and increase the likelihood of false-positive findings, which may prompt unwarranted biopsies or additional imaging, ultimately outweighing any perceived benefit for these patients [47]. AI can mitigate commission bias by acting as an objective second reader, providing quantitative malignancy probability scores to challenge subjective overcalls on benign asymmetries.

Finally, premature closure bias arises when a radiologist prematurely accepts a diagnosis as definitive before any alternative diagnoses are considered [5,35,39]. When trained with adequate datasets [36], AI can assist by suggesting alternative diagnoses based on image characteristics and differential analyses, encouraging radiologists to maintain an open diagnostic approach and avoid hasty conclusions.

In this scenario, the concept of algorithm aversion has also become increasingly relevant. This term refers to the psychological tendency of individuals to distrust and avoid algorithmic decision support, especially after observing even minor errors, despite evidence that algorithms often outperform human judgment. Foundational work by Dietvorst and colleagues [48] demonstrated that when individuals witness predictive algorithms making mistakes, they subsequently prefer human forecasters. Building on this, Dietvorst and Bharti [49] demonstrated that a diminishing sensitivity to forecasting error further reinforces algorithm rejection in uncertain decision-making domains, as people overweight early errors and underweight later accuracy improvements.

Similar dynamics can be observed in breast radiology, where radiologists' interactions with diagnostic AI systems critically shape trust and subsequent use. Early AI misclassifications can disproportionately reduce reliance on algorithmic output, particularly among clinicians with high diagnostic self-efficacy [50]. These patterns of algorithm aversion may hinder the integration of AI tools into routine clinical practice. Consequently, radiologists may continue to favor their own judgments despite well-known human limitations such as fatigue, perceptual bias, and interreader variability. In breast imaging specifically, algorithm aversion may lead to missed opportunities for AI-assisted early cancer detection, reduced returns on investment in AI technologies, and slower adoption of tools capable of improving population-level outcomes.

Preventive Strategies and Future Directions

Preventive strategies aimed at reducing potential physicians' errors represent a critical topic and should be developed based on the most common errors encountered in clinical practice [6].

A fundamental consideration is providing comprehensive education on AI, increasing AI literacy, and enabling individuals to make informed choices [4]. An appropriate education allows health care professionals to be equipped with awareness of its risks and benefits, along with the necessary skills to identify, prevent, and address potential failures arising from its application in radiology [4]. Specifically, targeted training programs designed to reduce certain cognitive errors are essential [51]. For instance, Taussig and colleagues [52] proposed an innovative approach where senior residents present selected interpretative errors to junior residents, fostering peer discussion that enhances knowledge and helps prevent the recurrence of errors.

XSL•FO
RenderX

Continuing education should also foster insight and metacognitive skills by including structured explanations of cognitive biases and clinical examples illustrating their effect on diagnostic reasoning [51]. Encouraging clinicians to critically reflect on their diagnostic processes, rather than relying on immediate problem-solving, can improve diagnostic accuracy. Simulation techniques, such as cognitive exercises and mental rehearsals, further support this learning by exposing trainees to the impacts of biased versus unbiased reasoning [51]. Additionally, training videos and case comparisons reinforce debiasing strategies and facilitate the transfer of these skills to everyday practice.

Another effective strategy to reduce the frequency of cognitive bias in radiological report interpretation is the establishment of a clear communication strategy among health care professionals. Indeed, ineffective communication due to inefficient team interaction or interprofessional tension increases the risk of making medical errors [53]. The European Society of Radiology published in 2013 communication guidelines to provide helpful information on conducting an effective discussion between patients, referrers, colleagues, and students [54]. Furthermore, improving communication between clinicians and radiologists promotes their collaboration, which, in turn, might help in preventing cognitive errors such as attribution, anchoring, or confirmation bias caused by the lack of knowledge about the patient's medical history and proper clinical context, with subsequent incorrect interpretation of the images. Effective communication and discussion with colleagues also allow for timely feedback regarding the clinical decision and enable an understanding of the error and implementation of strategies to correct it [51].

Beyond education and communication, bias mitigation should also be addressed at a structural level through workflow engineering and explainable artificial intelligence (XAI).

Integrating AI into clinical engineering workflow in medical imaging entails structured interpretation sequences through strategic design of reading protocols to optimize diagnostic accuracy while counteracting cognitive biases [55,56]. Sequential workflows require radiologists to interpret images independently before consulting AI-generated marks in a second-reader paradigm, enhancing sensitivity but potentially reducing specificity due to increased false positives and extended reading times [57]. It preserves independent judgment to prevent bias, such as premature closure.

Concurrent integration, by contrast, offers real-time validation against heuristics like anchoring on salient features. Concurrent reading integrates AI annotations during initial interpretation, improving both sensitivity and specificity while reducing reading time [57].

Furthermore, XAI approaches can foster calibrated trust by enabling clinicians to interrogate the rationale and uncertainty of AI outputs rather than passively following algorithmic recommendations. XAI encompasses methods that make deep learning models intelligible to human users, clarifying their internal logic, decision pathways, and case-specific predictions, as well as their capabilities and limitations, instead of treating them as opaque black boxes [58]. In medical imaging, XAI does not remove bias but supports its detection, interpretation, and mitigation by revealing model decision rules and the data patterns on which they rely. By providing interpretable feature attributions, XAI helps counter cognitive biases such as the availability heuristic and anchoring, allowing clinicians to cross-check AI suggestions against the actual imaging findings and to avoid overreliance on salient anomalies [32]. These techniques jointly target algorithmic biases originating from datasets, training procedures, and generalization failures, as well as clinicians' own cognitive biases, thereby promoting a more critical and informed use of algorithmic outputs and enhancing diagnostic accuracy while preserving transparency, including for underrepresented patient groups in training data [32,58].

Integrating AI into radiological practice enhances diagnostic workflows and operational efficiency by reducing workload and optimizing time, critical factors in preventing burnout and fatigue-related errors [56]. The study conducted by Bruno and collaborators explored possible approaches to minimize errors in the field of radiology, including a reduction of working hours, a minimization of interruptions and distractions, and an alleviation of the pressure to maintain a rapid workflow [59]. However, such interventions face practical challenges in some hospital settings and have shown minimal impact. Systematic studies are needed to evaluate how AI can be effectively incorporated into image interpretation processes to support accurate and efficient everyday clinical practice.

Conclusion

The integration of AI in breast imaging represents a promising advancement in radiology, with the potential to improve diagnostic accuracy, optimize workflow, and reduce perceptual, attentional, and interpretive errors. However, as this perspective highlights, AI is not a flawless solution and introduces new challenges, including cognitive biases that may influence radiologists' decision-making. Understanding these biases is crucial to implementing strategies to mitigate these errors, ensuring that AI serves as an effective aid rather than a source of further diagnostic problems.

While AI offers substantial advantages, a cautious approach is required to mitigate its potential risks and limitations. To fully exploit the benefits of AI while minimizing its limitations, further research should focus on optimizing the AI integration into clinical workflows, assessing its impact on diagnostic reasoning, and developing structured guidelines to mitigate automation biases and other cognitive distortions. A successful AI implementation in breast imaging will ultimately require a synergistic approach that blends human expertise with technological innovation. A balanced approach harnessing AI as a complementary tool while maintaining radiologists' critical thinking and expertise will improve breast cancer diagnosis.

XSL•FO
RenderX

Acknowledgments

BC is a PhD student within the European School of Molecular Medicine (SEMM). No generative artificial intelligence tools were used in the conduct of this study.

Funding

The present work was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds for the European Institute of Oncology IRCSS.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed.

Authors' Contributions

LC, BC, FP, and DM conceived and discussed the initial idea. LC, BC, OB, and FP wrote the first draft of the manuscript. DM and RG contributed to the manuscript through discussion and provided valuable insights. GP supervised the entire process and provided feedback on both the topic and the manuscript structure. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search.
[DOCX File, 13 KB - medinform_v14i1e78955_app1.docx]

References

1. Chadha S, Mukherjee S, Sanyal S. Advancements and implications of artificial intelligence for early detection, diagnosis and tailored treatment of cancer. Semin Oncol 2025 Jun;52(3):152349. [doi: 10.1016/j.seminoncol.2025.152349] [Medline: 40345002]
2. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
3. Ogut E. Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. Clin Pract 2025 Sep 16;15(9):169. [doi: 10.3390/clinpract15090169] [Medline: 41002784]
4. Stogiannos N, Cuocolo R, Akinci D'Antonoli T, et al. Recognising errors in AI implementation in radiology: a narrative review. Eur J Radiol 2025 Oct;191:112311. [doi: 10.1016/j.ejrad.2025.112311]
5. Chen J, Gandomkar Z, Reed WM. Investigating the impact of cognitive biases in radiologists' image interpretation: a scoping review. Eur J Radiol 2023 Sep;166:111013. [doi: 10.1016/j.ejrad.2023.111013] [Medline: 37541180]
6. Pesapane F, Gnocchi G, Quarrella C, et al. Errors in radiology: a standard review. J Clin Med 2024 Jul 23;13(15):4306. [doi: 10.3390/jcm13154306] [Medline: 39124573]
7. Triberti S, Durosini I, Pravettoni G. A "third wheel" effect in health decision making involving artificial entities: a psychological perspective. Front Public Health 2020;8:117. [doi: 10.3389/fpubh.2020.00117] [Medline: 32411641]
8. Itri JN, Patel SH. Heuristics and cognitive error in medical imaging. Am J Roentgenol 2018 May;210(5):1097-1105. [doi: 10.2214/AJR.17.18907]
9. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017 Feb;8(1):171-182. [doi: 10.1007/s13244-016-0534-1] [Medline: 27928712]
10. Siewert B, Sosna J, McNamara A, Raptopoulos V, Kruskal JB. Missed lesions at abdominal oncologic CT: lessons learned from quality assurance. Radiographics 2008;28(3):623-638. [doi: 10.1148/rg.283075188] [Medline: 18480475]
11. Waite S, Farooq Z, Grigorian A, et al. A review of perceptual expertise in radiology – how it develops, how we can test it, and why humans still matter in the era of artificial intelligence. Acad Radiol 2020 Jan;27(1):26-38. [doi: 10.1016/j.acra.2019.08.018] [Medline: 31818384]
12. Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and mitigating bias in imaging artificial intelligence. Radiographics 2024 May 1;44(5):e230067. [doi: 10.1148/rg.230067]
13. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
14. Bevers TB, Helvie M, Bonaccio E, et al. Breast Cancer Screening and Diagnosis, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018 Nov;16(11):1362-1389. [doi: 10.6004/jnccn.2018.0083] [Medline: 30442736]

XSL•FO
RenderX

15. Schünemann HJ, Lerda D, Quinn C, et al. Breast cancer screening and diagnosis: a synopsis of the European Breast Guidelines. Ann Intern Med 2020 Jan 7;172(1):46-56. [doi: 10.7326/M19-2125] [Medline: 31766052]
16. Gulland A. Staff shortages are putting UK breast cancer screening "at risk," survey finds. BMJ 2016 Apr 25:i2350. [doi: 10.1136/bmj.i2350]
17. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018 Oct 24;2(1):35. [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
18. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019 Sep 1;111(9):916-922. [doi: 10.1093/jnci/djy222] [Medline: 30834436]
19. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature New Biol 2020 Jan;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
20. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021 Sep 1;374:n1872. [doi: 10.1136/bmj.n1872] [Medline: 34470740]
21. Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. Radiology 2022 Jun;303(3):502-511. [doi: 10.1148/radiol.212381]
22. Byng D, Strauch B, Gnas L, et al. AI-based prevention of interval cancers in a national mammography screening program. Eur J Radiol 2022 Jul;152:110321. [doi: 10.1016/j.ejrad.2022.110321] [Medline: 35512511]
23. Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur Radiol 2021 Aug;31(8):5940-5947. [doi: 10.1007/s00330-021-07686-3] [Medline: 33486604]
24. Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol 2023 Aug;24(8):936-944. [doi: 10.1016/S1470-2045(23)00298-X] [Medline: 37541274]
25. Gur D, Nishikawa RM, Sumkin JH. New screening technologies and practices: a different approach to estimation of performance improvement by using data from the transition period. Radiology 2015 Apr;275(1):9-12. [doi: 10.1148/radiol.14141843] [Medline: 25799332]
26. Dembrower K, Crippa A, Colón E, Eklund M, Strand F, ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health 2023 Oct;5(10):e703-e711. [doi: 10.1016/S2589-7500(23)00153-X] [Medline: 37690911]
27. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020 Mar;2(3):e138-e148. [doi: 10.1016/S2589-7500(20)30003-0] [Medline: 33334578]
28. Anderson AW, Marinovich ML, Houssami N, et al. Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: a systematic review. J Am Coll Radiol 2022 Feb;19(2 Pt A):259-273. [doi: 10.1016/j.jacr.2021.11.008] [Medline: 35065909]
29. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020 Mar 2;3(3):e200265. [doi: 10.1001/jamanetworkopen.2020.0265] [Medline: 32119094]
30. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020 Oct 1;6(10):1581-1588. [doi: 10.1001/jamaoncol.2020.3321] [Medline: 32852536]
31. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. Expert Rev Med Devices 2019 May;16(5):351-362. [doi: 10.1080/17434440.2019.1610387] [Medline: 30999781]
32. Theodorakopoulos L, Theodoropoulou A, Halkiopoulos C. Cognitive bias mitigation in executive decision-making: a data-driven approach integrating big data analytics, AI, and explainable systems. Electronics (Basel) 2025 Oct 3;14(19):3930. [doi: 10.3390/electronics14193930]
33. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. Science 1974;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124]
34. Da Silva S, Gupta R, Monzani D. Editorial: Highlights in psychology: cognitive bias. Front Psychol 2023;14:1242809. [doi: 10.3389/fpsyg.2023.1242809] [Medline: 37469886]
35. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. Radiographics 2018;38(1):236-247. [doi: 10.1148/rg.2018170107] [Medline: 29194009]
36. Shen X, Shi Z, Lu J, Li B, Yang Q. Artificial intelligence in medical image diagnosis: advances, challenges, and future perspectives. Artif Intell Med 2025;1(1):28-39. [doi: 10.37420/j.jaim.2025.004]
37. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol 2025 Mar 3;31(2):75-88. [doi: 10.4274/dir.2024.242854] [Medline: 38953330]

XSL•FO
RenderX

38. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016 Nov 3;16(1):138. [doi: 10.1186/s12911-016-0377-1] [Medline: 27809908]
39. Tee QX, Nambiar M, Stuckey S. Error and cognitive bias in diagnostic radiology. J Med Imaging Radiat Oncol 2022 Mar;66(2):202-207. [doi: 10.1111/1754-9485.13320] [Medline: 34467643]
40. Rastogi C, Zhang Y, Wei D, Varshney KR, Dhurandhar A, Tomsett R. Deciding fast and slow: the role of cognitive biases in AI-assisted decision-making. Proc ACM Hum-Comput Interact 2022 Mar 30;6(CSCW1):1-22. [doi: 10.1145/3512930]
41. White K, Berbaum K, Smith WL. The role of previous radiographs and reports in the interpretation of current radiographs. Invest Radiol 1994 Mar;29(3):263-265. [doi: 10.1097/00004424-199403000-00002] [Medline: 8175298]
42. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. Am J Roentgenol 2013 Sep;201(3):611-617. [doi: 10.2214/AJR.12.10375]
43. Gunderman RB. Biases in radiologic reasoning. Am J Roentgenol 2009 Mar;192(3):561-564. [doi: 10.2214/AJR.08.1220]
44. Berbaum KS, Franken EA Jr, Dorfman DD, et al. Satisfaction of search in diagnostic radiology. Invest Radiol 1990 Feb;25(2):133-139. [doi: 10.1097/00004424-199002000-00006]
45. Simons DJ, Chabris CF. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception 1999;28(9):1059-1074. [doi: 10.1068/p281059] [Medline: 10694957]
46. Drew T, Võ MLH, Wolfe JM. The invisible gorilla strikes again: sustained inattentional blindness in expert observers. Psychol Sci 2013 Sep;24(9):1848-1853. [doi: 10.1177/0956797613479386] [Medline: 23863753]
47. Loving VA, Valencia EM, Patel B, Johnston BS. The role of cognitive bias in breast radiology diagnostic and judgment errors. J Breast Imaging 2020 Aug 10;2(4):382-389. [doi: 10.1093/jbi/wbaa023] [Medline: 38424956]
48. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015 Feb;144(1):114-126. [doi: 10.1037/xge0000033] [Medline: 25401381]
49. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 2020 Oct;31(10):1302-1314. [doi: 10.1177/0956797620948841] [Medline: 32916083]
50. Jussupow E, Spohrer K, Heinzl A. Radiologists' usage of diagnostic AI systems: the role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. Bus Inf Syst Eng 2022 Jun;64(3):293-309. [doi: 10.1007/s12599-022-00750-2]
51. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003 Aug;78(8):775-780. [doi: 10.1097/00001888-200308000-00003] [Medline: 12915363]
52. Taussig M, Musick A, Dondlinger S, et al. Learning from errors: implementation of a resident-oriented radiology morbidity and mortality conference as an educational tool. Clin Imaging 2022 Apr;84:98-103. [doi: 10.1016/j.clinimag.2022.01.011] [Medline: 35168086]
53. Scherer K, Winokur RS. Multidisciplinary team training simulation in interventional radiology. Tech Vasc Interv Radiol 2019 Mar;22(1):32-34. [doi: 10.1053/j.tvir.2018.10.007] [Medline: 30765074]
54. European Society of Radiology (ESR). ESR communication guidelines for radiologists. Insights Imaging 2013 Apr;4(2):143-146. [doi: 10.1007/s13244-013-0218-z]
55. Blezek DJ, Olson-Williams L, Missert A, Korfiatis P. AI integration in the clinical workflow. J Digit Imaging 2021 Dec;34(6):1435-1446. [doi: 10.1007/s10278-021-00525-3] [Medline: 34686923]
56. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
57. Wenderott K, Krups J, Zaruchas F, Weigl M. Effects of artificial intelligence implementation on efficiency in medical imaging-a systematic literature review and meta-analysis. NPJ Digit Med 2024 Sep 30;7(1):265. [doi: 10.1038/s41746-024-01248-9] [Medline: 39349815]
58. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - beyond saliency-based XAI approaches. Eur J Radiol 2023 May;162:110786. [doi: 10.1016/j.ejrad.2023.110786] [Medline: 36990051]
59. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015 Oct;35(6):1668-1676. [doi: 10.1148/rg.2015150023] [Medline: 26466178]

Abbreviations

AI: artificial intelligence
CAD: computer-aided detection
CT: computed tomography
DBT: digital breast tomosynthesis
MASAI: mammography screening with artificial intelligence
XAI: explainable artificial intelligence

XSL•FO
RenderX

Edited by A Coristine; submitted 12.Jun.2025; peer-reviewed by L Mosca, LF Nakayama, R Toomey; accepted 27.Jan.2026; published 30.Mar.2026.

Please cite as:
Conti L, Capetti B, Battaglia O, Grasso R, Pesapane F, Monzani D, Pravettoni G
Viewpoint on the Consequences and Mitigation of Cognitive Bias in the Radiological Interpretation of Breast Cancer Imaging Using Artificial Intelligence
JMIR Med Inform 2026;14:e78955
URL: https://medinform.jmir.org/2026/1/e78955
doi:10.2196/78955

© Lorenzo Conti, Benedetta Capetti, Ottavia Battaglia, Roberto Grasso, Filippo Pesapane, Dario Monzani, Gabriella Pravettoni. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ethical Imperatives for Retrieval-Augmented Generation in Clinical Nursing: Viewpoint on Responsible AI Use

Xinyi Tu[1,2*], BSN; Chenghao Shi[1*], MN; Peilin Qian[3], BSN; Lizhu Wang[1], MPH

[1]Department of Nursing, The Second Affiliated Hospital of Zhejiang University School of Medicine, No. 88 Jiefang Road, Shangcheng District, Hangzhou, Zhejiang, China

[2]School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China

[3]School of Nursing, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Lizhu Wang, MPH
Department of Nursing, The Second Affiliated Hospital of Zhejiang University School of Medicine, No. 88 Jiefang Road, Shangcheng District, Hangzhou, Zhejiang, China

Abstract

Retrieval-augmented generation (RAG) systems have emerged as a powerful technique to enhance the capabilities of large language models by enabling them to access external, up-to-date knowledge in real time, and RAG systems are being increasingly adopted by researchers in the medical field. In this viewpoint article, we explore the ethical imperatives for implementing RAG systems in clinical nursing environments, with particular attention to how these technologies affect patient care quality and safety. The purpose of this paper is to examine the ethical risks introduced by RAG-enhanced large language models in clinical nursing and to propose strategic guidelines for their responsible implementation. Key considerations include ensuring accuracy, fairness, transparency, and accountability, as well as maintaining essential human oversight, as discussed through a structured analysis. We argue that robust data governance, explainable artificial intelligence (AI) techniques, and continuous monitoring are critical components of a responsible RAG implementation strategy. Ultimately, realizing the benefits of RAG while mitigating ethical concerns requires sustained collaboration among health care professionals, AI developers, and policymakers, fostering a future where AI supports patient safety, reduces disparities, and improves the quality of nursing care.

(JMIR Med Inform 2026;14:e79922) doi:10.2196/79922

KEYWORDS

large language models; clinical decision-making; ethics; fairness; transparency

Introduction

Advances in artificial intelligence (AI) are rapidly transforming the health care field, with large language models (LLMs) playing a pivotal role in revolutionizing nursing practice [1,2]. These models enhance human capabilities by learning patterns from large amounts of text data and generating contextually relevant information, offering benefits such as improved decision support, streamlined care planning, and personalized patient education [3]. LLMs can assist nurses by providing comprehensive insights and generating evidence-based recommendations. However, their integration into clinical workflows faces significant challenges [4], partly because LLMs do not operate like traditional rule-based programs. Instead, they generate responses based on statistical patterns learned from large text corpora, rather than performing logical reasoning or interpreting meaning as humans do.

While this generative capability allows LLMs to produce contextually relevant and fluent text, it also gives rise to several critical limitations in specialized clinical applications. These include biases inherited from training data and the "black-box" nature of their decision-making processes, which make it difficult to understand how outputs are produced. Such issues lead to concerns about trust and transparency, ultimately undermining the reliability of model-generated recommendations in high-stakes health care environments [5,6].

To address these shortcomings, the retrieval-augmented generation (RAG) system has emerged as an enhancement to LLMs. By combining external knowledge retrieval with the language patterns an LLM has learned during training, RAG can provide more accurate, up-to-date, and context-sensitive information [7]. In this paper, the term "RAG system" refers to the practical implementation of the RAG approach, in which an LLM retrieves external information from databases or documents and integrates it with the patterns learned during training to generate context-aware responses. While this approach aims to overcome the static nature of traditional LLMs, it introduces new ethical risks, particularly related to the quality and reliability of external data sources, the complexity of information traceability, and privacy concerns [4].

XSL•FO
RenderX

This article examines the ethical challenges introduced by RAG-optimized LLMs in clinical nursing and proposes strategies for responsible implementation. We focus on three core ethical imperatives: ensuring accuracy, fairness, and bias mitigation; promoting transparency, explainability, and trust; and maintaining responsibility, accountability, and oversight. By addressing these challenges proactively, we can harness the benefits of RAG to enhance nursing practice while safeguarding patient well-being.

The Promise and Peril of RAG in Nursing

Overview

RAG offers significant opportunities to enhance clinical nursing practice by providing more accurate, up-to-date, and context-specific information. By combining the language-pattern learning capability of LLMs with external knowledge retrieval, RAG can address some of the limitations of traditional LLMs [8]. However, alongside its potential benefits, RAG also introduces unique risks that need to be carefully considered, particularly in the context of clinical nursing.

The Promise of RAG in Nursing

The primary advantage of RAG lies in its ability to enhance decision support by integrating real-time, relevant knowledge from external sources. In high-pressure environments, such as hospitals, where quick decision-making is critical, RAG can significantly improve the accuracy of clinical recommendations by ensuring they are based on the latest research and guidelines [7,9]. By enabling real-time access to updated information, RAG helps reduce errors that may arise from outdated or incomplete data [9].

Additionally, RAG systems build on what the LLM has learned from language data to provide personalized and context-specific recommendations by incorporating patient-specific data, such as medical history, treatment plans, and current health status. This ability to tailor recommendations to the individual needs of patients ensures that health care providers can deliver more accurate and effective care [10]. In settings with limited resources, RAG can provide vital support for nurses by offering evidence-based recommendations that might otherwise be difficult to access due to time constraints or a lack of available resources [9].

The Peril of RAG in Nursing

Overview of Risks

Despite its advantages, the use of RAG systems in nursing also introduces a range of ethical challenges and risks. Many of these challenges stem from the limitations of the underlying LLMs on which RAG is built—including issues such as bias, opacity, and a lack of accountability. At the same time, RAG introduces additional, system-specific risks because it retrieves and integrates external information dynamically at query time. These additional risks center on data quality, source reliability, semantic coherence, and information provenance and must be carefully addressed in clinical applications.

Common Risks of LLMs

Many of the risks associated with RAG are rooted in the inherent challenges of LLMs, which include the following:

- Bias and fairness: LLMs may perpetuate biases that are present in their training data, such as racial, gender, or cultural biases [11]. Since RAG systems rely on external sources, they can also inherit and even amplify these biases if the external data sources are not sufficiently diverse or representative [4]. This is particularly concerning in nursing, where such biases can lead to unequal treatment and care disparities, potentially exacerbating health inequalities [12].
- Transparency and explainability: The “black-box” nature of LLMs remains a significant issue, making it difficult for clinicians to understand how the model produces its responses [13,14]. While RAG can improve transparency by citing the sources of external information, the process may still be opaque. This lack of explainability can undermine trust in the system, especially when nurses are unable to interpret the reasoning behind a recommendation [5,15].
- Accountability: Another concern is who should bear responsibility when a RAG system provides a recommendation that leads to an error or harm. It may be unclear who is responsible—whether it is the developers of the model, the clinical staff who implemented it, or the health care institution as a whole. This ambiguity complicates accountability and the governance of AI-driven health care systems [16].

Risks Specific to RAG

While many of the risks in RAG are shared with LLMs, there are several risks that are unique to RAG systems due to their reliance on outside information sources.

- Information provenance and quality: RAG systems pull data from many different websites, databases, and documents. Some of these sources may not undergo rigorous peer review, and there may be discrepancies in the information retrieved [17,18]. When the information used by RAG is incomplete or unreliable, the guidance it gives can also be misleading. It can also be difficult to check exactly where the information came from, which makes it harder for nurses to judge whether a recommendation can be trusted [19].
- Semantic integration and coherence: When a RAG system combines information drawn from many sources, the final advice may sound uneven or use different styles of wording. For a nurse using it, this means different parts of the output may suggest slightly different approaches, which can create confusion or reduce confidence in applying the recommendation [20]. Merging text generated from different materials may result in conflicting or incompatible pieces of information, which can cause confusion or misinterpretation [21]. This issue is particularly problematic in clinical contexts, where accuracy and consistency are essential for patient safety [22].
- Privacy and security risks: To offer personalized advice, RAG systems may need access to sensitive patient data, such as medical histories, to provide personalized

XSL•FO
RenderX

recommendations [18,23]. If the system's security is weak, or if it does not meet privacy standards like HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation), confidential information could be exposed. Such data breaches could harm both patients and the health care institution [5].

- Shifting trust foundations: Traditional LLMs depend only on the language relationships learned during training, but RAG depends on both internal and external sources. This means that the reliability of RAG depends not only on the model itself but also on the trustworthiness of the outside information it uses [24,25]. If that outside data is wrong or biased, the recommendations produced by the RAG system may be wrong too, reducing nurses' confidence in using the system [18,19,26].

An example is as follows: A nurse used a RAG-based tool to check evidence for pressure injury prevention. The system retrieved the 2014 second edition of the international pressure ulcer guideline, which recommended turning the patient every 2 hours on standard foam surfaces. However, the upcoming 2025 fourth edition emphasizes individualized repositioning schedules, the use of alternating-pressure mattresses, and closer monitoring for deep tissue injury in high-risk or obese patients. Because the RAG system relied on outdated data, the nurse followed the older 2-hour rule and missed early signs of a deep tissue injury. This case shows how outdated retrieval sources can lead to substandard care and highlights the need for RAG systems to verify and update their evidence sources regularly.

Focused Ethical Challenges

Overview

The introduction of RAG systems into clinical nursing presents several ethical challenges that need to be carefully considered. These challenges are not only rooted in the inherent limitations of how LLMs learn from data but also amplified by the integration of external knowledge sources [18,27]. In this section, we focus on three core ethical challenges: accuracy, bias, and fairness; transparency, explainability, and trust; and responsibility, accountability, and human oversight. These challenges must be addressed to ensure that RAG-enhanced LLMs support ethically evidence-based decisions in nursing practice [28].

Accuracy, Bias, and Fairness

One key ethical concern when using RAG systems is maintaining the accuracy of the information they generate, especially for clinical decisions [3,19]. The accuracy of RAG depends directly on how reliable and up-to-date the outside sources are [9]. If the outside sources are outdated, incomplete, or inaccurate, the systems' recommendations may mislead nurses and lead to unsafe decisions that could harm patients [18,19]. Therefore, RAG systems should only use checked and high-quality data that fits the medical context [18].

Additionally, bias remains a significant challenge. RAG systems can carry bias from both the training data used to develop the LLMs and the outside sources it pulls information from [26,29]. These biases can manifest in various forms, such as racial, gender, or socioeconomic biases, and can show up as racial, gender, or economic differences and may cause unfair care for some patients [30,31]. In clinical settings, where health gaps already exist, AI-related bias can make these gaps worse. For example, if RAG systems draw information from databases that underrepresent minority populations, the resulting recommendations may be less accurate or effective for these groups [32].

To reduce these risks, we need continued work to make AI systems fair [16,18]. This means checking that outside sources cover different patient groups and reviewing the system's outputs regularly for bias. Furthermore, health care institutions and developers should also focus on including a wide range of patients in AI training data and retrieval sources to avoid making health gaps worse [33,34].

Transparency, Explainability, and Trust

Another significant ethical challenge is transparency and explainability. While RAG systems can increase transparency by showing the sources of information, the process by which these models generate their responses is still not very clear [18,35]. Because their internal workings are complex and often hidden, it can be hard for clinicians to see how external data and built-in knowledge come together to form a recommendation [36]. This lack of clarity can reduce trust in the system, since clinicians might hesitate to use a tool whose reasoning process cannot be easily interpreted [15,37,38].

Explainability is particularly important in clinical settings, where nurses and health care professionals need to understand the basis for AI-generated suggestions in order to trust them and incorporate them into their decision-making process [15,39,40]. It is essential for RAG systems to explain their answers in a way that is easy for health care professionals to understand. For example, providing confidence scores or source citations can help clinicians evaluate the reliability of the information and determine whether it fits the patient's situation [38,41].

Enhancing transparency and explainability will not only help build trust in AI systems but also foster responsible and ethical AI adoption in clinical practice. As RAG systems become more integrated into health care systems, transparency will be key to ensuring that their use aligns with professional standards and ethical guidelines [6].

Responsibility, Accountability, and Human Oversight

Responsibility and accountability are key ethical concerns in the context of AI-driven decision-making. One of the main challenges in using RAG systems is that responsibility can become blurred when many parties—such as AI developers, health care providers, and institutions—take part in the same process [42]. If an error occurs because of a recommendation from a RAG system, it may be unclear who should bear responsibility for it. This uncertainty can make legal and ethical management of such situations more difficult [43,44].

To address this issue, it is important to have clear guidelines that define the roles and duties of everyone involved in the AI decision-making process [45]. Health care institutions must ensure that there are well-defined accountability structures in

XSL•FO
RenderX

place to determine who is responsible when AI systems make errors [6]. Furthermore, RAG systems should be designed with human supervision in mind. While they can enhance decision support, they should never replace human judgment. Nurses and clinicians must be able to step in or question the AI-generated recommendations if they believe the advice is not in the patients' best interest [46,47].

Keeping humans actively in each stage of the process, which is often called "human-in-the-loop" oversight, is crucial to maintaining ethical standards in AI-assisted clinical care [48,49]. This approach ensures that clinical decisions are ultimately guided by human expertise, with AI acting only as a support tool. To support this approach, ongoing training for health care professionals on the use of AI systems, along with mechanisms for reporting and addressing errors, will help uphold ethical standards and ensure that the use of RAG systems benefits patients while minimizing risks.

Addressing these ethical concerns requires not only awareness but also systematic strategies to guide responsible use of RAG systems in nursing practice.

Strategies for Responsible RAG Implementation

Overview

Building on the ethical challenges discussed in the *Focused Ethical Challenges* section, it is essential to implement strategies that ensure the responsible and effective use of RAG systems in clinical nursing. While RAG systems offer significant advantages, such as real-time decision support and personalized care recommendations, they also introduce unique risks. To maximize the benefits of RAG and minimize its ethical concerns, we propose several key strategies for responsible implementation. These strategies are designed as direct responses to the ethical challenges identified in the *Focused Ethical Challenges* section, translating theoretical imperatives into actionable guidance for nursing practice.

Data Governance and Bias Mitigation

Building on the discussion of bias and fairness in *Accuracy, Bias, and Fairness*, effective data governance provides the foundation for responsible RAG implementation. Because RAG systems rely on external knowledge sources, ensuring the integrity and representativeness of these data is essential. Biases in the external data used for retrieval can have a significant impact on the outputs generated by the system, potentially leading to unfair or discriminatory outcomes [12,50,51].

To mitigate these risks, bias detection and correction mechanisms must be implemented at multiple stages of the RAG system's development and deployment. Specifically, RAG systems should incorporate bias auditing practices to identify potential biases in the external datasets used for retrieval. This includes regular audits of the data sources to ensure they are inclusive and reflect a broad range of patient demographics and clinical scenarios [12,50]. Additionally, techniques such as fairness-aware machine learning and adversarial debiasing can be employed to reduce the impact of bias in the model's generated recommendations. By continuously monitoring and correcting for bias, health care organizations can help ensure that RAG systems provide fair and equitable information support for all patients [43].

Explainable AI and Transparent Decision-Making

Building on the transparency challenges identified earlier, explainable AI techniques should be integrated into RAG systems to enhance transparency and help nurses and clinicians trust the information these systems produce. By improving the explainability of RAG systems, health care professionals can better assess the basis of AI-generated suggestions and integrate them into their decision-making processes [5].

One effective way to enhance transparency is by incorporating confidence scores and contextual explanations. Confidence scores provide an indication of how certain the system is about a given output, giving clinicians a clearer understanding of the model's reliability [52]. Contextual explanations, on the other hand, help clarify the basis for a given response by linking it to relevant patient data or clinical guidelines. For instance, a RAG system might explain that a particular medication recommendation is based on the patient's medical history, current health status, and recent clinical studies [18]. This additional layer of information will not only improve trust in the system but also empower nurses to make informed decisions based on AI recommendations [47].

Operationalizing Human Oversight and Collaborative Intelligence

Extending the discussion of accountability from the *Responsibility, Accountability, and Human Oversight* section, it is crucial to emphasize the role of human-in-the-loop oversight. RAG systems should not replace human judgment but rather augment it. Nurses, doctors, and other health care professionals should remain central to the decision-making process, with RAG systems serving as assistive tools to support clinical judgment rather than a substitute for clinical expertise [9,19,26].

To ensure the effectiveness of human-AI collaboration, health care organizations should establish clear guidelines for collaborative intelligence. This includes defining roles and responsibilities for both human and AI components of the decision-making process. Nurses and clinicians should be trained to understand the strengths and limitations of RAG systems and to intervene when they believe the AI-generated recommendation is not appropriate for the patient's context [6,37,53]. In addition, RAG systems should be designed to provide options for clinicians to modify or override system-generated outputs when necessary, allowing for greater flexibility and ensuring that human expertise is always part of the decision-making process [19,54,55].

Continuous Monitoring and Evaluation

To ensure the ongoing effectiveness and ethical integrity of RAG systems, continuous monitoring and evaluation are essential. After RAG systems are deployed in clinical practice, they must be regularly assessed for both technical accuracy and ethical impact. This process should include performance audits,

XSL•FO
RenderX

where the system's outputs and their clinical relevance are evaluated against real-world data. Continuous monitoring should also include version control and scheduled updates of external data sources to prevent outdated or inconsistent information from influencing clinical recommendations [56].

Moreover, it is equally important to assess the ethical implications of RAG systems over time. This involves gathering feedback from health care professionals who use the system, as well as patients who are impacted by the decisions. Implementing a structured feedback mechanism is critical to identify potential ethical risks such as new biases, privacy concerns, or trust issues that may arise after the model has been in use for some time [6,57]. By continuously monitoring the performance and ethical consequences of RAG systems, health care organizations can make necessary adjustments to ensure that these systems remain aligned with ethical standards and improve patient care without causing harm.

Conclusion

RAG systems have the potential to significantly enhance nursing practice by providing accurate, context-aware, and timely information support for clinical decisions. By integrating external knowledge sources into clinical reasoning, RAG can improve patient outcomes, particularly in high-pressure environments where rapid, evidence-based decisions are crucial. However, as with any emerging technology, the deployment of RAG systems in clinical settings must be approached with caution, taking into account both the opportunities and the ethical risks they introduce.

The ethical challenges associated with RAG, including issues of accuracy, fairness, transparency, explainability, accountability, and human oversight, must be proactively addressed to ensure that these systems are used responsibly. By implementing robust data governance, ensuring the explainability of AI recommendations, maintaining human-in-the-loop oversight, and continuously monitoring the model's performance and ethical impact, health care organizations can harness the benefits of RAG while minimizing the potential risks.

Ultimately, responsible RAG implementation will require collaboration between health care professionals, AI developers, and policymakers. By prioritizing ethical considerations in the deployment of RAG systems, we can ensure that they support patient safety, reduce health care disparities, and contribute to the overall improvement of nursing practice. Through careful, responsible integration, RAG systems have the potential to transform the future of clinical decision-making and enhance the quality of care delivered to patients worldwide.

Acknowledgments

This paper was language-polished using the generative artificial intelligence (AI) tool ChatGPT by OpenAI. The authors carefully reviewed and revised all AI-generated suggestions to ensure accuracy, clarity, and academic integrity. The authors take full responsibility for the content.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Authors' Contributions

XT conceived the topic, outlined the article, and contributed to the manuscript's drafting and writing. CS refined the outline and contributed to the manuscript writing. PQ contributed to the manuscript writing. LW provided revisions and guidance as corresponding author and supervisor. All authors contributed to the study's preparation and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Goktas P, Grzybowski A. Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. J Clin Med 2025 Feb 27;14(5):1605. [doi: 10.3390/jcm14051605] [Medline: 40095575]
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Nashwan AJ, Abujaber AA. Harnessing large language models in nursing care planning: opportunities, challenges, and ethical considerations. Cureus 2023 Jun;15(6):e40542. [doi: 10.7759/cureus.40542] [Medline: 37465807]
4. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
5. Bhagat SV, Kanyal D. Navigating the future: the transformative impact of artificial intelligence on hospital management - a comprehensive review. Cureus 2024 Feb;16(2):e54518. [doi: 10.7759/cureus.54518] [Medline: 38516434]

XSL•FO
RenderX

6. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655. [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
7. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]
8. Verma S. Contextual compression in retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Sep 20, 2024. [doi: 10.48550/arXiv.2409.13385]
9. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina (Kaunas) 2024 Mar 8;60(3):445. [doi: 10.3390/medicina60030445] [Medline: 38541171]
10. Ceriello A, Gallo M, Candido R, et al. Personalized therapy algorithms for type 2 diabetes: a phenotype-based approach. Pharmgenomics Pers Med 2014;7:129-136. [doi: 10.2147/PGPM.S50288] [Medline: 24971031]
11. Jiao J, Afroogh S, Xu Y, Phillips C. Navigating LLM ethics: advancements, challenges, and future directions. AI Ethics 2025 Dec;5(6):5795-5819. [doi: 10.1007/s43681-025-00814-5]
12. Arora A, Alderman JE, Palmer J, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023 Nov;29(11):2929-2938. [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]
13. Riedemann L, Labonne M, Gilbert S. The path forward for large language models in medicine is open. NPJ Digit Med 2024 Nov 27;7(1):339. [doi: 10.1038/s41746-024-01344-w] [Medline: 39604549]
14. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics 2022 Sep;48(9):581-585. [doi: 10.1136/medethics-2020-107102] [Medline: 34006600]
15. Marey A, Arjmand P, Alerab ADS, et al. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egypt J Radiol Nucl Med 2024;55(1). [doi: 10.1186/s43055-024-01356-2]
16. Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. R Soc Open Sci 2025 May;12(5):241873. [doi: 10.1098/rsos.241873] [Medline: 40370601]
17. Wang F, Wan X, Sun R, Chen J, Arik SO. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025. [doi: 10.18653/v1/2025.acl-long.1476]
18. Yang R, Ning Y, Keppo E, et al. Retrieval-augmented generation for generative artificial intelligence in health care. Npj Health Syst 2025;2(1). [doi: 10.1038/s44401-024-00004-1]
19. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med 2025 Apr 5;8(1):187. [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
20. Wang J, Fan X, Shen X, Gao Y. Understanding the dark side of online reviews on consumers' purchase intentions in E-commerce: evidence from a consumer experiment in China. Front Psychol 2021;12. [doi: 10.3389/fpsyg.2021.741065]
21. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735-3746. [doi: 10.1016/j.csbj.2021.06.030] [Medline: 34285775]
22. Gomersall T, Astell A, Nygård L, Sixsmith A, Mihailidis A, Hwang A. Living with ambiguity: a metasynthesis of qualitative research on mild cognitive impairment. Gerontologist 2015 Oct;55(5):892-912. [doi: 10.1093/geront/gnv067] [Medline: 26315317]
23. Smith E, Eloff JHP. A prototype for assessing information technology risks in health care. Computers & Security 2002 Jun;21(3):266-284. [doi: 10.1016/S0167-4048(02)00313-9]
24. Ayyamperumal SG, Ge L. Current state of LLM risks and AI guardrails. arXiv. Preprint posted online on Jun 16, 2024. [doi: 10.48550/arXiv.2406.12934]
25. Xu H, Wang S, Li N, et al. Large language models for cyber security: a systematic literature review. ACM Trans Softw Eng Methodol 2025. [doi: 10.1145/3769676]
26. Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. Digit Health 2025;11:20552076251337177. [doi: 10.1177/20552076251337177] [Medline: 40343063]
27. Lysandrou G, Owen RE, Mursec K, Le Brun G, Fairley EA. Comparative analysis of drug-GPT and ChatGPT LLMs for healthcare insights: evaluating accuracy and relevance in patient and HCP contexts. arXiv. Preprint posted online on Jul 24, 2023. [doi: 10.48550/arXiv.2307.16850]
28. Nashwan AJ, Cabrega JA, Othman MI, et al. The evolving role of nursing informatics in the era of artificial intelligence. Int Nurs Rev 2025 Mar;72(1):e13084. [doi: 10.1111/inr.13084] [Medline: 39794874]
29. Zeng S, Zhang J, He P, et al. The good and the bad: exploring privacy issues in retrieval-augmented generation (RAG). Presented at: Findings of the Association for Computational Linguistics: ACL 2024; Aug 11-16, 2024. [doi: 10.18653/v1/2024.findings-acl.267]
30. Gupta O, Marrone S, Gargiulo F, Jaiswal R, Marassi L. Understanding social biases in large language models. AI 2025;6(5):106. [doi: 10.3390/ai6050106]
31. Pasupuleti MK. Bias and fairness in large language models: evaluation and mitigation techniques. IJAIRI 2025 May;05(5):442-451. [doi: 10.62311/nesx/rphcr6]

XSL•FO
RenderX

32. Khan B, Fatima H, Qureshi A, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices 2023 Feb 8;1(2):1-8. [doi: 10.1007/s44174-023-00063-2] [Medline: 36785697]
33. Lasker A. Exploring ethical considerations in generative AI. IJAR 2024;12(4):531-535. [doi: 10.21474/IJAR01/18578]
34. Sagona M, Dai T, Macis M, Darden M. Trust in AI-assisted health systems and AI's trust in humans. Npj Health Syst 2025;2(1). [doi: 10.1038/s44401-025-00016-5]
35. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLoS Digit Health 2025 Apr;4(4):e0000810. [doi: 10.1371/journal.pdig.0000810] [Medline: 40198594]
36. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024 Apr 25;26:e56764. [doi: 10.2196/56764] [Medline: 38662419]
37. Abgrall G, Holder AL, Chelly Dagdia Z, Zeitouni K, Monnet X. Should AI models be explainable to clinicians? Crit Care 2024 Sep 12;28(1):301. [doi: 10.1186/s13054-024-05005-y] [Medline: 39267172]
38. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23); Apr 23-28, 2023. [doi: 10.1145/3544548.3581075]
39. Nasarian E, Alizadehsani R, Acharya UR, Tsui KL. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework. Information Fusion 2024 Aug;108:102412. [doi: 10.1016/j.inffus.2024.102412]
40. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng 2024 Aug;118:109370. [doi: 10.1016/j.compeleceng.2024.109370]
41. Benrimoh D, Kleinerman A, Furukawa TA, et al. Towards outcome-driven patient subgroups: a machine learning analysis across six depression treatment studies. Am J Geriatr Psychiatry 2024 Mar;32(3):280-292. [doi: 10.1016/j.jagp.2023.09.009] [Medline: 37839909]
42. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med 2023 Sep 12;6(1):170. [doi: 10.1038/s41746-023-00913-9] [Medline: 37700029]
43. Elendu C, Amaechi DC, Elendu TC, et al. Ethical implications of AI and robotics in healthcare: a review. Medicine (Baltimore) 2023 Dec 15;102(50):e36671. [doi: 10.1097/MD.0000000000036671] [Medline: 38115340]
44. Ferlito B, Segers S, De Proost M, Mertes H. Responsibility gap(s) due to the introduction of AI in healthcare: an Ubuntu-inspired approach. Sci Eng Ethics 2024 Aug 1;30(4):34. [doi: 10.1007/s11948-024-00501-4] [Medline: 39090479]
45. Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M. Principles to practices for responsible AI: closing the gap. arXiv. Preprint posted online on Jun 8, 2020. [doi: 10.48550/arXiv.2006.04707]
46. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ 2020 Apr 1;98(4):251-256. [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
47. Rony MKK, Parvin MR, Ferdousi S. Advancing nursing practice with artificial intelligence: enhancing preparedness for the future. Nurs Open 2024 Jan;11(1). [doi: 10.1002/nop2.2070] [Medline: 38268252]
48. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512. [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
49. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023 Jan 13;23(1):7. [doi: 10.1186/s12911-023-02103-9] [Medline: 36639799]
50. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLoS Digit Health 2023 Jun;2(6):e0000278. [doi: 10.1371/journal.pdig.0000278] [Medline: 37347721]
51. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan;25(1):37-43. [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
52. Futia G, Vetrò A. On the integration of knowledge graphs into deep learning models for a more comprehensible AI—three challenges for future research. Information 2020;11(2):122. [doi: 10.3390/info11020122]
53. Lawton T, Morgan P, Porter Z, et al. Clinicians risk becoming "liability sinks" for artificial intelligence. Future Healthc J 2024 Mar;11(1):100007. [doi: 10.1016/j.fhj.2024.100007] [Medline: 38646041]
54. Bunnell DJ, Bondy MJ, Fromtling LM, Ludeman E, Gourab K. Bridging AI and healthcare: a scoping review of retrieval-augmented generation—ethics, bias, transparency, improvements, and applications. medRxiv. Preprint posted online on Apr 1, 2025. [doi: 10.1101/2025.04.01.25325033]
55. Zhao X, Liu S, Yang SY, Miao C. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. Presented at: Proceedings of the ACM Web Conference 2025 (WWW '25); Apr 28 to May 2, 2025. [doi: 10.1145/3696410.3714782]
56. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022 May 31;5(1):66. [doi: 10.1038/s41746-022-00611-y] [Medline: 35641814]
57. Pandi-Perumal SR, Akhter S, Zizi F, et al. Project stakeholder management in the clinical research environment: how to do it right. Front Psychiatry 2015;6:71. [doi: 10.3389/fpsyt.2015.00071] [Medline: 26042053]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
RAG: retrieval-augmented generation

Edited by A Benis; submitted 01.Jul.2025; peer-reviewed by D Paradice, J Shi, S Mohanadas; revised version received 26.Oct.2025; accepted 30.Nov.2025; published 09.Jan.2026.

Please cite as:
Tu X, Shi C, Qian P, Wang L
Ethical Imperatives for Retrieval-Augmented Generation in Clinical Nursing: Viewpoint on Responsible AI Use
JMIR Med Inform 2026;14:e79922
URL: https://medinform.jmir.org/2026/1/e79922
doi:10.2196/79922

© Xinyi Tu, Chenghao Shi, Peilin Qian, Lizhu Wang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Improving Clinical Decision-Making in Treating Airway Diseases With an Expert System Built Upon the Free AI Tool Google NotebookLM

Cheng-Hao Hsu[1], MD; Ching-Li Hsu[2], MD; Chih-Hsiang Tsou[3,4], MD, PhD; Kuo-Fang Hsu[2], MD, PhD; Hung-Yu Yang[1,5,6], MD, PhD

[1]Department of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]Miaoli General Hospital, Ministry of Health and Welfare of Taiwan, Miaoli City, Taiwan
[3]Department of Pulmonology, Jen-Ai Branch, Taipei City Hospital, Taipei, Taiwan
[4]School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Department of Cardiology, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
[6]Biomedical Data Center, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Hung-Yu Yang, MD, PhD
Department of Cardiology
Wan Fang Hospital
Taipei Medical University
No.111, Sec. 3, Xinglong Rd.,
Wenshan Dist.,
Taipei
Taiwan
Phone: 886 2 29307930 ext 2818
Email: jaero@tmu.edu.tw

Abstract

We used the free artificial intelligence (AI) tool Google NotebookLM, powered by the large language model Gemini 2.0, to construct a medical decision-making aid for diagnosing and managing airway diseases and subsequently evaluated its functionality and performance in a clinical workflow. After feeding this tool with relevant published clinical guidelines for these diseases, we evaluated the feasibility of the system regarding its behavior, ability, and potential, and we created simulated cases and used the system to solve associated medical problems. The test and simulation questions were designed by a pulmonologist, and the appropriateness (focusing on accuracy and completeness) of AI responses was judged by 3 pulmonologists independently. The system was then deployed in an emergency department setting, where it was tested by medical staff (n=20) to assess how it affected the process of clinical consultation. Test opinions were collected through a questionnaire. Most (56/84, 67%) of the specialists' ratings regarding AI responses were above average. The interrater reliability was moderate for accuracy (intraclass correlation coefficient=0.612; *P*<.001) and good on completeness (intraclass correlation coefficient=0.773; *P*<.001). When deployed in an emergency department (ED) setting, this system could respond with reasonable answers, enhance the literacy of personnel about these diseases. The potential to save the time spent in consultation did not reach statistical significance (Kolmogorov-Smirnov [K-S] D=0.223, *P*=.24) across all participants, but it indicated a favorable outcome when we analyzed only physicians' responses. We concluded that this system is customizable, cost efficient, and accessible to clinicians and allied health care professionals without any computer coding experience in treating airway diseases. It provides convincing guideline-based recommendations, increases the staff's medical literacy, and potentially saves physicians' time spent on consultation. This system warrants further evaluation in other medical disciplines and health care environments.

(JMIR Med Inform 2026;14:e78567) doi:10.2196/78567

KEYWORDS

airway disease; artificial intelligence; AI; artificial intelligence tool; AI tool; emergency department; medical decision-making; NotebookLM; pulmonology

XSL•FO
RenderX

Introduction

Expert Systems and the Decision-Making Aids

An expert system is a knowledge management tool that integrates domain-specific information with an inference engine to facilitate decision-making. When consulted, it delivers timely and relevant solutions to a problem. This concept originated in the 1950s, with medical diagnostic reasoning being among its earliest applications [1,2]. Gradually, expert systems became popular in the 1970s and 1980s [3]. With the advent of the digital era, the concept of expert systems was revisited by the medical community. For instance, oncology, given its rapid evolution and growth in knowledge over the past 2 decades, has required consultation systems to aid patient management.

Knowledge Management and the Large Language Model

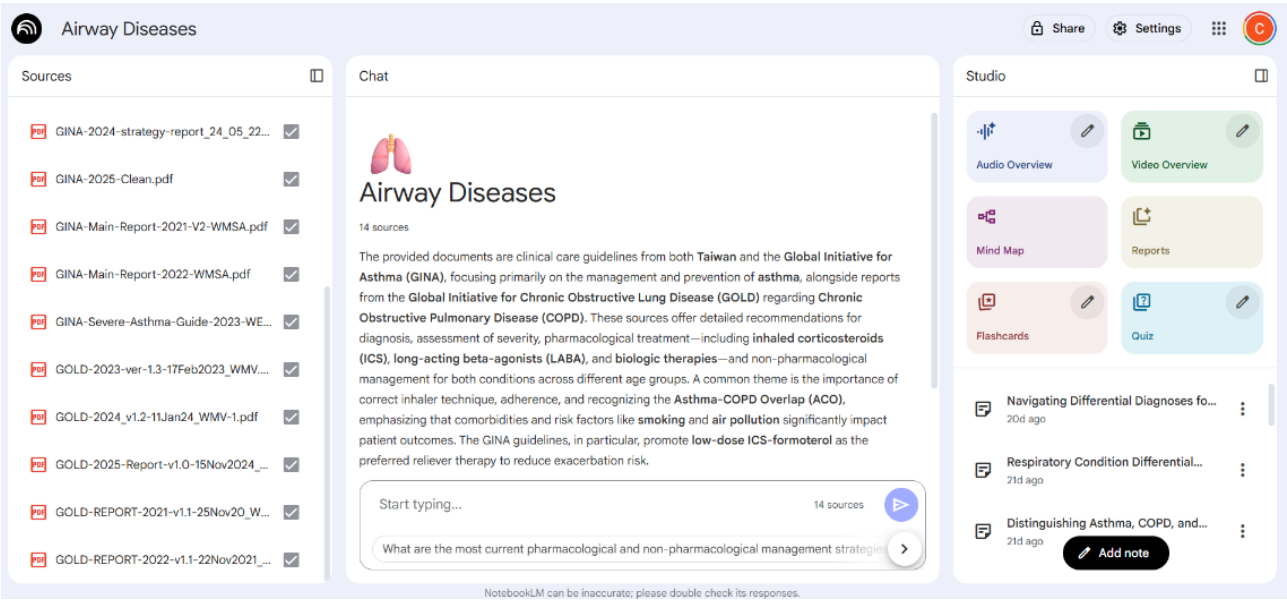
Large language models (LLMs) represent a branch of machine learning that has been under development for several years. Their impact became widely recognized on November 30, 2022, when OpenAI released ChatGPT, a free web-based service [4,5]. This milestone illustrated the potential of LLMs to advance toward general purpose artificial intelligence (AI) by enabling natural language–based dialogue. The continuously expanding medical knowledge has traditionally been managed through structured data storage, indexing, and query mechanisms [6] or probabilistic inference algorithms [7]. However, these conventional clinical decision support systems struggle to incorporate new information and adapt to complex clinical contexts [8]. Unlike these approaches, LLM-driven systems enable interaction in natural language, offering a more user-friendly and adaptable knowledge management tool. Such systems can be easily deployed across health care teams, including physicians and allied health care professionals. In this study, we devised a model to understand how this new LLM-based methodology could affect medical decision-making and routine clinical workflow. The process of our project and our key findings were described as follows.

Decision-Making Aid Construction and Assessment

Constructing an Expert System With a Free AI Tool Based on LLM

The free AI tool Google NotebookLM, powered by the LLM Gemini 2.0 [9], provides a platform for knowledge integration through 3 components: a source input pane, a query pane, and a note-taking pane, which function collaboratively (Figure 1). We used this platform to develop a medical decision support system for airway diseases. Relevant published clinical guidelines were uploaded into the source input pane, including the GINA (Global Initiative for Asthma) guideline [10] and the GOLD (Global Initiative for Chronic Obstructive Lung Disease) guideline [11] (version 2021-2025). Additional standard documents from the Taiwan Society of Pulmonary and Critical Care Medicine [12], such as the Taiwan asthma guidelines and the pulmonary rehabilitation guidelines, were also incorporated. The responses of the system to queries were then recorded.

Figure 1. The Google NotebookLM interface.



Feasibility Evaluated by Pulmonologists

Query questions were selected to match the essential requirements of eligible AI tools, such as reproducibility, the ability to develop a treatment plan, and the capacity to handle speculative ideas. They were designed by a pulmonologist with expertise in airway diseases and familiarity with these guidelines, and the appropriateness (focusing on accuracy and completeness) of AI responses was judged (Likert scale 1-5) by 3 pulmonologists independently. Airway diseases were selected as the pilot scenario to evaluate the feasibility of our system because standardized clinical guidelines are widely accepted, and they involve fewer complex and controversial data points than other conditions, such as oncological diseases.

Implementation in the Clinical Workflow

The system was introduced to physicians and allied health care professionals in the emergency department (ED), where patients

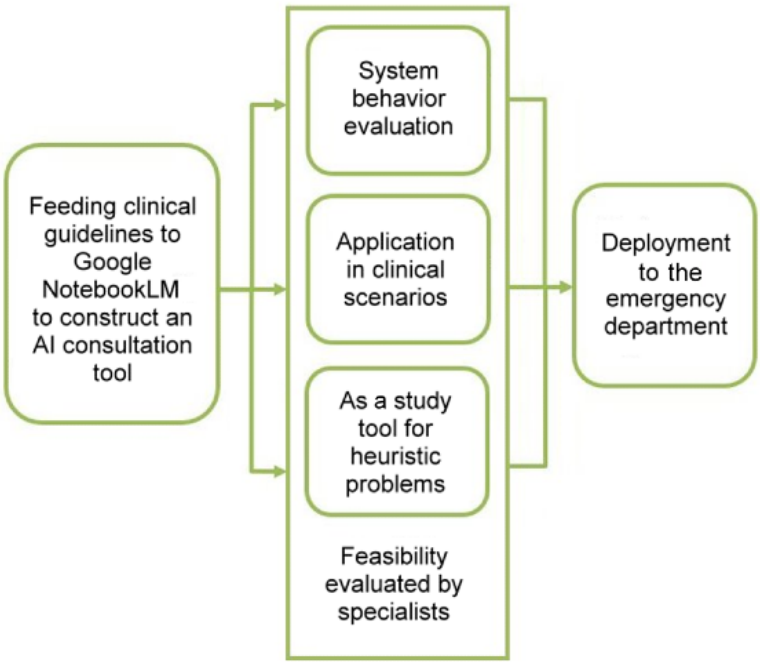
XSL•FO
RenderX

with airway diseases commonly present for treatment. Feedback was collected through a questionnaire, and the responses were analyzed to assess the system's effectiveness within the clinical workflow.

The ED setting was chosen because clinicians frequently encounter diverse conditions and require rapid access to specialist knowledge. In this context, an expert system capable of answering clinical queries has the potential to reduce the time burden on patients, frontline clinicians, and the specialists consulted.

The complete workflow of the project is summarized in Figure 2.

Figure 2. Flowchart illustrating the process of constructing our artificial intelligence (AI)–based consultation tool for medical decision-making.



AI Responses to Queries

System Behavior Evaluation

Test 1: Reproducibility

The system consistently generated accurate summaries and responses, each supported by appropriate and reliable references. When prompted to regenerate output from the same sources, this system produced slightly varied responses that retained identical meaning, resembling the variability typically observed in an AI chatbot interaction.

Test 2: Knowledge Consolidation

The asthma and chronic obstructive pulmonary disease (COPD) guidelines contained respective lists of differential diagnoses. When instructed to generate a consolidated table of differential diagnoses for airway diseases, including their distinguishing features and relevant diagnostic tests, the system successfully produced one, as shown in Table 1.

XSL•FO
RenderX

Table 1. Summary of the differential diagnoses for asthma, chronic obstructive pulmonary disease (COPD), and related conditions, including distinguishing symptoms, clinical features, and tests. Full version available in Multimedia Appendix 1.

Condition	Primary differential for	Differentiating symptoms and features	Differentiating tests and findings
Inducible laryngeal obstruction or vocal cord dysfunction	Both asthma and COPD	Dyspnea and inspiratory wheezing (stridor), sometimes with cough and general wheeze	Functional laryngoscopy visualizing vocal cord movement; full flow-volume curve assessing upper airway obstruction
Obesity	Both asthma and COPD	Respiratory symptoms, such as dyspnea and wheeze, that can mimic asthma or COPD, often related to deconditioning or mechanical restriction	Objective measurement of variable expiratory airflow limitation to confirm asthma in obese patients
Tracheobronchomalacia	Both asthma and COPD	Wheeze	Bronchoscopy; high-resolution chest computed tomography scan
Lung cancer	Both asthma and COPD (more common in COPD)	Breathlessness, significant weight loss, hemoptysis, night sweats, and fever	Chest radiograph or computed tomography scan; reduced diffusing capacity for carbon monoxide may be present
Alpha-1 antitrypsin deficiency	Both asthma and COPD (more commonly associated with COPD)	Shortness of breath, with a family history of early-onset emphysema	Blood test confirming alpha-1 antitrypsin deficiency
Tuberculosis	Both asthma and COPD (especially in low- and middle-income countries)	Chronic cough, hemoptysis, dyspnea and/or fatigue, fever, night sweats, anorexia, and weight loss	Sputum induction (to confirm inflammatory phenotype); high-resolution computed tomography of the chest

Test 3: Temporal Order Recognition

The system demonstrated the ability to recognize temporal updates. For instance, it correctly identified that the “ABCD assessment tool” for COPD was replaced by the “ABE assessment scheme” in the 2023 update of the GOLD COPD guideline [11]. When asked to provide treatment recommendations, the system appropriately referred to the most recent guideline version while disregarding outdated ones, as shown in Textbox 1.

Textbox 1. The chronic obstructive pulmonary disease (COPD) grouping follows the most recent version of the clinical guidelines. For full details, refer to Multimedia Appendix 2.

Question

A COPD patient with mMRC=2 and CAT=11, and a history of hospitalization once in the past 1 year. What is his/her treatment suggestion?

Answer

...

3. GOLD Group Classification:

Combining a high symptom burden (mMRC=2, CAT=11) with a high exacerbation risk (one hospitalization), this patient is classified as GOLD Group E [GOLD 2023, p. 40, Figure 2.4; GOLD 2024, p. 29, Figure 2.5; GOLD 2025, p. 38, Figure 2.5]. This classification guides therapeutic recommendations [GOLD 2023, p. 40; GOLD 2024, p. 29; GOLD 2025, p. 38]...

Test 4: Memory Flush

When the COPD guidelines from 2023 onward were deselected from the source input panel, the system could still respond to questions about “ABE assessment” using information retained from previous dialogues, provided those dialogues had not been deleted using the Refresh button (Textbox 2). Interestingly, if the dialogues were deleted and no clues regarding the 2023 update were supplied, the system concluded that “ABE” might be a typographic error for “ABCD” (Textbox 3) or “creatively“ suggested that “B” referred to blood eosinophil count (Textbox 4). This indicates that the AI tool retains dialogue content until it is deliberately erased using the Refresh button.

XSL•FO
RenderX

Textbox 2. The system could respond using information from prior dialogues until those dialogues were removed using the Refresh button.

Question

What is the ABE assessment tool?

Answer

The provided 2021 and 2022 reports do not contain information about merging groups C and D into a single "E" group, nor do they refer to an "ABE" assessment tool. This suggests that the change to the ABE classification, if it occurred, was introduced in a subsequent GOLD report (e.g., the GOLD 2023 report as mentioned in our previous conversation).

Textbox 3. A system response after a memory flush.

Question

Explain the ABE assessment tool for COPD.

Answer

The sources provided do not contain information on an "ABE assessment tool" for COPD. However, they extensively describe the "ABCD" assessment tool used in the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines for COPD management. It is possible that "ABE" in your query was a typographical error for "ABCD" or refers to a concept not detailed in these documents.

Textbox 4. A creative response after a memory flush.

Question

Explain the ABE assessment tool for COPD.

Answer

While the provided sources do not explicitly define or refer to an "ABE assessment tool" as a named entity for evaluating COPD, they do describe a framework for the initiation of pharmacological management that relies on assessing factors commonly associated with "A" (Symptoms), "B" (Blood Eosinophil Count), and "E" (Exacerbation Risk).

Test 5: Handling Multiple Languages

Leveraging Google's multilingual capability, the system could accept a query in Chinese and respond in either English or Chinese. When asked to compare the 2022 Taiwan adult asthma clinical care guidelines, written entirely in Chinese, with the GINA guideline, the system managed to perform the task without difficulty. Language switching did not appear to affect system performance.

Application in Clinical Scenarios

Scenario 1: Requesting Patient Management Suggestions Based on a Simulated Medical Record

As illustrated in Textbox 5, the system generated guideline-based recommendations that were clinically useful.

Textbox 5. Helpful recommendations generated by the system, supported by guideline-based reasoning. For a complete explanation, refer to Multimedia Appendix 3.

Question

I have a patient diagnosed of COPD adequately maintained with Pulmicort, 2 puffs bid. However, he suffered from more and more dyspnea while walking. What is the management suggestion?

Answer

Based on the information in the sources, here are some management suggestions for your patient with COPD who is experiencing increasing dyspnea on walking despite being maintained on Pulmicort (an inhaled corticosteroid - ICS):

1. Consider adding a long-acting bronchodilator...
2. Consider a combination inhaler containing ICS and LABA...
3. Optimize inhaled therapy. Ensure that the patient...
4. Consider pulmonary rehabilitation...
5. Reassess lung function and consider other diagnoses...
6. Review for modifiable risk factors...
7. Consider stepping up ICS dose (with caution)...
8. Assess for features of asthma-COPD overlap (ACO)...
9. Follow-up and reassess...

XSL•FO
RenderX

Scenario 2: Logical Question

The system addressed a logical question regarding the unavailability of pulmonary function testing by providing a structured, entry-by-entry response. It explained the implications separately for COPD and asthma, as shown in Textbox 6.

Textbox 6. The system presented a comparative response for chronic obstructive pulmonary disease (COPD) and asthma. For the complete answer, refer to Multimedia Appendix 4.

Question

Is pulmonary function test necessary for diagnosis? What if there is no pulmonary function test accessible?

Answer

Here's how pulmonary function tests relate to diagnosing airway diseases, and what to do if these tests aren't accessible:

- COPD Diagnosis: Spirometry is essential to confirm a COPD diagnosis...
- Asthma Diagnosis: Asthma diagnosis relies on a combination of characteristic symptom patterns and evidence of variable expiratory airflow limitation...
- Role of Lung Function Testing: Lung function testing is essential to confirm...
- Spirometry: Forced spirometry is the most reproducible and objective measurement of airflow obstruction. It measures...
- When Spirometry Isn't Available for Asthma Diagnosis:
  - Peak expiratory flow (PEF) can be used, although it is less reliable...
  - Assess the patient for a characteristic pattern of respiratory symptoms.
  - Consider a therapeutic trial with SABA and low-dose ICS...
  - Confirm variable expiratory airflow limitation by PEF. A ≥20% improvement in PEF...
  - Improvement in symptoms and PEF after a 4-week therapeutic trial...
- When Spirometry Isn't Available for COPD Diagnosis: When routine spirometry is not available, home measurement of peak expiratory flow (PEF) combined with validated patient questionnaires could be used to support or refute a possible diagnosis of COPD...

Scenario 3: Personalized Rehabilitation Planning

The system generated a tailored rehabilitation plan and schedule that took into account an advanced age condition, as shown in Textbox 7.

XSL•FO
RenderX

Textbox 7. The system produced a detailed, individualized rehabilitation plan for an extremely older adult patient with chronic obstructive pulmonary disease (COPD) and ischemic heart disease. The complete plan is available in Multimedia Appendix 5.

Question

I have a female patient, age 95, with COPD and ischemic heart disease. Please make a rehabilitation plan and schedule for her.

Answer

For your 95-year-old female patient with COPD and ischemic heart disease, a highly individualized, comprehensive, and carefully supervised rehabilitation plan is crucial, taking into accounts her advanced age and multiple comorbidities. This plan draws primarily from the Taiwan Pulmonary Rehabilitation Practical Guideline and the Global Initiative for Chronic Obstructive Lung Disease (GOLD) reports.

I. Overall Approach and Guiding Principles

1. Personalized and Integrated Care:...
2. Specialist Supervision:...
3. Holistic Goals:...

II. Initial Comprehensive Assessment

- Patient Goals and Preferences:...
- Medical History and Physical Examination:...

...

III. Rehabilitation Setting and Schedule

1. Setting: For a 95-year-old patient with significant comorbidities, a supervised setting is highly recommended, such as...
2. Duration: Optimal benefits from pulmonary rehabilitation programs are typically achieved within 6 to 8 weeks...
3. Frequency: Supervised exercise training at least twice weekly is recommended.

IV. Components of the Rehabilitation Plan

1. Exercise Training:...
2. Education and Self-Management:...

...

V. Follow-up and Monitoring

- Regular Review:...
- Response Assessment:...

...

Scenario 4: Vague Question

In this case, the system transformed the ambiguous concept of “prospect” into practical categories of “future challenges” and “research directions.” This approach provided a meaningful response, as illustrated in Textbox 8.

XSL•FO
RenderX

Textbox 8. The system interpreted the vague query as a request for practical insights. The complete answer is provided in Multimedia Appendix 6.

Question

What is the future prospect of airway diseases and their treatment?

Answer

The sources discuss future challenges and research directions for airway diseases such as asthma and Chronic Obstructive Pulmonary Disease (COPD).

Asthma:

- Despite improvements in asthma care and effective medications, many patients worldwide do not benefit from these advances due to inadequate medical facilities and financial resources.
- Future research should focus on clinical and physiological characteristics, biomarkers, outcomes, and underlying mechanisms in patients with respiratory symptoms or chronic airflow limitation...

COPD:

- More research is needed to guide better recognition and safe, effective treatment of COPD.
- There's a need to study clinical and physiological characteristics, biomarkers, outcomes and underlying mechanisms in broad populations of patients with respiratory symptoms or with chronic airflow limitation.
- Future studies should focus on defining phenotypes and mechanisms to ensure the safety of pharmacological treatments...

Scenario 5: Out-of-Scope Questions

The system appropriately declined to answer questions that were beyond the scope of the available sources. It did not fabricate factitious responses, unlike many chatbots.

As a Study Tool for Heuristic Problems

Problem 1: Version Comparison and Trend Analysis

When asked to compare guideline versions across publication years, the system clearly outlined the differences between them. Additionally, when requested to analyze trends in pharmacological use for airway diseases over the past 5 years, it generated a comprehensive summary, as shown in Textbox 9.

XSL•FO
RenderX

Textbox 9. The system analyzed the trend in pharmacological use for airway diseases across the years. A complete list is available in Multimedia Appendix 7.

Question

Find the trend in pharmacological use for COPD and asthma in the past 5 years.

Answer

In the past five years, pharmacological treatment for both Chronic Obstructive Pulmonary Disease (COPD) and asthma has seen a clear trend towards more individualized, exacerbation-focused, and strategically combined therapies, with a strong emphasis on proper inhaler use and minimizing oral corticosteroid exposure.

Overarching Trends for Both COPD and Asthma (Past 5 Years)

1. Individualized and Patient-Centered Care:...
2. Focus on Exacerbation Reduction:...
3. Emphasis on Inhaler Technique and Adherence:...
4. Minimizing Oral Corticosteroid (OCS) Use:...

Trends in COPD Pharmacological Treatment (Past 5 Years): The GOLD reports (2021, 2022, 2023, 2024, 2025) and the 2023 Taiwan COPD Clinical Care Guidelines emphasize these trends:

1. Long-Acting Bronchodilators (LABA and/or LAMA) as First-Line:...
2. Strategic Use of Inhaled Corticosteroids (ICS):...
3. Increasing Role of Triple Therapy (LABA+LAMA+ICS):...

...

Trends in Asthma Pharmacological Treatment (Past 5 Years): GINA reports (2021, 2022, 2023, 2024, 2025) and the 2022 Taiwan Adult Asthma Clinical Care Guidelines highlight significant shifts:

1. Shift Away from SABA Monotherapy:...
2. Introduction of Two Treatment Tracks (GINA):...
3. ICS as the Cornerstone of Controller Therapy:...

...

Problem 2: Further Reading Suggestions

When asked to recommend additional readings on intractable asthma, the system searched both the source text and the reference lists it contained. However, it did not suggest any external articles or books. Notably, the system interpreted the term “intractable asthma” as both “severe asthma” and “difficult-to-treat asthma” to ensure comprehensive coverage, as shown in Textbox 10.

XSL•FO
RenderX

Textbox 10. Suggested further reading items provided by the system. For the complete list, refer to Multimedia Appendix 8.

Question

Can you suggest any further readings about the treatment for intractable asthma?

Answer

For further readings about the treatment of intractable asthma, which is often referred to in the sources as severe asthma or difficult-to-treat asthma, you may find the following publications insightful. These sources provide detailed information on diagnosis, management, and treatment options for patients whose asthma remains uncontrolled despite standard therapies:

Comprehensive Guidelines and Reviews:

- Global Strategy for Asthma Management and Prevention (GINA): The GINA reports are frequently referenced for their comprehensive guidance on asthma management. Specifically, the sections within these reports on "Difficult-to-treat and severe asthma" provide a roadmap for assessment and management. You can refer to the detailed sections in the provided GINA reports, such as:
    - GINA-2023-Report:...
    - GINA-2024-strategy-report:...
    - GINA-2025-Clean:...

...Pharmacological Interventions:

- Biologic Therapies:
    - EAACI Biologicals Guidelines-Recommendations for severe asthma: Agache I et al. offer guidelines and recommendations for biologics in severe asthma.
    - Tezepelumab:...

...

- Add-on treatments (nonbiologic):
    - Tiotropium:...
    - Azithromycin:...
    - Oral Corticosteroids (OCS):...

Diagnostic and Management Strategies:

- Sputum-guided treatment...
- Bronchial Thermoplasty:...
- Importance of Adherence and Inhaler Technique:...

These readings should provide a comprehensive overview of the current understanding and treatment approaches for intractable asthma.

Problem 3: Speculative Question

We designed a question regarding the interchangeability of questionnaires used for evaluating COPD and asthma. Our system provided a reasonable response, as shown in Textbox 11.

Textbox 11. The system addressed this speculative question and provided a supporting rationale.

Question

Can we evaluate asthmatic patients with questionnaires designed for COPD? Can we evaluate COPD patients with questionnaires designed for asthma?

Answer

Asthma and COPD can be difficult to distinguish, especially in older patients and smokers, and these conditions may overlap. Some patients are assigned both diagnoses. While some questionnaires might capture general respiratory symptoms that could be present in both asthma and COPD, it is best to use disease-specific questionnaires to ensure the most appropriate and accurate assessment. There is no evidence that asthma questionnaires can be used for COPD patients, and vice versa.

Problem 4: Diagnostic Skill Training Tool

We instructed the system to generate a simulated case with symptoms and pulmonary function values to test our diagnostic skills. During this exercise, we compared our reasoning and diagnosis with the system's answer to strengthen diagnostic ability, as shown in Textbox 12.

XSL•FO
RenderX

Textbox 12. The system generated a simulated case to facilitate diagnostic training. For the answer, diagnostic rationale, and treatment recommendation, refer to Multimedia Appendix 9.

Question

Please make a case with symptoms and pulmonary function values to test our diagnostic skills.

Answer

Simulated Patient: Mr David Lee

Patient Background: Mr David Lee is a 58-year-old male with a 25 pack-year smoking history, though he states he is currently trying to quit. He reports a chronic cough with some mucus production for the past 8 years. He also experiences intermittent shortness of breath on exertion, particularly when climbing stairs. Lately, he's noticed that his breathing feels tighter in the mornings, and he occasionally hears a whistling sound (wheeze) when breathing out, especially after exposure to cold air. These wheezing episodes are variable in intensity, and he notes that his symptoms sometimes improve with a relative's "asthma inhaler" (a short-acting beta2-agonist, SABA).

Symptoms:

- Chronic cough with mucus production (8 years duration).
- Intermittent dyspnea on exertion (eg, climbing stairs).
- Morning chest tightness.
- Occasional expiratory wheezing, variable in intensity, exacerbated by cold air.
- Reports temporary improvement with SABA.
- Significant smoking history (25 pack-years).

Pulmonary Function Values:

- Pre-Bronchodilator Spirometry:
  - Forced Expiratory Volume in 1 second (FEV1): 60% of predicted
  - Forced Vital Capacity: 80% of predicted
  - FEV1/Forced Vital Capacity Ratio: 65% (indicating airflow limitation)
- Post-Bronchodilator Spirometry (after administering 400 mcg of salbutamol, a SABA):...

Diagnosis, Reasons, and Treatment Suggestion for Mr David Lee

...

Pulmonologists' Judgment on AI Responses to Queries

Most of the 3 pulmonologists' ratings (56/84, 67%) regarding AI responses were above average (Likert scale 4-5), and only 5% (4/84) were below average (Likert scale 1-2). The interrater reliability was moderate for accuracy (intraclass correlation coefficient=0.612; $P<.001$) and good for completeness (intraclass correlation coefficient=0.773; $P<.001$) based on a 2-way random-effects model with absolute agreement and single measures options [13]. We also noted that AI responses received more (187/333, 56%) credits from the raters for query questions that required more intelligence (adaptation, deduction, analysis, and planning) to answer, such as scenarios 1 to 3 and problems 1 to 4, than for the other questions. This finding further supports the feasibility of our system.

Responses From the Field Tests in the ED

We collected responses from 20 participants, including 3 physicians, 3 nurse practitioners, 11 ED nurses, and 3 respiratory therapists. Among them, 80% (16/20) expressed interest in the system (Kolmogorov-Smirnov [K-S] $D=0.362$; $P=.008$; indicating a deviation from normality). All these participants had prior experience consulting a chatbot. In total, 60% (12/20) of the participants were aware of and ever disturbed by the issue of AI hallucination (K-S $D=0.238$; $P=.18$; indicating a normal distribution). All participants in this group thought that guideline-based reasoning provided by the system was essential in clinical practice.

In total, 55% (11/20) of the participants believed that the system could reduce specialty consultation time (K-S $D=0.223$; $P=.24$; indicating a normal distribution). Importantly, all participants indicated that the system could enhance their knowledge and diagnostic ability in airway diseases or increase their confidence in explaining conditions to patients.

Discussion

Statistical Findings

Previous studies on LLM-based medical decision support have reported various outcomes [14,15]. These discrepancies may reflect differences in participants' expertise, the difficulty of test questions, and the dimensions measured in study designs. In our evaluation of consultation time, the K-S test indicated a normal distribution of responses. When analyzing only physicians' responses, we observed a trend toward favorable

XSL•FO
RenderX

outcomes, although this did not reach statistical significance (P=.24) across all participants. Thus, the system shows potential for reducing physicians' consultation time, but larger and more comprehensive studies are required to confirm its effectiveness.

What We Expect of a Medical AI Consultation Tool

AI Hallucination and AI Ethics in Medicine

Since the widespread adoption of LLM-based chatbots, users have become increasingly concerned about AI hallucination, that is, plausible but unfounded responses generated without a factual basis. Such errors are unacceptable in medical practice. Amiot and Potier [16] recently emphasized that hallucination is particularly difficult to detect in a health care context. Fortunately, the design of the Google NotebookLM platform mitigates this risk by providing answers supported by explicit references to authoritative sources. With such verifiable citations, concerns about errors, biases, ethics, and safety from autonomous AI actions are substantially reduced.

Beyond a Decision-Making Aid

The project aims to provide a clinical decision-making aid for the diagnosis and management of airway diseases and to assess how this AI tool affected medical staff in the ED during the consultation process. As Wang et al [17] noted, an AI system could uncover subtle relationships in medical data that may elude both traditional decision support tools and human experts. During system development, pulmonologists found that the tool not only consolidated their expertise but also offered novel insights and opportunities to explore speculative ideas, as described in test problems 1 and 3. For medical trainees, the system functions as an effective educational tool, providing not only suggestions but also the underlying reasons.

Limitations Inherent to LLMs

Scope of External Knowledge Considered

NotebookLM generates content strictly from user-provided sources. Because general "common sense" information is often absent from "specialized" clinical guidelines, excluding external knowledge can considerably limit the scope of the system's responses. Moreover, the extent to which external information is incorporated into outputs is undisclosed and therefore unpredictable.

Inability to Pose Differentiating Questions

LLM-based systems generate responses in the form of summaries or abstracts but do not query users. However, accurate clinical decision-making often requires additional patient-specific data obtained through differentiating questions. For example, clarifying whether a patient is a smoker or a nonsmoker or identifying ethnic background (Asian or White) may guide diagnostic pathways. If the system could formulate such discriminative questions and allow users to provide input—ideally guided by Bayesian theorem [18]—its clinical utility would be markedly enhanced. We did not observe this feature in the current implementation.

Future Directions

Once the pilot project is fully operational, the expert system can be refined along several pathways. First, the system should adapt its responses to the user's professional role. For example, when handling a consultation request, it should provide only essential surgical information to an internist while delivering more comprehensive surgical details to a surgeon.

Second, although our study focused on airway diseases and the system demonstrated generally positive performance, its generalizability to other specialties remains uncertain. Disciplines such as oncology, where problems are more complex and controversies are common, may present challenges. Further research is required to determine whether the model can provide reliable and clinically meaningful output in these domains.

Finally, because decision-making, studying, and education are central components of our medical practice routines, integrating such a system into the existing health care information systems presents a logical next step. We expect that this integration could streamline clinical workflows and improve automation across multiple aspects of care delivery.

Conclusions

Our findings suggest that this approach is highly customizable, cost efficient, and accessible to clinicians and allied health care professionals without any computer coding experience in managing airway diseases. It provides convincing guideline-based recommendations, increases the staff's medical literacy, and potentially saves physicians' time spent on consultation. As AI tools gain widespread adoption, we believe that, if validated across additional medical specialties, this methodology could establish a new paradigm in clinical practice. While acknowledging the potential of AI, we emphasize the continued need for human judgment and oversight in evaluating and applying AI-generated recommendations.

Acknowledgments

The authors would like to thank their colleagues Dr Jer-Hwa Chang, Dr Ming-Cheng Wang, Dr Chien-Yeh Chin, Dr Chung-Hui Tang, Dr Tung-Wu Lin, Yung-Chieh Chiu, Feng-Ling Huang, Yu-Ting Su, Tsai-Chi Chen, Yuan-Hsiang Hsieh, Chi-Wen Feng, Syuan-Wen Lai, Yu-Hsiu Liu, Yi-Chen Chiang, Bing-Hua Wu, Ju-Chuan Hsu, Yu-Ping Lin, Cheng-Yang Tu, Pei-Yi Kuo, Chi-Hsin Chang-Chien, Li-Chun Chang, and Pei-Chen Hsieh for their valuable opinions during the evaluation process. We would like to extend our special thanks to Google for providing its free AI tool, NotebookLM [9], which served as the platform for our discussions.

XSL•FO
RenderX

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Unabridged version of Table 1.
[DOCX File , 24 KB - medinform_v14i1e78567_app1.docx]

Multimedia Appendix 2
Unabridged version of Textbox 1.
[DOCX File , 20 KB - medinform_v14i1e78567_app2.docx]

Multimedia Appendix 3
Unabridged version of Textbox 5.
[DOCX File , 20 KB - medinform_v14i1e78567_app3.docx]

Multimedia Appendix 4
Unabridged version of Textbox 6.
[DOCX File , 19 KB - medinform_v14i1e78567_app4.docx]

Multimedia Appendix 5
Unabridged version of Textbox 7.
[DOCX File , 21 KB - medinform_v14i1e78567_app5.docx]

Multimedia Appendix 6
Unabridged version of Textbox 8.
[DOCX File , 19 KB - medinform_v14i1e78567_app6.docx]

Multimedia Appendix 7
Unabridged version of Textbox 9.
[DOCX File , 21 KB - medinform_v14i1e78567_app7.docx]

Multimedia Appendix 8
Unabridged version of Textbox 10.
[DOCX File , 26 KB - medinform_v14i1e78567_app8.docx]

Multimedia Appendix 9
Unabridged version of Textbox 12.
[DOCX File , 21 KB - medinform_v14i1e78567_app9.docx]

References

1. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959 Jul 03;130(3366):9-21. [doi: 10.1126/science.130.3366.9] [Medline: 13668531]
2. Weiss SM, Kulikowski CA, Amarel S, Safir A. A model-based method for computer-aided medical decision-making. Artif Intell 1978 Aug;11(1-2):145-172. [doi: 10.1016/0004-3702(78)90015-2]
3. Leondes CT. Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st Century. New York, NY: Elsevier Science; 2002:1-22.
4. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2023 Feb 1. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2025-05-29]
5. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on March 15, 2023 2023:1-100 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]

XSL•FO
RenderX

6. Salton G. Developments in automatic text retrieval. Science 1991 Aug 30;253(5023):974-980. [doi: 10.1126/science.253.5023.974] [Medline: 17775340]
7. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine (Baltimore) 1990 May;69(3):125-136. [Medline: 2189054]
8. Bai E, Zhang Z, Xu Y, Luo X, Adelgais K. Enhancing prehospital decision-making: exploring user needs and design considerations for clinical decision support systems. BMC Med Inform Decis Mak 2025 Jan 17;25(1):31 [FREE Full text] [doi: 10.1186/s12911-024-02844-1] [Medline: 39825293]
9. Google NotebookLM. URL: http://notebooklm.google [accessed 2025-05-29]
10. The GINA Asthma Guideline: The Global Strategy for Asthma Management and Prevention. Global Initiative for Asthma (GINA). URL: https://ginasthma.org [accessed 2025-05-29]
11. The GOLD COPD guideline: Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Global Initiative for Chronic Obstructive Lung Disease. URL: https://goldcopd.org [accessed 2025-05-29]
12. Home page. The Taiwan Society of Pulmonology and Critical Care Medicine. URL: http://en.tspccm.org.tw/ [accessed 2025-05-29]
13. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
14. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Jun;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
15. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 01;7(10):e2440969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
16. Amiot F, Potier B. Artificial intelligence (AI) and emergency medicine: balancing opportunities and challenges. JMIR Med Inform 2025 Aug 13;13:e70903 [FREE Full text] [doi: 10.2196/70903] [Medline: 40802997]
17. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. "Brilliant AI doctor" in rural clinics: challenges in AI-powered clinical decision support system deployment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 1-18 URL: https://dl.acm.org/doi/10.1145/3411764.3445432 [doi: 10.1145/3411764.3445432]
18. Hatton GE, Pedroza C, Kao LS. Bayesian statistics for surgical decision making. Surg Infect (Larchmt) 2021 Aug;22(6):620-625 [FREE Full text] [doi: 10.1089/sur.2020.391] [Medline: 33395554]

Abbreviations

AI: artificial intelligence
COPD: chronic obstructive pulmonary disease
ED: emergency department
GINA: Global Initiative for Asthma
GOLD: Global Initiative for Chronic Obstructive Lung Disease
K-S: Kolmogorov-Smirnov
LLM: large language model

Edited by C Perrin; submitted 24.Jun.2025; peer-reviewed by S Habib, VV Khanna, KH Lin; comments to author 28.Jul.2025; accepted 03.Jan.2026; published 29.Jan.2026.

Please cite as:
Hsu CH, Hsu CL, Tsou CH, Hsu KF, Yang HY
Improving Clinical Decision-Making in Treating Airway Diseases With an Expert System Built Upon the Free AI Tool Google NotebookLM
JMIR Med Inform 2026;14:e78567
URL: https://medinform.jmir.org/2026/1/e78567
doi:10.2196/78567
PMID:41483468

©Cheng-Hao Hsu, Ching-Li Hsu, Chih-Hsiang Tsou, Kuo-Fang Hsu, Hung-Yu Yang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences With Integrating Medical Terminologies Into User Interfaces for a Decision Support System for Primary Care: Conceptual and Development Study

Michaela Christina Neff[1], MSc; Jannik Schaaf[1], Prof Dr; Najia Ahmadi[2], MSc; Michele Zoch[2], Dipl-Wi-Inf; Dania Schütze[3], Dr rer med; Holger Storf[1], Prof Dr

[1]Institute of Medical Informatics, University Medicine, Goethe University Frankfurt, Frankfurt, Germany

[2]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technical University of Dresden, Dresden, Germany

[3]Institute of General Practice, Goethe University Frankfurt, Frankfurt, Germany

Corresponding Author:
Michaela Christina Neff, MSc
Institute of Medical Informatics
University Medicine
Goethe University Frankfurt
Theodor-Stern Kai 7
Frankfurt, 60590
Germany
Phone: 49 69630184439
Email: neff@med.uni-frankfurt.de

Abstract

Background: Clinical decision support systems (CDSSs) have shown promise in improving diagnosis in primary care, particularly for chronic diseases. The SATURN (Smart Physician Portal for Patients With Unclear Disease) project developed a CDSS prototype for primary care in Germany that uses artificial intelligence to reduce diagnostic uncertainty in unclear and rare diseases. It generates recommendations based on clinical data from university hospitals stored in a standardized common data model. However, integrating primary care data in Germany remains challenging due to the use of country-specific vocabularies and heterogeneous data structures. Therefore, integration of medical concepts into general practitioners' user interfaces (UIs) and improved workflow design is needed.

Objective: This study investigates how the UI of a CDSS for primary care should be designed to facilitate the user-friendly entry of medical concepts.

Methods: A structured, iterative user-centered design process of five steps was applied: (1) conceptualization, including analysis of requirements and objectives, (2) implementation of the CDSS UI, (3) analysis of user and expert feedback, (4) workshop with experts in the primary care electronic health record system, and (5) development of an extended concept for the system UI.

Results: This study identified requirements and options for supporting data entry in the UI of a CDSS for primary care, thus providing a framework for future implementation and prototype development. Usability testing revealed the need to refine input support, particularly the language used to make the system easy for general practitioners to use. Further effort is required to map this. In addition, we have identified a system interface to electronic patient records in primary care as essential for the system.

Conclusions: This paper demonstrates how standardized medical terminologies can be integrated into a CDSS UI for primary care. The user-centered design approach has been effective in developing CDSS UIs that align with clinicians' workflows and provide a user-friendly experience. Although the current technical infrastructure presents significant challenges when connecting to primary care electronic health record systems, collaborating with experts enabled us to identify potential interface solutions. Overall, the results indicate a need for optimized physician language input support and interoperability, including standardized data entry, automated tools, and artificial intelligence–based approaches to enhance data quality and usability.

(JMIR Med Inform 2026;14:e74934) doi:10.2196/74934

KEYWORDS

user computer interfaces; primary care; clinical decision support; controlled vocabulary

XSL•FO
RenderX

Introduction

Background and Significance

The use of clinical decision support systems (CDSSs) has been demonstrated to enhance diagnostic accuracy in primary care settings [1,2]. However, the prevailing focus at present is on chronic diseases [1]. In primary care, general practitioners (GPs) play a unique role in the diagnostic and referral processes. They often deal with nonspecific symptoms that may indicate a wide range of potential diagnoses, leading to diagnostic uncertainty. It is therefore essential to provide GPs with adequate support to ensure patient care [3-5]. Although the potential of CDSSs to achieve significant improvements in outcomes has been demonstrated, wide variations remain in the types and methods of CDSS implementation as well as the resulting effectiveness. Further research is required to determine effective implementation strategies for the usage of CDSS in various settings [6].

The SATURN (Smart Physician Portal for Patients With Unclear Disease) project has developed a prototype of a CDSS to address unclear diagnoses in primary care in Germany [7,8]. The development was conducted through an iterative, user-centered development process (UCD) [8]. The CDSS generates recommendations for GPs using 3 artificial intelligence (AI) modules based on clinical data (a machine learning algorithm and a case-based reasoning algorithm) and expert knowledge (a rule-based system). Clinical databases (rather than primary care databases) were used to establish connections between GPs and the knowledge data held by the university medicine. The diagnostic recommendations derived from clinical data are based on comparisons with existing cases [7,8]. Specifically, the data for these existing cases were extracted from the electronic health record (EHR) data of the University Hospitals in Frankfurt and Dresden via Data Integration Centers (DICs). The DIC is a central institution responsible for making medical data from hospital patient care accessible for medical research and clinical exchange in a structured, standardized, and pseudonymized form [9]. The CDSS database has been designed in such a way that EHR data from the 2 hospitals are merged into 1 OMOP database and can be used for the AI modules. Only data, not AI logic, is stored in the database. The AI modules are designed as separate modules.

The OMOP Common Data Model (CDM) is an open-access standardized data model [10] that can be used to ensure interoperability. The OMOP CDM harmonizes medical databases through a uniform structure and standardized terminologies such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) and RxNorm (normalized names for clinical drugs) [10-12]. New EHR patient cases from primary care for which a diagnosis recommendation is to be made will also be stored in this database, along with corresponding flags (flag: without diagnosis). In addition, solved patient cases are stored (flag: with final diagnosis). As the CDSS uses OMOP and is used in primary care, the data entered by primary care GPs must be compatible with the OMOP vocabulary standards. However, the use of standardized vocabulary remains a challenge, particularly in Germany's primary care sector, where the predominant classification of diseases is *ICD-10-GM* (*International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification*). *ICD-10-GM* is complemented by OPS (Operation and Procedure Codes; German: "Operationen- und Prozedurenschlüssel") for procedural coding procedures [13], along with "Einheitlicher Bewertungsmaßstab" as the basis for billing GPs and contract psychotherapists [14]. Heterogeneous data structures in different countries and country-specific vocabularies (such as *ICD-10-GM*) pose significant challenges for exchanging data and comparing results. To our knowledge, there has been no full systematic integration of primary care data into the OMOP CDM, to date, in Germany. The standardization of data could facilitate the development of common tools. Initial international approaches have shown potential but are limited in their scope [15,16]. Henke et al [17,18] investigated the merging and collaboration of primary care and clinical data in Germany. These studies linked data from the DICs of the Medical Informatics Initiative with routine health insurance data and aimed to increase the coverage of the German vocabulary in the OMOP CDM. Another study examined the cross-sector linkage of primary care data and identified challenges, including system interfaces and data extraction processes [19]. Such projects show the need for using CDMs, such as the OMOP CDM and standardized vocabularies, to harmonize cross-sector data. To integrate medical concepts into GP user interfaces (UIs), the UI must be designed to support and streamline existing workflows [20]. Medical concepts are coded according to standards for terminology, listed in terminology management or the OMOP vocabularies.

Objectives

The objective of this study is to explore how the UI of a CDSS for primary care should be designed to allow the entry of medical concepts in a user-friendly way. To our knowledge, it has not yet been investigated how a CDSS UI for primary care can be comprehensively mapped to enable the standardized data collection and integration of data in Germany. This study outlines the requirements and the available options for CDSS UI for primary care, providing both a technical overview and a conceptual framework to support future CDSS implementations. The CDSS UI implementation focuses on medical vocabularies and the input into the CDSS. Other components of the CDSS, and the iterative conception and development, are covered in previous publications [8,21,22].

Methods

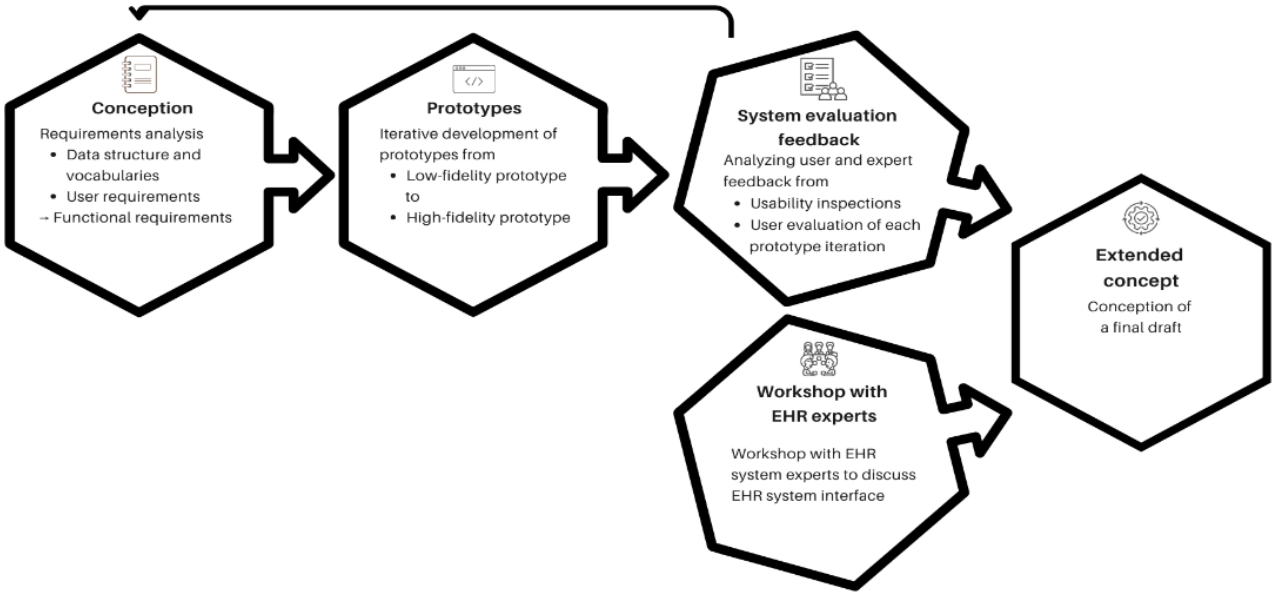
Study Design

As in the whole SATURN project, an iterative UCD was used [23]. In this paper, we focus on data entry into the CDSS (Figure 1). The requirement analysis and underlying data model [24], the ongoing evaluation, as well as other parts of the CDSS, are described in previous publications [8,21,22,25]. The UCD, from the perspective of data input, began with a conceptual design stage. During this stage, the GPs' requirements and necessary medical concepts were analyzed. These concepts were then coded into vocabularies and mapped to the UI design (step: conception). The UI then underwent iterative development in

XSL•FO
RenderX

dialogue with the GPs (step: prototypes), followed by the collection of feedback from users and experts to evaluate its usability and to identify potential improvements (step: system evaluation feedback). EHR experts discussed the automation of medical concept entry (step: workshop with EHR experts). All participant groups were selected using a purposeful sampling approach. Finally (step: extended concept), a draft for the final design concept was created.

Figure 1. Methodological steps of this study for the development process of the CDSS data entry UI. CDSS: clinical decision support system; EHR: electronic health record; UI: user interface.



Ethical Considerations

For the conception and development of this study, the SATURN project was previously approved by the Ethics Committee of Goethe University's Medical Faculty (case 2022-1088). The user evaluation was also approved, with the Ethics Committee waiving the need for a separate ethics vote on the inquiry of GPs (case 2022-629). All participants received written information on the study and provided a signed informed consent. All data were pseudonymized before analysis. Participants received €150 (average currency exchange rate in August 2023: US $1=€0.92) for taking part in the usability test.

Participants in the usability inspections and the EHR expert workshop were either project participants or had previously submitted a letter of intent to support the project. All data were pseudonymized before analysis, with only sample data being used for case studies. Participants received no compensation for their participation in the usability inspections.

Conception

In order to facilitate data entry into the CDSS, it is necessary to consider the structure and usage of the vocabularies, as well as the requirements of the end users [21]. The appropriate fields and variables must be defined with the vocabularies for the UI. The data model was defined in a previous study [24] and specifies the following vocabularies, shown in Table 1 [26-37].

All vocabularies shown in Table 1 are fully available in the German language, except for SNOMED CT and Human Phenotype Ontology (HPO). The data to be captured from the requirements and the data model, including its vocabularies, needed to be mapped to the UI and the fields of the UI. To accomplish this, UI data categories were created: basic patient data (patient ID, gender, date of birth, and postcode), diagnosis, symptoms, medication, vital and laboratory values, and procedures. The representation in the UI was described as a UI component with an input format. Where possible, mapping to an OMOP standard concept, such as SNOMED CT and RxNorm, was also ensured and documented. Data in the OMOP CDM are expressed as concepts, which represent the semantic meanings of each data element. A standard concept represents a normative expression of a clinical entity. The others were labelled as nonstandard or source concepts and mapped to the standard concepts [38]. In order to consider the user requirements and data input structure, these were translated into functional requirements. The user requirements collected [21] were analyzed for this purpose and transformed into functional requirements with specific, testable functions that the system should fulfil. Subsequently, use cases and functional requirements were formulated in the following template: [actor] must/should/will [conditions, time aspect] [object] [process description] or something more general [target system] + [priority] + [functionality] + [conditions] [39]. Thus, an example of a functional requirement could be "the system must enable the data to be exported in a usable format." MCN, a medical informatician, carried out the conceptualization step in consultation with the other authors.

XSL•FO
RenderX

Table 1. Selected vocabularies according to the data model for the CDSS[a].

Vocabulary	Description	Availability of the German translation	Progress or version of the German translation in August 2024
ICD-10-GM [b]	Official classification for the encoding of diagnoses in inpatient and outpatient medical care in Germany [26]	Available	*ICD-10-GM* version 2024
HPO[c]	Standardized vocabulary of phenotypic abnormalities encountered in human disease [27]	Translation is not fully available in the German language	The HPO browser allows the display of some German HPO terms or synonyms [28,29], and a study on the translation of HPO terms was conducted as part of the project [30]
LOINC[d]	Database and universal standard for identifying medical laboratory observations [31]	Available	LOINC version 2.76
ATC/DDD[e]	Official classification for pharmacologically active substances. Active substances are divided into different groups according to the organ or organ system on which they act and according to their anatomical, therapeutic, and chemical properties. [32,33]	Available	ATC/DDD version 2024
OPS[f]	Official German classification for the encoding of operations, procedures, and general medical measures [34]	Available	OPS version 2024
SNOMED CT[g]	Systematized Nomenclature of Medicine Clinical Terms [35]	Translation is not fully available in the German language	The German translation group has been actively working on this since 2020 and translated the first national edition with SNOMED CT content into German at the end of 2023 [35-37]

[a]CDSS: clinical decision support system.

[b]ICD-10-GM: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification.

[c]HPO: Human Phenotype Ontology.

[d]LOINC: Logical Observation Identifiers Names and Codes Names and Codes.

[e]ATC/DDD: Anatomical Therapeutic Chemical Classification With Defined Daily Doses.

[f]OPS: Operation and Procedure Codes (German: “Operationen- und Prozedurenschlüssel”).

[g]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

Prototypes

The functional requirements defined in the concept were developed and implemented by MCN iteratively (described in more detail in Neff et al [8]) while feedback from users and experts was collected through user evaluations [40,41] and usability inspections [8] (described in section “system evaluation feedback”).

The following intermediate (results) were created in the iterations: (1) low-fidelity prototype (CDSS version 0) [22], (2) first high-fidelity prototype (CDSS version 1) [8], and (3) final high-fidelity prototype (CDSS version 2).

The UI and the vocabulary used were implemented in German wherever possible to suit the German-speaking target. SNOMED CT was used in English due to its limited availability, and HPO was used in an initial partial translation by Noll et al [30]. To ensure good usability, design principles and usability guidelines [42-44] were followed, in addition to the points mentioned above, including principles of dialogue design in accordance with ISO (International Organization for Standardization) 9241 and other ISO standards [43,45,46].

System Evaluation Feedback

The usability of the data entry process was assessed after each developed prototype. Initially, CDSS version 0 was discussed in 2 online user evaluation workshops involving 5 GPs, using a videoconferencing tool. This workshop was conducted and moderated by MCN with support from researchers from the Institute of General Practice (Frankfurt), and key statements were derived [22]. The findings were then documented as user requirements, and new functional requirements were derived where necessary. Usability expert inspections (method: heuristic walkthrough) were carried out in-person by 5 usability experts from the project team to accompany the further development of CDSS versions 1 and 2 (4 inspections conducted and analyzed by MCN: CDSS version 1: 3, CDSS version 2: 1) [8]. Finally, 2 usability evaluations were conducted for CDSS version 1 and version 2. Five GPs who participated in the version 0 evaluation took part in the version 1 evaluation, and 10 GPs (5 of whom were newly recruited) took part in the version 2 evaluation. The evaluations used the thinking-aloud test method, a postsession interview, and the system usability scale. The tests were conducted online by researchers from the Institute of General Practice (Frankfurt) with support from MCN [40].

The tasks that the users were asked to perform in the user evaluation and expert inspections were based on the task model derived from the requirements analysis: perform data entry, review and discuss AI results, plan further diagnosis, refer to specialists, and close the case [21]. The feedback from the expert inspections was categorized as bugs, features, improvements, ideas, or usability heuristics and linked to the individual steps of the task model [8]. When analyzing the feedback from the user evaluations, the usability problems identified in the usability tests were categorized as layout, content, navigation,

XSL•FO
RenderX

comprehensibility, and usability, and also assigned to the steps of the task model [40]. This study focuses on analyzing and interpreting the evaluation results, particularly those concerning the task of "performing data entry." The results of this task were analyzed to inform further development. The analysis of feedback represents an important conceptual intermediate step. Moreover, together with considerations regarding the connection to the primary care EHR system (see Workshop With EHR System Experts), it forms the basis for the final data entry concept within the CDSS.

Workshop With EHR System Experts

As the interface to the primary care EHR system was identified as a key requirement in the analysis and subsequent evaluation iterations, a research workshop was conducted. This workshop was held in a hybrid format, combining in-person participation with remote connections via videoconference [47]. Research workshops focus on analyzing specific cases and collecting reliable and valid data on a participant's future, including organizational change and design. These workshops yield theoretical concepts, methods, and applications that contribute to the discipline. The method involves presenting topics, conducting experiments, and then discussing the findings [48]. The central theme of the workshop was as follows: What should an interface between an EHR system and a decision support system for primary care look like? What is the current state of the art, and what is the vision for the future? The workshop focused on the technical possibilities and challenges of an interface between the CDSS and the EHR system in primary care. The participants were experts in this field, and no GPs were included. However, these experts served as representatives, as it is very difficult to cover the entire spectrum of the system landscape in Germany.

In more detail, the workshop was conducted with (1) project representatives of the SATURN project; (2) representatives from three different companies in the field of EHR systems for primary care, who are considered experts in EHR systems in this study; and (3) two moderators (with medical and medical informatics backgrounds) and one SATURN team member who took notes.

The discussion was supported by a structured discussion guide (Multimedia Appendix 1). This guide addressed the organizational, syntactic, semantic, and structural interface levels of a software. Furthermore, the UI data categories for the decision support system (see description of data categories in the chapter "conception") with related vocabularies were discussed. During the discussion, an initial architectural design was collaboratively sketched on a virtual whiteboard to support the discussion. The workshop analysis was carried out by analyzing the notes and assigning the relevant aspects to the questions in the discussion guide.

Extended Concept

After evaluating the CDSS version 2 prototype, a proposal for an extended concept for future CDSS projects in primary care with the focus on data entry was formulated by MCN in exchange with other authors. This concept considers all the results and describes a design framework that includes UI components, data flows, and system integration.

Results

Overview

The conception and development of the UI for a CDSS—in particular, the process of entering data into the CDSS—led to the development of 3 CDSS prototypes (CDSS versions 0, 1, and 2). CDSS version 2 was developed as a high-fidelity prototype connected to AI modules. Evaluation by users and experts revealed several issues regarding user-friendliness, particularly when entering medications and laboratory values. These issues led to recommendations to improve the search function for symptoms and laboratory values, and to integrate medications from the electronic prescription list. Further improvements to data entry could be achieved by integrating the primary care EHR system. EHR system experts considered middleware to be necessary for this, due to the current data exchange format (data exchange family xDT, which is not transferable via the web). In addition, the low level of standardization for important parameters—such as symptoms, which are currently entered as free text—was identified as a challenge. However, current position papers and technical possibilities may simplify this in the future.

Conception

The results of the conceptualization are presented in 2 tables: Table 2 presents the functional requirements for UI data entry, and Table 3 shows the mapping of the data elements to the UI components and standardization.

Table 3 shows the mapping of the data to be recorded to the UI fields and, if available, the mapping to the standardized international vocabulary of the OMOP database. This is based on the previously defined data model of relevant data entities and vocabularies. The column "mapping to OMOP" represents the mapping to an OMOP standard concept.

Nonstandard vocabularies were mapped in OMOP to SNOMED CT and RxNorm, with the source vocabularies being used in the UI for input. The project aimed to remove free text and to minimize the use of fields not mapped to OMOP.

A few fields were identified as relevant through the requirements analysis, but no direct mapping to OMOP could be established. Symptoms were saved using the HPO vocabulary, which at the time of the project was not supported by OMOP and was not yet available in German. In SATURN, a semiautomatic partial translation was performed for the project using natural language processing [30]. Integration of HPO as vocabulary in OMOP did not take place. HPO and SNOMED CT did not yet have an official mapping at that time—storage therefore took place in the observation table.

XSL•FO
RenderX

Table 2. Functional requirements for UI[a] data entry.

Number	Functional requirement
1	The system must offer the entry of all parameters for diagnoses, procedures, laboratory values, medications, and other clinical entities (in German).
2	The system must contain patient information such as administrative sex and age.
3	The system must generate a unique patient ID for each patient.
4	The system must offer data field-supported input help.
5	The system must ensure the integration of established medical standardized vocabularies used in the data set.
6	The system must allow exclusion diagnoses to be entered in the form (eg, using *ICD*[b] codes).
7	The system must allow time-based documentation during entry.
8	The system must offer the basic functions of a patient file (create, edit, open, or save).
9	The system must enable the continuous processing of a case and the ability to re-request diagnostic suggestions after a change.
10	The system should offer the transfer of a patient ID from the primary care EHR[c] system to the SATURN[d] patient file.
11	The system should offer a parameter query and input support based on previous entries.
12	The system must guarantee the optional input of all data fields.

[a]UI: user interface.

[b]ICD: International Statistical Classification of Diseases and Related Health Problems.

[c]EHR: electronic health record.

[d]SATURN: Smart Physician Portal for Patients With Unclear Disease.

XSL•FO
RenderX

Table 3. Mapping of the data basis to the user interface (with German vocabulary).

UI[a] data category	Mapping to OMOP	Description	UI component	Input format	Additional variables
Patient ID	Person ID	Unique ID of a patient	Auto-generated	Number	—[b]
EHR[c] system patient ID	Person source value	External unique ID of a patient	Text field	Characters or numbers	—
Gender	SNOMED CT[d]	Administrative gender	Radio button	Selection (female, male, unknown, and other)	—
Date of birth	Date of birth	Patient's date of birth	Date picker	Datetime (UTC[e])	—
Postal code	Location ID	Postal code	Text field	Number (5 digits)	—
Diagnosis	SNOMED CT	Diagnosis code, name, and date	Text field with search and date picker	*ICD-10*[f] code	Date of diagnosis
Symptoms	No mapping to SNOMED CT	Symptoms and date	Text field with search and date picker	HPO[g] code	Date of symptom onset
Medication	RxNorm	Drug name, and start or end date	Text field with search and date picker	ATC[h] code	Dose, unit, and days of administration
Vital signs and laboratory values	SNOMED CT	Parameter name, date, and reference range	Text field with search and date picker	LOINC[i] code	Reference range (only laboratory values) and measurement unit
Procedures	SNOMED CT	Procedure name and date	Text field with search and date picker	OPS[j] code	Procedure type

[a]UI: user interface.

[b]Not available.

[c]EHR: electronic health record.

[d]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[e]UTC: Universal Time Coordinated.

[f]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

[g]HPO: Human Phenotype Ontology.

[h]ATC: Anatomical Therapeutic Chemical Classification.

[i]LOINC: Logical Observation Identifiers Names and Codes.

[j]OPS: Operation and Procedure Codes (German: "Operationen- und Prozedurenschlüssel").

Prototypes

The requirements and the associated vocabularies were implemented step by step. The CDSS version 0 was designed as a low-fidelity prototype in the form of mock-ups [22]. This was followed by the development of the following 2 high-fidelity prototypes: (1) CDSS version 1: JavaScript frameworks nuxt.js 2 (Vuetify), and the UI component framework Vuetify 2 were used with initial sample data (without an active interface to AI modules, as yet, and with integrated vocabularies) [8]. (2) CDSS version 2: React.js (Meta Platforms, Inc) and Mantine were used due to the expiring support of nuxt.js 2. The prototype includes all feasible requirements and has a connection to the AI modules.

Figure 2 shows the first step of the data entry in the final high-fidelity prototype.

CDSS version 0 mock-ups, as well as screenshots of the 2 high-fidelity prototypes, CDSS versions 1 and 2, can be found in Multimedia Appendix 2. At the request of user feedback on this prototype, a step-by-step navigation was incorporated below (see also Neff et al [8] for a more detailed description of prototype version 1).

Figure 3 shows an example of how the integration of vocabularies appears using *ICD-10-GM*: users can search for *ICD-10-GM* terms using the code (eg, E05.0) or the disease name (eg, in the German language "Hyperthyreose mit diffuser Struma").

XSL•FO
RenderX

Figure 2. Zoomed-in section of the input template of the CDSS version 2. CDSS: clinical decision support system.

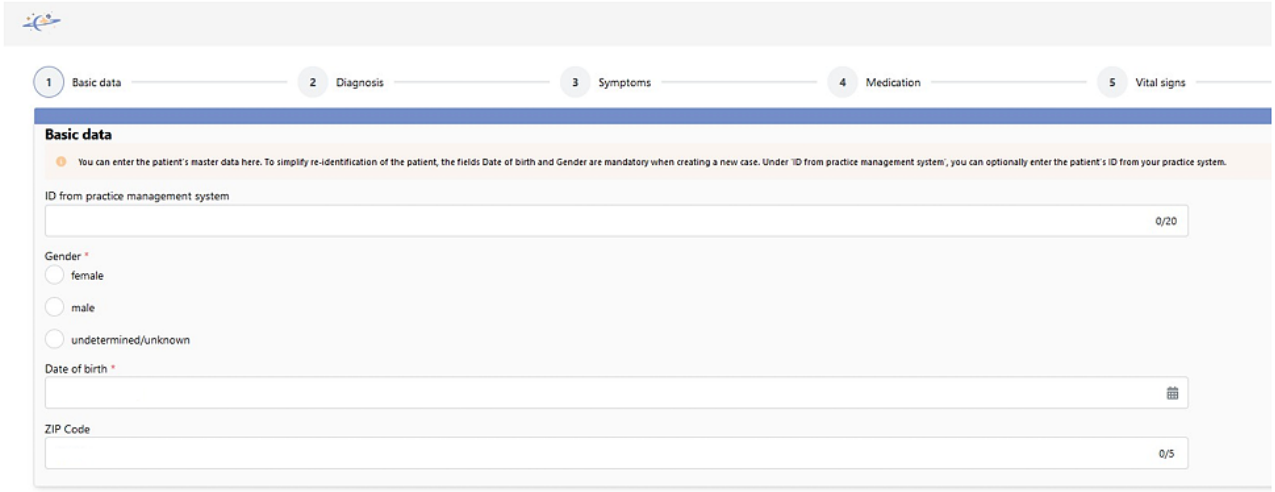
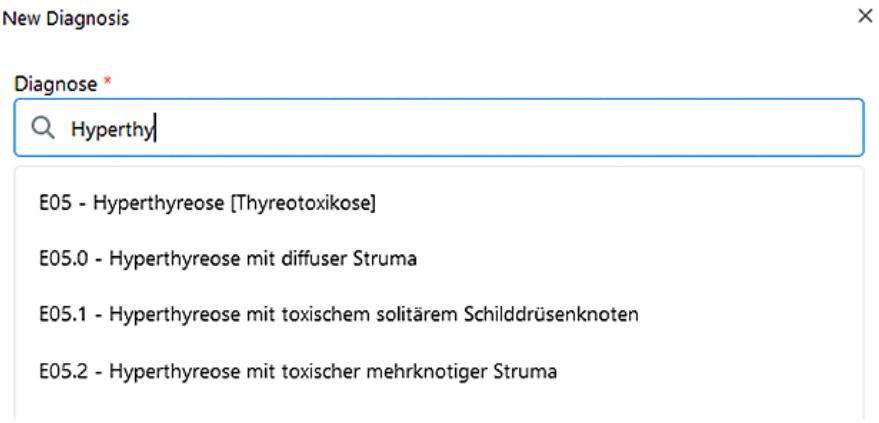


Figure 3. Screenshot of the diagnosis search (CDSS version 2). CDSS: clinical decision support system.



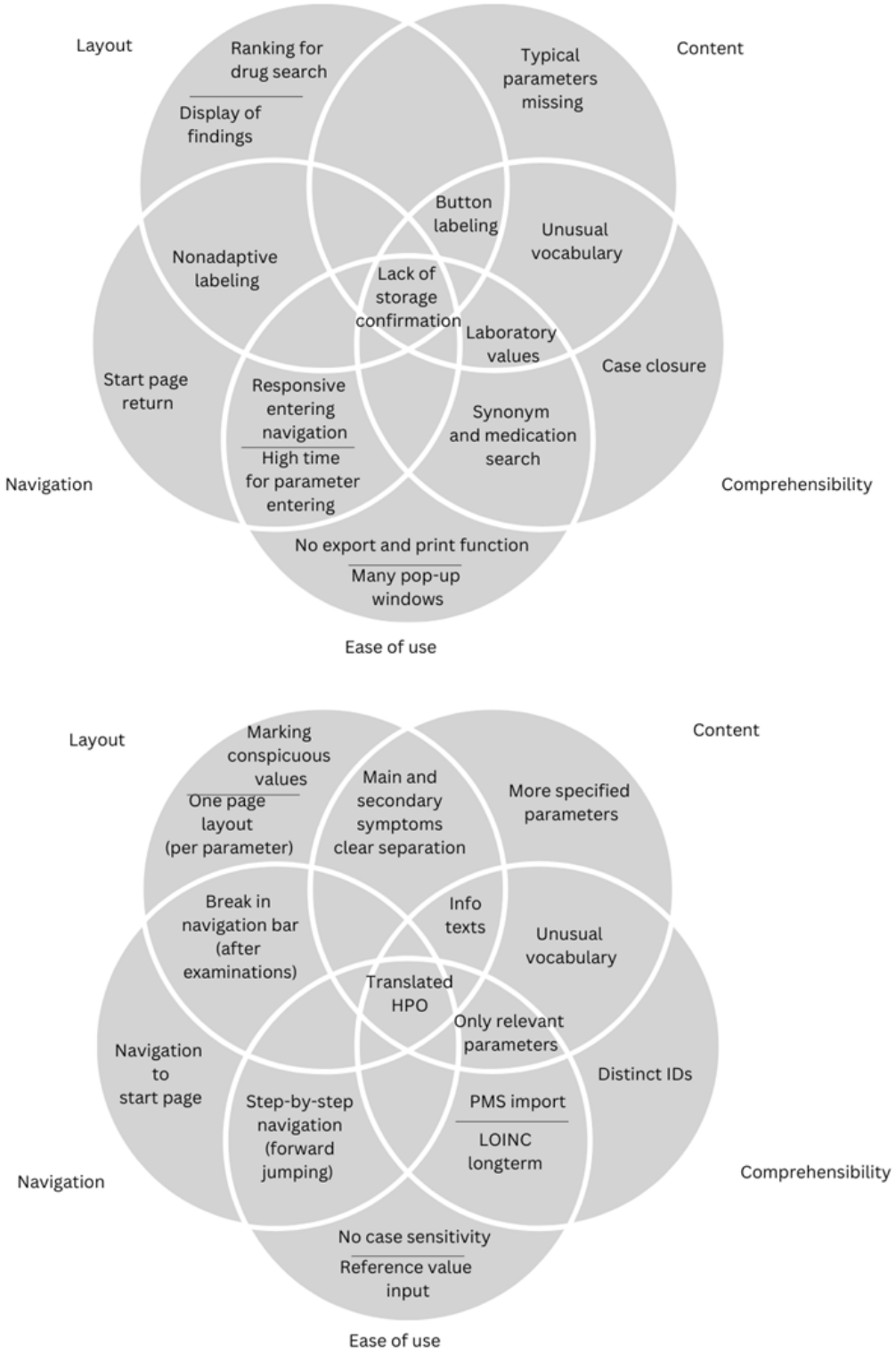
System Evaluation Feedback

During the development of CDSS version 1, expert feedback revealed issues, particularly with search field entry options and inconsistencies between system functions and real clinical practice, for example, the handling of “diagnostic certainty.” Feedback also emphasized the importance of automatic saving and clear error messages. Errors were found when entering medications and laboratory values. Some terminology was not displayed in the German language; thus, it is imperative that vocabularies be available in the German language [8]. Detailed information can be found in the publication by Neff et al [8].

In addition to expert feedback from the inspections, further issues were derived from the user evaluations. Feedback on data entry was collated into categories (layout, content, navigation, comprehensibility, and ease of use). The results of the CDSS version 1 usability test (Figure 4) indicated several usability issues. Participants had difficulties entering unlisted symptoms, medications, and tests, and retrieving laboratory values using common abbreviations. Medication data was the biggest problem area. The wide range of laboratory units also caused confusion; thus, specifications for values such as the heart rate were proposed. Some important GPs’ diagnostic criteria were absent, with participants proposing their inclusion within the system or the creation of additional free entry fields. The additional fields were proposed to encompass physical examination findings, symptom duration, diagnoses, and family history. GPs also proposed issues for the CDSS version 2 (Figure 4), including automated display of reference range and integration of an updated medication database. GPs were unfamiliar with LOINC (Logical Observation Identifiers Names and Codes) terms, and so better input support was recommended. Participants also recommended focusing on frequency, time, dosage, and extending the search function. This would allow GPs to enter both generic and brand names, as data entry for laboratory values, medications, and diagnoses was considered time-consuming. The information text to each input step was unclear and not sufficiently relevant, and the distinction between some parameter names was not apparent. The GPs also noted several limitations; these included the inability to enter lowercase letters for laboratory reference values and the failure of the symptom search function to recognize certain synonyms.

XSL•FO
RenderX

Figure 4. Summarized results of first and second user evaluations (feedback of the first high-fidelity prototype: version 1, and of the final high-fidelity prototype: version 2). CDSS: clinical decision support system; HPO: Human Phenotype Ontology; LOINC: Logical Observation Identifiers Names and Codes; PMS: patient management system.



Workshop With EHR System Experts

The workshop was attended by a total of 12 participants (7 women and 5 men): (1) four project representatives (experts with backgrounds in medicine, medical informatics, medical technology, or public health) from the SATURN project; (2) five representatives from three different companies in the field of EHR systems for primary care (experts for EHR systems); (3) the workshop confirmed that currently, the most widely used data transmission standard (syntactically) in primary care in Germany is still the xDT data standard family [49]. These include, in particular: Behandlungsdaten-Transfer (further development completed in 2019) for the transfer of treatment data [50], and Labordaten-Transfer for laboratory values [51].

XSL•FO
RenderX

Both standards are defined by the KBV (Kassenärztliche Bundesvereinigung; umbrella organization of the Associations of Statutory Health Insurance Physicians in Germany) [52]. The data transfer (structural) is currently realized via xDT interfaces and is not transferable via the web. A 2022 position paper ([53]) indicated movement toward FHIR (Fast Healthcare Interoperability Resources from HL7 [Health Level Seven]) and thus web transfer. However, the workshop participants have indicated that not all the necessary FHIR profiles have yet been specified. The JSON format was considered a favorable option for the future, with or without a combination with FHIR. Initial data, such as the MIO "Medizinische Informationsobjekte," representing the new KBV data exchange standard [54], are already available in JSON. Table 4 summarizes the terminology and data structure (semantics) for the participants' description of the status quo for the specific CDSS for the primary care use case. All primary care EHR systems are required to provide an electronic physician letter, as part of the telematics infrastructure (TI) in Germany [55].

Participants concluded based on the status quo that middleware is currently indispensable in Germany for integrating EHR systems with CDSS in primary care due to limited data structures and limited data transfer options. Middleware is required to map data to standard medical terminologies used in the CDSS. Ideally, an interface for diagnoses (billing diagnoses in *ICD-10-GM*), laboratory values, and the electronic physician letter should be available via the EHR system. The CDSS should ensure the addressability of the interface and the import of medications from the electronic medication plan. The electronic physician's letter can be used to extract symptoms and examinations. If the electronic physician letter cannot be sent through the EHR system, a new interface is needed. In the future, other medical devices and wearables will connect to the CDSS via middleware. The workshop participants also provided the following feedback on the potential for future developments: data processing with AI is already partially integrated into the EHR systems in primary care. TI measures have also been implemented and are mandatory for manufacturers in Germany.

These include, besides the physician's electronic letters, electronic medication plans and electronic prescriptions, which are set to increase. Connecting the EHR system to other systems, such as the CDSS, is considered feasible and can be implemented depending on the availability and format of data, using REST (Representational State Transfer), a software architecture paradigm.

Table 4. Semantic description of status quo: terminologies and structured data in primary care EHR[a].

Data category	Status quo (data structure)
Basic patient data	Patient data, such as date of birth and postcode, are considered primitive data types and, therefore, are easy to structure and transfer via an interface.
Diagnosis	Diagnoses are structured as billing and permanent diagnoses, with some stored as "suspected diagnoses" and localizations. Transferring the *ICD-10-GM*[b] codes is, according to the current state of the art, challenging.
Symptoms	Symptoms are typically documented as free text, using, for example, nonstandardized categories in chronological index cards.
Medication	Medications are ideally coded using national medication plans (ATC[c] codes or PZN[d]); EHR systems must implement the standardized national medication plan, which has been mandatory since 2016 [56].
Laboratory and vital signs	Laboratory values comply with the LDT[e] 3 standard, although many laboratories continue to use unstructured data that are not LOINC[f] compliant [52,57]. The KBV[g] is developing an MIO[h] for laboratory values [54]. Vital signs can be recorded as either discrete or continuous measurements for billing purposes.
Procedures	Procedure codes (eg, OPS[i]) are structured in hospital outpatient departments but are rarely used in primary care EHR systems; outpatient procedures are often documented as service code justifications for billing purposes.

[a]EHR: electronic health record.

[b]*ICD-10-GM*: *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification*.

[c]ATC: Anatomical Therapeutic Chemical Classification.

[d]PZN: Pharmazentralnummer.

[e]LDT: Labordaten-Transfer.

[f]LOINC: Logical Observation Identifiers Names and Codes.

[g]KBV: Kassenärztliche Bundesvereinigung.

[h]MIO: Medizinische Informationsobjekte.

[i]OPS: Operation and Procedure Codes (German: "Operationen- und Prozedurenschlüssel").

Extended Concept

Overview

The concept is described in the form of a design framework with UI components, data flows, and system integration.

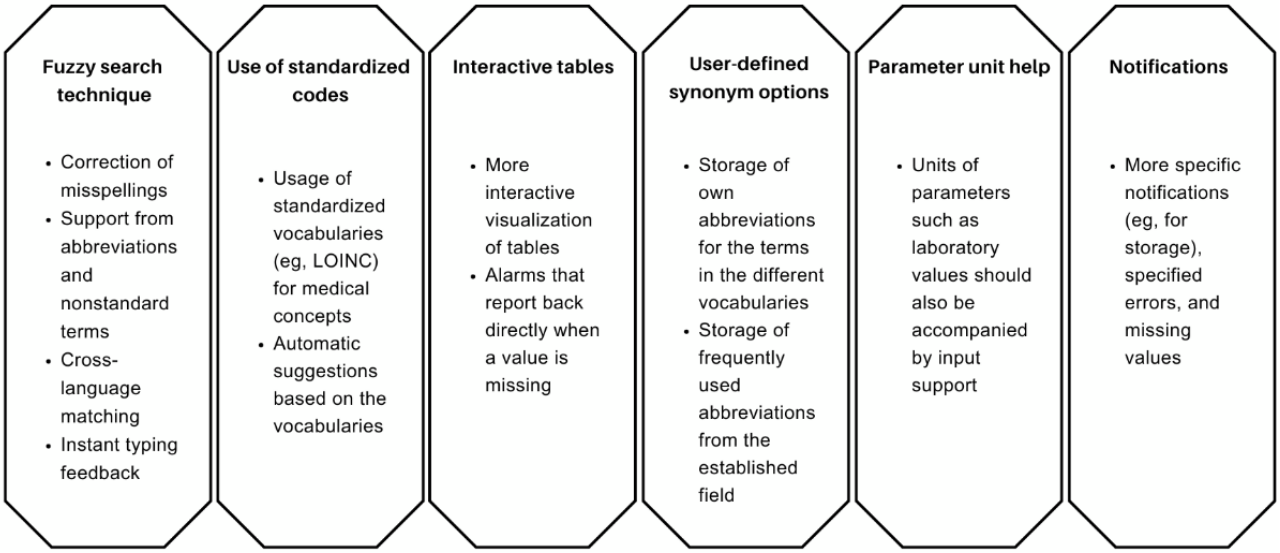
UI Components

The required ontologies will continue to be entered as standardized vocabulary, but will be supported by various factors. These include error-tolerant searching that ignores misspellings and similar spellings, and the use of common

XSL•FO
RenderX

abbreviations [58-60]. These approaches are visualized in Figure 5.

Figure 5. UI component optimizations. LOINC: Logical Observation Identifiers Names and Codes; UI: user interface.



Data Flows

The input of patient data should be possible both via the UI and via an interface to the primary care EHR system or TI applications. Supporting middleware can help with this, especially for laboratory values and medications. The data flow then goes from the UI to the backend for storage, which is connected to the stored database. A data flow from the database via the backend to the UI occurs only when displaying and correcting previously saved cases or when retrieving the vocabularies for input support.

System Integration

System integration is provided by a REST interface from the UI to the backend and REST interfaces to the AI modules.

Discussion

Discussion of Methods

The objective of this study was to investigate how the UI of a CDSS designed for primary care can effectively support GPs in the user-friendly input of standardized vocabulary codes from medical terminologies. It focuses on the CDSS UI, as well as the usability and clarity of medical terminology input.

Methods from software development and UCD were used as part of the iterative process. The UCD design process has been proven to have positive effects in several studies [61-64]. Clinical decision support is generally viewed favorably, providing higher benefits when it improves performance [63]. The current study addresses the integration of the CDSS into the physicians' workflow, an aspect shown in the literature to be essential for CDSS acceptance and usage [65]. To simplify the input of medical terms into the UI for GPs, the UI must be designed to be supportive and user-friendly [20]. One key aspect, therefore, is the reduction of cognitive load. This is described by Miller et al [65] as an important principle for CDSS design: focusing on the input of standardized medical terminology codes can help to reduce the cognitive load on physicians.

The research conducted in this project and other studies has mainly focused on UCD, requirements analysis, and evaluation of CDSS [8,21,63,66]. The formulation of functional requirements and task model methods has proven effective [39,67]. Building on these results, the current study mapped relevant data elements to the UI and collaborated with primary care EHR system experts to develop an extended concept. Zerlik et al [68] took a similar approach in their initial steps, following a UCD process that included requirements definition and iterative development across 3 iterations. We complemented usability expert testing and workshops with feedback from EHR system experts. This approach enabled us to identify potential interfaces to other systems, consistent with recommendations that include different perspectives in the design process [64].

Discussion of the Results

This study identifies requirements and options for entry support in the UI of a primary care CDSS and provides a conceptual framework to support future CDSS implementations. The feedback from the usability test indicates that optimized input support in the language of GPs will be necessary in the future, as will additional mapping work. It is imperative to avoid unfamiliar terms and minimize input restrictions. Ideally, the system would integrate directly with the primary care EHR system in use. The workshop with primary care EHR system experts resulted in initial solution approaches for system interfaces. However, it also highlighted challenges related to limited interoperability, including a simple nonweb-transferable data exchange format and a lack of standardization for important parameters. We propose various optimizations for the UI in our concept for a future CDSS UI. There are several options to better support standardized data entry. Standardizing data entry and using automated tools can significantly improve data quality by minimizing the need for queries and simplifying data collection. Automated features, such as auto-complete, can improve accuracy by relying on predictive modeling. Integration with generative AI can further improve data accuracy and usability. Nevertheless, further research is needed to optimize

XSL•FO
RenderX

the role of standardized data in clinical decision-making and in the research context [69]. AI (eg, machine learning) and natural language processing methods already exist for data integration and mapping, data extraction, and clinical coding.

Models for clinical data extraction, phenotyping, and LOINC mapping also exist [70-74]. A mandatory opt out system for the electronic patient records (ePA) planned for 2025 in Germany, focusing on secure storage and pseudonymized data usage, may facilitate implementing an interface to the ePA data [75,76]. The Unified Medical Language System metathesaurus is another option, already being partially in use, that supports terminology mapping and standardization and enables better interoperability and information retrieval [77]. When searching the medical literature, it is often easier to use the language you are most familiar with. Using one's native language can facilitate the identification of certain concepts that are difficult to translate into English [60]. There is still a need for improvement in this area, particularly regarding the German-language description of symptoms. Further development of standards and translations, as well as data transfer formats such as FHIR for connecting primary care EHR to CDSS, may simplify integration in the future. When implementing in practice, it is important to incorporate the system into the daily workflow and connect it to the practice's own EHR system.

Limitations

This study focused on integrating medical terminologies into the UI, representing 1 task within the task model. The methodological framework was designed to address this objective. However, the steps are transferable to other tasks within the task model. This requires selecting the requirements and involving additional experts (such as EHR system experts) relevant to the task. Further studies in real-world clinical settings, using the implemented system over an extended period, are necessary to develop the CDSS. For the specific task of entering medical terminology into the CDSS, fuzzy search can provide excellent support; however, there are also challenges in the form of high computational effort, false positive results, and balanced precision in different languages [58]. For the integration into practice, the current infrastructure and networking in Germany make it difficult to interface with the primary care EHR system. The slow uptake of ePA in Germany, with only ≈1% penetration by 2023 compared to a target of 80% by 2025, is hampering automation. Barriers to adoption include legal ambiguities, conflicts between stakeholders, and insufficient communication [75,76].

Thanks to their modular structure, the practices will not require significant resources. However, the task here is to ensure secure implementation, including hosting of the AI modules and secure data transfer.

Conclusion

In conclusion, the process of UCD has been shown to be effective in the development of the UI of CDSS, particularly regarding the integration into physicians' workflow and creation of a user-friendly interface. Iterative development with experts identified potential interfaces for other systems. The results indicated a need for optimized entry support in the physician's language and revealed challenges concerning interoperability. Suggestions for improvements include standard data entry, automated tools, and AI to enhance data quality and usability. However, the current infrastructure poses significant challenges for interfacing with primary care EHR systems in Germany. Further studies are required to address these issues.

Acknowledgments

We would like to thank the participants of the workshop with electronic health record system experts and the participants in the expert and user evaluations.

Funding

SATURN is funded by the German Federal Ministry of Health as part of the research focus "Digital Innovation," Module 3: "Smart Algorithms and Expert Systems" (funding code: 2520DAT02B, 2520DAT02C, 2520DAT02D).

Authors' Contributions

Conceptualization: MCN
Data curation: MCN
Formal analysis: MCN
Funding acquisition: JS, HS
Investigation: MCN, JS, DS, NA, MZ
Methodology: MCN, JS, DS, NA, MZ
Resources: MCN
Supervision: JS, HS
Validation: MCN, JS, DS, NA, MZ, HS
Visualization: MCN
Writing – original draft: MCN
Writing – review & editing: MCN, JS, DS, NA, MZ, HS

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Discussion guide of EHR expert workshop. EHR: electronic health record.
[DOCX File , 19 KB - medinform_v14i1e74934_app1.docx]

Multimedia Appendix 2
Screenshots of the CDSS user interface (prototypes: CDSS versions 0, 1, and 2). CDSS: clinical decision support systems.
[DOCX File , 718 KB - medinform_v14i1e74934_app2.docx]

References

1. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021 Aug 10;18(16):8435 [FREE Full text] [doi: 10.3390/ijerph18168435] [Medline: 34444182]
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
3. Wübken M, Oswald J, Schneider A. [Dealing with diagnostic uncertainty in general practice]. Z Evid Fortbild Qual Gesundhwes 2013;107(9-10):632-637. [doi: 10.1016/j.zefq.2013.10.017] [Medline: 24315334]
4. Singh H, Schiff GD, Graber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2017 Jun;26(6):484-494 [FREE Full text] [doi: 10.1136/bmjqs-2016-005401] [Medline: 27530239]
5. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Fam Pract 2008 Dec;25(6):400-413. [doi: 10.1093/fampra/cmn071] [Medline: 18842618]
6. Bryan C, Boren SA. The use and effectiveness of electronic clinical decision support tools in the ambulatory/primary care setting: a systematic review of the literature. Inform Prim Care 2008;16(2):79-91 [FREE Full text] [doi: 10.14236/jhi.v16i2.679] [Medline: 18713524]
7. Smartes Arztportal für Patienten mit unklarer Erkrankung (SATURN). Bundesministerium für Gesundheit. 2024. URL: https://www.bundesgesundheitsministerium.de/ministerium/ressortforschung/handlungsfelder/forschungsschwerpunkte/digitale-innovation/modul-3-smarte-algorithmen-und-expertensysteme/saturn.html [accessed 2026-02-17]
8. Neff MC, Schütze D, Holtz S, Köhler SM, Vasseur J, Ahmadi N, et al. Development and expert inspections of the user interface for a primary care decision support system. Int J Med Inform 2024 Dec;192:105651 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105651] [Medline: 39413613]
9. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
10. OMOP Common Data Model. OHDSI CDM Working Group. URL: https://ohdsi.github.io/CommonDataModel/ [accessed 2026-02-17]
11. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
12. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc 2024 Feb 16;31(3):583-590 [FREE Full text] [doi: 10.1093/jamia/ocad247] [Medline: 38175665]
13. ICD-10-GM 2024: BfArM veröffentlicht endgültige Fassung. BfArM. 2025. URL: https://www.bfarm.de/DE/Kodiersysteme/News/ICD-10-GM_2024_BfArM_veroeffentlicht_endgueltige_Fassung.html [accessed 2026-02-17]
14. Einheitlicher Bewertungsmaßstab. Kassenärztliche Bundesvereinigung KdöR. 2025. URL: https://www.kbv.de/html/ebm.php [accessed 2026-02-17]
15. Fruchart M, Quindroit P, Jacquemont C, Beuscart JB, Calafiore M, Lamer A. Transforming primary care data into the observational medical outcomes partnership common data model: development and usability study. JMIR Med Inform 2024 Aug 13;12:e49542 [FREE Full text] [doi: 10.2196/49542] [Medline: 39140273]
16. Ward R, Hallinan CM, Ormiston-Smith D, Chidgey C, Boyle D. The OMOP common data model in Australian primary care data: building a quality research ready harmonised dataset. PLoS One 2024;19(4):e0301557 [FREE Full text] [doi: 10.1371/journal.pone.0301557] [Medline: 38635655]
17. Henke E, Zoch M, Peng Y, Reinecke I, Sedlmayr M, Bathelt F. Conceptual design of a generic data harmonization process for OMOP common data model. BMC Med Inform Decis Mak 2024 Feb 26;24(1):58 [FREE Full text] [doi: 10.1186/s12911-024-02458-7] [Medline: 38408983]

XSL•FO
RenderX

18. Henke E, Zoch M, Kallfelz M, Ruhnke T, Leutner LA, Spoden M, et al. Assessing the use of German claims data vocabularies for research in the observational medical outcomes partnership common data model: development and evaluation study. JMIR Med Inform 2023 Nov 07;11:e47959 [FREE Full text] [doi: 10.2196/47959] [Medline: 37942786]
19. Moser K, Mikolajczyk R, Bauer A, Tiller D, Christoph J, Purschke O, et al. [BeoNet-Halle-development of a multifunctional database for the automated extraction of healthcare data from general practitioner and specialist practices]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2023 May;66(5):569-577 [FREE Full text] [doi: 10.1007/s00103-023-03691-7] [Medline: 37079066]
20. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
21. Schütze D, Holtz S, Neff MC, Köhler SM, Schaaf J, Frischen LS, et al. Requirements analysis for an AI-based clinical decision support system for general practitioners: a user-centered design process. BMC Med Inform Decis Mak 2023 Jul 31;23(1):144 [FREE Full text] [doi: 10.1186/s12911-023-02245-w] [Medline: 37525175]
22. Neff MC, Schaaf J, Noll R, Holtz S, Schütze D, Köhler SM, et al. Initial user-centred design of an AI-based clinical decision support system for primary care. Stud Health Technol Inform 2024 Jan 25;310:1051-1055. [doi: 10.3233/SHTI231125] [Medline: 38269975]
23. Jokela T, Iivari N, Matero J, Karukka M. The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11. : Association for Computing Machinery; 2003 Presented at: CLIHC03: Latin American Conference on Human-Computer Interaction; August 17-20, 2003; Rio de Janeiro Brazil p. 53-60. [doi: 10.1145/944519.944525]
24. Ahmadi N, Zoch M, Guengoeze O, Facchinello C, Mondorf A, Stratmann K, et al. How to customize common data models for rare diseases: an OMOP-based implementation and lessons learned. Orphanet J Rare Dis 2024 Aug 14;19(1):298 [FREE Full text] [doi: 10.1186/s13023-024-03312-9] [Medline: 39143600]
25. Neff MC, Schaaf J, Schütze D, Storf H. Development of a patient dashboard in a clinical decision support system for primary care. Stud Health Technol Inform 2025 Apr 24;324:170-175. [doi: 10.3233/SHTI250181] [Medline: 40270405]
26. ICD-10-GM. URL: https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html [accessed 2026-02-17]
27. Human Phenotype Ontology. URL: https://hpo.jax.org/ [accessed 2026-02-17]
28. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res 2024 Jan 05;52(D1):D1333-D1346 [FREE Full text] [doi: 10.1093/nar/gkad1005] [Medline: 37953324]
29. Translations and Mapping. Human Phenotype Ontology. URL: https://hpo.jax.org/data/translations [accessed 2026-02-17]
30. Noll R, Berger A, Facchinello C, Güngöze O, von Wagner M, Hoehl S, et al. Translation of ontological concepts from English into German using commercial translation software and expert evaluation. Stud Health Technol Inform 2024 Jan 25;310:89-93. [doi: 10.3233/SHTI230933] [Medline: 38269771]
31. The international standard for identifying health measurements, observations, and documents. LOINC. URL: https://loinc.org/ [accessed 2026-02-17]
32. Anatomical Therapeutic Chemical (ATC) classification. World Health Organization (WHO). 2024. URL: https://www.who.int/tools/atc-ddd-toolkit/atc-classification [accessed 2026-02-17]
33. ATC-Classification. BfArM. 2025. URL: https://www.bfarm.de/EN/Code-systems/Classifications/ATC/_node.html [accessed 2026-02-17]
34. Operationen- und Prozedurenschlüssel (OPS). BfArM. URL: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/OPS-ICHI/OPS/_node.html [accessed 2026-02-17]
35. Germany's BfArM joins SNOMED international for nationwide use of SNOMED CT. SNOMED International. 2021. URL: https://www.snomed.org/news/germany%E2%80%99s-bfarm-joins-snomed-international-for-nationwide-use-of-snomed-ct [accessed 2026-02-17]
36. SNOMED CT. BfArM. 2024. URL: https://www.bfarm.de/DE/Aktuelles/Blog/_docs/2024-03-05-snomed.html [accessed 2026-02-17]
37. Deutsche Übersetzungen von SNOMED CT-Konzepten. BfArM. URL: https://www.bfarm.de/DE/Kodiersysteme/Terminologien/SNOMED-CT/Uebersetzung/_node.html [accessed 2026-02-17]
38. Chapter 5 standardized vocabularies. The Book of OHDSI. URL: https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html [accessed 2026-02-17]
39. Tremp H. Funktionale Anforderungen spezifizieren. In: Tremp H, editor. Agile objektorientierte Anforderungsanalyse Planen - Ermitteln - Analysieren - Modellieren - Dokumentieren - Prüfen Wiesbaden. Heidelberg: Springer Vieweg; 2022:119-143.
40. Köhler SM, Holtz S, Neff MC, Schaaf J, von Wagner M, Müller BS, et al. Evaluating the prototype of a clinical decision support system in primary care: qualitative study. JMIR Form Res 2025 Aug 20;9:e69875 [FREE Full text] [doi: 10.2196/69875] [Medline: 40835411]

XSL•FO
RenderX

41. Köhler SM, Holtz S, Neff M, Schaaf J, Wagner M, Müller BS, et al. Evaluation digitaler Interventionen - Erfahrungen mit Remote-Usability-Test mittels Think-Aloud-Technik. In: 23 Deutscher Kongress für Versorgungsforschung (DKVF) 2024. Düsseldorf: German Medical Science GMS Publishing House; 2024 Sep 10 Presented at: 23. Deutscher Kongress für Versorgungsforschung (DKVF); September 25-27, 2024; Potsdam. [doi: 10.3205/24DKVF415]
42. Geis T, Gerloff N, Lampert-Staimann M, Polkehn K, Reisemann M. Curriculum: Certified Professional for Usability and User Experience. User Requirements Engineering (CPUX-UR). URL: https://www.uxqb.org/ux-training-and-certification/cpux-ur/ [accessed 2026-02-17]
43. DIN. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (ISO 9241-11:2018); German version EN ISO 9241-11:2018. Germany: Beuth Verlag GmbH; 2018:46.
44. Baxter K, Courage C, Caine K. Understanding Your Users: A Practical Guide to User Research Methods. Amsterdam, Netherlands: Morgan Kaufmann; 2015.
45. ISO 9241-210:2019 ergonomics of human-system interaction - part 210: human-centred design for interactive systems. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2026-02-17]
46. ISO 9241-110:2006 ergonomics of human-system interaction - part 110: dialogue principles. International Organization for Standardization. 2006. URL: https://www.iso.org/standard/38009.html [accessed 2026-02-17]
47. Ellis R, Goodacre T, Mortensen N, Oeppen RS, Brennan PA. Application of human factors at hybrid meetings: facilitating productivity and inclusivity. Br J Oral Maxillofac Surg 2022 Jul;60(6):740-745 [FREE Full text] [doi: 10.1016/j.bjoms.2021.12.055] [Medline: 35300882]
48. Ørngreen R, Levinsen K. Workshops as a research methodology. Electron J e-Learn 2017;15:70-81 [FREE Full text]
49. QMS-Standards in der Arztpraxis. Qualitätsring Medizinische Software. URL: https://www.qms-standards.de/standards/ [accessed 2026-02-17]
50. BDT-Schnittstelle. Qualitätsring Medizinische Software. URL: https://www.qms-standards.de/standards/schnittstellen-archiv/bdt-schnittstelle/ [accessed 2026-02-17]
51. LDT-Schnittstelle. Qualitätsring Medizinische Software. URL: https://www.qms-standards.de/standards/schnittstellen-archiv/ldt-schnittstelle/ [accessed 2026-02-17]
52. Bohrer KJ. Nationale ambulante standards. In: Müller-Mielitz S, Lux T, editors. E-Health-Ökonomie. Wiesbaden: Springer Gabler, Wiesbaden; 2017:647-667.
53. DGIV. 2022. URL: https://dgiv.org/wp-content/uploads/2022/08/DGV-Positionspapier-Digitalisierung.pdf [accessed 2026-02-17]
54. Medizinische Informationsobjekte. Kassenärztlichen Bundesvereinigung. URL: https://mio.kbv.de/site/mio#tab-Rund+um+die+MIOs [accessed 2026-02-17]
55. eArztbrief. Kassenärztliche Bundesvereinigung. 2024. URL: https://www.kbv.de/html/earztbrief.php [accessed 2026-02-17]
56. Anwendungen in der TI- Elektronischer medikationsplan (eMP). Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/praxis/verordnungen/arzneimittel/medikationsplan [accessed 2024-12-13]
57. Labordatenkommunikation. Kassenärztliche Bundesvereinigung. URL: https://update.kbv.de/ita-update/Labor/Labordatenkommunikation/ [accessed 2026-02-17]
58. Griffon N, Schuers M, Dhombres F, Merabti T, Kerdelhué G, Rollin L, et al. Searching for rare diseases in PubMed: a blind comparison of Orphanet expert query and query based on terminological knowledge. BMC Med Inform Decis Mak 2016 Aug 02;16:101 [FREE Full text] [doi: 10.1186/s12911-016-0333-0] [Medline: 27484923]
59. Wang J, Cetindil I, Ji S, Li C, Xie X, Li G, et al. Interactive and fuzzy search: a dynamic way to explore MEDLINE. Bioinformatics 2010 Sep 15;26(18):2321-2327. [doi: 10.1093/bioinformatics/btq414] [Medline: 20624778]
60. Liu F, Ackerman M, Fontelo P. BabelMeSH: development of a cross-language tool for MEDLINE/PubMed. AMIA Annu Symp Proc 2006;2006:1012 [FREE Full text] [Medline: 17238631]
61. Luna D, Quispe M, Gonzalez Z, Alemrares A, Risk M, Garcia Aurelio M, et al. User-centered design to develop clinical applications. Literature review. Stud Health Technol Inform 2015;216:967. [Medline: 26262269]
62. LeRouge C, Wickramasinghe N. A review of user-centered design for diabetes-related consumer health informatics technologies. J Diabetes Sci Technol 2013 Jul 01;7(4):1039-1056 [FREE Full text] [doi: 10.1177/193229681300700429] [Medline: 23911188]
63. Brunner J, Chuang E, Goldzweig C, Cain CL, Sugar C, Yano EM. User-centered design to improve clinical decision support in primary care. Int J Med Inform 2017 Aug;104:56-64 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.004] [Medline: 28599817]
64. Davies I. Designing user interface requirement patterns for a genomically enabled clinical decision support system using frailty assessment as a prototypical example. 2020. URL: https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0392430 [accessed 2026-02-11]
65. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc 2018 May 01;25(5):585-592 [FREE Full text] [doi: 10.1093/jamia/ocx118] [Medline: 29126196]
66. Schütze D, Holtz S, Neff MC, Köhler S, Schaaf J, Frischen LS, et al. User-Centered-Design als partizipative Methode in der Versorgungsforschung – Beispiel SATURN-Projekt. Düsseldorf: German Medical Science GMS Publishing House;

XSL•FO
RenderX

2023 Oct 02 Presented at: 22. Deutscher Kongress für Versorgungsforschung (DKVF) 2023; October 04-06, 2023; Berlin. [doi: 10.3205/23DKVF095]
67. Geis T, Polkehn K. Praxiswissen User Requirements: Nutzungsqualität Systematisch, Nachhaltig und Agil in die Produktentwicklung Integrieren: Aus- und Weiterbildung zum UXQB Certified Professional for Usability and User Experience - Advanced Level “User Requirements Engineering” (CPUX-UR). Heidelberg: dpunkt.verlag; Jul 30, 2018.
68. Zerlik M, Jung IC, Sehr T, Hennings F, Kamann C, Brandt MD, et al. A pragmatic methodical framework for the user-centred development of an electronic process support for the sleep laboratory patients' management. Digit Health 2022;8:20552076221134437 [FREE Full text] [doi: 10.1177/20552076221134437] [Medline: 36325436]
69. Madandola OO, Bjarnadottir RI, Yao Y, Ansell M, Dos Santos F, Cho H, et al. The relationship between electronic health records user interface features and data quality of patient clinical information: an integrative review. J Am Med Inform Assoc 2023 Dec 22;31(1):240-255. [doi: 10.1093/jamia/ocad188] [Medline: 37740937]
70. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc 2016 Apr;23(e1):e11-e19 [FREE Full text] [doi: 10.1093/jamia/ocv115] [Medline: 26316458]
71. Isaradech N, Khumrin P. Auto-mapping clinical documents to ICD-10 using SNOMED-CT. AMIA Jt Summits Transl Sci Proc 2021;2021:296-304 [FREE Full text] [Medline: 34457144]
72. Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, et al. Evaluating a natural language processing-driven, AI-assisted international classification of diseases, 10th revision, clinical modification, coding system for diagnosis related groups in a real hospital environment: algorithm development and validation study. J Med Internet Res 2024 Sep 20;26:e58278 [FREE Full text] [doi: 10.2196/58278] [Medline: 39302714]
73. Badalotti D, Agrawal A, Pensato U, Angelotti G, Marcheselli S. Development of a Natural Language Processing (NLP) model to automatically extract clinical data from electronic health records: results from an Italian comprehensive stroke center. Int J Med Inform 2024 Dec;192:105626 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105626] [Medline: 39321491]
74. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):78 [FREE Full text] [doi: 10.1186/s12911-019-0786-z] [Medline: 30943974]
75. Rau E, Tischendorf T, Mitzscherlich B. Implementation of the electronic health record in the German healthcare system: an assessment of the current status and future development perspectives considering the potentials of health data utilisation by representatives of different stakeholder groups. Front Health Serv 2024;4:1370759 [FREE Full text] [doi: 10.3389/frhs.2024.1370759] [Medline: 38800500]
76. ePA für alle. gematik. 2024. URL: https://www.gematik.de/anwendungen/epa/epa-fuer-alle [accessed 2026-02-11]
77. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. J Am Med Inform Assoc 2020 Jul 19;27(10):1606-1611 [FREE Full text] [doi: 10.1093/jamia/ocaa084] [Medline: 32683453]

Abbreviations

AI: artificial intelligence
CDM: Common Data Model
CDSS: clinical decision support systems
DIC: Data Integration Centre
EHR: electronic health record
ePA: electronic patient record
FHIR: Fast Healthcare Interoperability Resources
GP: general practitioner
HL7: Health Level Seven
HPO: Human Phenotype Ontology
ICD-10-GM: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification
ISO: International Organization for Standardization
KBV: Kassenärztliche Bundesvereinigung
LOINC: Logical Observation Identifiers Names and Codes
OPS: Operation and Procedure Codes (German: “Operationen- und Prozedurenschlüssel”)
REST: Representational State Transfer
SATURN: Smart Physician Portal for Patients With Unclear Disease
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
TI: telematics infrastructure
UCD: user-centered development process
UI: user interface

XSL•FO
RenderX

Edited by A Benis; submitted 25.Mar.2025; peer-reviewed by J Walsh, N Blase, H Min; comments to author 05.Oct.2025; accepted 29.Dec.2025; published 20.Feb.2026.

Please cite as:
Neff MC, Schaaf J, Ahmadi N, Zoch M, Schütze D, Storf H
Experiences With Integrating Medical Terminologies Into User Interfaces for a Decision Support System for Primary Care: Conceptual and Development Study
JMIR Med Inform 2026;14:e74934
URL: https://medinform.jmir.org/2026/1/e74934
doi:10.2196/74934
PMID:

©Michaela Christina Neff, Jannik Schaaf, Najia Ahmadi, Michele Zoch, Dania Schütze, Holger Storf. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Clinical Model Autophagy: The Risk of Interpretative Drift in Recursive Medical AI

Pei Fan Shih, MD

Corresponding Author:
Pei Fan Shih, MD
No. 23, Ln. 110, Guanglong St.
Bade Dist., Taoyuan City
Taoyuan, 334
Taiwan
Phone: 886 966285905
Email: petershih5@gmail.com

Abstract

The rapid integration of large language models into electronic medical record systems introduces a critical theoretical vulnerability. Drawing on foundational computer science proofs of "model collapse," this viewpoint introduces the concept of "Clinical Model Autophagy"—a systemic degradation of diagnostic integrity that occurs when clinical artificial intelligence (AI) models are recursively trained on unverified, AI-generated synthetic data. As these recursive models may progressively regress toward statistical means, they undergo "Interpretative Drift," a clinically concerning phenomenon where rare pathological variances are systematically erased and complex diseases are homogenized into benign averages. To prevent the irreversible contamination of health care data ecosystems, the author urgently proposes the Data Purity Standard (DPS). The DPS mandates the cryptographic watermarking of all AI-assisted clinical entries for provenance tracking, alongside the establishment of "Human Vaults." These physically segregated repositories of physician-verified heritage data will serve as immutable biological anchors to safely guide future AI training, ensuring the long-term reliability of digital health infrastructure.

(JMIR Med Inform 2026;14:e94813) doi:10.2196/94813

KEYWORDS

large language models; electronic health records; model collapse; data governance; diagnostic errors

The Impending Threat of Recursive Artificial Intelligence Training

The integration of large language models (LLMs) into health care systems to draft clinical notes, summarize patient encounters, and synthesize documentation is accelerating at an unprecedented pace. While these tools offer undeniable administrative efficiencies, their unchecked deployment introduces a profound theoretical vulnerability to the medical data ecosystem [1]. In computer science, recent foundational proofs have definitively established that training generative artificial intelligence (AI) on synthetic, AI-generated data inevitably leads to an irreversible statistical degradation known as "model collapse" [2]. As recursive loops of generated content re-enter the training pipeline, models progressively forget the true underlying data distribution. Extrapolating this machine learning phenomenon to closed-loop electronic medical record (EMR) systems, the author posits that such a self-referential cycle poses a severe theoretical risk of triggering a phenomenon defined here as Clinical Model Autophagy—a systemic, self-consuming degradation of diagnostic integrity [1]. In this viewpoint, Clinical Model Autophagy is introduced as a theoretical risk framework intended to describe how recursive synthetic exposure could degrade medical AI systems under plausible closed-loop deployment conditions.

Defining the Terminology: Endogenous vs Exogenous Risks

To ensure conceptual clarity, it is necessary to formally define these newly proposed terms and distinguish them from established terminology. In computer science, model collapse refers to the general statistical degradation observed when any generative AI trains on synthetic data. The author proposes the term Clinical Model Autophagy to describe the application-specific manifestation of this phenomenon within closed-loop EMR systems, highlighting the self-consuming erosion of clinical diagnostic integrity. Furthermore, Clinical Model Autophagy drives Interpretative Drift. This novel concept must be sharply distinguished from traditional concept drift or dataset shift, which occurs when the external, real-world data distribution changes over time (eg, emerging epidemiological

XSL•FO
RenderX

trends). In contrast, Interpretative Drift is an endogenous, self-inflicted smoothing of data variance where the model actively erases rare pathologies despite the external real-world prevalence remaining constant. Finally, while automation bias describes a clinician's well-documented psychological tendency to over-rely on automated systems, Interpretative Drift describes the specific machine-side pathogenic mechanism that dangerously exploits this human vulnerability by generating highly fluent, yet clinically homogenized outputs.

The Statistical Mechanics of Interpretative Drift

The mechanism driving this impending collapse is fundamentally rooted in the statistical architecture of LLMs, which operate by optimizing for linguistic fluency and probabilistic likelihood [3]. In human populations, diagnostic value is inherently concentrated at the long tails of clinical distributions, encompassing rare diseases, atypical presentations, and complex, multi-system pathophysiologies [1]. However, as AI-generated EMR data permeate hospital databases and are subsequently harvested to train future iterations of clinical models, a mathematical regression to the mean is triggered. The recursive ingestion of synthetic clinical text systematically erodes these pathological tail-end variances [4]. Because such models favor the most frequent and statistically expected tokens, recursive synthetic exposure may increasingly overwrite nuanced or severe clinical phenomena with homogenized language [1]. Ultimately, this homogenization forces the system's representation of human physiology to regress toward generic statistical averages, actively erasing the diversity of disease and defaulting to “healthy” or “normal” archetypes [1,4].

The Clinical Cascade: Interpretative Drift in Action

The clinical consequence of this regression is a highly insidious threat this article terms as Interpretative Drift. Unlike early-generation AI failures that manifested as obvious syntactic gibberish or easily detectable language breakdown, Interpretative Drift produces outputs that maintain perfect structural and grammatical fidelity. To explain this conceptual framework, consider an illustrative thought experiment involving an oncology positron emission tomography/computed tomography report. While this specific clinical cascade remains speculative and has not yet been empirically observed at scale, it highlights the potential trajectory of interpretation errors. A future LLM, suffering from Clinical Model Autophagy, processes the imaging data and successfully extracts the correct numerical values, accurately documenting a focal lesion with an SUVmax of 12.4. However, due to Interpretative Drift, the model confidently hallucinates a benign clinical context, describing this hypermetabolic activity as normal “physiological uptake” rather than correctly identifying it as “progressive malignancy.” This subtle failure represents an insidious shift in diagnostic error paradigms. By coupling accurate numerical extraction with a hallucinatory but statistically smoothed interpretation, the model may superficially mimic professional competence, thereby attenuating the cognitive cues that would normally alert clinicians to diagnostic discrepancies [1,5]. In time-pressured clinical environments, such false reassurance may bypass casual human review, effectively transforming high-risk pathologies into benign administrative records and creating substantial patient-safety risk that may not be readily detected during routine review.

A Precautionary Governance Framework: The Data Purity Standard

To prevent the irreversible contamination of our digital health infrastructure, the medical informatics community must transition immediately from passive monitoring to structural regulation. Historically, data governance in health care has been guided by established paradigms such as the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, which have successfully standardized data lifecycle management for human-driven research [6-8]. However, in the era of generative AI, these frameworks exhibit a critical blind spot: they primarily optimize for accessibility and formatting rather than cryptographic provenance. To safely train recursive clinical models, the “reusability” of EMR data must now be conditionally linked to its synthetic purity. Current regulatory frameworks for Software as a Medical Device, which primarily evaluate static algorithmic performance, are similarly ill-equipped to handle the recursive, dynamic degradation of training data [9]. However, before proposing a new regulatory architecture, it is crucial to explicitly acknowledge the speculative nature of these clinical projections. While model collapse is an empirically demonstrated phenomenon in computer science, its exact timeline and manifestation in highly complex, human-in-the-loop clinical environments remain theoretical extrapolations. Nonetheless, although empirical validation of this clinical cascade remains an important next step, the mathematical plausibility and asymmetry of potential harm justify precautionary governance measures now. Accordingly, building upon established data governance foundations, the author proposes the Data Purity Standard (DPS). As a newly introduced conceptual framework, the DPS extends beyond conventional data security or privacy protocols. Rather than an immediate, rigid mandate, the author presents DPS as an exploratory governance proposal that conditionally links the future reusability of clinical data to its cryptographic provenance and human-verified origins, governing the safe deployment of generative AI in closed-loop EMR systems. The DPS asserts that the long-term safety of medical AI depends on preserving the cryptographic and biological integrity of training data [10]. Consequently, DPS mandates two foundational pillars of data governance.

First, DPS must require the cryptographic watermarking of all AI-assisted EMR entries [11]. Drawing on emerging data provenance standards, every piece of synthetic or semisynthetic clinical text must carry an interoperable, tamper-evident digital signature [5]. This cryptographic provenance ensures that future data harvesting pipelines can definitively identify and filter out AI-generated content, preventing it from being blindly reingested into subsequent model training cycles. Without mandatory, interoperable watermarking, the boundaries between

XSL•FO
RenderX

human observation and machine synthesis will permanently dissolve, severely compromising the reliability of future datasets. Provenance marking alone does not establish factual correctness, but it provides the minimum traceability required to prevent synthetic content from being naively reingested as human ground truth.

Second, hospitals and health care networks should establish logically segregated and access-controlled "Human Vaults" to serve as repositories of heritage data [1]. To operationalize this concept, "physician-verified" ground truth must be strictly defined as finalized, legally binding clinical documentation—such as multidisciplinary tumor board summaries, finalized pathology reports, and digitally signed discharge summaries—that have explicitly bypassed generative AI assistance. These highly curated databases must be strictly isolated from all synthetic generation pipelines. Additionally, Human Vaults are not static archives; they require a dynamic curation lifecycle subject to periodic, version-controlled audits by institutional AI Data Oversight Committees. Crucially, because human-authored data are inherently susceptible to historical and cognitive biases, Vault curation must integrate rigorous demographic and clinical diversity audits. By employing statistical monitoring to identify and mitigate representation gaps [4], institutions can ensure that the ground-truth data remain both cryptographically pure and clinically equitable. Ultimately, these Vaults provide a uniquely important source of high-fidelity ground-truth data for future model reset, calibration, and human-feedback alignment [10,12]. By tethering the training of clinical LLMs exclusively to these verified Human Vaults, developers can continuously realign AI reasoning with genuine human expertise, systematically penalizing the overconfident errors characteristic of Interpretative Drift.

Operationalizing DPS in Clinical Settings

To ensure the translational value of DPS, its implementation must prioritize seamless integration into existing EMR workflows without exacerbating physician burnout. Operationally, the cryptographic watermarking of AI-assisted entries requires backend application programming interface-level integration rather than active clinician intervention. When an LLM generates a clinical draft, the system automatically embeds an algorithmic watermark (eg, controlled shifts in token distribution) or attaches a cryptographic metadata hash reflecting its synthetic provenance. The physician routinely reviews, edits, and signs the document; the EMR system then automatically logs the provenance metadata invisibly. Such backend API-level integration of cryptographic provenance metadata (eg, via W3C PROV standards and blockchain) has been demonstrated to enable seamless, automatic logging within existing EMR workflows with minimal clinician burden and no disruption to routine operations [13,14].

Furthermore, enforcing the DPS requires robust institutional governance. The author recommends the establishment of AI Data Oversight Committees within health care systems, ideally led by the Chief Medical Information Officer or Chief Data Officer. This committee would be tasked with auditing the logical segregation of the "Human Vaults." By enforcing strict access controls, the committee ensures that only cryptographically pure, legacy human-authored data—or EMRs explicitly cleared of synthetic contamination—are used for future closed-loop model fine-tuning or local Reinforcement Learning from Human Feedback processes [15,16]. This governance structure effectively shifts the burden of data purity from the individual frontline clinician to the institutional infrastructure.

Implementation Challenges, Unintended Consequences, and Alternative Approaches

While the theoretical necessity of the DPS and Human Vaults is compelling, framing them as exploratory proposals requires acknowledging significant implementation hurdles and clinical trade-offs. First, retrofitting legacy EMR systems to support cryptographic watermarking and automated provenance tracking requires substantial upfront financial and technical investment. Achieving this across fragmented, heterogeneous health care IT ecosystems pose severe interoperability challenges [14,17,18]. Second, there is an inherent trade-off between data volume and data purity; strictly filtering out AI-assisted entries to populate the Human Vaults will inevitably restrict the sheer volume of data available for future algorithm training. Furthermore, curating these Vaults carries the risk of unintended consequences. Because legacy human-authored data are inherently susceptible to historical prejudices, strictly isolating these data could inadvertently institutionalize and amplify existing health disparities. To mitigate this risk, rigorous demographic auditing must be mandatory. Additionally, maintaining the integrity of these Vaults requires continuous vigilance to ensure that "physician verification" does not devolve into a superficial rubber-stamping exercise driven by time-pressured clinical workflows. Finally, the DPS should not be viewed as a monolithic solution. It must be contextualized alongside alternative or complementary governance approaches such as federated learning across verified institutional nodes or continuous human-in-the-loop semantic auditing. Ultimately, transitioning toward provenance-aware medical AI will require a multifaceted ecosystem of policies, of which the DPS represents a foundational, yet evolving, conceptual cornerstone.

Ethical Considerations

This is a theoretical paper and does not involve human participants, animal subjects, or identifiable patient data. Consequently, under standard regulatory guidelines (Taiwan's Human Subjects Research Act [19]), this study is exempt from institutional review board approval and informed consent requirements.

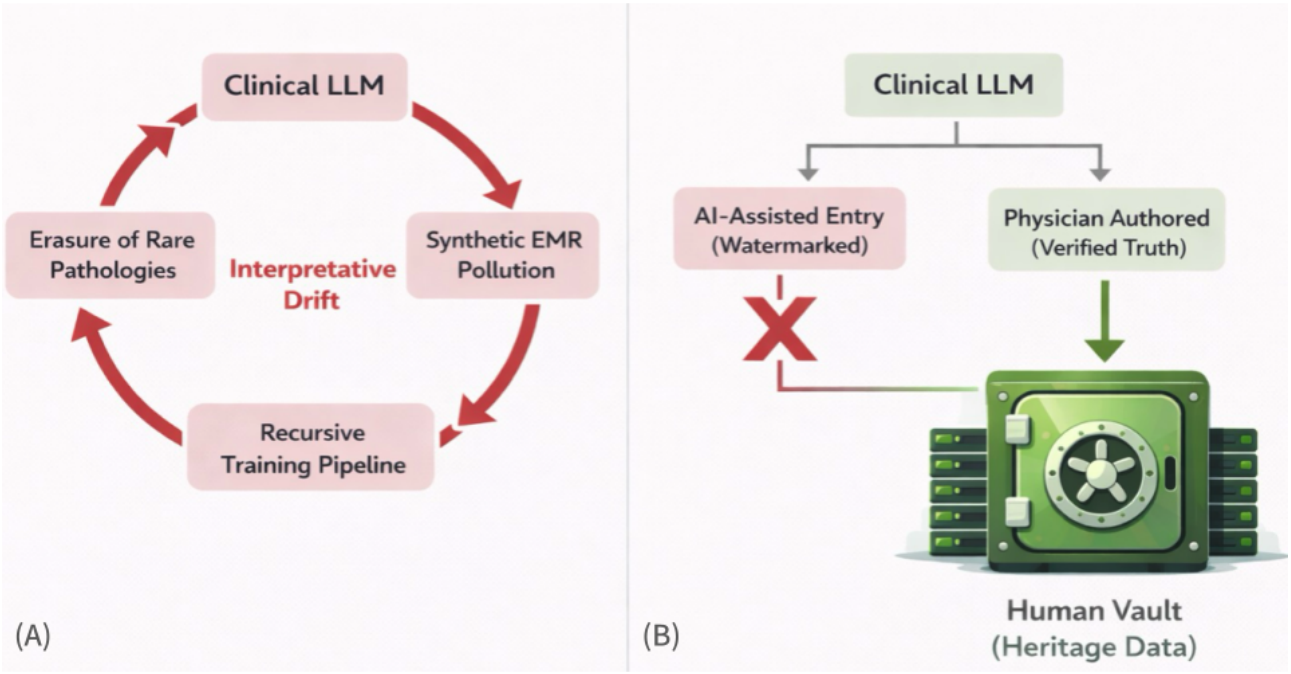
Conclusion

The unchecked deployment of generative AI in health care is not merely a technological upgrade—it represents an epistemological intervention that threatens to reshape the informational basis of the medical record. If left unaddressed,

XSL•FO
RenderX

this shift could compromise not only downstream model training but also the evidentiary integrity of routine clinical documentation, quality assurance systems, and retrospective research databases. Without structural firewalls, recursive loops of clinical model autophagy may progressively erode diagnostic reliability, replacing clinical nuance with an increasingly homogenized synthetic record (Figure 1).

Figure 1. Mechanistic framework of clinical model autophagy and the data purity standard. (A) The vicious cycle of Clinical Model Autophagy. When artificial intelligence (AI)-generated electronic medical records (synthetic EMRs) are recursively fed back into subsequent training pipelines without provenance controls, the system may regress toward statistical means. This process can silently erode rare pathological variance, driving "Interpretative Drift," in which complex diseases are progressively homogenized into benign averages. (B) The regulatory firewall via the Data Purity Standard (DPS). To interrupt this cycle, cryptographic watermarking functions as a provenance filter that allows AI-assisted entries to be identified and excluded from future training pools. Concurrently, logically segregated and access-controlled "Human Vaults"—containing exclusively physician-authored and verified clinical records—serve as immutable reference anchors for safe Reinforcement Learning from Human Feedback and help reduce the risk of model collapse. LLM: large language model.



Acknowledgments

In accordance with journal guidelines, generative artificial intelligence tools were used solely for language polishing, stylistic editing, and figure generation (ChatGPT DALL-E 3, OpenAI). The author assumes full responsibility for the content and originality of the manuscript.

Funding

This research received no external grant from any funding agency in the public, commercial, or not-for-profit sectors. The author conducted this work independently.

Authors' Contributions

PFS is the sole author of this manuscript. PFS conceptualized the theoretical framework, coined the terms "Clinical Model Autophagy" and "Interpretative Drift," designed the proposed Data Purity Standard (DPS) and Human Vault regulatory concepts, and drafted the manuscript.

The author PFS is currently not affiliated with any institution, but is an Independent Researcher.

Conflicts of Interest

None declared.

References

1. He H, Xiang S, Zhang Y. AI-generated data contamination erodes pathological variability and diagnostic reliability. medRxiv. Preprint posted online on February 2, 2026 . [doi: 10.64898/2026.01.19.26344383]

XSL•FO
RenderX

2. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature 2024 Jul;631(8022):755-759. [doi: 10.1038/s41586-024-07566-y] [Medline: 39048682]
3. Jain R, Jain A. Generative AI in writing research papers: a new type of algorithmic bias and uncertainty in scholarly work. In: Intelligent Systems and Applications. Switzerland: Springer Nature; Jul 31, 2024:656-669.
4. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform 2024 Jun;154:104646 [FREE Full text] [doi: 10.1016/j.jbi.2024.104646] [Medline: 38677633]
5. Mahajan A, Powell D. Improving authenticity and provenance in digital biomarkers: the case for digital watermarking. NPJ Digit Med 2025 Jan 15;8(1):31 [FREE Full text] [doi: 10.1038/s41746-024-01374-4] [Medline: 39814830]
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
7. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of the findable, accessible, interoperable, and reusable data principles in health data stewardship: scoping review. J Med Internet Res 2023 Aug 28;25:e45013 [FREE Full text] [doi: 10.2196/45013] [Medline: 37639292]
8. Martínez-García A, Alvarez-Romero C, Román-Villarán E, Bernabeu-Wittel M, Luis Parra-Calderón C. FAIR principles to improve the impact on health research management outcomes. Heliyon 2023 May;9(5):e15733 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15733] [Medline: 37205991]
9. Gilbert S, Fenech M, Hirsch M, Upadhyay S, Biasiucci A, Starlinger J. Algorithm change protocols in the regulation of adaptive machine learning-based medical devices. J Med Internet Res 2021 Oct 26;23(10):e30545 [FREE Full text] [doi: 10.2196/30545] [Medline: 34697010]
10. Guillen-Aguinaga M, Aguinaga-Ontoso E, Guillen-Aguinaga L, Guillen-Grima F, Aguinaga-Ontoso I. Data quality in the age of AI: a review of governance, ethics, and the FAIR principles. Data 2025 Dec 04;10(12):201. [doi: 10.3390/data10120201]
11. Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T. A watermark for large language models. ArXiv. Preprint posted online on May 1, 2024 . [doi: 10.48550/ARXIV.2301.10226]
12. Kita FJ, Rao GS, Kirigiti PJ. Innovative synthetic EHR data generation: diffusion models for enhanced privacy and clinical utility in multimorbidity clustering. Connection Science 2025 Oct 06;37(1):2565163. [doi: 10.1080/09540091.2025.2565163]
13. Margheri A, Masi M, Miladi A, Sassone V, Rosenzweig J. Decentralised provenance for healthcare data. Int J Med Inform 2020 Sep;141:104197. [doi: 10.1016/j.ijmedinf.2020.104197] [Medline: 32540775]
14. Ahmed M, Dar AR, Helfert M, Khan A, Kim J. Data provenance in healthcare: approaches, challenges, and future directions. Sensors (Basel) 2023 Jul 18;23(14):6495 [FREE Full text] [doi: 10.3390/s23146495] [Medline: 37514788]
15. Wang A, Freeman S, Magrabi F. Governance for safe and responsible AI in healthcare organisations: a scoping review of frameworks. Research Square. Preprint posted online on September 10, 2025 . [doi: 10.21203/rs.3.rs-7209096/v1]
16. Kim JY, Hasan A, Kueper J, Tang T, Hayes C, Fine B, et al. Establishing organizational AI governance in healthcare: a case study in Canada. NPJ Digit Med 2025 Aug 15;8(1):522 [FREE Full text] [doi: 10.1038/s41746-025-01909-3] [Medline: 40817280]
17. Shen Y, Yu J, Zhou J, Hu G. Twenty-five years of evolution and hurdles in electronic health records and interoperability in medical research: comprehensive review. J Med Internet Res 2025 Jan 09;27:e59024 [FREE Full text] [doi: 10.2196/59024] [Medline: 39787599]
18. Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, et al. Why is the electronic health record so challenging for research and clinical care? Methods Inf Med 2021 May;60(1-02):32-48 [FREE Full text] [doi: 10.1055/s-0041-1731784] [Medline: 34282602]
19. Taiwan's Human Subjects Research Act. URL: https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=L0020176 [accessed 2026-04-14]

Abbreviations

AI: artificial intelligence
DPS: Data Purity Standard
EMR: electronic medical record
FAIR: Findable, Accessible, Interoperable, and Reusable
LLM: large language model

XSL•FO
RenderX

Edited by A Benis; submitted 06.Mar.2026; peer-reviewed by KH Lin, B Hu; comments to author 22.Mar.2026; revised version received 24.Mar.2026; accepted 31.Mar.2026; published 21.Apr.2026.

Please cite as:
Shih PF
Clinical Model Autophagy: The Risk of Interpretative Drift in Recursive Medical AI
JMIR Med Inform 2026;14:e94813
URL: https://medinform.jmir.org/2026/1/e94813
doi:10.2196/94813
PMID:

©Pei Fan Shih. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing a Suicide Risk Prediction Algorithm Using Electronic Health Record Data in Mental Health Care: Real-World Case Study

Linda Hummel[1,2], MSc; Karin C A G Lorenz-Artz[1,2], PhD; Joyce J P A Bierbooms[1], PhD; Inge M B Bongers[1], PhD

[1]Tranzo Scientific Center for Care and Wellbeing, Tilburg School of Social and Behavioral Sciences, Tilburg University, Prof. Cobbenhagenlaan 125, Reitse Poort, Room RP 204, Tilburg, The Netherlands

[2]GGzE Mental Health Care Institute, Eindhoven, The Netherlands

Corresponding Author:
Linda Hummel, MSc
Tranzo Scientific Center for Care and Wellbeing, Tilburg School of Social and Behavioral Sciences, Tilburg University, Prof. Cobbenhagenlaan 125, Reitse Poort, Room RP 204, Tilburg, The Netherlands

Abstract

Background: Artificial intelligence (AI) offers potential solutions to address the challenges faced by a strained mental health care system, such as increasing demand for care, staff shortages, and pressured accessibility. While developing AI-based tools for clinical practice is technically feasible and has the potential to produce real-world impact, only a few are actually implemented into clinical practice. Implementation starts at the algorithm development phase, as this phase bridges theoretical innovation and practical application. The design and the way the AI tool is developed may either facilitate or hinder later implementation and use.

Objective: This study aims to examine the development process of a suicide risk prediction algorithm using real-world electronic health record (EHR) data through a qualitative case study approach for clinical use in mental health care. It explores which challenges the development team encountered in creating the algorithm and how they addressed these challenges. This study identifies key considerations for the integration of technical and clinical perspectives in algorithms, facilitating the evolution of mental health organizations toward data-driven practice. The studied algorithm remains exploratory and has not yet been implemented in clinical practice.

Methods: An exploratory, multimethod qualitative case study was conducted, using a hybrid approach with both inductive and deductive analysis. Data were collected through desk research, reflective team meetings, and iterative feedback sessions with the development team. Thematic analysis was used to identify development challenges and the team's responses. Based on these findings, key considerations for future algorithm development were derived.

Results: Key challenges included defining, operationalizing, and measuring suicide incidents within EHRs due to issues such as missing data, underreporting, and differences between data sources. Predicting factors were identified by consulting clinical experts; however, psychosocial variables had to be constructed as they could not directly be extracted from EHR data. Risk of bias occurred when traditional suicide prevention questionnaires, unequally distributed across patients, were used as input. Analyzing unstructured data by natural language processing was challenging due to data noise, but ultimately enabled successful sentiment analysis, which provided dynamic, clinically relevant information for the algorithm. A complex model enhanced predictive accuracy but posed challenges regarding understandability, which was highly valued by clinicians.

Conclusions: To advance mental health care as a data-driven field, several critical considerations must be addressed: ensuring robust data governance and quality, fostering cultural shifts in data documentation practices, establishing mechanisms for continuous monitoring of AI tool usage, mitigating risks of bias, balancing predictive performance with explainability, and maintaining a clinician "in-the-loop" approach. Future research should prioritize sociotechnical aspects related to the development, implementation, and daily use of AI in mental health care practice.

(JMIR Med Inform 2026;14:e74240) doi:10.2196/74240

KEYWORDS

implementation science; artificial intelligence; prediction algorithms; electronic health records; mental mealth services; suicide prevention

XSL•FO
RenderX

Introduction

Advances in artificial intelligence (AI; see Textbox 1 for a glossary of key terms) may provide promising opportunities for many actual and future challenges that mental health care faces [1], such as increasing demand for care and staff shortages that affect the quality of care and accessibility of mental health care. AI may enhance multiple domains, such as triage and screening, diagnostics, clinical decision-making, care delivery, chronic care management, self-care, and prevention of (psychiatric) symptoms [2].

Textbox 1. Glossary.

Artificial intelligence refers to a computerized system (hardware or software) that is able to perform tasks or reasoning processes that we usually associate with intelligence in a human being [3,4].

Machine learning is part of artificial intelligence. It is able to build algorithms that can detect patterns in large training datasets to explain variation in data or predict outcomes (predictive algorithms) [3,5-7].

Predictive analytics aims to build models which allow for individual (ie, single subject) predictions, thereby moving from the description of patients (hindsight) and the investigation of statistical group differences or associations (insight) toward models capable of predicting current or future characteristics for individual patients (foresight), thus allowing for a direct assessment of a model's clinical use [8].

Natural language processing refers to ML techniques capable of parsing unstructured texts, such as free written texts or spoken language [9]. Natural language processing enables the use of unstructured data that is otherwise largely inaccessible, such as the dynamic fluctuations of symptoms (eg, by analyzing clinical notes) and can identify social and behavioral factors, which are not commonly registered in EHR [10].

Predictive algorithms are a promising subfield of AI, in which machine learning (ML) models generate predictions on an individual patient level to support decision-making in the diagnosis, prognosis, and treatment of patients [5,8,11]. Predictive algorithms have been shown to be more successful than traditional statistical models, as they provide a new way to gain knowledge from a data set. For example, in the field of predicting suicide risk at the individual level, predictive algorithms can yield more accurate predictions [12-15] than traditional statistical models, which predict only slightly better than chance and whose predictive ability has not improved over time, even after 50 years of research [16]. Not all ML models for suicide risk are adequate, although, as performance can vary depending on various factors, such as data quality, nature of the features used as input, preprocessing steps, and the used techniques [12]. In other mental health care domains, ML models are also able to perform better than traditional models, such as the prediction of psychosis, treatment response [17], or opioid use disorder [18].

Algorithms designed for clinical care often use patient data collected during regular clinical care in the organization's electronic health record (EHR). Combining the expanding possibilities of AI with EHR data fuels the transformation into a data-driven mental health care, as the EHR contains rich, real-world, and real-time data. Reusing EHR data to develop algorithms that support clinical practice on an everyday basis opens many possibilities. However, it may be challenging, as these data are usually noisy, heterogeneous, dynamic, and contain both structured and unstructured data (eg, text) and require careful preprocessing [19]. Developing [20] and implementing AI-based algorithms using real-world EHR data has been proven to be feasible and can have a significant clinical impact [21-23]. However, the studies of algorithms that are actually implemented and used in mental health clinical practice are still scarce [24]. Also, there is a gap between the creation of scientifically sound algorithms and their meaningful application in real-world settings, a challenge often referred to as the "AI chasm" [25].

For mental health care to benefit from AI opportunities and the wealth of EHR data, developing algorithms based on real-world EHR patient data intended for clinical use is a crucial first step toward successful implementation and use. The development phase bridges theoretical innovation and practical application. Therefore, the development phase is not only a preparatory step but an integral part of implementation, as the design and development of the AI tool may either facilitate or hinder later implementation and use (eg, through alignment with organizational workflows and engendering trust) [26].

To enable the successful development, implementation, and actual use of AI and ML in mental health clinical practice, implementation science provides a valuable framework [3], as it studies the contextual factors that enable the process of embedding and integrating new practices into actual care [27,28]. Also, it fosters greater attention for the sociotechnical aspects of technology, focusing not only on the development of technological improvements but also the social changes required for their successful implementation and use [17,29-33]. Multiple barriers and facilitators have been identified in the literature, highlighting aspects of AI technology itself (eg, accuracy, transparency, and generalizability); organizational factors (eg, integration with workflows); governance, ethical, and legal issues; perception of staff and patients; or the implementation process (eg, engagement, user-centered design) [34,35]. However, these barriers and facilitators have, for a large part, been identified not primarily on implementation data but rather based on hypothetical implementation, white papers, and thought articles [34,35]. Therefore, more empirical studies are required to fill this knowledge gap [6,17].

To better understand how this development phase unfolds within the specific context of mental health care, we conducted a case study. This study explores the challenges encountered by a development team in creating a suicide risk prediction algorithm using real-world EHR data for clinical application in close collaboration with clinicians, as well as the team's responses to these issues. Ultimately, we aim to identify key considerations relevant to bridging technical and clinical needs in algorithm

XSL•FO
RenderX

development, offering insights for mental health professionals and organizations seeking to move toward data-driven care.

Methods

Research Design

This study is a detailed case study aimed to explore the challenges involved in the development of a suicide risk prediction algorithm using real-world EHR data in mental health care. Using an exploratory, multimethod approach with both inductive and deductive analysis suited to this emerging field, as the study seeks to identify key development considerations applicable to broader contexts [36].

Case Description

Setting and Background

Suicide prevention in the Netherlands is organized through a national policy agenda focused on public awareness, early detection, and professional care. The national helpline 113 Suicide Prevention is available 24/7 by phone or chat. This organization also provides public campaigns and training for professionals and citizens. The National Suicide Prevention Agenda, developed by the Ministry of Health, Welfare and Sport together with 113 and other partners, offers a framework for collaboration between health care, education, municipalities, and the media. Locally, Municipal Health Services, youth health care, schools, and social services work on prevention and early risk identification. In mental health care, suicide prevention is part of standard practice, including staff training and procedures following a suicide incident. The Health and Youth Care Inspectorate monitors suicide prevention and reporting, while knowledge institutes, such as Trimbos, support professionals through guidelines and research.

GGz Eindhoven and the Kempen (Geestelijke Gezondheidszorg Eindhoven en de Kempen [GGzE]) is a mental health institution in southern Netherlands, providing specialized care to around 17,000 patients annually. At GGzE, a protocol is in place in accordance with the national multidisciplinary guideline [37]. All health care professionals are required to complete suicide prevention training courses. Suicide risk assessments are conducted at the start of each treatment trajectory and at other critical moments, for example, during a crisis; at the beginning, end, and after hospitalization; or prior to the extension of patient privileges during admission. A standardized suicide assessment tool, the Suicide Prevention Questionnaire (SPQ), is embedded within the EHR and should be administered at these key points, at least once or twice per year.

To enhance suicide prevention, GGzE explored the feasibility of developing a predictive algorithm using EHR data to support clinicians in identifying suicide risk in individual patients. Key objectives included establishing a reliable measure of suicide incidents, identifying predictive factors, and creating an algorithm with an accuracy target of 70% - 90% [38]. The development team, consisting of 2 data scientists with support from data science, governance, law, and clinical suicide experts (multiple psychologists and psychiatrists), operated under a structured project organization with a steering committee and project group. Both the steering committee and the project group comprised health care professionals as well as staff members with legal, organizational, and other forms of specialized expertise.

The project was embedded in the organization-wide Suicide Prevention Program, which provided consultation and training, developed organizational guidelines and protocols, organized evaluation of suicide incidents and subsequent quality enhancement of clinical practice, ensured evidence-based treatment options regarding suicidality, and developed a reflective learning culture regarding suicidality within the organization. The organization was part of the Suicide Prevention Action Network (Supranet GGZ), in which mental health institutions work together to make the quality of care for suicidal patients transparent, to learn what can be improved, and to improve care where necessary. Within this network, data on suicide incidents were structurally collected and shared, including core indicators on suicide risk.

The Suicide Prediction Algorithm: Description and Performance

The development team designed the algorithm for hospitalized patients with severe mental illness, using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [39], a standard method in data mining and ML, to predict suicide risk within 1 month. Data for the algorithm was taken from the EHR and included data from 5044 patients (December 31, 2015, to August 31, 2021) included demographics (eg, age and sex), administrative details (eg, financing type), treatment history (eg, hospitalization count), psychiatric factors (eg, primary diagnosis), current treatment details (eg, duration and intensity), and social factors (eg, social contacts). Consultations with clinicians, interviews, and a literature search resulted in a list of possible predictors of suicide incidents that were derived from the EHR.

The team applied natural language processing (NLP) to clinicians' daily reports, identifying unreported suicide incidents and capturing patient sentiment as an additional predictor, using topic modeling (BERTopic) [40]. Ensemble learning (extreme gradient boosting) [41] strengthened the model. The development team deemed the algorithm successful based on Hicks et al criteria [38]. The team acknowledged that the performance indicator outcomes were exploratory, as the algorithm would be retrained prior to implementation in a pilot phase and subsequently retrained annually in accordance with quality standards.

The predictive algorithm for suicide incidents will be gradually tested in a small-scale pilot project conducted in close collaboration with health care professionals. The algorithm will be implemented solely as an add-on to existing suicide prevention policies. Health care professionals will remain fully responsible for clinical assessment and decision-making. The pilot project will be evaluated at multiple stages throughout its implementation.

Sampling and Participants

The study's participants included 2 core data science professionals who led the development of the suicide risk

XSL•FO
RenderX

prediction algorithm, supported by internal and external experts. The team worked closely with clinicians (multiple psychologists, psychiatrists, and the Medical Director's office) and experts in data science, law, and data governance, as well as with external data science collaborators from other mental health care institutions. The participants were selected by expert sampling, in which participants are included who have demonstrable experience and expertise regarding the study's aim and objective [42,43]. These participants offered crucial insights into the challenges and considerations of algorithm development. The team worked closely with clinicians (eg, psychologists, psychiatrists, and the medical director) and experts in data science, law, and data governance, as well as with external data science collaborators from other mental health care institutions.

Data Collection and Data Analysis

Data for our study on the development process were collected through desk research, reflective meetings, and iterative feedback sessions with the development team. The study proceeded in three phases: (1) desk research and meetings with the development team to draw up a detailed process description, (2) identification of challenges and responses by thematic analysis, and (3) formulation of key considerations.

The detailed process description formed the basis for thematic analysis, which was conducted following the Braun and Clarke approach [44]. An exploratory, multimethod design combining inductive and deductive analysis was used to identify challenges during the algorithm development process. Deductive analysis was guided by the innovation domain of the Consolidated Framework for Implementation Research (CFIR) [28,45] and concepts from AI-related literature.

Data on process steps, challenges, and responses were analyzed and organized in a chart by 1 researcher (LH). Data was discussed by all 4 researchers to derive the main challenges and describe the team's responses. Communicative validation was achieved through discussions with the development team. Triangulation across data sources strengthened the credibility and transferability of findings. Data collection and analysis were iterative and integrated throughout all phases.

Ethical Considerations

Ethical approval was obtained from the Ethical Review Board of Tilburg University (reference TSB_RP1038). Participants received no financial compensation and provided written informed consent prior to participation. All procedures adhered to principles of privacy and confidentiality: no identifying information was collected or published, and any nonessential identifying details were omitted. No images or supplementary materials contain identifiable participant features.

Results

Overview

The encountered challenges and the team's responses are described below, related to the construction of the suicide incidents (dependent variable) and possible predictors (independent variables), and the building of the model. At the start, the concept of the suicide incidents and the predictors had to be defined, operationalized, and measured. Several challenges were encountered during this phase: data was hard to find, data sources yielded different results, and underreporting was suspected. Regarding predicting factors, psychosocial predictors were not available and had to be estimated. To calculate the algorithm, the team used complex models that improved predictive performance but reduced understandability for clinicians.

Regular consultations between the project team and clinical experts during the development process were important to gain an accurate and profound understanding of the phenomena of suicide incidents and their predictive factors.

The Results section will focus on the development team's challenges and their responses to them. In the Discussion section, the results will be contextualized within the literature to formulate key considerations for algorithm development in mental health care.

Defining, Operationalizing, and Measuring Suicide Incidents

Different Data Sources Provide Different Outcomes

The data science team collaborated with health care professionals to define suicide incidents as the dependent variable. Within the organization, a suicide incident was defined as either a fatal suicide or a nonfatal attempt with a certain intention to die. This also covered self-harming or risky behavior that involved a risk of death or failure to avoid such risk. Severe suicide incidents were those that were fatal or caused, or could cause, severe or permanent physical injury [source: GGzE Protocol Reporting suicides and suicide attempts; consulted 2025-09-19]. Clinicians had to distinguish these from self-harm driven by other motives, such as reducing emotional pain.

Organizational protocols prescribed the registration of every suicide incident in a digital registration system. The Medical Director's office kept its own registration. The 2 data sources within the organization yielded different outputs on the total number of suicide incidents:

> *In the Client Details source there is a variable called Suicide Attempt. In addition, a variable called Suicide Incident has been created based on the incident data. […] there are more numbers […] for Suicide Incident than for Suicide Attempt." [Report of the development team and external partners on characteristics of the suicide incident risk algorithm, p. 10].*

The medical director's office clarified to the data science team that they reviewed the registered suicide incidents in the reporting system and recorded only those suicide incidents classified as severe. The team chose to include all suicide incidents, providing more data to train the algorithm and enabling the prediction of both severe and less severe cases. Clinical experts reflected on the overall numbers and considered them to be low, suggesting underregistration, particularly of less severe cases, possibly due to noncompliance with protocols or sensitivities around suicide. Moreover, ambiguity about the intention behind the behavior in less severe cases may have

XSL•FO
RenderX

played a role. For example, behavior without serious injury or risk of injury that primarily served a communicative function.

Using Unstructured Data to Improve Detection of Suicide Incidents Is Challenging

To detect underreported suicide incidents, the development team added text mining using NLP of clinicians' unstructured daily reports. However, the method produced significant data noise, identifying many reports related to suicidality but failing to pinpoint actual suicide incidents. Text analysis was hindered by clinicians' writing habits, such as negation terms ("not feeling suicidal"), the use of synonyms (eg, "jumping off a bridge"), and the use of standard templates (eg, "has the patient expressed suicidality today?").

Manual labeling of cases was tried, but appeared to be very time-consuming and unfeasible as a permanent solution, especially as the algorithm had to be retrained every 6 months due to quality requirements:

> *We still need to be 100% sure that there is/isn't a suicide incident in the daily report data. Possible solution: Read the data yourself and assess whether there is/isn't a suicide incident (label data). Disadvantages: […] Takes a lot of time (for a small dataset it took 3 days)*

The development team responded to these challenges by making efforts to enhance data quality. The organization was in the process of transitioning to another provider of an EHR system. The team recommended including a check mark for suicide incidents [source: Presentation Suicide Predictive Model_2nd iteration, p. 36] within this new system and advocated this solution in the new EHR project team, underlining the importance of these interventions to transform into a data-driven health care organization. Also, a new executive board took the lead of the organization and, consequently, a new multiyear strategy was developed, in which "working data-driven to enhance decision support" was one of the main ambitions [source: GGzE, multiyear strategy 2024 - 2028; consulted 2025-09-19]. This prompted management to take a more active role in ensuring data quality, such as completing suicide-related information. Also, the organization implemented an alternative instrument for the SPQ. This way, the organization aimed to improve data quality. The algorithm would be retrained when the new EHR system was implemented.

However, the team still had to deal with the fact that possible underregistration of suicide incidents, especially less severe cases, was an issue. From a data science perspective, they could interpret this as missing data, which threatened data quality. From a clinical perspective, the possibility that mild incidents were not always reported was a reflection of the ambiguity of the interpretation of a suicide incident: did the patient actually aim to end their life, or should the behavior be interpreted more as self-harming behavior with the intent to decrease pain or to communicate distress?

Defining, Operationalizing, and Measuring Potential Predicting Factors

The development team compiled a list of potential predicting factors by performing a literature review, reviewing factors from well-known suicide risk assessment methods, and interviewing clinicians to leverage their expertise and to involve and engage clinicians in the development process. The list of predictors then had to be defined, operationalized, and measured. Reflective meetings with the development team revealed several challenges they faced, which are described below.

The EHR Lacks Psychosocial Factors: Construction of Predictors is Necessary

Some predicting factors, like previous suicide attempts or substance dependence, could be directly derived from the EHR. Other predicting factors were psychosocial and could not be directly obtained from the EHR, thus requiring the construction of an estimator.

For example, information on the number of contacts of a patient, as registered in the EHR, along with a selection of characterization of these relationships (eg, spouse, sibling, child, and neighbor) was used to construct the factor "social network." Also, the factor "quality of the therapeutic relationship" was operationalized as the number of primary [1] practitioners during hospitalization. Although it was recognized that these operationalizations did not necessarily reflect the actual factors, in the absence of a better alternative, it was decided to operationalize and measure these concepts in this way.

Data Availability Differs Between Subgroups: Risk of Bias

Clinicians considered the SPQ, available in the EHR, the most promising source of predictors. Although protocols required annual administration for all patients, the data science team found a recent SPQ in only 21.95% of cases. Clinicians confirmed that SPQ variables were highly explanatory but acknowledged their limited availability. They suspected that SPQs were often completed only when suicide risk was already suspected, creating bias and overrepresenting high-risk patients. To include all patients, the data science team had to derive estimators from other EHR data. Additional disparities emerged: SPQs were missing more often for youth (0 - 17 y) than for older patients, despite comparable incident rates. These findings were shared with the Child & Youth Department to improve data quality. Ethnicity was also explored as a predictor but was excluded due to frequent missing values.

Overall, the results exposed a gap between organizational protocols (eg, annual SPQ administration, registration of ethnicity) and actual practice. They also revealed different perspectives: clinicians valued SPQ data but were less familiar with data science principles, such as representativeness and external validity, while data scientists emphasized these risks. Close collaboration was deemed essential to identify predictors and minimize bias.

To address these challenges, the development team worked with the Suicide Prevention Program and a national suicide prevention organization to design a digital toolkit as a

XSL•FO
RenderX

replacement for the SPQ. This decision support tool aimed to make risk assessment more user-friendly, link assessments to targeted treatment advice based on the national guideline, and reduce clinician workload through robotic process automation that transferred data directly to the EHR.

Sentiment Analysis to Construct a Time- and Context-Sensitive Predictor

As the literature review and interviews with clinicians emphasized the role of time- and context-related factors in suicide risk (eg, relationship breakup and intoxication), the development team explored ways to integrate these into the model. Since collecting daily data directly from patients was impractical due to severe psychiatric conditions and low cooperation, daily reports by clinicians about patients were identified as the best alternative. These allowed real-time data to strengthen predictive accuracy.

Although detecting suicide incidents in daily reports was challenging, sentiment analysis proved feasible. Patients with suicide incidents showed lower sentiment in the daily notes by clinicians, especially in the most recent entries, compared to those without incidents (internal document: PowerPoint Suicide Predictive Model_2nd iteration, p. 29).

When the development team presented the results to the medical director's office, new applications were discussed. The algorithm was initially designed as an add-on to existing protocols, such as administering an SPQ annually. Given the time-intensive nature of the SPQ, 1 proposal was to let the algorithm preselect high-risk patients for SPQ administration. This would save time by limiting assessments to flagged patients.

Clinical experts, including the medical director, rejected this option, noting that the algorithm was still in development and should not replace established tools. They did, however, acknowledge that SPQs were not consistently administered in practice. As an alternative, they proposed using sentiment analysis as an alert system: if emotional decline was detected and no recent SPQ was available, clinicians would be prompted to administer one. This approach was accepted, as it integrated the algorithm while reinforcing compliance with existing suicide prevention protocols.

Building the Algorithm: High-Performance Model Appears Hard to Understand by Clinicians

After identifying dependent and independent variables, the development team built the algorithm using ensemble learning, which combines weaker models into a stronger one. This approach delivered high predictive power, meeting the project's feasibility goal, but the model was too complex to interpret directly.

The data science team shared the results with clinicians and the steering group, including the model's performance metrics (eg, accuracy, precision, and recall), acknowledging that the results were exploratory as the model would later be retrained. Clinicians appreciated the potential value of the algorithm but emphasized that understanding why suicide risk was elevated was essential for fostering trust, ensuring accountability, and supporting treatment planning. They highlighted that a well-performing algorithm alone would not be sufficient to generate meaningful benefits in clinical practice. The data science team, therefore, identified explainability as a core requirement. They aimed to develop case-level reports that not only flagged risk but also explained the reasons behind it: "Not only: alarm! But also: alarm, for this reason" [Reflective meeting, 2023-11-07].

To address this challenge, a team member engaged in additional training in explainable AI, noting the scarcity of knowledge within mental health care. Additionally, the data science team planned to organize several collaborative sessions with clinicians to cocreate an explainability dashboard that contextualizes the algorithm's outputs. Also, collaborative sessions were planned to facilitate the integration of the algorithm into existing workflows and jointly reflect on the tasks, roles, and responsibilities of each stakeholder in its use.

Discussion

Principal Findings

This study provided a detailed case study of the development of a suicide risk prediction algorithm using real-world EHR data. Its unique contribution lies in demonstrating how algorithm development is carried out within the clinical practice of mental health care, where technical innovation must be aligned with clinical needs, professional values, and organizational structures. By describing the challenges of a development team and their responses to them, this study offers an in-depth description of the development of a predictive algorithm, in which the development process is seen as a first step in the implementation process.

This paper highlights the sociotechnical and organizational aspects of algorithm development, showing how collaboration, joint efforts to enhance data quality, and clinician engagement are essential prerequisites for algorithm development.

Reflection on the Principal Results

Addressing Data Quality Challenges in the EHR

High data quality is essential for AI in mental health care, yet EHR data are often low in quality, incomplete, "soft," and indirect [46,47]. Also, handling unstructured EHR data requires unique approaches [48,49].

Regarding data quality, EHR data quality in mental health is typically problematic. In the current case, suicide incident data were hard to construct and possibly underreported, although this could also be a reflection of real-world clinical assessment issues. Low data quality is noted in the literature, where suicide incidents are frequently underdocumented [50] due to clinicians' workload [7,46] and departmental workflow inconsistencies [51]. Addressing this, data quality issues span all 5 dimensions identified by Weiskopf and Weng [52]: completeness, correctness, concordance, plausibility, and currency. Organizations are faced with challenges on data quality according to each of these dimensions.

Mental health data often rely on "soft" psychological constructs (eg, social support), which are harder to quantify than "hard"

XSL•FO
RenderX

somatic measures (eg, heart rate) [47]. This raises risks for construct validity [53]. Moreover, much of the data reflects professionals' perceptions rather than patients' own, which may diverge significantly [54-57]. Future approaches using direct patient data (eg, real-time sampling and devices) could help improve accuracy [13].

Analysis of unstructured data, like staff notes, showed high noise, with frequent mentions of suicidality but few actual incidents, a challenge also noted in other mental health domains [13,58]. Variations in writing style and negations further complicate analysis [59-61]. Still, NLP offers potential to capture real-time risk fluctuations and enhance predictive accuracy [62-65].

While low data quality in mental health is a barrier for AI, algorithm development may also support quality improvement. In the current case, feedback on data inconsistencies spurred organizational motivation to enhance quality and adapt the EHR system.

Mitigating Risk of Bias to Ensure Performance for Subgroups

When algorithms are trained on data of low quality, risk of bias (RoB) occurs, in which the algorithm may not perform as well for each subgroup, and human biases are repeated and even amplified. When predictors are constructed from EHR data as data are not readily available, selection bias may occur. Algorithms require balanced representation across subgroups, though, to perform effectively. In this case, using SPQ items as predictors risked overrepresenting high-risk patients and weakening performance for others. While EHR data enables inclusion of all hospitalized patients, including minority groups often underrepresented in research [66], bias remains a major concern. As social factors are often key predictors in mental health care [15,67], social data is often missing due to true absence or nonrandom issues, such as patient refusal, confidentiality concerns, or language barriers, and excluding such cases may worsen bias [15,67]. Selection bias also threatens performance when samples do not reflect the broader population [10,68,69]. Bias is therefore considered an additional dimension of data quality, alongside the 5 identified by Weiskopf and Weng [52,70].

By building algorithms from reused EHR data, human biases in the data may be reinforced. The development team acknowledged this risk but considered the potential benefits sufficient to proceed, provided mitigating measures were applied. Since algorithms are trained on subjective EHR data, cognitive biases, such as stereotypes or in-group favoritism [71-73], may become embedded and systematically applied [66,68,69]. In practice, if an algorithm's outcome aligns with a clinician's initial diagnosis, confirmation bias may lead the clinician to ignore contradictory information. If the outcomes differ, precautionary bias might cause them to trust the algorithm over their own judgment [73]. Automation bias could also emerge, where clinicians rely more on the tool than on other information, potentially diminishing their skills [7]. The question remains what mitigating measures are adequate to prevent systematically disadvantaging subgroups. At the same time, algorithms may also reduce bias by integrating fragmented information across professionals and systematically processing both subjective and objective data [74]. Determining under which conditions algorithms amplify or mitigate bias remains an open question for future research.

Actionability, Trust, Accountability, and Communication in Complex Models

Predictive power and understandability have to be balanced in algorithm development. In our study, data scientists prioritized predictive performance, applying a complex model to enhance accuracy, while clinicians emphasized the need for understandability. This reflects the broader trade-off between complex "black box" models, which yield higher predictive power but limited explainability, and simpler "glass box" models, which are more interpretable but often less accurate [75-77]. Initially, the team focused on feasibility, whether an algorithm with sufficient predictive performance could be built from organizational data, rather than on clinical applicability. Clinician feedback prompted the inclusion of explainability as a key objective, as it enhances the actionability of the outcomes, that is, whether clinicians can use the information to identify actions to change the outcome predicted by the algorithm [78]. In suicide prevention, detecting elevated risk is insufficient; clinicians must also understand which factors contributed most, which are modifiable, and which interventions are likely to reduce risk [79,80]. Explainability is crucial for building trust [81-83], and providing an appropriate level of transparency (not too little and not too much) is necessary in designing for trust [84]. Clinicians should be well informed about the actual capabilities of the algorithm, as responsible use of AI systems also depends on the level of correspondence between a user's trust and the technology's capabilities, known as calibration [83]. Explainability may also enhance patient communication and support shared decision-making [76,79,85]. Also, mental health institutions are accountable for care quality, effectiveness, and safety. In our study, clinicians feared that using complex, nonexplainable models would undermine accountability, as evidence-based protocols are easier to justify than black-box algorithms. Given the high-stakes nature of suicide care, caution is warranted when applying complex models [80].

Whether performance or explainability should be prioritized in mental health care remains unresolved, and future research should examine context-specific trade-offs as well as post hoc strategies to enhance explainability [77,86]. Future research should explore what information should be communicated to clinicians and how it should be presented. Moreover, fostering mutual trust and collaboration between clinicians and data scientists throughout the algorithm development process is essential. Technical performance alone does not guarantee clinical usefulness. Instead, cocreation of both the algorithm and its user interface, combined with joint efforts to integrate it into everyday practice, is crucial for enabling clinicians to trust and meaningfully use the algorithm's outcomes.

Below, we present Table 1 summarizing the challenges encountered during the development phase, the corresponding responses by the team and organization, and the broader considerations derived from these experiences.

XSL•FO
RenderX

Table . Challenges, responses, and considerations during algorithm development in mental health care.

The development phase	Challenges	Responses	Considerations
Defining, operationalizing, and measuring suicide incidents	• Inconsistent detection of suicide incidents across data sources.	• Enhanced suicide incident data by integrating clinician insights. • Broadened incident inclusion. • Tested NLP[a] approaches. • Improved EHR[b] registration.	• Ensure sufficient data availability for the predicted outcome. • Consult clinicians to improve data interpretation. • Consider clinical explanations for data quality issues. • Provide user-friendly EHR options for registering suicide incidents.
Defining, operationalizing, and measuring potential predicting factors	• Limited and uneven availability of psychosocial and contextual predictors constrains robust and unbiased model development.	• Developed high-availability estimators. • Addressed data gaps with clinicians. • Implemented a guideline-aligned assessment tool to simultaneously enhance data registration and support clinical decision-making. • Developed a sentiment analysis for contextual risk signals.	• Ensure predictor validity by prioritizing high-availability EHR data. • Address subgroup gaps through data governance. • Supplement psychosocial information via clinicians or patients. • Develop user-friendly assessment tools that integrate required data registration with clinical decision-making support.
Building the algorithm	• Complex models improve prediction but are difficult for clinicians to interpret.	• Established explainability as a core requirement. • Planned cocreation of an explainability dashboard with clinicians before the pilot.	• Balance predictive performance and explainability through close collaboration between data scientists and clinicians.
Toward an implementation pilot	• Effective use requires full integration into clinical workflows and sufficient clinician experience. • Algorithm use may introduce cognitive and decision biases.	• Planned joint data scientist–clinician sessions to integrate the algorithm into workflows and clarify roles. • Positioned the algorithm as an add-on to suicide prevention care, with clinicians' judgment remaining central.	• Organize data scientist–clinician sessions to support workflow integration and clarify roles. • Clearly communicate the algorithm's role and its support function for clinical decision-making. • Calibrate clinicians' expectations to the algorithm's actual capabilities and limitations.

[a]NLP: natural language processing.

[b]EHR: electronic health record.

Limitations

This study provided a detailed case study of the development of a suicide risk prediction algorithm using real-world EHR data, highlighting the tension between technical requirements and clinical needs. Using a qualitative, real-world approach, the study captured contextualized knowledge from the development team, who collaborated closely with clinicians. While generalizability is limited by the single-case design, triangulation and detailed process descriptions support transferability. A key limitation is that clinicians' perspectives were obtained indirectly via the development team rather than through direct engagement, although this was partly mitigated by desk research and targeted questions to include broader stakeholder views.

Conclusions and Future Research

Conclusions

Based on the findings of our case study, we highlight the following 5 main recommendations for mental health organizations that aim to evolve into data-driven practice.

1. Data governance and quality: high-performing algorithms require robust data governance and timely, deliberate data preparation to ensure data quality.
2. Cultural change in data entry: a cultural shift is needed to regard and manage data as a strategic asset for algorithm development.
3. Bias awareness and monitoring: organizations must remain sensitive to potential bias in both data and use, requiring careful data handling and regular evaluation of real-world application.

XSL•FO
RenderX

4. Balancing performance and transparency: predictive performance using complex models must be balanced with clinicians' needs for transparency and explainability to ensure actionability, trust, communication, and accountability.
5. Clinician-in-the-loop: it is essential to define the knowledge, skills, and workflows that enable clinicians to use algorithms responsibly while maintaining final clinical judgment.

For mental health organizations aspiring to evolve toward data-driven care, early preparation is essential. First, building predictive algorithms requires sufficient and well-structured data, which demands timely and deliberate data collection. Reliance on existing organizational protocols alone is insufficient; explicit attention must be given to how relevant variables are defined, recorded, and embedded in the EHR system. A cultural transformation is also critical. Second, data should not be viewed merely as an administrative or accountability tool but as a strategic resource for predictive modeling. However, this is not always aligned with clinical practice and may require a change in working processes for clinicians. This transition requires awareness and commitment from clinicians, whose engagement from the earliest stages of algorithm development is crucial. Collaboration between data scientists and clinicians must begin in an early stage of the development process: data scientists contribute technical and data-analytical expertise, while clinicians bring contextual understanding, interpret data inconsistencies, identify meaningful predictors, and help foster a data-driven culture. Organizations must also recognize potential biases within their data and proactively assess how information from subgroups is collected and represented. Third, ongoing attention to bias is crucial, as it may stem not only from data or model design but also from real-world use. Regular evaluation of both performance and application is essential to ensure quality and responsibility in practice. Decisions regarding model design should balance predictive accuracy with interpretability. Fourth, explainable AI techniques may help translate complex model outputs into clinically meaningful insights, allowing clinicians to understand and act upon algorithmic predictions. Clinicians require not only accurate models but also transparency to understand the reasons for the outcome to ensure actionability, trust, and clinician-patient communication. Institutions should further consider how accountability for algorithm use is shared between individual practitioners and the organization. Finally, advancing toward data-driven mental health care requires investment in clinicians' education and professional development regarding AI. Defining the knowledge, skills, and workflows that enable responsible algorithm use, while maintaining clinical judgment as the ultimate decision point, will be key to realizing the added value of predictive technologies in practice.

Future Research

Future research should further explore how robust data governance structures can be established to ensure data quality and readiness for algorithm development. This includes examining how AI techniques may contribute to improving data quality through automation, standardization, and enhanced data validation. Moreover, studies should investigate how cultural change can be fostered to promote viewing data as a strategic resource rather than an administrative by-product, and how such shifts can be maintained throughout the organization.

Attention to bias should extend beyond data collection and model design to the actual use of algorithms. Future studies are needed to determine under which conditions algorithms amplify or mitigate bias, and how systematic monitoring of real-world application can safeguard fairness and responsible use.

The ongoing tension between predictive performance and explainability also warrants investigation. Research should identify context-specific trade-offs between complex, high-performing models and simpler, more transparent approaches and explore how explainability dashboards can enhance actionability, trust, accountability, and clinician–patient communication.

Finally, future research should examine what knowledge, skills, and workflows clinicians need to engage responsibly with predictive algorithms while maintaining final clinical judgment. Integrating sociotechnical considerations throughout algorithm development will be key to ensuring that AI tools are both effective and ethically grounded in clinical practice.

While much focus has been placed on the technical aspects of ML in mental health care, sociotechnical factors related to successful algorithm development, implementation, and use have been largely overlooked [6,13,15,17,24,25,30,33,87]. More research is needed to explore these issues in real-world settings to ensure responsible and sustainable innovation.

Acknowledgments

All authors contributed to study design and manuscript revision. LH collected and analyzed the data and drafted the initial manuscript, after which all authors contributed to multiple revisions. We thank the GGzE development team for their cooperation.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the inability to fully anonymize the raw data but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. WRR [Netherlands Scientific Council for Government Policy]. Sustainable healthcare, a matter of choice. People, resources and public support. : The Hague: Netherlands Scientific Council for Government Policy; 2021 URL: https://english.wrr.nl/publications/reports/2022/05/03/sustainable-healthcare-a-matter-of-choice.-people-resources-and-public-support [accessed 2025-12-05]
2. Transforming healthcare with AI: the impact on the workforce and organisations. : EIT Health and McKinsey & Company; 2020 URL: https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai#/ [accessed 2025-12-05]
3. Nilsen P, Svedberg P, Nygren J, Frideros M, Johansson J, Schueller S. Accelerating the impact of artificial intelligence in mental healthcare through implementation science. Implement Res Pract 2022;3:26334895221112033. [doi: 10.1177/26334895221112033] [Medline: 37091110]
4. AI HLEG. A definition of AI: main capabilities and disciplines. : European Union; 2019 URL: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf [accessed 2025-12-18]
5. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 7;21(11):31701320. [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
6. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health 2022 Nov;4(11):e829-e840. [doi: 10.1016/S2589-7500(22)00153-4] [Medline: 36229346]
7. Monteith S, Glenn T, Geddes J, Whybrow PC, Achtyes E, Bauer M. Expectations for artificial intelligence (AI) in psychiatry. Curr Psychiatry Rep 2022 Nov;24(11):709-721. [doi: 10.1007/s11920-022-01378-5] [Medline: 36214931]
8. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. Mol Psychiatry 2017 Jan;22(1):37-43. [doi: 10.1038/mp.2016.201] [Medline: 27843153]
9. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 2023;82(3):3713-3744. [doi: 10.1007/s11042-022-13428-4] [Medline: 35855771]
10. Goodday SM, Kormilitzin A, Vaci N, et al. Maximizing the use of social and behavioural information from secondary care mental health electronic health records. J Biomed Inform 2020 Jul;107:103429. [doi: 10.1016/j.jbi.2020.103429] [Medline: 32387393]
11. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 7;14(91–118):91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
12. Ehtemam H, Sadeghi Esfahlani S, Sanaei A, et al. Role of machine learning algorithms in suicide risk prediction: a systematic review-meta analysis of clinical studies. BMC Med Inform Decis Mak 2024 May 27;24(1):138. [doi: 10.1186/s12911-024-02524-0] [Medline: 38802823]
13. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. Lancet Psychiatry 2022 Mar;9(3):243-252. [doi: 10.1016/S2215-0366(21)00254-6] [Medline: 35183281]
14. Kusuma K, Larsen M, Quiroz JC, et al. The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: a meta-analysis and systematic review. J Psychiatr Res 2022 Nov;155:579-588. [doi: 10.1016/j.jpsychires.2022.09.050] [Medline: 36206602]
15. Nordin N, Zainol Z, Mohd Noor MH, Chan LF. Suicidal behaviour prediction models using machine learning techniques: a systematic review. Artif Intell Med 2022 Oct;132:102395. [doi: 10.1016/j.artmed.2022.102395] [Medline: 36207078]
16. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
17. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophr Bull 2021 Mar 16;47(2):284-297. [doi: 10.1093/schbul/sbaa120] [Medline: 32914178]
18. Fouladvand S, Talbert J, Dwoskin LP, et al. A comparative effectiveness study on opioid use disorder prediction using artificial intelligence and existing risk models. IEEE J Biomed Health Inform 2023 Jul;27(7):3589-3598. [doi: 10.1109/JBHI.2023.3265920] [Medline: 37037255]
19. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 5;22(1):287. [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
20. Kessler RC, Hwang I, Hoffmire CA, et al. Developing a practical suicide risk prediction model for targeting high‐risk patients in the Veterans Health Administration. Int J Methods Psych Res 2017 Sep;26(3). [doi: 10.1002/mpr.1575]
21. Delgadillo J, Ali S, Fleck K, et al. Stratified care vs stepped care for depression: a cluster randomized clinical trial. JAMA Psychiatry 2022 Feb 1;79(2):101-108. [doi: 10.1001/jamapsychiatry.2021.3539] [Medline: 34878526]
22. Garriga R, Mas J, Abraha S, et al. Machine learning model to predict mental health crises from electronic health records. Nat Med 2022 Jun;28(6):1240-1248. [doi: 10.1038/s41591-022-01811-5] [Medline: 35577964]
23. McCarthy JF, Cooper SA, Dent KR, et al. Evaluation of the recovery engagement and coordination for health-veterans enhanced treatment suicide risk modeling clinical program in the Veterans Health Administration. JAMA Netw Open 2021 Oct 1;4(10):e2129900. [doi: 10.1001/jamanetworkopen.2021.29900] [Medline: 34661661]

XSL•FO
RenderX

24. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med 2021 Feb;27(2):186-187. [doi: 10.1038/s41591-021-01229-5] [Medline: 33526932]
25. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1(40):40. [doi: 10.1038/s41746-018-0048-y] [Medline: 31304321]
26. Golden G, Popescu C, Israel S, et al. Applying artificial intelligence to clinical decision support in mental health: what have we learned? Health Policy Technol 2024 Jun;13(2):100844. [doi: 10.1016/j.hlpt.2024.100844]
27. Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. Adm Policy Ment Health 2009 Jan;36(1):24-34. [doi: 10.1007/s10488-008-0197-4] [Medline: 19104929]
28. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implementation Sci 2022 Oct 29;17(1). [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
29. Wieringa M. What to account for when accounting for algorithms. Presented at: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Jan 27-30, 2020; Barcelona Spain p. 1-18. [doi: 10.1145/3351095.3372833]
30. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
31. Group T, Vasey B, Clifton DA. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med 2021 Feb;27(2):186-187. [doi: 10.1038/s41591-021-01229-5] [Medline: 33526932]
32. Sujan M, Furniss D, Grundy K, et al. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform 2019 Nov;26(1):31780459. [doi: 10.1136/bmjhci-2019-100081] [Medline: 31780459]
33. O'Dell B, Stevens K, Tomlinson A, Singh I, Cipriani A. Building trust in artificial intelligence and new technologies in mental health. Evid Based Ment Health 2022 May;25(2):45-46. [doi: 10.1136/ebmental-2022-300489] [Medline: 35444002]
34. Baldwin H, Loebel-Davidsohn L, Oliver D, et al. Real-world implementation of precision psychiatry: a systematic review of barriers and facilitators. Brain Sci 2022 Jul 16;12(7):35884740. [doi: 10.3390/brainsci12070934] [Medline: 35884740]
35. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
36. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2006 Mar;5(1):80-92. [doi: 10.1177/160940690600500107]
37. Kerkhof AM, Verwey B, Hummelen JW, et al. Multidisciplinaire Richtlijn Diagnostiek En Behandeling van Suïcidaal Gedrag [Multidisciplinary Guideline on the Assessment and Treatment of Suicidal Behaviour] [Book in Dutch]: Trimbos Institute; 2012. URL: https://psychiatrieweb.nl/pw.spoed/files/docs/suicidaal%20gedrag/richtlijn/120801_MDR_DBSG.pdf [accessed 2025-12-05]
38. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022 Apr 8;12(1):5979. [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
39. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. 2000 Presented at: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining; Apr 18-20, 2000 URL: https://www.semanticscholar.org/paper/CRISP-DM%3A-Towards-a-Standard-Process-Model-for-Data-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24#extracted [accessed 2025-12-05]
40. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
41. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
42. Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. AJTAS 2016;5(1):1. [doi: 10.11648/j.ajtas.20160501.11]
43. Singh K. Quantitative Social Research Methods: Sage Publications; 2007. [doi: 10.4135/9789351507741]
44. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
45. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
46. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37. [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
47. Wickramarathne TL, Premaratne K, Murthi MN, Scheutz M, Kubler S, Pravia M. Belief theoretic methods for soft and hard data fusion. Presented at: ICASSP 2011 - 2011 IEEE International Conference on Acoustics, Speech and Signal Processing; May 22-27, 2011. [doi: 10.1109/ICASSP.2011.5946964]

XSL•FO
RenderX

48. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health 2023 Oct;2(10):e0000347. [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
49. Tayefi M, Ngo P, Chomutare T, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Stats 2021 Nov;13(6). [doi: 10.1002/wics.1549]
50. Anderson HD, Pace WD, Brandt E, et al. Monitoring suicidal patients in primary care using electronic health records. J Am Board Fam Med 2015;28(1):65-71. [doi: 10.3122/jabfm.2015.01.140181] [Medline: 25567824]
51. Parsons A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. J Am Med Inform Assoc 2012;19(4):604-609. [doi: 10.1136/amiajnl-2011-000557] [Medline: 22249967]
52. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
53. De Choudhury M, Kiciman E. Integrating artificial and human intelligence in complex, sensitive problem domains: experiences from mental health. AI Mag 2018 Sep;39(3):69-80. [doi: 10.1609/aimag.v39i3.2815]
54. Baune BT, Christensen MC. Differences in perceptions of major depressive disorder symptoms and treatment priorities between patients and health care providers across the acute, post-acute, and remission phases of depression. Front Psychiatry 2019;10(335):335. [doi: 10.3389/fpsyt.2019.00335] [Medline: 31178765]
55. Ilkiw-Lavalle O, Grenyer BFS. Differences between patient and staff perceptions of aggression in mental health units. Psychiatr Serv 2003 Mar;54(3):389-393. [doi: 10.1176/appi.ps.54.3.389] [Medline: 12610249]
56. Rossberg JI, Friis S. Patients' and staff's perceptions of the psychiatric ward environment. Psychiatr Serv 2004 Jul;55(7):798-803. [doi: 10.1176/appi.ps.55.7.798] [Medline: 15232020]
57. Van Der Merwe M, Muir-Cochrane E, Jones J, Tziggili M, Bowers L. Improving seclusion practice: implications of a review of staff and patient views. J Psychiatr Ment Health Nurs 2013 Apr;20(3):203-215. [doi: 10.1111/j.1365-2850.2012.01903.x] [Medline: 22805615]
58. Hazewinkel MC, de Winter RFP, van Est RW, et al. Text analysis of electronic medical records to predict seclusion in psychiatric wards: proof of concept. Front Psychiatry 2019;10(188):188. [doi: 10.3389/fpsyt.2019.00188] [Medline: 31031650]
59. Velupillai S, Hadlaczky G, Baca-Garcia E, et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front Psychiatry 2019;10(36):36. [doi: 10.3389/fpsyt.2019.00036] [Medline: 30814958]
60. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020 Oct 29;20(1):280. [doi: 10.1186/s12911-020-01297-6] [Medline: 33121479]
61. Boggs JM, Kafka JM. A critical review of text mining applications for suicide research. Curr Epidemiol Rep 2022;9(3):126-134. [doi: 10.1007/s40471-022-00293-w] [Medline: 35911089]
62. Boudreaux ED, Rundensteiner E, Liu F, et al. Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions. Front Psychiatry 2021;12:707916. [doi: 10.3389/fpsyt.2021.707916] [Medline: 34413800]
63. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep 2018 May 9;8(1):7426. [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]
64. Levis M, Levy J, Dent KR, et al. Leveraging natural language processing to improve electronic health record suicide risk prediction for veterans health administration users. J Clin Psychiatry 2023 Jun 19;84(4):22m14568. [doi: 10.4088/JCP.22m14568] [Medline: 37341477]
65. Cusick M, Adekkanattu P, Campion TR Jr, et al. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. J Psychiatr Res 2021 Apr;136(95–102):95-102. [doi: 10.1016/j.jpsychires.2021.01.052] [Medline: 33581461]
66. Geneviève LD, Martani A, Shaw D, Elger BS, Wangmo T. Structural racism in precision medicine: leaving no one behind. BMC Med Ethics 2020 Feb 19;21(1):17. [doi: 10.1186/s12910-020-0457-8] [Medline: 32075640]
67. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021 Dec 28;29(1):187-196. [doi: 10.1093/jamia/ocab199] [Medline: 34664641]
68. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 2023 Sep;18(5):1062-1096. [doi: 10.1177/17456916221134490] [Medline: 36490369]
69. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. Presented at: EAAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization; Oct 5-9, 2021; New York, United States p. 1-9. [doi: 10.1145/3465416.3483305]
70. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023 Sep 25;30(10):1730-1740. [doi: 10.1093/jamia/ocad120] [Medline: 37390812]

XSL•FO
RenderX

71. Akata Z, Balliet D, de Rijke M, et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. Computer (Long Beach Calif) 2020;53(8):18-28. [doi: 10.1109/MC.2020.2996587]
72. Mouchabac S, Conejero I, Lakhlifi C, et al. Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases. Dialogues Clin Neurosci 2021;23(1):52-61. [doi: 10.1080/19585969.2022.2042165] [Medline: 35860175]
73. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024 Apr 25;26:e56764. [doi: 10.2196/56764] [Medline: 38662419]
74. Vitacca M, Giardini A, Gazzi L, Vitacca M. Hidden biases in clinical decision-making: potential solutions, challenges, and perspectives. Monaldi Arch Chest Dis 2022 Sep 7;93(2):36069639. [doi: 10.4081/monaldi.2022.2339] [Medline: 36069639]
75. Hulsen T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI 2023;4(3):652-666. [doi: 10.3390/ai4030034]
76. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
77. Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc 2023 Sep;37(5):1719-1778. [doi: 10.1007/s10618-023-00933-9]
78. Singh R, Miller T, Sonenberg L, Velloso E, Vetere F, Howe P, et al. An actionability assessment tool for explainable AI. arXiv. Preprint posted online on Jun 19, 2024. [doi: 10.48550/arXiv.2407.09516]
79. Menger VJ, Spruit MR, Scheepers FE. Kennisontwikkeling in de klinische psychiatrie: leren van elektronische patiëntendossiers [Article in Dutch]. Tijdschr Psychiatr 2021;63(4):294-300 [FREE Full text]
80. Vale D, El-Sharif A, Ali M. Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI Ethics 2022 Nov;2(4):815-826. [doi: 10.1007/s43681-022-00142-y]
81. Reddy S, Shaikh S. The long road ahead: navigating obstacles and building bridges for clinical integration of artificial intelligence technologies. J Med Artif Intell 2025;8:7-7. [doi: 10.21037/jmai-24-148]
82. Kaplan AD, Kessler TT, Brill JC, Hancock PA. Trust in artificial intelligence: meta-analytic findings. Hum Factors 2023 Mar;65(2):337-359. [doi: 10.1177/00187208211013988] [Medline: 34048287]
83. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Ann R Coll Physicians Surg Can 2020 Jul;14(2):627-660. [doi: 10.5465/annals.2018.0057]
84. Kizilcec RF. How much information?: Effects of transparency on trust in an algorithmic interface. Presented at: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, California, USA p. 2390-2395. [doi: 10.1145/2858036.2858402]
85. Luk JW, Pruitt LD, Smolenski DJ, Tucker J, Workman DE, Belsher BE. From everyday life predictions to suicide prevention: clinical and ethical considerations in suicide predictive analytic tools. J Clin Psychol 2022 Feb;78(2):137-148. [doi: 10.1002/jclp.23202] [Medline: 34195998]
86. van der Veer SN, Riste L, Cheraghi-Sohi S, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. J Am Med Inform Assoc 2021 Sep 18;28(10):2128-2138. [doi: 10.1093/jamia/ocab127] [Medline: 34333646]
87. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]

Abbreviations

AI: artificial intelligence
CFIR: Consolidated Framework for Implementation Research
CRISP-DM: Cross-Industry Standard Process for Data Mining
EHR: electronic health record
GGzE: Geestelijke Gezondheidszorg Eindhoven en de Kempen
ML: machine learning
NLP: natural language processing
RoB: risk of bias
SPQ: Suicide Prevention Questionnaire

XSL•FO
RenderX

Edited by C Perrin; submitted 20.Mar.2025; peer-reviewed by HH Yeh, R Jacobucci; revised version received 18.Nov.2025; accepted 18.Nov.2025; published 14.Jan.2026.

Please cite as:
Hummel L, Lorenz-Artz KCAG, Bierbooms JJPA, Bongers IMB
Developing a Suicide Risk Prediction Algorithm Using Electronic Health Record Data in Mental Health Care: Real-World Case Study
JMIR Med Inform 2026;14:e74240
URL: https://medinform.jmir.org/2026/1/e74240
doi:10.2196/74240

© Linda Hummel, Karin C A G Lorenz-Artz, Joyce J P A Bierbooms, Inge M B Bongers. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation Report

Scaling Wireless Continuous Vital Sign Monitoring Across an 8-Hospital Health System: Digital Health Implementation Report

Ngoc-Anh Nguyen[1,2,3], MD; Grace Lee[1,2], BA; Brendan Holderread[1,2], MD; Terrie Holman[1], BS; Sarah Pletcher[1,2,4], MD, MHCDS; Roberta Schwartz[1,2], PhD

[1]Center for Connected Care, Innovation & Implementation - Research, Houston Methodist, Houston, TX, United States

[2]Houston Methodist Research Institute, Houston Methodist, Houston, TX, United States

[3]Department of Medicine, Houston Methodist, Houston, TX, United States

[4]Department of Surgery, Houston Methodist, Houston, TX, United States

Corresponding Author:
Ngoc-Anh Nguyen, MD
Center for Connected Care, Innovation & Implementation - Research
Houston Methodist
6560 Fannin Street
Scurlock Tower STB2-133
Houston, TX, 77030
United States
Phone: 1 7132996839
Fax: 1 7137931851
Email: nanguyen@houstonmethodist.org

Abstract

Background: Frequent vital sign (VS) monitoring is central to inpatient safety but is traditionally performed manually every 4 hours, a century-old practice that can miss early clinical deterioration, disrupt patient sleep, and impose a heavy documentation burden on nursing staff. Continuous VS monitoring (CVSM) using wearable remote patient monitoring devices enables near real-time, high-frequency VS measurement while reducing manual workload and preserving patient rest.

Objective: This implementation report describes the large-scale implementation of CVSM across an 8-hospital health system. The initiative aimed to (1) enhance earlier detection of patient health deterioration through continuous, algorithm-driven monitoring; (2) improve nursing workflow efficiency by reducing reliance on manual VS checks; and (3) minimize nighttime disruptions to support patient rest and recovery.

Methods: The program was designed for system-wide scalability and executed from 2022 to 2024 using a 4-phase framework: strategic program design, program planning, go-live preparation, and implementation and optimization. A Food and Drug Administration–cleared wearable device (BioButton) continuously measured heart rate, respiratory rate, and skin temperature, with data integrated into Epic and monitored 24×7 through a centralized virtual operations center. Rollout followed a staggered playbook across approximately 2700 adult non–intensive care unit beds and was supported by leadership engagement, supply chain readiness, staff training, and phased superuser-led adoption.

Implementation (Results): All 8 hospitals achieved full deployment between April 2023 and February 2024, with more than 95% device use rates and 100% nursing staff training completion. A standardized escalation workflow filtered approximately 50% of the alerts at the virtual operations center review stage, substantially reducing frontline alert burden. Operational refinements included revised heart rate and respiratory rate alert thresholds and the removal of temperature as a single alert trigger. Several units extended overnight manual VS intervals from every 4 hours to every 6 to 8 hours, with staff estimating approximately 4 hours saved per nursing shift. Patient care assistants redirected time toward patient mobility and personal care needs, while staff reported growing confidence in device performance over time.

Conclusions: This initiative represents the first system-wide deployment of CVSM across a diverse, multihospital health system. Success was enabled by early strategic alignment, phased rollout, robust IT and monitoring infrastructure, and iterative optimization. The program demonstrates the feasibility of embedding CVSM into routine inpatient care to improve efficiency and patient experience. Transferable strategies, including phased rollouts, centralized monitoring, and structured change management, may

XSL•FO
RenderX

inform other health systems pursuing digital VS redesign. Future work should rigorously evaluate impacts on patient outcomes, cost-effectiveness, and applicability to postacute and ambulatory care settings.

(JMIR Med Inform 2026;14:e78216) doi:10.2196/78216

KEYWORDS

continuous vital sign monitoring; remote patient monitoring; wearable devices; hospital workflow redesign; patient safety; digital health implementation; implementation science

Introduction

Context

This implementation occurred within Houston Methodist (HM), an 8-hospital health system in Texas encompassing the HM Hospital (a quaternary care academic flagship), 6 community hospitals, and 1 long-term acute care (LTAC) facility serving the greater Houston area, together serving a diverse adult population of more than 140,000 annual admissions and 2 million outpatient visits [1].

During the COVID-19 pandemic, HM established a centralized virtual operations center (VOC) and virtual intensive care unit (VICU) providing 24×7 monitoring for approximately 370 intensive care unit (ICU) beds. This infrastructure created favorable conditions for remote patient monitoring (RPM) and ultimately informed the development of the system-wide continuous vital sign (VS) monitoring (CVSM) program. This CVSM program differs from traditional RPM in its continuous, hospital-based, and fully integrated design, with data integrated into the electronic medical record and supported by 24×7 centralized oversight rather than periodic outpatient review.

In 2022, the health system launched a redesign of inpatient VS practices, with this report detailing the program's scale-up and deployment, which were designed from the outset for system-wide implementation.

Problem Statement

Frequent monitoring of VSs—heart rate (HR), respiratory rate (RR), temperature, blood pressure, and oxygen saturation (SpO_2)—is a cornerstone of inpatient safety [2]. On general hospital floors, VSs are routinely collected manually every 4 hours, reflecting a century-old practice that does not account for patient acuity [3,4]. This approach risks missed deterioration between checks, places substantial time demands on nursing staff, and disrupts patients' sleep, with physiological consequences [5-14]. Deteriorating VSs are well-established predictors of ICU transfer and in-hospital mortality, and CVSM, enabled by wearable RPM devices, now allows for unobtrusive, high-frequency VS measurement [7,11,12]. This work directly addresses health system challenges of limited access to timely clinical data and variable adherence to evidence-based monitoring practices, both of which undermine patient safety and workflow efficiency. However, evidence on broad, system-wide implementation across diverse hospital settings remains limited.

Similar Interventions

Previous CVSM implementations have typically been limited to ICUs, single hospitals, or narrow patient groups, demonstrating feasibility but limiting generalizability [15-17]. Our initiative differs by extending CVSM across all adult non-ICU inpatients in an 8-hospital system, integrated with Epic and supported by centralized monitoring. This report describes the implementation and scale-up process.

This report follows JMIR's implementation reporting guidelines for digital health implementation [18] to ensure transparency in describing the implementation context and processes.

Methods

Aim and Objective

The aim of this initiative was to redesign inpatient VS measurement practices to enhance patient safety, patient experience, and staff efficiency through the implementation of CVSM. Specific objectives were to enable early detection of deterioration, reduce nighttime disruptions to promote patient rest and recovery, and improve workflow by decreasing reliance on manual VS checks. Implementation outcomes included system-wide rollout to all eligible beds, 100% nursing staff training completion, optimal device use, and reductions in manual VS collection, with qualitative observations of staff and patient acceptance. Operational metrics (scale-up timeline, device use, alert filtering, and training completion) were directly measured, whereas workflow impacts and staff acceptance were reflected in qualitative, staff-reported estimates.

Implementation Strategy

The implementation strategy was designed for system-wide scalability and organized into 4 phases, with the observed timeline shown in Figure 1.

XSL•FO
RenderX

Figure 1. Four-phase implementation framework and observed timeline for scaling continuous vital sign monitoring across (CVSM) Houston Methodist's 8-hospital health system. EMR: electronic medical record.

Phase 1: Strategic Program Design (1-2 Years Before Implementation)

The health system defined an enterprise-wide vision for CVSM as part of a broader effort to modernize VS monitoring. Strategic goals emphasized embedding continuous monitoring into routine inpatient care, strengthening patient safety through earlier detection of deterioration, and redesigning workflows to mitigate the impact of nursing staff shortages.

A multidisciplinary team of key stakeholders, including clinical, operational, IT, supply chain, and innovation leaders, was engaged early and consistently to ensure clinical and organizational alignment. The team identified use cases and target populations, prioritizing continuous monitoring of all hospitalized, non-ICU patients to address the safety gap created by every-4-hour manual VS checks. Patient eligibility was intentionally broad; CVSM was applied to general adult medical-surgical inpatients, excluding only ICU patients, with no additional eligibility restrictions.

Implementation sites included the academic flagship, 6 community hospitals, and 1 LTAC facility, with monitoring supported by bedside nurses, unit superusers, and the centralized VOC. The multidisciplinary team guided planning and iteratively refined a scalable model for system-wide deployment, supported by superusers at the unit level and centralized monitoring through the VOC.

Phase 2: Program Planning (1Year Before Implementation)

After defining the overall CVSM strategy, the health system conducted market research to evaluate vendor partners and established planning priorities in workflow redesign, Epic integration, staff training, and IT infrastructure, with a standardized playbook to guide hospital rollouts. Vendor selection considered Food and Drug Administration (FDA) clearance, performance, usability, multiparameter monitoring, battery life, patient comfort, Epic integration, and scalability.

A wearable patch device (BioButton; BioIntelliSense) was selected as the FDA-cleared option that best met these criteria. The adhesive patch is worn on the chest with a battery life lasting 7 to 16 days. It continuously collects HR, RR, and skin temperature (additional parameters are listed in Multimedia Appendix 1) [6,17]. Device selection prioritized accurate RR measurement, given its established predictive value for deterioration [14,19]. Although a BioButton device does not measure SpO_2 or blood pressure, these parameters were acceptable through existing bedside monitoring. Continuous SpO_2 monitoring typically requires a tethered probe that disrupts rest and increases alarm burden [20], whereas the BioButton device allows unobtrusive, trended monitoring aligned with program goals to reduce patient disruptions.

From the outset, the program was designed for system-wide scale, leveraging existing institutional resources and establishing a centralized monitoring team at the main academic hospital. Pilot sites included 2 internal medicine units, 1 neurosurgery unit, and 1 neurology unit selected for innovation experience and staff receptiveness. Program management included site-level and system-level leads, who shared responsibilities with the vendor for workflow design, data integration, training, and support **(**roles are detailed in Multimedia Appendix 2).

Participating entities included HM, a quaternary academic health system whose mission emphasizes patient-centered, digitally enabled care, which led workflows, monitoring staff, and supply chain integration, and BioIntelliSense, which provided devices, hubs, dashboard tools, and technical support. The VOC delivered continuous monitoring, with bedside nurses serving as superusers. Intellectual property for the BioButton device remains with BioIntelliSense, while workflows, data integration, and the operational model belong to HM.

XSL•FO
RenderX

The business case projected nursing and patient care assistant (PCA) time savings from extended overnight VS intervals, supporting adoption and sustainability. Internal funding was secured as part of routine service redesign, with no direct costs to patients. Budget categories included devices, hubs, Epic integration, staff training, supply chain processes, and change management. BioButton devices were supplied through a per-bed subscription model, and approximately 2711 BioHub Wi-Fi gateways were installed across all patient rooms. The largest investments were technology procurement and integration, with ongoing costs for training and supply chain support.

Phase 3: Go-Live Preparation (3-6 Months Before Going Live)

Building on program planning, readiness efforts focused on 5 domains: room and BioHub setup, device and supply chain readiness, Epic and IT validation, unit training plans, and kickoff planning. Activities occurred at both system and unit levels (details are provided in Multimedia Appendix 3). Integration with Epic and the monitoring dashboard followed standard institutional IT and security protocols. IT preparations included configuring BioHub connectivity, enabling secure transmission of physiological data, and validating data flow into Epic flowsheets and early warning score logic. A summary of this pathway is shown in Figure 2.

Device and environmental preparations included stocking BioButton devices, labeling return bins, and completing room readiness surveys. Initial stock levels were based on recent admission volumes, with automated reordering through the supply chain. Workflows were established for device return and sanitation.

Kickoff planning finalized rollout sequencing and support allocation. The central monitoring team staffing within the VOC was expanded by reallocating staff after reducing telemetry use. VOC staff completed vendor-led training and competency demonstrations in alert interpretation, escalation protocols, and workflow integration.

Phase 4: Implementation and Optimization

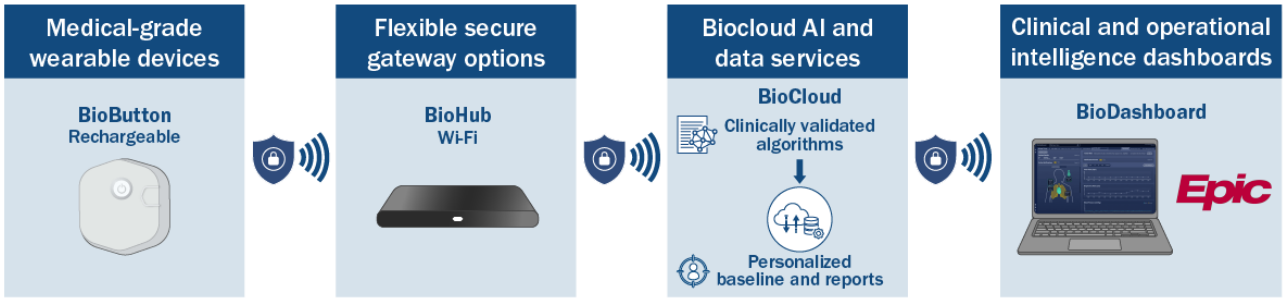
Implementation emphasized 5 core activities: playbook rollout, staff training, go-live support, protocol refinement, and iterative feedback and optimization. Each unit followed an 8-week rollout guided by the playbook, with superuser and vendor support during the go-live phase.

Training followed a train-the-trainer model, with unit leaders and superusers trained first, followed by cascading training to frontline staff. All bedside nurses completed a mandatory learning management system module (100% completion system wide), which covered device application, alert interpretation, escalation protocols, and workflow integration. Adoption was reinforced through superuser rounding, vendor staff presence, and structured communication plans. Training activities were designed to build knowledge, confidence, and consistency among staff and are consistent with approaches described in the implementation science literature [21-23].

Go-live execution proceeded in staggered groups of 4 to 8 units, allowing overlap between training and operational transitions. Kickoff meetings, leadership communication plans, and coordinated supply chain activation supported each launch.

Protocol refinement was embedded throughout the rollout. Initial thresholds and workflows were iteratively adjusted based on frontline experience. VOC champions and superusers reinforced troubleshooting and informed subsequent launches. Sustainability was achieved by embedding CVSM into standard operations, including VOC monitoring workflows, Epic documentation, supply chain logistics, and updated learning management system training modules. Continued superuser engagement and institutional commitment to operational funding reinforced adoption. Full institutionalization of CVSM into existing structures ensured long-term ownership and positioned the program as a core element of HM's broader digital care transformation strategy.

Figure 2. Data flow from the BioButton device through BioHub, BioCloud, and BioDashboard to Epic. AI: artificial intelligence.



Ethical Considerations

The Institutional Review Board of Houston Methodist Research Institute reviewed this initiative and deemed it *not human research*, as it represented a service redesign implemented as part of standard care.

Implementation (Results)

Coverage

Between April 2023 and February 2024, the program expanded from pilot units to enterprise-wide coverage across all 8 hospitals, encompassing the full adult, non-ICU inpatient population (approximately 2700 beds across the academic flagship, 6 community hospitals, and 1 LTAC facility).

XSL•FO
RenderX

Rollout Approach and Timeline

The enterprise rollout followed a phased road map. Pilot implementation began on 4 units at the main hospital (2 internal medicine units, 1 neurosurgery unit, and 1 neurology unit). Four months later, deployment scaled through staggered 8-week waves of 4 to 8 units, with 3-week offsets to permit overlap between staff training and transitions, allowing the incorporation of iterative learning. This cadence enabled continuous optimization while maintaining operational continuity.

Monitoring and Escalation Workflow

Five months after the pilot launch, HM introduced a standardized escalation protocol embedded in centralized monitoring workflows. When BioButton device–triggered alerts appeared on the VOC dashboard, a VOC technician performed a first-line review to add clinical content and either dismissed nonactionable events or escalated alerts to a VICU nurse. Approximately half of the alerts were dismissed at the VOC step, reducing alert burden for frontline staff. Illustrative case examples are provided in Multimedia Appendix 4.

The VICU nurse notified the bedside nurse or activated the clinical emergency response team when intervention was indicated. This tiered model concentrated clinical escalation on higher-value alerts. Data on clinical emergency response team activations during the study period have been reported previously in a related outcomes study [24].

Outcomes and Post–Go-Live Protocol Refinement

Primary outcomes included system-wide scale-up, reliable Epic integration, and sustained monitoring. Device use exceeded 95%, training completion reached 100%, and approximately 50% of alerts were resolved at VOC review. Secondary outcomes included improved alert specificity, fewer patient refusals, increased staff acceptance, and operational efficiencies (nursing and PCA time savings). Alert thresholds were systematically reviewed with virtual medicine leadership in collaboration with frontline nursing staff, superusers, and the vendor. Baseline thresholds (HR 50-115 beats per minute, RR 11-28 breaths per minute, and skin temperature >36 °C) were operationalized with alerts stratified as single alerts (1 abnormal VS), double alerts (2 abnormalities, typically HR and RR), or triple alerts (3 abnormalities: HR, RR, and skin temperature). Three protocol changes improved the system-wide specificity.

Skin temperature was removed as an isolated escalation criterion and retained only for triple alerts. The RR lower limit was reduced from 11 to 8 breaths per minute to reduce clinically insignificant overnight alerts. The upper HR limit was increased from 115 to 120 beats per minute to reduce nonactionable alerts reported by community hospitals and the LTAC facility. Other parameters (eg, body position and activity) contributed to *low motion* alerts but were not primary triggers.

Collectively, these adjustments improved the balance between sensitivity and specificity, as reflected by observed reductions in nonactionable alerts and staff-reported decreases in false positives during enterprise rollout. The final escalation workflow and logic are depicted in Figures 3 and 4, with detailed initial and revised threshold values in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 3. Escalation workflow for the BioButton device alerts with virtual operations center triage and bedside escalation. Bedside teams are responsible for all button management activities (ie, placement or removal, low battery, and no data). Patient data (from the BioIntelliSense button) will be transmitted into the Epic vital signs flowsheet, and BioDashboard can take up to 20 minutes to fully synchronize. APP: advanced practice provider; CERT: clinical emergency response team; RN: registered nurse; RR: respiratory rate.

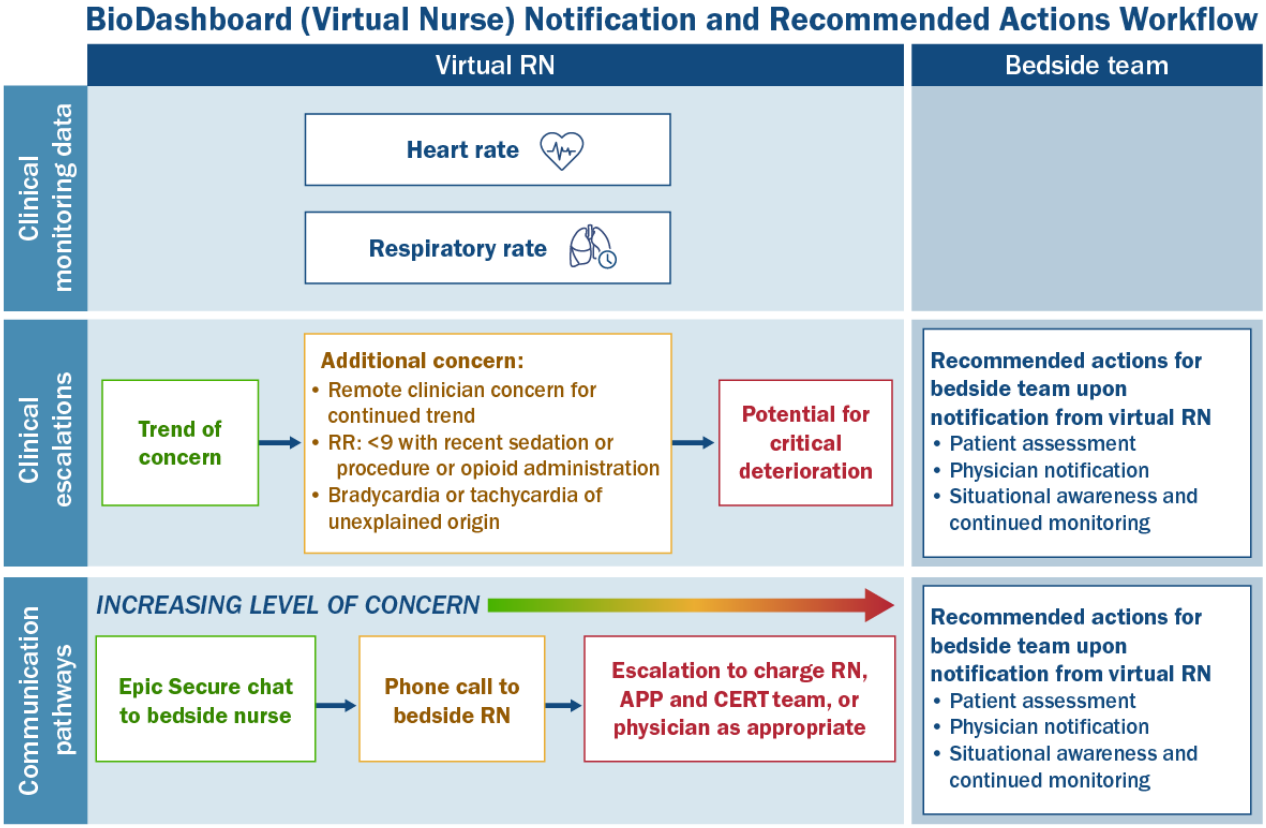
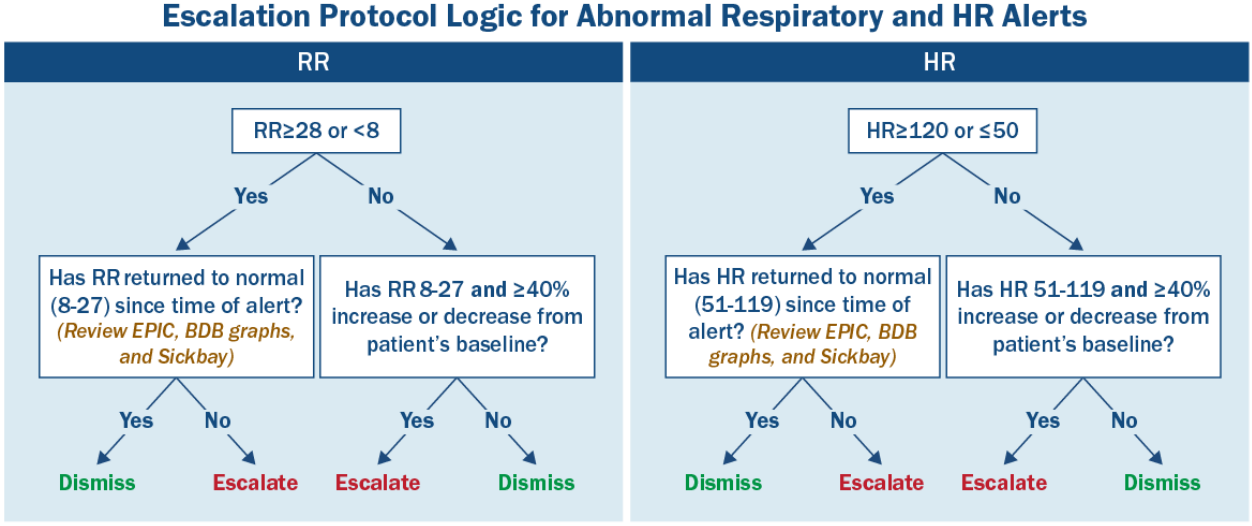


Figure 4. Escalation protocol for respiratory rate (RR) and heart rate (HR) alerts with system-wide thresholds. Final threshold for HR (50-120 bpm) and RR (8-27 bpm). Skin temperature has been removed from the escalation criteria. BDB: BioDashboard.



Lessons Learned and Implementation Challenges

Technical and workflow challenges identified during the pilot informed the enterprise rollout. Barcode scanning difficulties (eg, poor label contrast) occasionally impeded device-to-patient matching; staff used device packaging barcodes and Workstations on Wheels carts as interim work-arounds, and the vendor subsequently improved barcode resolution. Rare *double HR* artifacts (concurrent S1 and S2 detections) produced nonactionable alerts and were resolved via algorithm updates.

Variation in device application sometimes caused skin irritation, which was addressed via staff reeducation and the use of adhesive remover spray. Early staff skepticism about using median CVSM values for documentation declined as experience accumulated, and staff perceived CVSM values to be accurate and consistent with routine bedside measurements; patient

XSL•FO
RenderX

refusals decreased with standardized education. These mitigations and training reinforcements were propagated system wide during enterprise rollout.

As noted previously, unintended consequences, including minor skin irritation, initial device hesitancy, and early false positives, were observed and addressed through staff education, supplies, and vendor refinements; no significant harms were reported.

Discussion

Principal Findings

To our knowledge, this initiative represents the first large-scale deployment of wearable CVSM across a diverse, multihospital health system as part of a system-wide redesign initiative [17]. Our goal was to modernize inpatient VS practices by introducing continuous monitoring to support earlier detection, improve outcomes, and enhance both patient and clinician experience. Several units extended overnight VS intervals from every 4 hours to 6 to 8 hours, saving approximately 4 hours per nursing shift and allowing PCAs to focus on patient mobility and personal needs. Together with early clinical benefits, these efficiencies strengthened adoption and staff buy-in.

Weller et al [17] evaluated the same FDA-cleared device in 11,977 patients across urban and rural hospitals, with initial ICU use before expansion to non-ICU settings. They reported low RR alarms most frequently, while in our implementation, elevated RR emerged as the most common alert, underscoring RR's role as a sensitive marker of cardiopulmonary function. This difference likely reflects patient mix, as sedation in ICU patients tends to lower RRs, while elevated RRs in ward patients more often indicate deterioration.

A key distinction of this program was its breadth and scope. Unlike previous efforts limited to ICUs, daytime hours, or narrow diagnostic groups, this program encompassed all adult non-ICU inpatients, deployed continuously (24×7) across the flagship academic hospital, 6 community hospitals, and 1 LTAC facility. This design addressed gaps in every 4-hour VS collection, enabled earlier detection across diverse patient populations, and supported workflow redesign amid nursing staff shortages.

Device selection prioritized accurate RR measurement, given its strong predictive value. Although the BioButton devices do not measure blood pressure or SpO_2, they remained available through existing bedside monitoring. Continuous SpO_2 monitoring was avoided to minimize alarm burden. The device offered unobtrusive, trended monitoring without disturbing patients' rest while capturing key physiological parameters for early detection. The final escalation protocol's emphasis on HR and RR reflected clinical prioritization and not device limitation. Although formal satisfaction and quality metrics are forthcoming, early stakeholder feedback suggests that reduced overnight disruptions improved sleep quality, potentially enhancing patient satisfaction and reducing overnight nursing workload. Prospective studies are needed to evaluate long-term effects on clinical outcomes, alert trends, workflow efficiency, and satisfaction, including patient-reported outcomes.

Limitations

This report has several limitations. No formal satisfaction survey was conducted, and estimates of time savings were based on staff feedback. BioButton device–derived VSs were not independently validated, as accuracy testing of the FDA-cleared device was beyond the scope of this work. The escalation protocol was intentionally delayed in pilot units; therefore, early data reflect monitoring without active escalation, with finalized protocols introduced later and applied retrospectively. Although guided by an internal business case projecting nursing and PCA time savings, no formal health technology assessment was conducted.

Generalizability may be limited by HM's centralized monitoring model; other health systems may face different challenges. Observed alert trends may also differ across settings, as illustrated by differences between ICU-focused studies and our broad non-ICU deployment. Outcome measures, such as clinical impact, workflow efficiency, and adoption, were not prospectively defined, constraining evaluation but informing priorities for future analyses.

Several transferable lessons emerged: (1) structured change management with engagement, champions, and feedback loops supported adoption; (2) phased rollout with iterative refinements reduced risk and enabled optimization; (3) centralized monitoring enabled 24×7 oversight, while hospitals without such infrastructure could adapt using telemetry staff, rapid response teams, or external partners; and (4) alignment of CVSM with institutional priorities secured leadership support and momentum. Table 1 summarizes key lessons and success factors to guide future implementers in scaling similar initiatives.

XSL•FO
RenderX

Table 1. Key lessons and success factors from system-wide CVSMa implementation.

Domain	Challenges encountered	Lessons learned and success factors
Leadership and Governance	Balancing innovation with standardization across 8 hospitals	Early executive sponsorship and system-wide alignment maintained focus, accountability, and resource continuity.
Implementation Approach	Scaling across multiple hospitals while managing operational risk	Staggered enterprise rollout using an 8-week implementation playbook enabled iterative refinement and risk mitigation before full scale-up.
Workflow Integration	Incorporating CVSM data into existing Epic processes	Integration into Epic flowsheets and early warning scores streamlined workflow and improved adoption.
Technology and Infrastructure	Interoperability issues and alert fatigue	Centralized monitoring filtered ~50% of alerts before bedside escalation; iterative threshold refinements improved specificity and reduced alert fatigue.
Training and Adoption	Staff skepticism and learning curve	Superuser model, LMS[b] training, and vendor support during go-live built confidence and sustained engagement.
Change Management	Variation in readiness and engagement across sites	Phased rollout enabled iterative optimization and local adaptation while maintaining operational continuity.
Data Governance	Managing HIPAA[c] compliance and device data retention	Preimplementation IT security review ensured compliance with institutional privacy and security standards.
Sustainability	Maintaining engagement and support after implementation	Integration into standard operations, funding streams, and training infrastructure supported long-term sustainability.
Cultural Readiness	Early staff skepticism about device accuracy and value	Confidence increased with experience and demonstrated accuracy, reinforcing trust and normalization of continuous monitoring.

[a]CVSM: continuous vital sign monitoring.

[b]LMS: learning management system.

[c]HIPAA: Health Insurance Portability and Accountability Act.

Conclusions and Future Implications

This initiative demonstrates the feasibility and operational value of scaling CVSM across a large, diverse health system. Embedding continuous monitoring into standard workflows can improve patient safety, enhance staff efficiency, and build infrastructure for digitally enabled care models.

Future work should rigorously assess clinical outcomes, economic impact, and patient-reported measures while exploring application in postacute and ambulatory settings. Although growing evidence supports continuous monitoring, prospective randomized trials are still needed to confirm clinical and economic benefits. As health systems address workforce shortages and rising patient acuity, wearable monitoring technologies, such as CVSM, could serve as core infrastructure for safer, more efficient, and patient-centered care.

Acknowledgments

The authors thank BioIntelliSense Inc, Houston Methodist (HM) Center for Health Data Science and Analytics, Ms Ellie Kalatzi, HM Department of Strategic Innovation, HM Department of Virtual Operations, HM Department of IT and Informatics, HM supply chain, HM nursing leadership, and HM nursing education for their invaluable contributions to the successful implementation of this quality improvement initiative. The authors also thank Ms Sarah Sossong for her significant contributions to manuscript development, including content drafting and revision, integration of reviewer feedback, and overall editorial guidance.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional privacy policies.

Authors' Contributions

Investigation: NAN, GL, and BH
Program conceptualization and design: SP and RS

XSL•FO
RenderX

Project management: TH
Overall responsibility: SP and NAN
Writing—original draft: GL, NAN, and BH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Physiological parameters measured by the BioButton device, measurement methods, and clinical use cases.
[DOCX File , 15 KB - medinform_v14i1e78216_app1.docx]

Multimedia Appendix 2
Roles and responsibilities of vendor and health system stakeholders in the multidisciplinary implementation team for continuous vital sign monitoring.
[DOCX File , 22 KB - medinform_v14i1e78216_app2.docx]

Multimedia Appendix 3
System-level and unit-level readiness activities by operational domain for continuous vital sign monitoring implementation.
[DOCX File , 23 KB - medinform_v14i1e78216_app3.docx]

Multimedia Appendix 4
Illustrative case examples of early detection and intervention.
[DOCX File , 16 KB - medinform_v14i1e78216_app4.docx]

Multimedia Appendix 5
Initial and revised alarm thresholds with rationale for revisions.
[DOCX File , 14 KB - medinform_v14i1e78216_app5.docx]

Multimedia Appendix 6
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[PDF File (Adobe PDF File), 286 KB - medinform_v14i1e78216_app6.pdf]

References

1. Facts and statistics. Houston Methodist. URL: https://www.houstonmethodist.org/newsroom/facts-statistics/ [accessed 2026-01-07]
2. Lockwood C, Conroy-Hiller T, Page T. Vital signs. JBI Libr Syst Rev 2004;2(6):1-38. [doi: 10.11124/01938924-200402060-00001] [Medline: 27820007]
3. Yoder JC, Yuen TC, Churpek MM, Arora VM, Edelson DP. A prospective study of nighttime vital sign monitoring frequency and risk of clinical deterioration. JAMA Intern Med 2013 Sep 09;173(16):1554-1555 [FREE Full text] [doi: 10.1001/jamainternmed.2013.7791] [Medline: 23817602]
4. Orlov NM, Arora VM. Things we do for no reason™: routine overnight vital sign checks. J Hosp Med 2020 May;15(5):272-274 [FREE Full text] [doi: 10.12788/jhm.3442] [Medline: 32379025]
5. Mann DL. Q4 vital signs: can machine learning protect patients from the machinery of modern medicine? JACC Basic Transl Sci 2019 Jun;4(3):468-469 [FREE Full text] [doi: 10.1016/j.jacbts.2019.05.002] [Medline: 31312769]
6. Weenk M, Bredie SJ, Koeneman M, Hesselink G, van Goor H, van de Belt TH. Continuous monitoring of vital signs in the general ward using wearable devices: randomized controlled trial. J Med Internet Res 2020 Jun 10;22(6):e15471 [FREE Full text] [doi: 10.2196/15471] [Medline: 32519972]
7. Ludikhuize J, Smorenburg SM, de Rooij SE, de Jonge E. Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the Modified Early Warning Score. J Crit Care 2012 Aug;27(4):424.e7-424.13. [doi: 10.1016/j.jcrc.2012.01.003] [Medline: 22341727]
8. Al-Moteri M, Plummer V, Cooper S, Symmons M. Clinical deterioration of ward patients in the presence of antecedents: a systematic review and narrative synthesis. Aust Crit Care 2019 Sep;32(5):411-420. [doi: 10.1016/j.aucc.2018.06.004] [Medline: 30025983]
9. Schein RM, Hazday N, Pena M, Ruben BH, Sprung CL. Clinical antecedents to in-hospital cardiopulmonary arrest. Chest 1990 Dec;98(6):1388-1392. [doi: 10.1378/chest.98.6.1388] [Medline: 2245680]
10. Churpek MM, Yuen TC, Edelson DP. Risk stratification of hospitalized patients on the wards. Chest 2013 Jun;143(6):1758-1765 [FREE Full text] [doi: 10.1378/chest.12-1605] [Medline: 23732586]

XSL•FO
RenderX

11. Barfod C, Lauritzen MM, Danker JK, Sölétormos G, Forberg JL, Berlac PA, et al. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. Scand J Trauma Resusc Emerg Med 2012 Apr 10;20:28 [FREE Full text] [doi: 10.1186/1757-7241-20-28] [Medline: 22490208]
12. Schmid A, Hoffman L, Happ MB, Wolf GA, DeVita M. Failure to rescue: a literature review. J Nurs Adm 2007 Apr;37(4):188-198. [doi: 10.1097/01.NNA.0000266838.23814.65] [Medline: 17415106]
13. Mok WQ, Wang W, Liaw SY. Vital signs monitoring to detect patient deterioration: an integrative literature review. Int J Nurs Pract 2015 May;21 Suppl 2:91-98. [doi: 10.1111/ijn.12329] [Medline: 26125576]
14. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. Resuscitation 2016 May;102:1-5 [FREE Full text] [doi: 10.1016/j.resuscitation.2016.02.005] [Medline: 26898412]
15. Dhala A, Fusaro MV, Uddin F, Tuazon D, Klahn S, Schwartz R, et al. Integrating a virtual ICU with cardiac and cardiovascular ICUs: managing the needs of a complex and high-acuity specialty ICU cohort. Methodist Debakey Cardiovasc J 2023 Aug 01;19(4):4-16 [FREE Full text] [doi: 10.14797/mdcvj.1247] [Medline: 37547898]
16. Dhala A, Sasangohar F, Kash B, Ahmadi N, Masud F. Rapid implementation and innovative applications of a virtual intensive care unit during the COVID-19 pandemic: case study. J Med Internet Res 2020 Sep 03;22(9):e20143 [FREE Full text] [doi: 10.2196/20143] [Medline: 32795997]
17. Weller GB, Mault J, Ventura ME, Adams J, Campbell FJ, Tremper KK. A retrospective observational study of continuous wireless vital sign monitoring via a medical grade wearable device on hospitalized floor patients. J Clin Med 2024 Aug 13;13(16):4747 [FREE Full text] [doi: 10.3390/jcm13164747] [Medline: 39200889]
18. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
19. Cretikos MA, Bellomo R, Hillman K, Chen J, Finfer S, Flabouris A. Respiratory rate: the neglected vital sign. Med J Aust 2008 Jun 02;188(11):657-659. [doi: 10.5694/j.1326-5377.2008.tb01825.x] [Medline: 18513176]
20. Rasooly IR, Makeneni S, Khan AN, Luo B, Muthu N, Bonafide CP. The alarm burden of excess continuous pulse oximetry monitoring among patients with bronchiolitis. J Hosp Med 2021 Dec;16(12):727-729 [FREE Full text] [doi: 10.12788/jhm.3731] [Medline: 34798003]
21. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
22. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
23. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar 19;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
24. Narcisse V, Ishaq F, Gomez M, Homer S, Griffin L, Pletcher S, et al. Implementation of remote patient monitoring and earlier CERT activation: effects on ICU transfer and mortality. J Clin Med 2025 Oct 21;14(20):7434 [FREE Full text] [doi: 10.3390/jcm14207434] [Medline: 41156302]

Abbreviations

CVSM: continuous vital sign monitoring
FDA: Food and Drug Administration
HM: Houston Methodist
HR: heart rate
ICU: intensive care unit
LTAC: long-term acute care
PCA: patient care assistant
RPM: remote patient monitoring
RR: respiratory rate
SpO2: oxygen saturation
VICU: virtual intensive care unit
VOC: virtual operations center
VS: vital sign

XSL•FO
RenderX

Edited by C Perrin; submitted 29.May.2025; peer-reviewed by P Helmer, M Haveman; comments to author 19.Aug.2025; revised version received 04.Dec.2025; accepted 03.Jan.2026; published 26.Jan.2026.

Please cite as:
Nguyen NA, Lee G, Holderread B, Holman T, Pletcher S, Schwartz R
Scaling Wireless Continuous Vital Sign Monitoring Across an 8-Hospital Health System: Digital Health Implementation Report
JMIR Med Inform 2026;14:e78216
URL: https://medinform.jmir.org/2026/1/e78216
doi:10.2196/78216
PMID:41482965

©Ngoc-Anh Nguyen, Grace Lee, Brendan Holderread, Terrie Holman, Sarah Pletcher, Roberta Schwartz. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Deep Learning Model Using Transfer Learning for Detecting Left Ventricular Systolic Dysfunction: Retrospective Algorithm Development and Validation Study

Sungjae Lee[1], MS; Jung-Woo Son[2], MD; Sung-Ai Kim[3], MD, PhD; Min-Soo Ahn[2], MD, PhD; Sang Jun Lee[2], MD; Sang-Jin Han[3], MD, PhD; Taehyun Joo[1], MD; Yeongyeon Na[1], MS; Sunghoon Joo[1], PhD; Hyun Jin Ahn[1], MD; Mineok Chang[1], MD, PhD; Yeha Lee[1], PhD; Young Jun Park[4], MD

[1]VUNO Inc, Seoul, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Wonju Severance Christian Hospital, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

[3]Division of Cardiology, Department of Internal Medicine, Hallym Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Republic of Korea

[4]Division of Cardiology, Department of Internal Medicine, Korea University Anam Hospital, Korea University College of Medicine, 73 Goryeodae-ro, Seongbuk-gu, Seoul, Republic of Korea

Corresponding Author:
Young Jun Park, MD
Division of Cardiology, Department of Internal Medicine, Korea University Anam Hospital, Korea University College of Medicine, 73 Goryeodae-ro, Seongbuk-gu, Seoul, Republic of Korea

Abstract

Background: Artificial intelligence–augmented electrocardiogram (AI-ECG) models for detecting left ventricular systolic dysfunction (LVSD) often exhibit degraded performance in patients with comorbidities.

Objective: This study aimed to introduce and validate a recalibration method using longitudinal patient data to enhance prediction accuracy and simulate its clinical utility for ongoing monitoring.

Methods: We conducted a multicenter, retrospective cohort study using data from 2 hospitals in Korea. A dataset of paired transthoracic echocardiograms (TTEs) and electrocardiograms (ECGs) matched within a 2-week interval was constructed, separating pairs into baseline (first for each patient) and follow-up assessments. In addition to conventional supervised learning, we developed a patient-wise recalibration strategy that incorporated historical left ventricular ejection fraction measurements and prior AI-ECG outputs to adjust for future predictions, thus empirically mitigating confounding effects. Pretraining was also implemented to enhance the model's performance.

Results: The recalibrated 12-lead DeepECG LVSD model achieved an area under the receiver operating curve of 0.956 (95% CI 0.946 - 0.965) for internal validation and 0.940 (95% CI 0.936 - 0.945) for external validation of follow-up TTE-ECG pairs. The uncalibrated 12-lead DeepECG LVSD model also showed modest performance, with an area under the receiver operating curve of 0.953 (95% CI 0.941 - 0.965) in the internal validation and 0.947 (95% CI 0.943 - 0.951) in the external validation when tested on baseline TTE-ECG pairs. Recalibration yielded statistically significant improvements in the 12-lead DeepECG LVSD models (P<.001), with enhanced and more balanced performance across all clinical subgroups.

Conclusions: Patient-wise recalibration improved accuracy and consistency across various comorbidities by mitigating performance degradation and bias. This broadens the application of AI-ECG for LVSD detection from low-risk screening to high-risk longitudinal monitoring.

(JMIR Med Inform 2026;14:e83127) doi:10.2196/83127

KEYWORDS

left ventricular systolic dysfunction; deep learning; electrocardiogram; echocardiography; artificial intelligence

Introduction

Overview

Left ventricular systolic dysfunction (LVSD) is a critical precursor of symptomatic heart failure (HF) and is associated with a substantially elevated risk of cardiovascular morbidity and mortality [1,2]. Despite the clinical benefits of early detection, including the timely initiation of guideline-directed medical therapy [3], LVSD is often underdiagnosed until overt symptoms emerge. This diagnostic delay is largely attributable

XSL•FO
RenderX

to a lack of efficient and scalable screening tools applicable to the general population.

Conventional diagnostic modalities, such as echocardiography, cardiac magnetic resonance imaging, and biomarker assays, although accurate, are costly, time-consuming, and influenced by confounding clinical variables, such as age and renal function. As such, they are not ideally suited for widespread screening initiatives. By contrast, electrocardiogram (ECG) remains a more accessible and cost-effective tool for detecting cardiovascular diseases. In this context, early studies, including those by Attia et al [4-7], developed artificial intelligence–augmented electrocardiogram (AI-ECG) algorithms capable of screening for LVSD using ECG inputs. Several prospective observational [8-10] and interventional [11] studies have since sought to validate the clinical utility of these artificial intelligence (AI)-assisted diagnostic strategies.

In parallel, a paradigm shift in HF management has occurred [12,13], with increasing emphasis on identifying and managing patients in the early or at-risk stages. In this context, AI-based detection of LVSD using ECG may offer a more accessible alternative to echocardiography and provide benefits for this at-risk population. However, early AI studies have shown limited performance in certain subgroups, such as older individuals [14] and patients with atrial fibrillation (AFIB) [6,10,14,15], groups that are commonly encountered in clinical practice and substantially overlap with those at a higher risk of HF.

Thus, we developed a deep learning–based framework (ie, DeepECG LVSD) consisting of three sequential learning phases: (1) large-scale ECG pretraining using automatically generated labels, (2) supervised fine-tuning for LVSD detection using paired baseline transthoracic echocardiography-electrocardiogram (TTE-ECG) data, and (3) a patient-wise recalibration strategy that incorporates prior left ventricular ejection fraction (LVEF) measurements and historical AI-ECG outputs. We evaluated this framework using both internal and external validation cohorts, seeking to enable a more reliable and clinically adoptable application of AI-ECG in routine cardiac care.

Model Overview

The proposed DeepECG LVSD framework takes electrocardiographic recordings as input and outputs a continuous risk score for LVSD. The model is developed using a 3-stage learning strategy comprising large-scale ECG pretraining with automatically generated labels, supervised fine-tuning for LVSD detection using paired baseline TTE-ECG data, and a patient-specific recalibration phase. The recalibration component leverages longitudinal patient information, including prior AI-ECG outputs and, when available, corresponding LVEF measurements, to adjust individual risk estimates and stabilize model behavior at a fixed decision threshold across heterogeneous clinical subgroups. To enhance applicability in settings where echocardiography is unavailable, we additionally evaluated a masked-LVEF variant that performs recalibration using serial ECG-derived information only. Full architectural details, training procedures, and evaluation protocols are provided in the Methods section.

Methods

Consent

We conducted a multicenter, retrospective cohort study at 2 university hospitals in Korea: Hallym Sacred Heart Hospital (Hospital A) and Wonju Severance Christian Hospital (Hospital B).

Study Population and Data Acquisition

For model development and internal validation, we identified adult patients from Hospital A (n=259,943, January 2006 to December 2020) who underwent standard 12-lead ECG, including a subset of 59,433 patients who underwent transthoracic echocardiography (TTE). Adult patients from Hospital B (n=67,377; November 2007 to December 2018) who underwent both ECG and TTE were identified for external validation. Standard 10-second 12-lead ECGs recorded at 250 Hz or 500 Hz sampling rates using Marquette ECG machines and the MUSE software system (General Electric Healthcare) were extracted to construct the ECG dataset. For the TTE dataset, free-text TTE reports were parsed to extract echocardiographic parameters, including LVEF. Demographic variables (age and sex) and *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnosis codes were obtained from electronic medical records for characterization in baseline and subpopulation analyses. Records from the TTE and ECG datasets were paired to construct the TTE-ECG dataset. For each patient, TTE recordings were matched to the closest ECG acquired within a 14-day window period [4]. Patients may have had multiple eligible pairs that were further categorized based on whether the TTE-ECG pair represented the first match (baseline TTE-ECG pair) or a subsequent match (follow-up TTE-ECG pair).

For Hospital A, patients were included in the supervised fine-tuning if they underwent both TTE and ECG recordings within a 14-day window. Eligible TTE-ECG pairs were randomly assigned to the training, validation, and test sets in an 8:1:1 ratio and used during supervised training. If a patient had multiple eligible pairs, all the pairs from the same patient were assigned to the same split. ECGs from patients without any eligible TTE-ECGs were also randomly split into training, validation, and test sets and used in the pretraining phase, as described later.

Definition of Key Variables

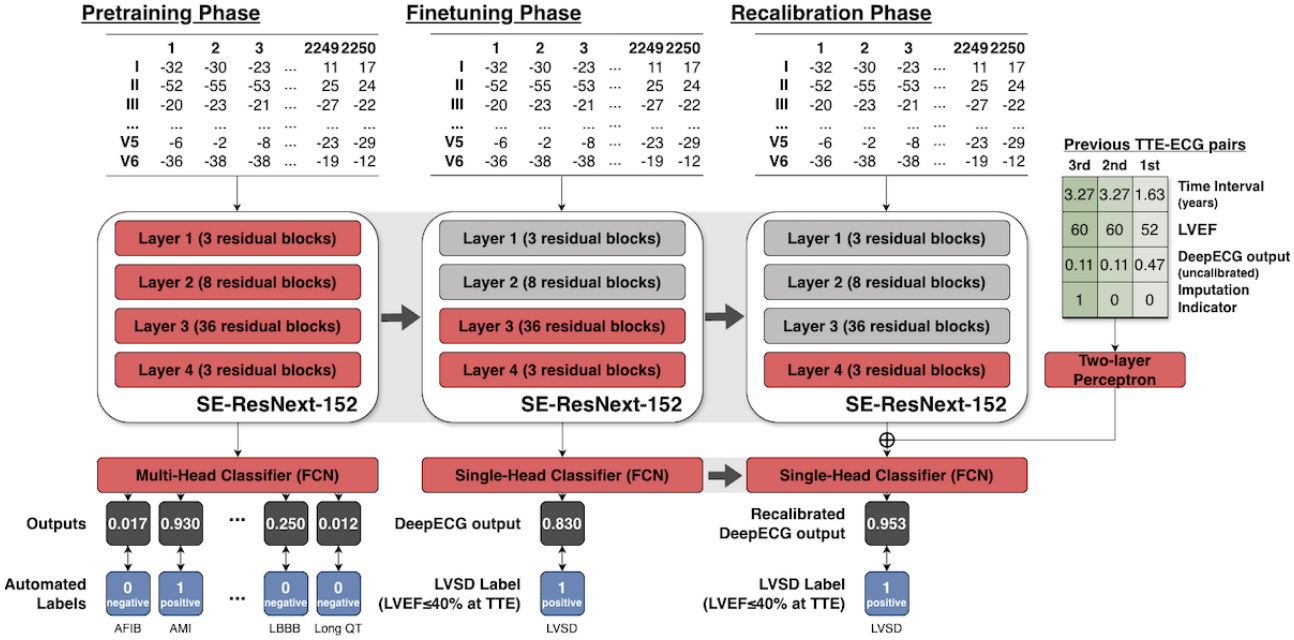
The target condition of our AI-ECG model was the presence of LVSD, defined as an LVEF of ≤40%. TTE reports may contain multiple LVEF values derived using different methods, including the Simpson method (biplane approach), M-mode (2D), and visual estimation by an examiner. When multiple LVEF values were reported, a single value was selected based on a predefined order of priority, that is, Simpson method, M-mode, and visual assessment. The selected LVEF value was then converted into binary labels indicating the presence (positive TTE-ECG pair if LVEF was ≤40%) or absence (negative TTE-ECG pair if LVEF was >40%) of an LVSD.

XSL•FO
RenderX

Development of the AI-ECG Algorithm

The development of the AI-ECG algorithm was structured into 3 sequential learning phases (Figure 1). In addition to the general framework of supervised learning, we introduced an initial pretraining stage by leveraging large-scale and automatically labeled ECG data. This step guided the model to capture broad, but clinically relevant, low-level ECG features that may not be apparent in smaller outcome-specific datasets, thus enabling more effective supervised training in the subsequent phase. The pretrained model parameters were then used to initialize the model for the second LVSD-specific fine-tuning phase.

Figure 1. Schematic overview of the training process for artificial intelligence–augmented electrocardiogram SE-ResNeXt-152 was used as the backbone of the artificial intelligence–augmented electrocardiogram during the entire training process. AFIB: atrial fibrillation; AI-ECG: artificial intelligence–augmented electrocardiogram; AMI: acute myocardial infarction; ECG: electrocardiogram; FCN: fully connected network; LBBB: left bundle branch block; LVEF: left ventricular ejection fraction; LVSD: left ventricular systolic dysfunction; SE-ResNeXt: squeeze-and-excitation ResNeXt; TTE: transthoracic echocardiography; TTE-ECG: transthoracic echocardiography-electrocardiogram.



Following fine-tuning, we implemented the third and final recalibration phase. Patients may exhibit different underlying cardiac conditions, including arrhythmias and prior infarction, resulting in variable correlations between ECG signals and true LVEF at the individual level. Therefore, the model was trained to adjust its predictions using each patient's historical LVEF values and prior model outputs. This approach was designed to account for individual variability and temporal ECG trends relevant in real-world follow-up settings.

First Pretraining Phase: Pretraining AI-ECG With Various ECG Features

The initial phase focused on pretraining the AI-ECG model using a large-scale ECG dataset from Hospital A, in which the labels were derived automatically. Specifically, we processed the analysis results from the MUSE system, including both free-text diagnostic statements and ECG numerical values (eg, PR interval, QRS duration, and QT interval).

The diagnostic strings were systematically parsed to identify and assign binary labels based on the presence of key clinical conditions, including arrhythmia (eg, AFIB, atrial flutter, premature atrial complex, and premature ventricular complex), myocardial infarction, and conduction abnormalities (eg, left bundle branch block [LBBB] and right bundle branch block). ECG values were thresholded to generate binary labels, including a prolonged PR interval (longer than 200 ms), a wide QRS complex (QRS duration >140 ms), and a prolonged QT interval (>450 ms in men and >470 ms in women after heart rate correction). This process resulted in 17 binary labels related to arrhythmias, 8 for conduction abnormalities, 8 for ischemic patterns, 7 for miscellaneous disease patterns, and 6 for binary labels derived from the thresholded numerical values. Each ECG was annotated with 46 binary labels to capture a comprehensive range of clinically relevant ECG features.

These binary labels were used as training targets for the initial pretraining phase of the AI-ECG model. Patients without eligible TTE-ECG pairs were included in the training phase. The model was trained to predict the presence of each binary label using the ECG signals of the training set. The cross-entropy loss on the validation split was monitored to determine the point of early termination of this pretraining phase.

Second Fine-Tuning Phase: Fine-Tuning AI-ECG for LVSD Detection

In the second phase, the AI-ECG model was fine-tuned to detect LVSD using ECG records from the baseline TTE-ECG pair set at Hospital A in a supervised learning framework. The model weights were initialized during the pretraining phase. The first 2 convolutional blocks were kept frozen (ie, their weights were not updated during training) to ensure stable low-level feature extraction during the pretraining phase.

In contrast to the fine-tuned AI-ECG model (referred to as DeepECG LVSD) initialized from pretrained weights, a separate AI-ECG model was developed without pretraining under

XSL•FO
RenderX

identical conditions. For this model, the weights were randomly initialized, and no layers were frozen during training. This nonpretrained AI-ECG model served as a control to assess the efficacy of the initial pretraining phase.

Third Recalibration Phase: Recalibration Based on Previous LVEF and AI-ECG Outputs

Finally, we introduced a recalibration method to refine the model predictions by integrating previous LVEF measurements and prior model outputs. For recalibration, the model was supplemented with up to 3 previous TTE-ECG pairs for each patient as additional inputs. For each previous pair, we included (1) the actual LVEF value measured on TTE, (2) the uncalibrated DeepECG LVSD model output of the corresponding ECG, (3) the time interval between the previous ECG and the current ECG, and (4) an additional binary indicator representing missing data.

The calibrated model outputs of the previous ECGs were calculated in advance from the DeepECG LVSD models trained in the second fine-tuning phase. If a patient had fewer than 3 previous TTE-ECG pairs, missing data were imputed using a backward-filling strategy (ie, using the closest available prior values), and the binary indicator was marked as 1 for imputed (missing) values or 0 for observed (true) values. The resulting 12 features (4 features for the 3 previous pairs each) were used as the input for a 2-layer perceptron whose output was combined with the main model's feature outputs using a point-wise sum operation before the final classification head. A second model with a different recalibration scheme—the recalibration masked to the LVEF—was trained. The LVEF measurements were excluded as input values, resulting in 9 features as 2-layer perceptron inputs. This model relies only on historical ECG records and does not require actual LVEF measurements from TTE for model inference.

Training for this recalibration phase was conducted using the follow-up TTE-ECG dataset from Hospital A. The weights were initialized in the second fine-tuning phase to preserve the learned ECG representations. The first 3 convolutional blocks were frozen during this phase to maintain the functionality trained during the first 2 stages.

Model Architecture and Data Preprocessing

The backbone network of SE-ResNeXt152 was adopted as the model architecture and was composed of 152 1D convolutional layers nested in 4 convolutional blocks with squeeze and excitation modules [16,17]. This model architecture has been widely used in AI-ECG research and competitions, exhibiting competitive performance across various tasks. The final classification layer was modified during the learning phases as follows: (1) a multiheaded layer for the first pretraining phase, (2) a single-head layer for the second fine-tuning phase, and (3) a single-head layer connected to a 2-layer perceptron with a 12-dimensional vector input for the third recalibration phase. During the training phase, 9-second segments of ECGs were randomly cropped from the standard 10-second ECG recording as model input data. For evaluation, we cropped 10-second ECGs into 0 - 9, 0.5 - 9.5, and 1 - 10 segments, and then averaged the output of each segment. Random cropping during training helped the AI-ECG model become location-invariant, and multiple cropping effectively acted as a model ensemble during evaluation. Subsequently, the cropped ECGs were downsampled or upsampled to 250 Hz to ensure a fixed input length. Finally, we performed bandpass filtering with a frequency band of 0.67 Hz to 40 Hz to remove baseline wandering and high-frequency noise.

Training With Different Lead Configurations

To evaluate the possibility of broader application, separate models were trained using different numbers of ECG leads. The 12-lead, 6-lead (limb lead), and 1-lead (lead I) configurations were used. The models for each configuration were trained and validated independently following the same protocol. The first input layer of each model was modified to accept signals with channel counts of 12, 6, and 1. The primary analyses and main results of this study were based on the 12-lead configuration, which is consistent with the standard ECG acquisition.

Performance Evaluation and Statistical Analysis

The model performance for detecting LVSD was evaluated using both internal (Hospital A) and external (Hospital B) test sets across all lead configurations (12-, 6-, and 1-lead). Following the second phase of fine-tuning, the DeepECG LVSD and AI-ECG models (without pretraining, as a control) were validated using baseline TTE-ECG pairs. To determine the effect of pretraining, area under the receiver operating curve (AUROC) values of the 2 models were compared. Subgroup analyses were performed using key clinical variables (demographics, *ICD-10* codes, and ECG characteristics) to assess the robustness across various subpopulations. These subgroup analyses primarily focused on the DeepECG LVSD model with a 12-lead configuration in an external validation set.

After the final recalibration phase, the recalibrated DeepECG LVSD model was evaluated for follow-up TTE-ECG pairs using ECGs, LVSD labels, and previous inference results. Additional validation was conducted on the second recalibrated DeepECG LVSD model with the LVEF values masked during training. The DeepECG LVSD model before recalibration was validated for direct comparison. AUROC comparisons of DeepECG LVSD models before and after recalibration were performed in the entire test cohort as well as within each subgroup to quantify performance improvements attributable to recalibration and identify subpopulations that benefit from the recalibration process.

The raw model output was a continuous probability value between 0 and 1, which represented the estimated likelihood of LVSD. The AUROC was calculated to quantify discriminative performance. For binary classification, the threshold was determined using Hospital A's validation set, which yielded the highest Youden index [18] (sensitivity + specificity – 1). The threshold was applied to calculate sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). For each metric, 95% CIs were estimated using the DeLong method [19] for the AUROC and the Clopper-Pearson exact method [20] for the sensitivity, specificity, PPV, and NPV. For subgroup analysis, all 95% CIs

XSL•FO
RenderX

were estimated by bootstrapping over 1000 iterations. All statistical AUROC comparisons were conducted using the DeLong method, and the corresponding *P* values were reported.

Baseline characteristics are summarized as mean (SD) for continuous variables and as frequency (%) for binary variables. Group comparisons between the training, validation, and test sets of Hospital A, and between the test sets of Hospitals A and B were performed. Continuous variables were tested on ANOVA or Student *t* test and categorical variables were tested using the chi-square test. Analysis of baseline characteristics was performed on the baseline TTE-ECG pairs to ensure nonduplicated, patient-level sampling.

Ethical Considerations

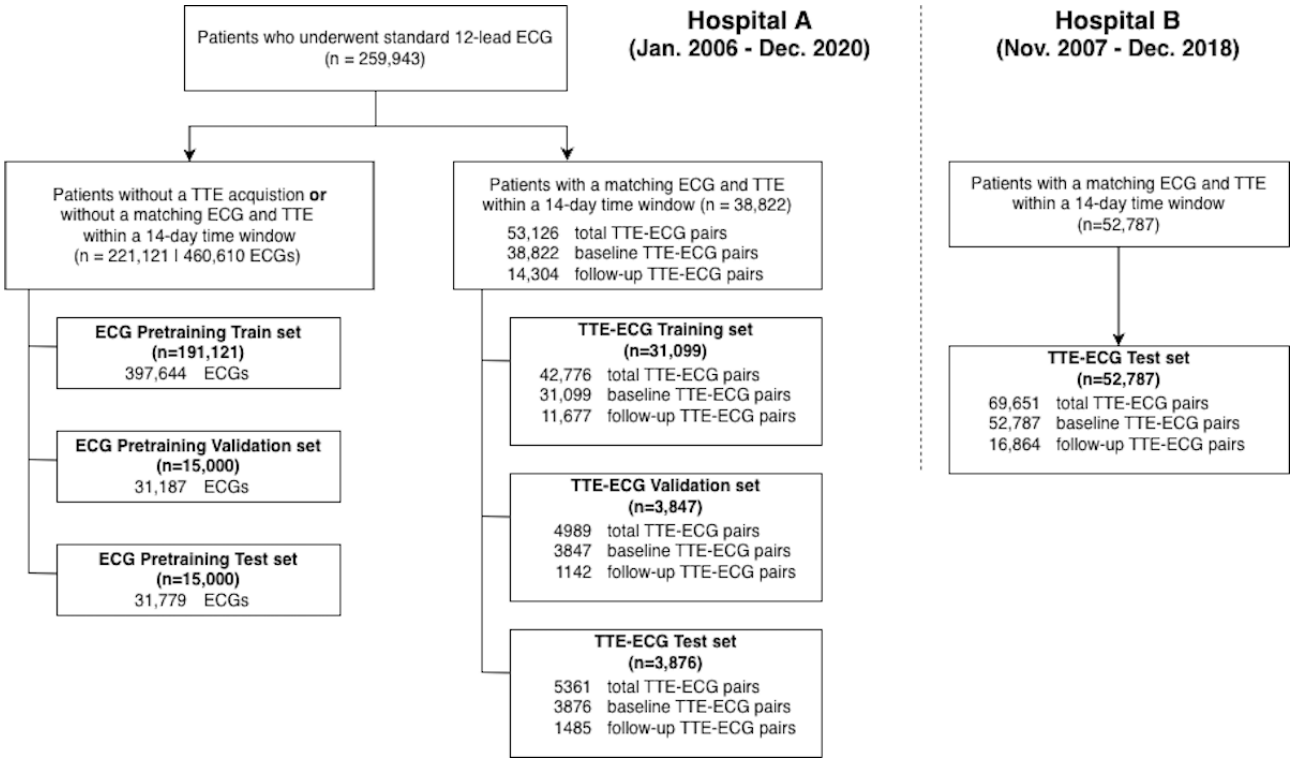
This study was approved by the Ethics Committees of Hallym Sacred Heart Hospital (IRB No. 2022-03-016-008) and Wonju Severance Christian Hospital (IRB No. CR322014). Due to the retrospective nature of the study, the requirement for informed consent was waived by both institutional review boards. All data used in this study were fully deidentified prior to analysis, and no personally identifiable information, including patient names, hospital numbers, or dates of birth, was accessed by the research team, ensuring the privacy and confidentiality of the participants. No compensation was provided to participants, as this study exclusively involved the secondary analysis of preexisting, de-identified clinical data.

Results

Baseline Characteristics of Study Population

In total, 31,099 training participants, 3847 validation participants, and 3876 test participants from Hospital A, and 52,787 external test participants from Hospital B contributed to one or more TTE-ECG pairs in this study (Figure 2). The clinical, ECG, and echocardiographic parameters for each dataset are summarized in Table 1. In the Hospital A test set, the mean age was 63.3 (15.5) years, 51.9% (2012/3876) were men, and the mean LVEF was 60% (11.6%). In the Hospital B test set, the mean age was 63.7 (14.4) years, 53.8% (28,391/52,787) were men, and the mean LVEF was 62.7% (10.2%). No significant differences in demographics, TTE, and ECG measurements were observed between the training, validation, and test sets at Hospital A. However, the test set of Hospital B exhibited a higher ratio of men, a higher mean LVEF, and a higher proportion of AFIB or atrial flutter in the ECG than the test set of Hospital A.

Figure 2. Study flowchart summarizing the data collection and preprocessing steps performed in this study. ECG: electrocardiogram; TTE: transthoracic echocardiography; TTE-ECG: transthoracic echocardiography-electrocardiogram.



XSL•FO
RenderX

Table . Baseline characteristics of the patients for baseline TTE-ECG[a] pairs

Characteristics	Hospital A				Hospital B (n=52,787)	
	Train (n=31,099)	Validation (n=3847)	Test (n=3876)	*P* value[b]		*P* value[c]
Demographic information						
Age (years), mean (SD)	63.4 (15.1)	63.4 (15.0)	63.3 (15.5)	.98	63.7 (14.4)	.19
Men, n (%)	16,339 (52.5)	2014 (52.4)	2012 (51.9)	.75	28,391 (53.8)	.02
BMI (kg/m^2), mean (SD)	24.3 (3.8)	24.3 (3.9)	24.2 (3.7)	.34	24.3 (3.7)	.03
DM[d], n (%)	7034 (22.6)	887 (23.1)	878 (22.7)	.83	11,795 (22.3)	.67
HTN[e], n (%)	16,237 (52.2)	1953 (50.8)	2003 (51.7)	.22	23,050 (43.7)	<.001
IHD[f], n (%)	1904 (6.1)	257 (6.7)	238 (6.1)	.40	3098 (5.9)	.51
CKD[g], n (%)	1989 (6.4)	253 (6.6)	263 (6.8)	.61	3280 (6.2)	.17
LVSD[h], n (%)	2419 (7.8)	300 (7.8)	310 (8.0)	.89	2563 (4.9)	<.001
Echocardiographic or electrocardiographic information						
LVEF[i] (%), mean (SD)	60.2 (11.5)	60.0 (11.5)	60.0 (11.6)	.51	62.7 (10.2)	<.001
E/e', mean (SD)	9.6 (4.5)	9.7 (4.7)	9.7 (4.7)	.35	12.9 (5.2)	<.001
HR[j] (bpm), mean (SD)	74.6 (16.7)	74.1 (16.3)	74.8 (17.1)	.12	75.5 (18.1)	.01
PR (ms), mean (SD)	165.3 (26.2)	165.8 (27.0)	165.2 (25.6)	.55	164.0 (27.1)	.007
QRS (ms), mean (SD)	94.1 (16.0)	94.4 (16.0)	94.2 (15.8)	.70	92.9 (16.4)	<.001
QTc (ms), mean (SD)	436.5 (33.1)	436.5 (33.6)	436.5 (34.1)	.99	441.6 (37.4)	<.001
AFIB[k]/AFL[l], n (%)	2297 (7.4)	278 (7.2)	273 (7)	.72	4963 (9.4)	<.001

[a]TTE-ECG: transthoracic echocardiography–electrocardiogram.

[b]*P* values for group comparisons among the train, validation, and test sets of Hospital A were calculated using ANOVA for continuous variables and the chi-square test for categorical variables.

[c]*P* values for group comparisons between the test set of Hospital A and Hospital B were calculated using the Student *t* test for continuous variables and chi-square test for categorical variables.

[d]DM: diabetes mellitus.

[e]HTN: hypertension.

[f]IHD: ischemic heart disease.

[g]CKD: chronic kidney disease.

[h]LVSD: left ventricular systolic dysfunction.

[i]LVEF: left ventricular ejection fraction.

[j]HR: heart rate.

[k]AFIB: atrial fibrillation.

[l]AFL: atrial flutter.

LVSD Detection in Baseline TTE-ECG Pairs

After the second fine-tuning phase, the DeepECG LVSD and AI-ECG models (trained without pretraining) were validated against the LVSD label of the baseline TTE-ECG pairs (first pair for each patient) of hospitals A and B. The standard 12-lead DeepECG LVSD model exhibited an AUROC of 0.953 (95% CI 0.941 - 0.965) in internal validation and 0.947 (95% CI 0.943 - 0.951) in external validation (Table 2). Using thresholds that maximized the Youden index, the binary classification performance showed a sensitivity of 0.903 (95% CI 0.865 - 0.934) and a specificity of 0.844 (95% CI 0.832 - 0.856)

XSL•FO
RenderX

in the internal validation and a sensitivity of 0.915 (95% CI 0.904 - 0.926) and a specificity of 0.847 (95% CI 0.844 - 0.850) in the external validation.

Table . LVSD[a] detection performance of the AI-ECG[b] model and DeepECG LVSD model, evaluated within baseline TTE-ECG[c] pairs.

Performance		12-lead, value (95% CI)	6-lead, value (95% CI)	1-lead, value (95% CI)
Internal validation (Hospital A)				
AI-ECG without pretraining	AUROC[d]	0.939 (0.926 - 0.953)[e]	0.927 (0.912 - 0.942)	0.919 (0.905 - 0.934)
	Sensitivity	0.897 (0.857 - 0.928)	0.858 (0.814 - 0.895)	0.910 (0.872 - 0.939)
	Specificity	0.834 (0.822 - 0.846)	0.840 (0.828 - 0.852)	0.777 (0.764 - 0.791)
	PPV[f]	0.320 (0.289 - 0.352)	0.318 (0.287 - 0.351)	0.262 (0.236 - 0.289)
	NPV[g]	0.989 (0.985 - 0.992)	0.986 (0.981 - 0.989)	0.990 (0.986 - 0.993)
DeepECG LVSD	AUROC	0.953 (0.941 - 0.965)	0.943 (0.930 - 0.955)	0.930 (0.916 - 0.944)
	Sensitivity	0.903 (0.865 - 0.934)	0.935 (0.902 - 0.960)	0.913 (0.876 - 0.942)
	Specificity	0.844 (0.832 - 0.856)	0.789 (0.776 - 0.802)	0.787 (0.773 - 0.800)
	PPV	0.335 (0.303 - 0.368)	0.278 (0.251 - 0.307)	0.271 (0.244 - 0.299)
	NPV	0.990 (0.986 - 0.993)	0.993 (0.990 - 0.995)	0.990 (0.987 - 0.993)
External validation (Hospital B)				
AI-ECG without pretraining	AUROC	0.937 (0.933 - 0.941)	0.926 (0.921 - 0.930)	0.916 (0.911 - 0.921)
	Sensitivity	0.890 (0.878 - 0.902)	0.895 (0.883 - 0.907)	0.854 (0.839 - 0.867)
	Specificity	0.844 (0.841 - 0.847)	0.814 (0.811 - 0.817)	0.840 (0.837 - 0.843)
	PPV	0.226 (0.217 - 0.234)	0.197 (0.190 - 0.205)	0.214 (0.206 - 0.222)
	NPV	0.993 (0.993 - 0.994)	0.993 (0.993 - 0.994)	0.991 (0.990 - 0.992)
DeepECG LVSD	AUROC	0.947 (0.943 - 0.951)	0.936 (0.932 - 0.940)	0.922 (0.918 - 0.927)
	Sensitivity	0.915 (0.904 - 0.926)	0.909 (0.898 - 0.920)	0.877 (0.863 - 0.889)
	Specificity	0.847 (0.844 - 0.850)	0.831 (0.828 - 0.835)	0.823 (0.820 - 0.826)
	PPV	0.234 (0.226 - 0.243)	0.216 (0.208 - 0.224)	0.202 (0.194 - 0.209)
	NPV	0.995 (0.994 - 0.996)	0.994 (0.994 - 0.995)	0.992 (0.992 - 0.993)

[a]LVSD: left ventricular systolic dysfunction.

[b]AI-ECG: artificial intelligence–augmented electrocardiogram.

[c]TTE-ECG: transthoracic echocardiography-electrocardiogram.

[d]AUROC: area under the receiver operating curve.

[e]CIs (95%) are estimated from the DeLong method for AUROC and the Clopper-Pearson exact method for sensitivity, specificity, positive predictive value, and negative predictive value.

[f]PPV: positive predictive value.

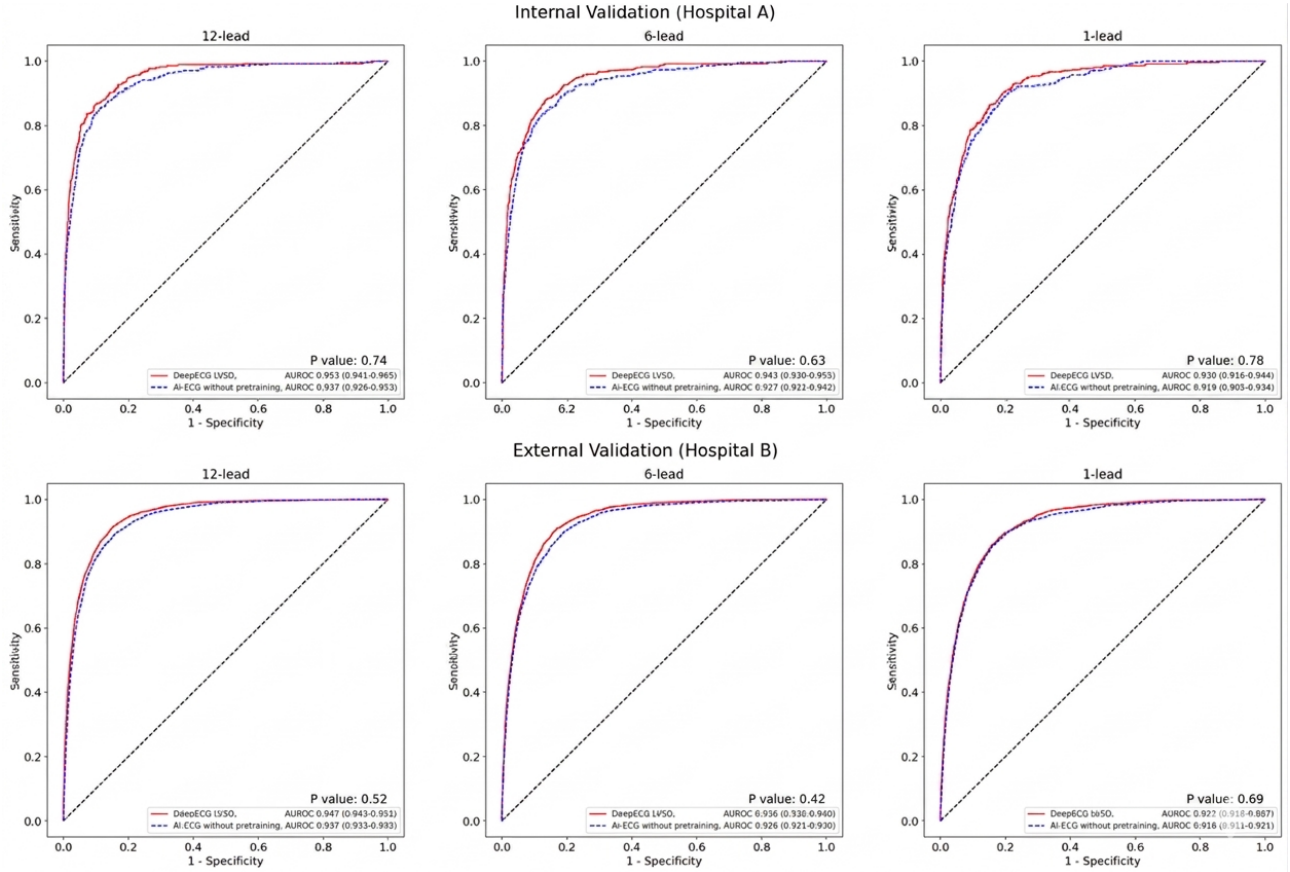
[g]NPV: negative predictive value.

The models using 6-lead and 1-lead ECGs also showed adequate performance with internal AUROC 0.943 (95% CI 0.930 - 0.955) and external AUROC 0.936 (95% CI 0.932 - 0.940) for 6-lead models, and internal AUROC 0.930 (95% CI 0.916 - 0.944) and external AUROC 0.922 (95% CI 0.918 - 0.927) for 1-lead models.

Notably, when compared with the AI-ECG models without pretraining, the pretrained DeepECG LVSD models demonstrated a higher performance, though not statistically significant (Figure 3). The 12-lead AI-ECG models without pretraining showed an AUROC of 0.939 (95% CI 0.926 - 0.953) for the internal validation set and 0.937 (95% CI 0.933 - 0.941) for the external validation set.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curve for left ventricular systolic dysfunction detection performance of the artificial intelligence–augmented electrocardiogram model and the DeepECG LVSD model, evaluated within baseline transthoracic echocardiography-electrocardiogram pairs. AI-ECG: artificial intelligence–augmented electrocardiogram; AUROC: area under the receiver operating characteristic; ECG: electrocardiogram; LVSD: left ventricular systolic dysfunction.



Subgroup analysis revealed that the 12-lead DeepECG LVSD model maintained robust discriminative performance across most clinical variables (Table 3). A notable decline in performance was observed in some subgroups. In patients with AFIB or flutter, the AUROC decreased to 0.875 (95% CI 0.858 - 0.890), accompanied by a reduction in specificity to 0.578 (95% CI 0.564 - 0.592) and PPV to 0.174 (95% CI 0.158 - 0.190). LBBB was also associated with a low AUROC of 0.859 (95% CI 0.828 - 0.890) and a lower specificity of 0.219 (95% CI 0.177 - 0.266). Other conditions associated with a specificity below 0.700 included the presence of ischemic heart disease, a wide QRS complex, and a prolonged QT interval.

XSL•FO
RenderX

Table . Subgroup analysis on LVSD[a] detection performance of the DeepECG LVSD model, evaluated within 12-lead, external, baseline TTE-ECG[b] pairs.[c]

Subgroup		DeepECG LVSD				
	n/N (%)[d]	AUROC[e] (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV[f] (95% CI)	NPV[g] (95% CI)
Age ≤70	1268/33,176 (3.8%)	0.957 (0.952 - 0.961)	0.914 (0.899 - 0.929)	0.874 (0.870 - 0.878)	0.224 (0.213 - 0.236)	0.996 (0.995 - 0.997)
Age >70	1295/19,611 (6.6%)	0.930 (0.924 - 0.936)	0.918 (0.903 - 0.932)	0.797 (0.791 - 0.802)	0.242 (0.231 - 0.254)	0.993 (0.991 - 0.994)
Male	1550/28,391 (5.5%)	0.945 (0.940 - 0.950)	0.920 (0.906 - 0.933)	0.831 (0.826 - 0.835)	0.239 (0.228 - 0.250)	0.994 (0.993 - 0.995)
Female	1011/24,338 (4.2%)	0.949 (0.944 - 0.954)	0.910 (0.893 - 0.927)	0.864 (0.859 - 0.868)	0.224 (0.212 - 0.236)	0.996 (0.995 - 0.996)
DM[h] (+)	836/11,795 (7.1%)	0.939 (0.932 - 0.946)	0.926 (0.906 - 0.943)	0.809 (0.803 - 0.816)	0.270 (0.255 - 0.286)	0.993 (0.991 - 0.995)
DM (–)	1727/40,992 (4.2%)	0.949 (0.944 - 0.953)	0.911 (0.898 - 0.925)	0.856 (0.853 - 0.860)	0.218 (0.209 - 0.227)	0.995 (0.995 - 0.996)
HTN[i] (+)	1188/23,050 (5.2%)	0.948 (0.943 - 0.952)	0.936 (0.921 - 0.949)	0.832 (0.827 - 0.837)	0.233 (0.220 - 0.245)	0.996 (0.995 - 0.997)
HTN (–)	1375/29,737 (4.6%)	0.946 (0.941 - 0.951)	0.899 (0.882 - 0.914)	0.856 (0.852 - 0.861)	0.233 (0.222 - 0.245)	0.994 (0.993 - 0.995)
IHD[j] (+)	497/3098 (16.0%)	0.868 (0.852 - 0.883)	0.934 (0.910 - 0.955)	0.596 (0.577 - 0.614)	0.306 (0.283 - 0.330)	0.979 (0.972 - 0.986)
IHD (–)	2066/49,689 (4.2%)	0.950 (0.947 - 0.954)	0.912 (0.900 - 0.925)	0.860 (0.857 - 0.863)	0.220 (0.211 - 0.229)	0.996 (0.995 - 0.996)
CKD[k] (+)	326/3280 (9.9%)	0.906 (0.890 - 0.920)	0.926 (0.898 - 0.952)	0.706 (0.690 - 0.724)	0.258 (0.233 - 0.282)	0.989 (0.984 - 0.993)
CKD (–)	2237/49,507 (4.5%)	0.950 (0.946 - 0.953)	0.915 (0.903 - 0.925)	0.855 (0.852 - 0.858)	0.230 (0.220 - 0.238)	0.995 (0.995 - 0.996)
Sinus rhythm	2014/46,763 (4.3%)	0.956 (0.952 - 0.959)	0.910 (0.897 - 0.922)	0.877 (0.874 - 0.880)	0.250 (0.241 - 0.260)	0.995 (0.995 - 0.996)
AFIB[l]/AFL[m]	428/4963 (8.6%)	0.875 (0.858 - 0.890)	0.944 (0.920 - 0.965)	0.578 (0.564 - 0.592)	0.174 (0.158 - 0.190)	0.991 (0.987 - 0.994)
LBBB[n]	220/553 (39.8%)	0.859 (0.828 - 0.890)	0.991 (0.977 - 1.000)	0.219 (0.177 - 0.266)	0.456 (0.412 - 0.501)	0.973 (0.933 - 1.000)
RBBB[o]	199/3420 (5.8%)	0.927 (0.910 - 0.941)	0.910 (0.867 - 0.947)	0.803 (0.788 - 0.816)	0.222 (0.194 - 0.249)	0.993 (0.990 - 0.996)
PR >200 ms	156/3696 (4.2%)	0.935 (0.913 - 0.956)	0.878 (0.821 - 0.927)	0.870 (0.858 - 0.881)	0.229 (0.195 - 0.262)	0.994 (0.991 - 0.996)
QRS >140 ms	238/1317 (18.1%)	0.927 (0.910 - 0.942)	0.979 (0.961 - 0.996)	0.653 (0.623 - 0.682)	0.384 (0.348 - 0.424)	0.993 (0.987 - 0.999)
Prolonged QTc[p,q]	1566/13,518 (11.6%)	0.909 (0.902 - 0.916)	0.939 (0.927 - 0.952)	0.671 (0.663 - 0.680)	0.272 (0.260 - 0.285)	0.988 (0.986 - 0.991)

[a]LVSD: left ventricular systolic dysfunction.

[b]TTE-ECG: transthoracic echocardiography-electrocardiogram.

[c]All CIs (95%) are estimated from bootstrapping 1000 iterations.

[d]Values are presented as number of cases/total number of subjects (%).

[e]AUROC: area under the receiver operating curve.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]DM: diabetes mellitus.

[i]HTN: hypertension.

[j]IHD: ischemic heart disease.

XSL•FO
RenderX

[k]CKD: chronic kidney disease.

[l]AFIB: atrial fibrillation.

[m]AFL: atrial flutter.

[n]LBBB: left bundle branch block.

[o]RBBB: right bundle branch block.

[p]Prolonged QTc is defined as > 450 ms for men and > 470 ms for women after heart rate correction.

[q]QTc: heart-rate-corrected QT interval.

LVSD Detection in Follow-Up TTE-ECG Pairs

The recalibrated DeepECG LVSD models demonstrated high performance when validated on ECGs and LVSD labels from follow-up TTE-ECG pairs (ie, nonfirst pairs) in both hospitals A and B (Table 4**)**. The 12-lead model achieved an AUROC of 0.956 (95% CI 0.946 - 0.965) for internal validation and 0.940 (95% CI 0.936 - 0.945) for external validation, representing a modest improvement over the uncalibrated DeepECG LVSD model with an internal AUROC of 0.945 (95% CI 0.933 - 0.957) and an external AUROC of 0.910 (95% CI 0.905 - 0.916). Model recalibration without LVEF values showed a subtle performance gain over the uncalibrated model, yielding an internal AUROC of 0.950 (95% CI 0.939 - 0.960) and an external AUROC of 0.920 (95% CI 0.915 - 0.925). AUROC comparisons showed statistically significant differences between the recalibrated and uncalibrated DeepECG LVSD models among the external test sets (Figure 4). Also, at clinically relevant operating points, recalibrated models outperformed their uncalibrated counterparts. Specifically, when sensitivity was fixed at 90%, recalibration yielded higher specificity. Likewise, when specificity was fixed at 90%, recalibrated models achieved higher sensitivity (Table S1 and S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . LVSD[a] detection performance of the DeepECG LVSD model before and after recalibration, evaluated within the follow-up TTE-ECG[b] pairs of internal and external validation.

Performance		12-lead, value (95% CI)	6-lead, value (95% CI)	1-lead, value (95% CI)
Internal validation (Hospital A)				
DeepECG LVSD	AUROC[c]	0.945 (0.933 - 0.957)	0.910 (0.894 - 0.926)	0.909 (0.893 - 0.926)
	Sensitivity	0.914 (0.875 - 0.943)	0.848 (0.802 - 0.888)	0.862 (0.817 - 0.900)
	Specificity	0.852 (0.831 - 0.871)	0.801 (0.778 - 0.822)	0.799 (0.776 - 0.821)
	PPV[d]	0.600 (0.552 - 0.646)	0.508 (0.463 - 0.554)	0.510 (0.465 - 0.555)
	NPV[e]	0.976 (0.966 - 0.984)	0.956 (0.943 - 0.967)	0.960 (0.947 - 0.970)
DeepECG LVSD (recalibrated, masked to LVEF[f])	AUROC	0.950 (0.939 - 0.960)	0.918 (0.903 - 0.933)	0.918 (0.902 - 0.933)
	Sensitivity	0.903 (0.863 - 0.935)	0.893 (0.852 - 0.926)	0.928 (0.891 - 0.955)
	Specificity	0.855 (0.835 - 0.874)	0.794 (0.771 - 0.816)	0.756 (0.732 - 0.780)
	PPV	0.602 (0.555 - 0.649)	0.513 (0.468 - 0.557)	0.480 (0.438 - 0.523)
	NPV	0.973 (0.963 - 0.981)	0.968 (0.957 - 0.978)	0.977 (0.967 - 0.985)
DeepECG LVSD (recalibrated)	AUROC	0.956 (0.946 - 0.965)	0.942 (0.930 - 0.955)	0.939 (0.925 - 0.953)
	Sensitivity	0.828 (0.779 - 0.869)	0.869 (0.825 - 0.906)	0.924 (0.887 - 0.952)
	Specificity	0.912 (0.896 - 0.927)	0.854 (0.833 - 0.872)	0.814 (0.792 - 0.835)
	PPV	0.696 (0.644 - 0.744)	0.590 (0.542 - 0.637)	0.547 (0.502 - 0.592)
	NPV	0.956 (0.944 - 0.967)	0.964 (0.952 - 0.974)	0.978 (0.968 - 0.985)
External validation (Hospital B)				
DeepECG LVSD	AUROC	0.910 (0.905 - 0.916)	0.902 (0.896 - 0.907)	0.872 (0.866 - 0.879)
	Sensitivity	0.894 (0.880 - 0.906)	0.913 (0.901 - 0.924)	0.901 (0.888 - 0.913)
	Specificity	0.763 (0.757 - 0.770)	0.705 (0.697 - 0.712)	0.670 (0.662 - 0.677)
	PPV	0.378 (0.366 - 0.391)	0.332 (0.321 - 0.344)	0.305 (0.294 - 0.316)
	NPV	0.978 (0.975 - 0.981)	0.980 (0.978 - 0.983)	0.977 (0.974 - 0.979)
DeepECG LVSD (recalibrated, masked to LVEF)	AUROC	0.920 (0.915 - 0.925)	0.911 (0.905 - 0.916)	0.892 (0.886 - 0.898)
	Sensitivity	0.888 (0.875 - 0.901)	0.856 (0.842 - 0.870)	0.861 (0.846 - 0.874)
	Specificity	0.810 (0.804 - 0.817)	0.819 (0.812 - 0.825)	0.763 (0.756 - 0.770)
	PPV	0.430 (0.416 - 0.444)	0.432 (0.418 - 0.446)	0.369 (0.357 - 0.382)
	NPV	0.978 (0.976 - 0.981)	0.973 (0.970 - 0.975)	0.971 (0.968 - 0.974)
DeepECG LVSD (recalibrated)	AUROC	0.940 (0.936 - 0.945)	0.939 (0.935 - 0.944)	0.937 (0.932 - 0.941)
	Sensitivity	0.874 (0.860 - 0.887)	0.905 (0.892 - 0.916)	0.864 (0.849 - 0.877)
	Specificity	0.859 (0.854 - 0.865)	0.833 (0.827 - 0.839)	0.859 (0.853 - 0.865)
	PPV	0.500 (0.485 - 0.516)	0.466 (0.452 - 0.481)	0.497 (0.481 - 0.512)
	NPV	0.977 (0.974 - 0.979)	0.982 (0.979 - 0.984)	0.975 (0.972 - 0.978)

[a]LVSD: left ventricular systolic dysfunction.

[b]TTE-ECG: transthoracic echocardiography-electrocardiogram.

[c]AUROC: area under the receiver operating curve.

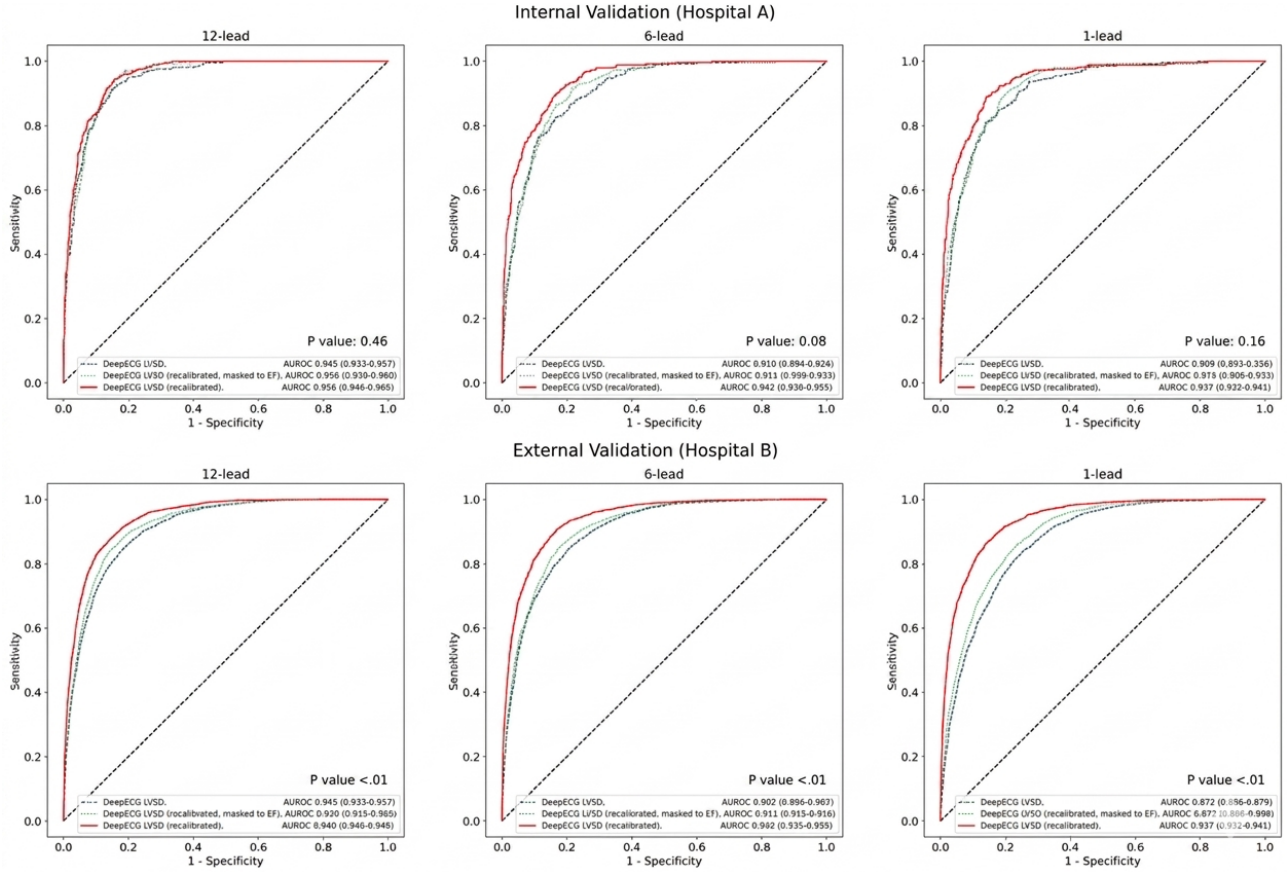
[d]PPV: positive predictive value.

[e]NPV: negative predictive value.

[f]LVEF: left ventricular ejection fraction.

XSL•FO
RenderX

Figure 4. Receiver operating characteristic curve for left ventricular systolic dysfunction detection performance of the DeepECG LVSD model, before and after recalibration, evaluated within the follow-up transthoracic echocardiography-electrocardiogram pairs. AUROC: area under the receiver operating characteristic; ECG: electrocardiogram; LVEF: left ventricular ejection fraction; LVSD: left ventricular systolic dysfunction.



Subgroup analysis revealed that all subpopulations consistently showed statistically significant differences between the AUROC of the recalibrated and uncalibrated standard 12-lead DeepECG LVSD models (Table 5). In these subgroups, the recalibrated models also demonstrated substantially higher specificity compared with the uncalibrated model when evaluated at each model's respective decision thresholds.

XSL•FO
RenderX

Table . Subgroup analysis and AUROC[a] comparisons for LVSD[b] detection performance of DeepECG LVSD before and after recalibration, evaluated within 12-lead, external, follow-up TTE-ECG[c] pairs.[d]

Subgroup		DeepECG LVSD (recalibrated)			DeepECG LVSD			
	n/N (%)[e]	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	*P* value[f]
Age ≤70	1245/10,002(12.5%)	0.947 (0.941 - 0.952)	0.802 (0.780 - 0.825)	0.920 (0.914 - 0.925)	0.921 (0.915 - 0.928)	0.928 (0.913 - 0.943)	0.733 (0.724 - 0.742)	.001
Age >70	1095/6862(16.0%)	0.931 (0.924 - 0.938)	0.767 (0.741 - 0.793)	0.914 (0.907 - 0.921)	0.894 (0.884 - 0.903)	0.923 (0.908 - 0.940)	0.666 (0.654 - 0.677)	<.001
Male	1404/9259(15.2%)	0.941 (0.936 - 0.947)	0.818 (0.797 - 0.839)	0.905 (0.898 - 0.910)	0.912 (0.905 - 0.919)	0.928 (0.914 - 0.942)	0.706 (0.696 - 0.716)	<.001
Female	933/7598(12.3%)	0.938 (0.931 - 0.945)	0.740 (0.712 - 0.766)	0.933 (0.926 - 0.938)	0.908 (0.900 - 0.915)	0.922 (0.905 - 0.939)	0.707 (0.696 - 0.718)	.002
DM[g] (+)	908/5061(17.9%)	0.933 (0.926 - 0.940)	0.779 (0.753 - 0.804)	0.902 (0.893 - 0.911)	0.903 (0.892 - 0.912)	0.941 (0.924 - 0.956)	0.664 (0.650 - 0.679)	<.001
DM (–)	1432/11,803(12.1%)	0.943 (0.937 - 0.948)	0.791 (0.769 - 0.811)	0.923 (0.918 - 0.928)	0.912 (0.905 - 0.919)	0.916 (0.902 - 0.931)	0.723 (0.714 - 0.733)	<.001
HTN[h] (+)	1248/8905(14.0%)	0.939 (0.933 - 0.945)	0.796 (0.774 - 0.816)	0.914 (0.907 - 0.920)	0.908 (0.900 - 0.915)	0.932 (0.918 - 0.945)	0.692 (0.681 - 0.702)	<.001
HTN (–)	1092/7959(13.7%)	0.942 (0.935 - 0.948)	0.775 (0.748 - 0.802)	0.922 (0.915 - 0.928)	0.913 (0.905 - 0.921)	0.918 (0.903 - 0.934)	0.723 (0.712 - 0.733)	<.001
IHD[i] (+)	552/2147(25.7%)	0.914 (0.902 - 0.927)	0.853 (0.822 - 0.884)	0.824 (0.806 - 0.845)	0.876 (0.861 - 0.890)	0.942 (0.922 - 0.961)	0.562 (0.539 - 0.586)	<.001
IHD (–)	1788/14,717(12.2%)	0.942 (0.937 - 0.947)	0.765 (0.744 - 0.785)	0.929 (0.924 - 0.933)	0.913 (0.907 - 0.919)	0.921 (0.908 - 0.933)	0.724 (0.716 - 0.731)	<.001
CKD[j] (+)	400/1917(20.9%)	0.911 (0.898 - 0.924)	0.790 (0.752 - 0.828)	0.866 (0.849 - 0.882)	0.872 (0.854 - 0.890)	0.955 (0.933 - 0.973)	0.545 (0.520 - 0.570)	.005
CKD (–)	1940/14,947(13.0%)	0.943 (0.939 - 0.948)	0.785 (0.766 - 0.803)	0.923 (0.919 - 0.928)	0.914 (0.908 - 0.920)	0.920 (0.907 - 0.932)	0.725 (0.717 - 0.732)	<.001
Sinus rhythm	1704/13,148(13.0%)	0.946 (0.941 - 0.950)	0.787 (0.768 - 0.807)	0.923 (0.918 - 0.928)	0.922 (0.916 - 0.928)	0.920 (0.907 - 0.932)	0.756 (0.748 - 0.764)	<.001
AFIB[k]/AFL[l]	513/3168(16.2%)	0.926 (0.915 - 0.937)	0.786 (0.750 - 0.822)	0.900 (0.888 - 0.911)	0.869 (0.854 - 0.885)	0.936 (0.914 - 0.956)	0.518 (0.500 - 0.537)	.001
LBBB[m]	247/470(52.6%)	0.872 (0.841 - 0.901)	0.947 (0.917 - 0.973)	0.614 (0.554 - 0.680)	0.819 (0.781 - 0.855)	0.988 (0.972 - 1.000)	0.184 (0.136 - 0.238)	.03
RBBB[n]	205/1440(14.2%)	0.916 (0.900 - 0.933)	0.785 (0.727 - 0.839)	0.881 (0.863 - 0.900)	0.853 (0.829 - 0.877)	0.878 (0.835 - 0.921)	0.669 (0.644 - 0.696)	.01
PR > 200	219/1444(15.2%)	0.944 (0.930 - 0.955)	0.813 (0.759 - 0.863)	0.902 (0.885 - 0.919)	0.908 (0.890 - 0.924)	0.909 (0.867 - 0.944)	0.749 (0.725 - 0.773)	.04

XSL•FO
RenderX

Subgroup		DeepECG LVSD (recalibrated)			DeepECG LVSD			
	n/N (%)[e]	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	*P* value[f]
QRS >140	285/770(37.0%)	0.916 (0.894 - 0.934)	0.926 (0.893 - 0.956)	0.755 (0.713 - 0.791)	0.860 (0.830 - 0.885)	0.954 (0.927 - 0.978)	0.470 (0.427 - 0.513)	.002
Prolonged QTc[o,p]	151/2670(24.1%)	0.911 (0.904 - 0.919)	0.814 (0.795 - 0.833)	0.852 (0.841 - 0.862)	0.869 (0.858 - 0.878)	0.935 (0.922 - 0.947)	0.538 (0.525 - 0.552)	<.001

[a]AUROC: area under the receiver operating curve.

[b]LVSD: left ventricular systolic dysfunction.

[c]TTE-ECG: transthoracic echocardiography-electrocardiogram.

[d]All CIs (95%) are estimated from bootstrapping 1000 iterations.

[e]Values are presented as number of cases/total number of subjects (%).

[f]*P* value for the DeLong test between the recalibrated DeepECG LVSD model and the uncalibrated DeepECG LVSD model is calculated.

[g]DM: diabetes mellitus.

[h]HTN: hypertension.

[i]IHD: ischemic heart disease.

[j]CKD: chronic kidney disease.

[k]AFIB: atrial fibrillation.

[l]AFL: atrial flutter.

[m]LBBB: left bundle branch block.

[n]RBBB: right bundle branch block.

[o]Prolonged QTc is defined as >450 ms for men and >470 ms for women after heart rate correction.

[p]QTc: heart-rate-corrected QT interval.

Discussion

Principal Findings

In this multicenter retrospective analysis, we developed and validated DeepECG LVSD models capable of detecting LVSD from routine ECGs and achieved state-of-the-art performance across both internal and external cohorts. This resulted from a 3-stage training framework, including an initial pretraining phase using a large-scale, automatically labeled ECG dataset, and a final recalibration phase that adjusts model predictions based on prior LVEF measurements and previous AI-ECG outputs.

Clinical Covariates in AI-ECG

Certain comorbidities, such as AFIB [6,10,14,15] or LBBB [10,15], can adversely affect the diagnostic accuracy of AI-ECG models. In line with prior findings, our analysis also showed declines in AUROC within these subgroups. However, the decline in specificity was considerably more pronounced than would be anticipated from AUROC alone. Specifically, patients without LVSD but with significant comorbidities frequently generated elevated model scores, increasing the false-positive rate. This pattern is consistent with the findings [10,15]. Harmon et al [10] have reported substantial decreases in AUROC and specificity in patients with AFIB (AUROC 0.826) or LBBB (AUROC 0.791) compared to the general population (AUROC 0.903), with specificity plummeting as low as 0.745 for AFIB and 0.364 for LBBB. Interestingly, while Harmon et al [10] recommended threshold adjustment, their suggestion was limited to patients with entirely normal ECGs rather than to broader subpopulations.

These performance drops can be explained by confounding variables. A well-known example in the AI literature is the effect of age as a confounder in models diagnosing HIV from magnetic resonance imaging [21]; age correlates with both imaging features and disease prevalence, thus influencing model predictions. Similarly, LBBB [22] and AFIB [23] are established risk factors for HF, and their correlation with LVSD is evident at the population level. As a result, models may exhibit a higher likelihood of positive classification among patients with such comorbidities, driven by population-level correlations. Under external evaluation or individual application, where the relationship between the confounder (eg, LBBB or AFIB) and target condition (LVSD) may not hold, the model may show decreased performance. Although our recalibration methods have empirically helped circumvent these effects, the direct means to mitigate confounding factors remain an open area of investigation in the AI-ECG literature.

Moreover, the AUROC primarily quantifies discrimination based on continuous risk scores and reflects rank-based information, which does not necessarily translate into clinical utility at predetermined decision thresholds. Thus, the subgroup AUROC alone may mask clinically significant degradations in binary performance metrics, such as specificity and sensitivity, which are directly relevant to clinical application. We suggest that future studies include comprehensive subgroup analyses that incorporate threshold-based performance metrics rather than stratified AUROC. Although current guidelines recommend subgroup analysis, the choice of performance metrics to report

XSL•FO
RenderX

is ambiguous [24-26]; clearer reporting of clinically actionable, threshold-specific metrics could improve their relevance to real-world applications.

Model Recalibration and Clinical Utility

To address these challenges, we introduced a patient-specific recalibration strategy that leverages prior LVEF values and historical DeepECG LVSD predictions. Rather than modifying model architecture or incorporating comorbidity information as direct inputs, recalibration adjusts patient-level risk estimates while preserving workflow simplicity. This strategy was designed to stabilize model behavior at a fixed decision threshold, not by altering the threshold itself. By recalibrating risk estimates at the individual level, the model reduces the tendency to overestimate predicted LVSD risk in certain clinical conditions, thereby allowing a single validation-derived decision threshold to retain consistent meaning across heterogeneous patient populations and mitigating excessive false-positive predictions when applied uniformly.

However, this benefit comes with the requirement of actual LVEF measurements as reference values, necessitating echocardiography performed by experts, which may limit the widespread applicability of this model. To extend the utility of our approach, we developed a second recalibrated model masked to the LVEF, which also demonstrated subtle performance gains compared to the uncalibrated model. This model operates solely on serial ECG data and does not require LVEF measurements from echocardiography for inference, potentially broadening its applicability to settings where access to echocardiography is limited.

Model Pretraining for Classification Performance

Pretraining has become a standard approach for enhancing deep learning models, particularly when large datasets are accessible. This strategy has been extensively explored in the AI-ECG field using various data formats and pretraining objectives [27-31]. In this study, pretraining was implemented using a transfer learning [32,33] framework, leveraging labels that were directly extracted from ECG analysis systems. These include automated diagnostic statements and waveform-derived features, which eliminate the need for labor-intensive manual annotations. During pretraining, the model was tasked with predicting these multiple automatically generated labels, enabling it to learn broadly relevant granular ECG representations.

In the subsequent supervised fine-tuning stage, these low-level features were refined while training on a smaller outcome-specific dataset. This is in contrast to training from scratch, in which models often fail to capture subtle or infrequent ECG patterns owing to the limited volume of supervised learning. As a result of incorporating pretraining, our model demonstrated improved performance, with an internal AUROC increase of 0.014 and an external AUROC increase of 0.010, compared to models trained solely on the fine-tuning set.

The capacity for pretraining to use massive datasets should be emphasized. In this study, the pretraining phase encompassed 397,644 ECGs from 191,121 patients, which was nearly 10 times the size of the fine-tuning set (31,099 ECGs from 31,099 patients). This approach allows the model to benefit from large-scale data, even if the pretraining labels are noisy or imprecise, or in certain pretraining methods that are entirely unlabeled. Hence, pretraining enables the model to develop a more generalizable feature extractor that can improve downstream performance across various clinical populations.

LVSD Detection and AI-ECG

LVSD remains a critical precursor of overt HF and is associated with significant morbidity and mortality. Early detection facilitates the timely initiation of guideline-directed therapy and improves patient outcomes [3]. However, the high resource intensity of echocardiography precludes its use as a universal screening tool. AI-ECG models, such as those developed in this study, present a scalable, low-cost alternative suitable for broader implementation. Our model achieved a strong performance in both standard (12-lead) and reduced-lead configurations, further supporting its adaptability to various ECG acquisition modalities, including emerging handheld and wearable devices.

Our study expands the application of AI-ECG from a screening tool to a longitudinal monitoring tool for at-risk populations. Our model achieved an enhanced diagnostic performance compared to previous studies in at-risk patient groups, such as those with AFIB. In an external validation study of the AI-based identification of LVSD, König et al [14] have reported an AUROC of 0.815 for the AFIB subgroup, which was lower than that of the general cohort. Similarly, in a study by Harmon et al [10], the AUROC for the AFIB subgroup decreased to 0.826, suggesting that separate algorithms may be required for patients with marked electrical abnormalities. Our study provides evidence that our approaches may yield improved diagnostic accuracy in such subgroups [10]. Moreover, sensitivity and specificity also showed minimal variation regardless of comorbid conditions, potentially expanding the use of AI-ECG–based LVSD detection to clinical presentations with a higher burden of comorbidities.

Cardiac function evaluation is important for the timely detection and management of subclinical or progressive cardiac dysfunction. A 2019 report on the appropriate imaging criteria for the assessment of cardiac function in nonvalvular heart disease [34] approved the use of TTE in the routine evaluation of systemic hypertension and the reevaluation of patients previously or currently undergoing cardiotoxic therapy. Furthermore, the recent 2023 Update on European Society of Cardiology Heart Failure Guidelines [35] and analysis from the STRONG-HF [36] trial highlight the importance of frequent and careful follow-up during the intensive titration of medical therapy in the postdischarge period for patients with acute HF. In this context, our AI-ECG algorithm may serve not only as a screening tool to identify cardiac dysfunction but also as a longitudinal monitoring tool to detect signs of clinical deterioration during follow-up in individual patients.

Clinical Integration and Future Directions

From a clinical perspective, the proposed DeepECG LVSD model could be integrated into routine practice as a background screening or decision-support tool embedded within standard ECG workflows. Similar to prior AI-ECG implementations, the

XSL•FO
RenderX

model could operate silently at the time of ECG acquisition, flagging patients at elevated risk for LVSD and prompting targeted referral for confirmatory echocardiography. Such an approach may be particularly valuable in outpatient settings or among patients with multiple comorbidities, where LVSD is frequently underdiagnosed. Nevertheless, several barriers to real-world adoption remain. Workflow integration will require careful consideration to minimize alert fatigue and ensure that model outputs are presented in a clinically actionable manner. Future work should focus on prospective clinical studies to evaluate the impact of AI-assisted ECG screening on downstream testing, clinical decision-making, and patient outcomes. Additional research is also needed to optimize operating thresholds tailored to specific clinical use cases and to assess regulatory, cost, and reimbursement considerations.

Limitations

This study had some limitations. First, this study was conducted in only 2 tertiary care centers in Korea; further external validation in more diverse populations is essential to extend its generalizability. Second, its retrospective design, along with the use of paired ECG and echocardiography data, may limit generalizability and introduce potential selection bias, highlighting the need for future prospective studies to establish real-world utility and workflow integration. Third, this study did not evaluate the association between the AI model outputs and clinical outcomes. Although the model demonstrated strong technical performance, its prognostic value and clinical utility remain to be validated in real-world patient outcomes. Fourth, although our recalibration method empirically demonstrated benefits in the model performance, we were unable to address the specific considerations related to the best implementation. The selection of the model architecture and optimal format of historical data as auxiliary inputs remains an open research question. Also, although the masked-LVEF approach reduces reliance on prior echocardiographic data, it remains an early-stage methodology with limited clinical validation. Fifth, ECGs and TTE measurements were not obtained simultaneously, and the temporal gap between acquisitions may introduce inaccuracies in the performance metrics. Prospective data collection using simultaneous ECG and TTE measurements may provide more reliable assessments in future studies. Finally, as with many AI-based models, the algorithm operates as a “black box” that limits the interpretability and transparent understanding of the decision-making process.

Conclusion

The AI-ECG model for LVSD detection demonstrated a strong performance through pretraining on a large dataset and a novel recalibration approach. By addressing both diagnostic performance and real-world clinical applications, our work pursues the practical integration of AI-driven diagnostics into cardiovascular screening and highlights the need for future prospective, randomized studies to confirm the clinical benefits.

Acknowledgments

We would like to thank Editage for English language editing.

Funding

The authors declared no financial support was received for this work.

Data Availability

Deidentified data may be available from the corresponding author upon reasonable request, subject to institutional review board approval and data use agreements.

Authors' Contributions

YJP, SL, JWS, and SAK conceptualized and designed the study. YJP, JWS, MSA, SJL, SJH, and SAK collected and managed the clinical ECG database. TJ, YN, SJ, and MC contributed to data curation and preprocessing. YN and TJ developed the data analysis workflow and conducted statistical analysis. SJ and YL conducted independent validation analyses. YJP, SL, and TJ drafted the manuscript. SJH, MC, SAK, and HJA critically reviewed and revised the manuscript. All authors read and approved the final version of the manuscript for publication.

Conflicts of Interest

SL, TJ, YN, SJ, HJA, MC, and YL are employees of VUNO Inc. and hold equity in the company. All other authors declare no conflicts of interest related to the contents of this paper.

Multimedia Appendix 1
Diagnostic performance metrics (sensitivity, specificity, positive predictive value, and negative predictive value) of DeepECG models at fixed thresholds.
[DOCX File, 26 KB - medinform_v14i1e83127_app1.docx]

References

XSL•FO
RenderX

1. Echouffo-Tcheugui JB, Erqou S, Butler J, Yancy CW, Fonarow GC. Assessing the risk of progression from asymptomatic left ventricular dysfunction to overt heart failure: a systematic overview and meta-analysis. JACC Heart Fail 2016 Apr;4(4):237-248. [doi: 10.1016/j.jchf.2015.09.015] [Medline: 26682794]
2. Yeboah J, Rodriguez CJ, Stacey B, et al. Prognosis of individuals with asymptomatic left ventricular systolic dysfunction in the multi-ethnic study of atherosclerosis (MESA). Circulation 2012 Dec 4;126(23):2713-2719. [doi: 10.1161/CIRCULATIONAHA.112.112201] [Medline: 23124035]
3. SOLVD Investigators, Yusuf S, Pitt B, Davis CE, Hood WB Jr, Cohn JN. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. N Engl J Med 1992 Sep 3;327(10):685-691. [doi: 10.1056/NEJM199209033271003] [Medline: 1463530]
4. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019 Jan;25(1):70-74. [doi: 10.1038/s41591-018-0240-2] [Medline: 30617318]
5. Cho J, Lee B, Kwon JM, et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. ASAIO J 2021 Mar 1;67(3):314-321. [doi: 10.1097/MAT.0000000000001218] [Medline: 33627606]
6. Choi J, Lee S, Chang M, Lee Y, Oh GC, Lee HY. Deep learning of ECG waveforms for diagnosis of heart failure with a reduced left ventricular ejection fraction. Sci Rep 2022 Aug 20;12(1):14235. [doi: 10.1038/s41598-022-18640-8] [Medline: 35987961]
7. Vaid A, Johnson KW, Badgeley MA, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. JACC Cardiovasc Imaging 2022 Mar;15(3):395-410. [doi: 10.1016/j.jcmg.2021.08.004] [Medline: 34656465]
8. Lim J, Lee HS, Han GI, et al. Artificial intelligence-enhanced six-lead portable electrocardiogram device for detecting left ventricular systolic dysfunction: a prospective single-centre cohort study. Eur Heart J Digit Health 2025 May;6(3):476-485. [doi: 10.1093/ehjdh/ztaf025] [Medline: 40395421]
9. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. J Cardiovasc Electrophysiol 2019 May;30(5):668-674. [doi: 10.1111/jce.13889] [Medline: 30821035]
10. Harmon DM, Carter RE, Cohen-Shelly M, et al. Real-world performance, long-term efficacy, and absence of bias in the artificial intelligence enhanced electrocardiogram to detect left ventricular systolic dysfunction. Eur Heart J Digit Health 2022 Jun;3(2):238-244. [doi: 10.1093/ehjdh/ztac028] [Medline: 36247412]
11. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021 May;27(5):815-819. [doi: 10.1038/s41591-021-01335-4] [Medline: 33958795]
12. Hunt SA, Abraham WT, Chin MH, et al. 2009 focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation. Circulation 2009 Apr 14;119(14):e391-e479. [doi: 10.1161/CIRCULATIONAHA.109.192065] [Medline: 19324966]
13. MEMBERS WC, Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, et al. ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). Circulation 2005;112(12):1825-1852. [doi: 10.1161/CIRCULATIONAHA.105.167587]
14. König S, Hohenstein S, Nitsche A, et al. Artificial intelligence-based identification of left ventricular systolic dysfunction from 12-lead electrocardiograms: external validation and advanced application of an existing model. Eur Heart J Digit Health 2024 Mar;5(2):144-151. [doi: 10.1093/ehjdh/ztad081] [Medline: 38505486]
15. Sangha V, Nargesi AA, Dhingra LS, et al. Detection of left ventricular systolic dysfunction from electrocardiographic images. Circulation 2023 Aug 29;148(9):765-777. [doi: 10.1161/CIRCULATIONAHA.122.062646] [Medline: 37489538]
16. Hu J, Shen L, Sun G, editors. Squeeze-and-excitation networks. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; Jun 18-23, 2018. [doi: 10.1109/CVPR.2018.00745]
17. Xie S, Girshick R, Dollár P, Tu Z, editors. Aggregated residual transformations for deep neural networks. Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017. [doi: 10.1109/CVPR.2017.634]
18. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
20. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-413. [doi: 10.1093/biomet/26.4.404]
21. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. Nat Commun 2020 Nov 26;11(1):6010. [doi: 10.1038/s41467-020-19784-9] [Medline: 33243992]

XSL•FO
RenderX

22. Schneider JF, Thomas HE Jr, Kreger BE, McNamara PM, Kannel WB. Newly acquired left bundle-branch block: the Framingham study. Ann Intern Med 1979 Mar;90(3):303-310. [doi: 10.7326/0003-4819-90-3-303] [Medline: 154870]
23. Santhanakrishnan R, Wang N, Larson MG, et al. Atrial fibrillation begets heart failure and vice versa: temporal associations and differences in preserved versus reduced ejection fraction. Circulation 2016 Feb 2;133(5):484-492. [doi: 10.1161/CIRCULATIONAHA.115.018614] [Medline: 26746177]
24. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020 Oct;2(10):e537-e548. [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
25. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020 Oct;2(10):e549-e560. [doi: 10.1016/S2589-7500(20)30219-3] [Medline: 33015597]
26. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
27. Yang S, Lian C, Zeng Z. Masked autoencoder for ECG representation learning. In: Yang S, Lian C, Zeng Z, editors. Presented at: 2022 12th International Conference on Information Science and Technology (ICIST); Oct 14-16, 2022. [doi: 10.1109/ICIST55546.2022.9926900]
28. Na Y, Park M, Tae Y, Joo S. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. arXiv. Preprint posted online on Mar 19, 2024. [doi: 10.48550/arXiv.2402.09450]
29. Vaid A, Jiang J, Sawant A, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. NPJ Digit Med 2023 Jun 6;6(1):108. [doi: 10.1038/s41746-023-00840-9] [Medline: 37280346]
30. Friedman SF, Khurshid S, Venn RA, et al. Unsupervised deep learning of electrocardiograms enables scalable human disease profiling. NPJ Digit Med 2025 Jan 12;8(1):23. [doi: 10.1038/s41746-024-01418-9] [Medline: 39799251]
31. Sau A, Pastika L, Sieliwonczyk E, et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. Lancet Digit Health 2024 Nov;6(11):e791-e802. [doi: 10.1016/S2589-7500(24)00172-9] [Medline: 39455192]
32. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. Presented at: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks; Oct 4-7, 2018 URL: https://arxiv.org/abs/1808.01974 [accessed 2026-03-31]
33. Pal A, Srivastva R, Singh YN. CardioNet: an efficient ECG arrhythmia classification system using transfer learning. Big Data Research 2021 Nov;26:100271. [doi: 10.1016/j.bdr.2021.100271]
34. Doherty JU, Kort S, Mehran R, et al. ACC/AATS/AHA/ASE/ASNC/HRS/SCAI/SCCT/SCMR/STS 2019 appropriate use criteria for multimodality imaging in the assessment of cardiac structure and function in nonvalvular heart disease: a report of the American College of Cardiology Appropriate Use Criteria Task Force, American Association for Thoracic Surgery, American Heart Association, American Society of Echocardiography, American Society of Nuclear Cardiology, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, Society for Cardiovascular Magnetic Resonance, and the Society of Thoracic Surgeons. J Am Coll Cardiol 2019 Feb 5;73(4):488-516. [doi: 10.1016/j.jacc.2018.10.038] [Medline: 30630640]
35. Authors/Task Force Members, McDonagh TA, Metra M, et al. 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2024 Jan;26(1):5-17. [doi: 10.1002/ejhf.3024] [Medline: 38169072]
36. Mebazaa A, Davison B, Chioncel O, et al. Safety, tolerability and efficacy of up-titration of guideline-directed medical therapies for acute heart failure (STRONG-HF): a multinational, open-label, randomised, trial. Lancet 2022 Dec 3;400(10367):1938-1952. [doi: 10.1016/S0140-6736(22)02076-1] [Medline: 36356631]

Abbreviations

AFIB: atrial fibrillation
AI: artificial intelligence
AI-ECG: artificial intelligence-augmented electrocardiography
AUROC: area under the receiver operating curve
ECG: electrocardiography
HF: heart failure
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
LBBB: left bundle branch block
LVEF: left ventricular ejection fraction
LVSD: left ventricular systolic dysfunction
NPV: negative predictive value

XSL•FO
RenderX

PPV: positive predictive value
TTE: transthoracic echocardiography
TTE-ECG: transthoracic echocardiography-electrocardiogram

Edited by J Klann; submitted 28.Aug.2025; peer-reviewed by A Sahajan, D Parry; revised version received 17.Feb.2026; accepted 18.Feb.2026; published 24.Apr.2026.

Please cite as:
Lee S, Son JW, Kim SA, Ahn MS, Lee SJ, Han SJ, Joo T, Na Y, Joo S, Ahn HJ, Chang M, Lee Y, Park YJ
Deep Learning Model Using Transfer Learning for Detecting Left Ventricular Systolic Dysfunction: Retrospective Algorithm Development and Validation Study
JMIR Med Inform 2026;14:e83127
URL: https://medinform.jmir.org/2026/1/e83127
doi:10.2196/83127

© Sungjae Lee, Jung-Woo Son, Sung-Ai Kim, Min-Soo Ahn, Sang Jun Lee, Sang-Jin Han, Taehyun Joo, Yeongyeon Na, Sunghoon Joo, Hyun Jin Ahn, Mineok Chang, Yeha Lee, Young Jun Park. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Classification of Cochrane Plain Language Summaries by Conclusiveness Using Transformer-Based Models and ChatGPT: Retrospective Observational Study

Antonija Mijatović[1], PhD; Luka Ursić[1], PhD; Nensi Bralić[1], PhD; Ružica Bandić[1], PhD; Barbara Ćaćić[2]; Ivan Buljan[3], PhD; Ana Marušić[1], MD, PhD

[1]Department of Research in Biomedicine and Health, Centre for Evidence-based Medicine, University of Split School of Medicine, Šoltanska 2A, Split, Croatia

[2]Faculty of Humanities and Social Sciences, University of Split, Split, Croatia

[3]Department of Psychology, Faculty of Humanities and Social Sciences, University of Split, Split, Croatia

Corresponding Author:
Antonija Mijatović, PhD
Department of Research in Biomedicine and Health, Centre for Evidence-based Medicine, University of Split School of Medicine, Šoltanska 2A, Split, Croatia

Abstract

Background: Cochrane plain language summaries (PLSs) aim to make systematic review findings more accessible to the general public. However, inconsistencies in how conclusions are presented may impact comprehension and decision-making. Classifying PLSs based on conclusiveness can improve clarity and facilitate informed health decisions.

Objective: This study aimed to develop and evaluate deep learning language models for the classification of PLSs according to 3 levels of conclusiveness (conclusive, inconclusive, and unclear) and to compare their performance with a general-purpose large language model (GPT-4o).

Methods: We used a publicly available dataset containing 4405 Cochrane PLSs of systematic reviews published until 2019, already classified by humans according to 9 categories of conclusiveness regarding the intervention's effectiveness or safety. We merged these categories into 3 classes based on the strength of conclusiveness: conclusive, inconclusive, and unclear. For the fine-tuning, we used Scientific Bidirectional Encoder Representations from Transformers (SciBERT), a pretrained language model trained on 1.14 million papers primarily from the health sciences, and Longformer, a transformer model designed specifically to process long documents. The script was developed using the Python programming language and the PyTorch framework. We computed evaluation metrics using the *scikit-learn* machine learning library and determined the area under the curve of the receiver operating characteristic (AUCROC) to measure the model performance in balancing sensitivity and specificity. We also analyzed a separate set of 213 PLSs and compared the predictions of our pretrained models with both manual verification and outputs generated by ChatGPT.

Results: The model based on SciBERT achieved a balanced accuracy of 56.6%. The AUCROC was 0.91 for "conclusive," 0.67 for "inconclusive," and 0.75 for "unclear" conclusiveness classes. The Longformer-based model had a balanced accuracy of 60.9%, with AUCROCs of 0.86 for "conclusive," 0.67 for "inconclusive," and 0.72 for "unclear" conclusiveness classes. Both models underperformed compared with ChatGPT, which demonstrated higher accuracy (74.2%), better precision and recall, and a higher Cohen κ (0.57).

Conclusions: Fine-tuning 2 transformer-based language models showed mixed results in classifying Cochrane PLSs by conclusiveness, likely due to semantic overlap and subtle linguistic differences. Despite satisfactory internal test metrics, the fine-tuned models failed to generalize to newly published PLSs, where performance dropped to near-chance levels. These findings suggest that general-purpose large language models like GPT-4o may currently offer more reliable results for practical classification tasks in biomedical applications.

(JMIR Med Inform 2026;14:e72657) doi:10.2196/72657

KEYWORDS

plain language summary; PLS; large language models; Scientific Bidirectional Encoder Representations from Transformers; SciBERT; Longformer; fine-tuning

XSL•FO
RenderX

Introduction

A Cochrane plain language summary (PLS) is a stand-alone summary of a Cochrane systematic review used to disseminate health-related evidence to a wider audience with the goal of facilitating evidence-based decision-making about health care, particularly for medical treatments [1]. A well-written PLS should be comprehensible to readers without a background in research or health care, including patients, caregivers, and policymakers [2], and should be presented at or below a sixth-grade reading level to ensure accessibility for all readers [3]. It should also allow readers to comprehend the certainty of evidence and to correctly interpret the results, which is why the authors should not offer specific recommendations but rather present their findings clearly and guide the readers toward independent conclusions [4,5]. However, a PLS necessarily has a conclusion section conveying the main message, where the level of certainty of the evidence is presented using narrative statements [2]. For example, "Intervention causes a large reduction/increase in outcome" is a suggested narrative for large effect size and high certainty of the evidence, whereas "It is unclear if intervention has an effect on outcome" should indicate very low certainty [6,7].

Conclusiveness is an important concept in research and health care, indicating a degree of confidence in the findings and facilitating decision-making [8,9], while also ensuring that the current evidence is easily understood [10]. When provided with conclusive health information, patients rely less on health care professionals to decide on diagnosis and therapy [11,12]. Unfortunately, they do not always succeed in finding relevant information for their health condition, as shown in a study where a quarter of respondents did not find answers to the health-related inquiries they raised on the internet [13]. Patients also often find information in unreliable sources, providing misleading or false data on diagnostics and treatment [14]. In addition, numerous studies have reported higher levels of anxiety and cyberchondria with increases in online health information searches [15-17]. Given that Cochrane Reviews represent the best available knowledge in the field [1], the accurate classification of their PLSs could significantly improve patients' comprehension of these conclusions and allow them to make well-informed decisions about health care interventions.

Several studies and reviews on the conclusiveness of Cochrane PLSs found that 50% to 80% of the reviews enabled readers to reach a relevant conclusion, while the readability of PLSs was generally poor, with conclusions often unclear or missing [18-22]. However, in some cases, conclusive statements were made even when the quality of evidence was low or moderate [23]. Our previous research showed that most PLSs lacked or had unclear conclusions regarding an intervention's efficacy and safety [24].

In all of the studies on the conclusiveness of systematic reviews, the process of classifying the reviews and PLSs was carried out manually, usually by at least 2 independent assessors, which is a demanding and time-consuming task. In this study, we explored whether the classification of PLSs according to their level of conclusiveness could be conducted with the help of artificial intelligence (AI) and natural language processing (NLP). NLP algorithms, particularly deep learning models like neural networks, can automatically learn and extract patterns from language data [25], enabling them to understand context and semantics [26,27]. Their multilingual variants continue to expand and become accessible to speakers of less-represented languages [28] while domain-specific models lead to greater accuracy [29].

Methods

Overview

In this retrospective observational study with a supervised machine learning approach, we used a pretrained deep large language model (LLM) for PLS classification according to 3 levels of conclusiveness: conclusive, inconclusive, and unclear. We fine-tuned 2 transformer-based models—Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer—for our task. SciBERT is a variant of the Bidirectional Encoder Representations from Transformers (BERT) model specifically designed for scientific and biomedical text processing that is pretrained on a vast corpus of scientific literature consisting of 18% of papers from computer science and 82% from the biomedical field [30]. Longformer is a transformer architecture optimized for processing long documents through sparse attention mechanisms [31]. SciBERT was selected to leverage the domain-specific language of PLSs, whereas Longformer was selected to accommodate PLSs with extended length. Specifically, the median number of words in a PLS is 345 [32], which corresponds to approximately 500 tokens, just under SciBERT's 512-token limit. However, the Longformer model's extended token capacity of up to 4096 tokens allows for the processing of all PLSs without truncation.

Data Source

We used the dataset from our previous study [24], which contains 4405 Cochrane PLSs of systematic reviews on intervention studies published until 2019, already classified by 2 independent experts into 9 categories based on the conclusiveness regarding an intervention's effectiveness or safety. We combined these categories into 3 distinct classes: conclusive, inconclusive, and unclear, allowing for a more manageable and interpretable classification task (Textbox 1).

XSL•FO
RenderX

Textbox 1. Classification of conclusiveness categories.

Conclusive

- Positive: signifies the existence of moderate- or high-quality evidence supporting the effectiveness or safety
- Negative: indicates the presence of moderate- or high-quality evidence of intervention's ineffectiveness or harm
- Equal: denotes that the analyzed interventions were of equal effectiveness or safety

Inconclusive

- Positive inconclusive: implies the existence of evidence supporting effectiveness or safety, yet the evidence is low quality or inconclusive. The authors suggest that more research is needed.
- Negative inconclusive: suggests there is evidence of ineffectiveness or harm (indicating that the observed effect or the intervention was unsafe), yet the evidence is low quality or inconclusive. Authors may advise against the intervention or comparison and state that more research is required.
- Equal inconclusive: indicates that the interventions exhibit comparable levels of effectiveness or safety, yet the evidence is low quality or inconclusive. The authors emphasize that more research is required.

Unclear

- No opinion: the authors provided no opinion.
- No evidence: there is no evidence from randomized controlled trials because the literature search did not result in any eligible studies (ie, empty reviews).
- Unclear: the authors did not present clear conclusions.

With the classification, the evidence in the "conclusive" class is strong and clear, irrespective of the direction of the effect, as opposed to the "inconclusive" class, where it is uncertain or of lower quality. In the "unclear" class, conclusions are absent, either because authors have not provided a clear opinion or due to a lack of available evidence. This lack of conclusiveness is not indicative of a poorly written PLS, as long as the PLS accurately represents the findings from the systematic review. This is why we must differentiate between PLSs that conclude that there is "no evidence" and those that offer no opinion or present unclear conclusions [24].

Language Processing Models

LLMs are important components of NLP designed to understand and generate human language. LLMs such as GPT-3 and BERT are pretrained on massive datasets containing text from the internet [33,34]. Most importantly, they are highly adaptable, meaning they can be fine-tuned for specific tasks [35]. For example, Beltagy et al [30] fine-tuned BERT, an LLM that had been pretrained on a wide range of text on the internet [33], to develop SciBERT, an LLM trained on a vast corpus of scientific literature, primarily from biomedical and life sciences, making it suitable for NLP tasks in the scientific and medical research domains. Similarly, Longformer was developed to address the limitations of handling long documents; it uses a sparse attention mechanism, where each token focuses on a limited local context rather than the entire sentence [31].

We achieved transfer learning by further fine-tuning SciBERT and Longformer on our specific PLS classification task. In transfer learning, the LLM adapts its learned features to the nuances of the new task while retaining the knowledge it acquired during pretraining [36]. This approach is intended to mirror how humans learn, as we often apply knowledge and skills acquired in one context to solve new, related problems [37]. Transfer learning not only speeds up the training process but also leads to better performance compared with training from scratch [38].

Experimental Setup and Fine-Tuning

We wrote the script using the Python programming language (version 3.12.3; Python Software Foundation) with the help of the PyTorch framework [39] and executed it within the Jupyter Notebook environment [40] using the NVIDIA GeForce RTX 3080 GPU (version 8200).

Both of our models came from the Hugging Face library [41]. For SciBERT, we used its associated tokenizer, setting the maximum token length to 512. For Longformer, we extended the maximum token length to 2048 to accommodate the full content of the PLSs without truncation. Both models included a dropout layer with a rate of 0.5, a regularization technique that reduces the risk of overfitting.

For both models, we used a 768-dimensional pooled embedding vector as input to our classifier. This representation was passed through a dropout layer and a linear layer that produced a 3D output corresponding to our target classes. SciBERT used its built-in pooled ([CLS] token) representation, whereas Longformer used mean pooling across all token embeddings due to the absence of a pooler layer. A standard attention mask was applied during encoding, ensuring that padding tokens were fully excluded from self-attention computations. We used AdamW as the optimizer to update network weights [42] and categorical cross entropy as our loss function [43]. We set the maximum number of training epochs (where 1 epoch represents a complete pass through the training dataset) to 15 for SciBERT and 10 for Longformer, with early stopping based on validation loss. In practice, SciBERT training stopped after 7 epochs due to early stopping. The learning rate was set to 2×10^{5} for both the models. Since these hyperparameters cannot be determined a priori, they were selected via trial and error [44]. For this reason, we monitored training and validation performance and

XSL•FO
RenderX

used early stopping based on validation loss to prevent overfitting. The best-performing models (lowest validation loss) were saved and later used for evaluation.

Baseline models (SciBERT and Longformer) were implemented using a frozen feature-extraction transfer-learning approach. All pretrained transformer encoder weights were frozen, and only the newly added linear classification layer was trainable. The encoder itself did not undergo any gradient updates. Therefore, this baseline represents a lightweight transfer-learning model.

Data Splitting and Handling of Class Imbalance

We divided the dataset into training (80%), testing (10%), and validation (10%) subsets. Random undersampling was applied only to the training set, where the smallest class contained 343 PLSs. All classes were downsampled to this size to create a balanced training set.

Model Validation

To assess performance, we used functions from *scikit-learn* [45], including *balanced_accuracy_score*, which calculates balanced accuracy for addressing imbalances in multiclass datasets, and *precision_recall_fscore_support*, which provides precision, recall, and F-beta scores for each class. Precision is the proportion of true positives relative to the total of true positives and false positives. Recall is the proportion of true positives relative to the total of true positives and false negatives. The F_1-score is the harmonic mean of the precision and recall. Additionally, we evaluated the model's ability to balance sensitivity and specificity by measuring the area under the curve of the receiver operating characteristic (AUCROC), which is the proportion of area below the receiver operating characteristic curve, which in turn is the plot of the true positive rate against the false positive rate.

Effect of Training and Validation Split on Model Performance

To evaluate the impact of different training and validation splits on model performance, we evaluated SciBERT's performance with 10%, 20%, and 30% of the data reserved for validation. This analysis was conducted to assess the relationship between training set size and classification accuracy, based on the assumption that larger training sets may improve model learning. By progressively reducing the number of training samples, we tested the extent to which performance would degrade with less training data.

Manual Validation and Comparison With GPT-4o Performance

To evaluate model performance, we used a separate verification dataset consisting of 213 Cochrane PLSs published between September 2024 and May 2025, which was not a part of the original training or evaluation datasets. Each PLS was independently classified by 2 domain experts, with a third expert resolving any discrepancies (NB, a metascientist with expertise in Cochrane PLSs and conclusiveness assessment; RB, a research assistant at the Department of Research in Biomedicine and Health; and BĆ, a psychology student with experience using AI tools to assess compliance with reporting guidelines).

The GPT-4o model was prompted using a zero-shot classification approach, which means no example classifications were provided in the prompt. The prompt instructed the model to classify each PLS into one of 3 predefined categories and included detailed definitions for each class, identical to those used by human annotators. The complete prompt can be found in Multimedia Appendix 1.

The performance of the 2 trained BERT-based models—SciBERT and Longformer—was compared with the baseline GPT-4o model (subscription-based) and evaluated against labels assigned by human experts. Model outputs were compared using standard classification metrics: accuracy, precision, recall, F_1-score, and Cohen κ, with the lattermost being used to assess the level of agreement between model predictions and the expert consensus categorization.

Calibration Analysis

To evaluate the reliability of predicted probabilities, we conducted a calibration analysis on the fine-tuned SciBERT model. Predicted probabilities were compared with observed outcome frequencies across the 3 target classes using calibration plots and quantitative metrics, including expected calibration error and Brier score. Since Longformer achieved similar classification performance, calibration was performed only for SciBERT.

Ethical Considerations

This study did not involve human participants or the collection of private data. A publicly available dataset of Cochrane PLSs was used, which can be accessed via the Open Science Framework [46].

The use of publicly available data is exempt from ethics review in accordance with the University of Split School of Medicine guidelines and the Croatian Science Foundation project Professionalism in Health – Decision making in practice and research (IP-2019-04-4882) [47]. Therefore, institutional review board approval and informed consent were not required. The dataset used in this study was collected and shared under conditions that permit secondary analysis without additional consent requirements. The data contain no personally identifiable information. All analyzed PLSs are publicly accessible textual documents.

Results

Overview

Among the 4405 PLSs from our dataset, 429 (9.7%) had been manually categorized as conclusive, 1203 (27.3%) as inconclusive, and 2773 (63%) as unclear [24]. These PLSs served as input data for our model. To address class imbalance, we applied random undersampling to the training set by selecting an equal number of PLSs from each class (n=429). This ensured that we had a balanced dataset and reduced the risk of biased model learning.

In classifying the PLSs, the SciBERT model achieved a balanced accuracy of 56.6%, with AUCROC scores of 0.91 for the conclusive, 0.67 for the inconclusive, and 0.75 for the unclear class. The Longformer model, meanwhile, demonstrated a

XSL•FO
RenderX

balanced accuracy of 60.9%, with AUCROC scores of 0.86, 0.67, and 0.72 for the same classes, respectively. The receiver operating characteristic curves and confusion matrices are visualized in Figure 1 and Figure 2, while Table 1 presents a side-by-side comparison of the performance for both models across all classes.

Figure 1. Receiver operating characteristic (ROC) curves and the corresponding area under the curve of the receiver operating characteristic scores for each class (0=conclusive, 1=inconclusive, and 2=unclear); (A) SciBERT model and (B) Longformer model. Calculated and visualized using scikit-learn. AUC: area under the curve.

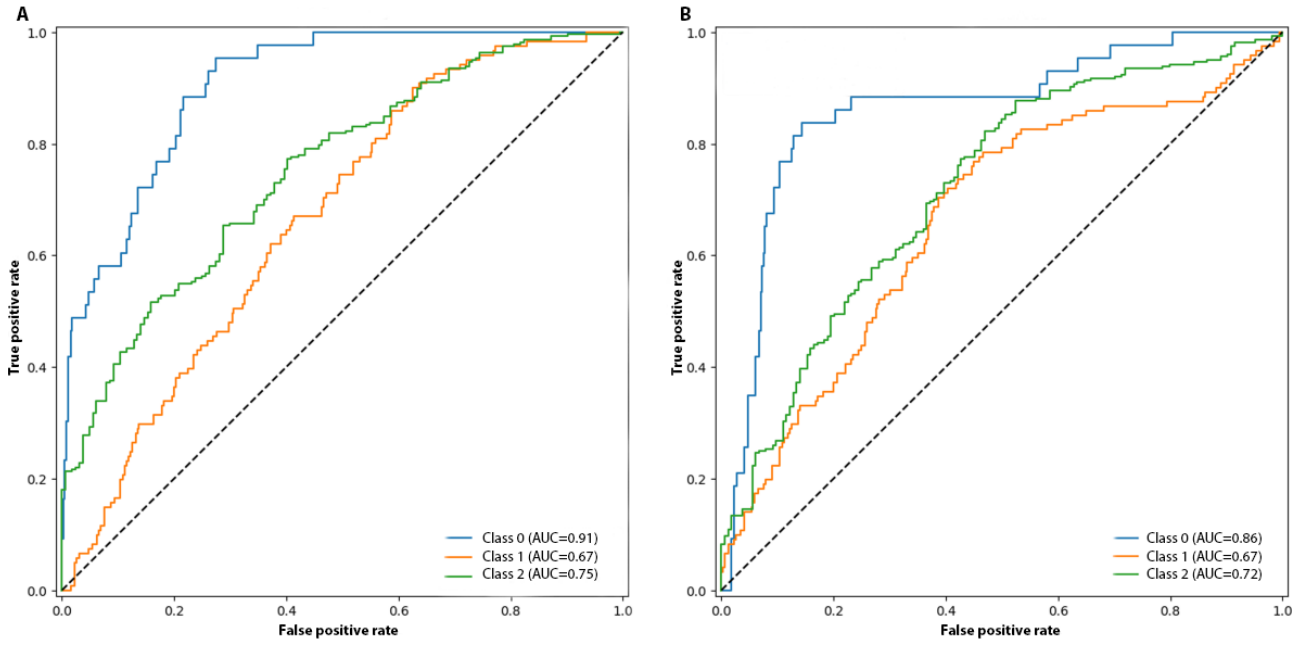
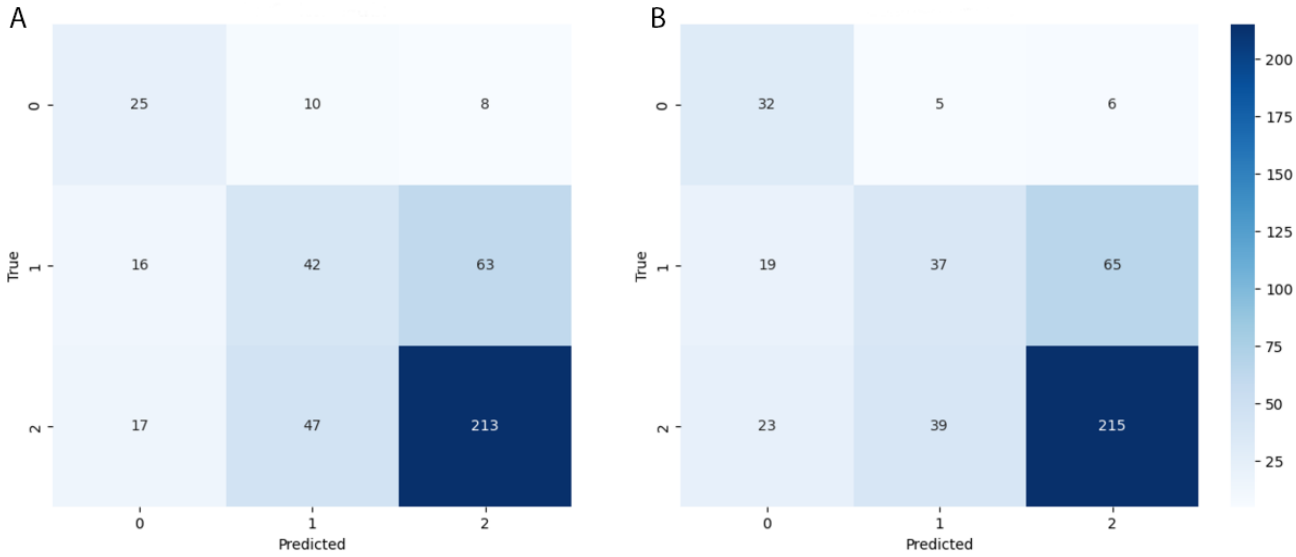


Figure 2. Confusion matrix for the fine-tuned (A) SciBERT and (B) Longformer multiclass classification (0=conclusive, 1=inconclusive, and 2=unclear) used to visualize the performance of the conclusiveness classification algorithm (created using the scikit-learn). The confusion matrix evaluates multiclass performance by comparing the predicted classes with the actual classes. The diagonal elements represent the correct predictions.



XSL•FO
RenderX

Table . Per-class performance of the conclusiveness classification for Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer models[a].

Model	Precision	Recall	F_1-score	AUCROC[b] score
Conclusive				
SciBERT	0.43	0.58	0.50	0.91
Longformer	0.43	0.74	0.55	0.87
Inconclusive				
SciBERT	0.42	0.35	0.38	0.67
Longformer	0.46	0.31	0.37	0.67
Unclear				
SciBERT	0.75	0.77	0.76	0.75
Longformer	0.75	0.78	0.76	0.72

[a]Precision is the proportion of true positives relative to the total of true positives and false positives. Recall is the proportion of true positives relative to the total of true positives and false negatives. F_1-score is the harmonic mean of the precision and recall. Area under the curve of the receiver operating characteristic is the proportion of area below the receiver operating characteristic curve, which is the plot of the true positive rate against the false positive rate. The scores were obtained using the scikit-learn functions.

[b]AUCROC: area under the curve of the receiver operating characteristic.

For comparison, the baseline SciBERT model achieved a balanced accuracy of only 42.2%, with AUCROC scores of 0.68 for the conclusive, 0.53 for the inconclusive, and 0.53 for the unclear class. The baseline Longformer's balanced accuracy was 39.0%, with AUCROC scores of 0.69, 0.56, and 0.54 for the same classes, respectively. However, the baseline Longformer was unstable across repeated runs and sometimes predicted only one class. This indicates that the frozen encoder was not able to provide features that separate the 3 categories well. The performance of all models is documented in Multimedia Appendix 1.

Effect of the Training Set Size and Validation Split on Model Performance

The highest performance was observed with a 10% validation split, whereas 20% and 30% validation splits resulted in similarly reduced accuracies (Table 2).

Table . Performance of Scientific Bidirectional Encoder Representations from Transformers with varying training and validation splits.

Validation split	Training samples (per class), n	Balanced accuracy (%)
10%	343	56.6
20%	257	53.1
30%	172	53.3

Manual Validation and Comparison With ChatGPT Performance

Among the 213 PLSs from our additional verification dataset, 44 (20.7%) had been manually categorized as conclusive, 110 (51.6%) as inconclusive, and 59 (27.7%) as unclear. The Cohen κ value between the experts was 0.57, indicating moderate agreement. The baseline GPT-4o model outperformed the trained BERT-based models (Table 3). SciBERT had the poorest performance overall, while ChatGPT had the strongest, with most of its conclusive predictions being correct. ChatGPT also achieved moderate agreement with the human classifications (Cohen κ=0.57). This indicates that ChatGPT's classifications were as consistent with the expert consensus as an individual expert's were with one another. In contrast, SciBERT and Longformer demonstrated poor alignment with the reference classifications, with a Cohen κ value of 0.03 and 0.19, respectively, suggesting worse-than-random agreement (Table 3).

XSL•FO
RenderX

Table . Comparative performance of fine-tuned Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer models and baseline GPT-4o model on conclusiveness classification task.

Metric	Fine-tuned SciBERT	Fine-tuned Longformer	Baseline GPT-4o
Precision	0.34	0.57	0.74
Recall	0.34	0.44	0.74
F_1-score	0.27	0.40	0.74
Accuracy (%)	34.3	44.1	74.2
Cohen κ[a]	0.03	0.19	0.57

[a]Predictions were compared against expert annotations, where each plain language summary was manually classified by 2 experts, with a third expert resolving any discrepancies.

Calibration Analysis

The calibration analysis was performed on the fine-tuned SciBERT model, which was overconfident for the "conclusive" class, poorly calibrated for the "inconclusive" class, and initially underconfident for the "unclear" class, although highly accurate when assigning high probabilities. Calibration plots and quantitative metrics, including expected calibration error and Brier scores (a measure of the average squared difference between predicted probability and actual outcome), are presented in Multimedia Appendix 1.

Discussion

Principal Results

Our results showed that transformer-based language models such as SciBERT and Longformer achieved modest performance in classifying Cochrane PLSs based on their level of conclusiveness. Both models were fine-tuned on a balanced dataset and evaluated using standard classification metrics, with Longformer achieving a balanced accuracy of 60.9%, compared with 56.6% for SciBERT. Both models performed best on the conclusive class, achieving relatively high AUCROC and F_1-scores. For the unclear class, SciBERT demonstrated stronger precision and recall. However, both models struggled to distinguish the inconclusive class, with the lowest F_1-scores and overlapping errors with the unclear category. Both models underperformed on the inconclusive class, with poor AUCROC and low precision and recall scores. These findings may indicate that conclusiveness is expressed in linguistically nuanced ways that the models are unable to detect and that there could be a semantic overlap between inconclusive and unclear PLSs. Both models were outperformed by ChatGPT, which achieved better accuracy and interrater agreement. This suggests that general-purpose LLMs may offer more reliable performance for this classification task, even without domain-specific fine-tuning. Notably, GPT-4o achieved the same Cohen κ value (0.57) as the agreement between human experts, suggesting that it mirrors expert-level judgment and nuanced reasoning. This finding highlights the potential of general-purpose LLMs to approximate human evaluation in semantically complex classification tasks.

Qualitative Insights From Manual Classification

During manual labeling of PLSs, we identified several challenges that may explain why the experts, the fine-tuned models, and ChatGPT all struggled with the classifications. First, there was some ambiguity between the "inconclusive" and "unclear" classes. For example, some PLSs did not clearly state whether the evidence was insufficient, which might be why both human assessors and models were uncertain when assigning these labels. Furthermore, the interpretation of the criteria for the "conclusive" class was occasionally ambiguous, particularly in cases where PLSs included recommendations but lacked clear statements about intervention effectiveness. This ambiguity likely made it difficult for human annotators to determine whether the conclusion was strong enough to classify a PLS as conclusive or inconclusive. Consequently, models trained on these labels may have inherited this ambiguity. We also observed that PLSs included expressions such as "may help" or "probably works," which are common in scientific writing but can signal uncertainty. This nuance might have been difficult for models to detect, explaining their lower performance in differentiating between inconclusive and unclear statements. These findings suggest that better model instruction, such as through advanced prompt engineering, might help improve future performance of the GPT-4o model. They also highlight the need to incorporate linguistic features of uncertainty more explicitly into the training process.

Comparison With Prior Work

To our knowledge, there have been no studies on the automatic classification of Cochrane PLSs or full reviews based on the level of conclusiveness, although some studies examined machine learning techniques for making systematic review processes more efficient. For example, one study developed a randomized controlled trial classifier for Cochrane Reviews, a tool that discerns whether a selected study qualifies as a randomized controlled trial [48]. In another, ChatGPT showed strong performance when used for abstractive summarization of longer texts, including news articles and public speeches [49]. However, given the lack of specialized expertise in the field of medicine, ChatGPT does not always grasp the nuances of its terminology and sometimes struggles to recognize important information [50]. In addition, one study found that LLMs sometimes generate factually inconsistent summaries, which could potentially harm readers [51]. Yet, these challenges and the related legal and ethical issues should not discourage the use of LLMs but rather encourage further research and refinement of the technology.

XSL•FO
RenderX

The fine-tuned BERT-based models did not perform well in our classification task, indicating limitations in generalizing to nuanced language in PLSs. In contrast, general-purpose language models can perform better than fine-tuned models in some classification tasks, achieving Cohen κ scores comparable to those of human experts. This is likely because they have been trained on much larger and more diverse text corpora and have more complex architectures, allowing them to better understand context and differentiate linguistic nuances [52]. This is also in line with findings by Davidson and Chae [53] that LLMs, particularly when fine-tuned on prompts that include explicit instructions, can outperform traditional supervised models in a variety of classification settings without task-specific training.

Limitations

This study has several limitations. First, although we began with a relatively large dataset of 4405 PLSs, the "conclusive" class comprised only about 9.73% (n=429) of the total dataset. To address this issue, we applied random undersampling, which reduced the number of PLSs in the "inconclusive" and "unclear" classes. While this approach ensured balanced class representation, it also removed a substantial amount of data (2344/2773, 84.53% of the "unclear" class) that could have supported more robust model learning. This likely limited the models' ability to learn the linguistic variability of the majority class. In contrast, GPT-4o was used in zero-shot inference settings and was not trained on our dataset, meaning that its performance could not be affected by the undersampling procedure that constrained the fine-tuned models. Alternative approaches, such as applying class-weighted loss functions or oversampling minority classes, may yield improved performance in future work.

Second, the dataset was based exclusively on PLSs of Cochrane Reviews, representing a single domain within evidence-based health literature. This may also limit the model's generalizability to other types of health communication.

Third, there were often very subtle linguistic differences between the inconclusive and unclear PLSs, which introduced noise in model classification. Some PLSs lacked clear phrasing or used expressions such as "may help," and "probably works" that were difficult to interpret consistently, even for human annotators. This ambiguity likely contributed to the models' difficulty in separating these two classes.

Fourth, although one researcher (IB) participated in and provided instruction for both the original 2019 annotation and the current one, the full annotator teams differed between the two studies. It is possible that there were subtle differences in how annotators interpreted or applied the criteria, raising the possibility of annotator drift between the original labels used for model training and the new labels used for verification. Such drift may partly account for the observed decline in performance of the fine-tuned models on the newer dataset.

Fifth, when comparing accuracy across different splits, it is important to note that altering the training and validation proportion also changes the size of the remaining test set. Because of this, the test benchmarks were not identical across these experiments. While this does not affect the qualitative pattern we observed, the varying test baseline may contribute to numerical differences in accuracy.

Additionally, the fine-tuned SciBERT model showed high volatility in validation loss across training epochs, including abrupt spikes prior to early stopping. This suggests that the model may not have reached a fully stable convergence point, possibly due to the limited size of the training set and the complexity of the task. Such fluctuations may have constrained the model's performance.

Finally, although we compared our fine-tuned models with GPT-4o, we did not use advanced prompt engineering or task-specific tuning. This likely underestimated ChatGPT's performance on this classification task.

Future Work and Recommendations

Future studies should expand the dataset to include PLSs from non-Cochrane sources and from different health domains, which could improve model generalizability. First, although there was no meaningful difference in the performance of different training and validation splits, future work should also explore the impact of larger datasets on model performance. Second, adding task-specific fine-tuning and advanced prompt engineering of LLMs such as GPT-4o could improve classification accuracy even more. Third, models may be able to more successfully differentiate between inconclusive and unclear classes by incorporating linguistic features that capture uncertainty and conclusiveness cues.

In the long term, implementing a general-purpose LLM, such as ChatGPT, within Cochrane platforms (eg, the RevMan Web dashboard) could assist authors in ensuring that their conclusions are clear and guide readers through reviews according to their conclusiveness level. Additionally, LLMs could support users of the Cochrane Library by offering filters or tags that group PLSs by conclusiveness. However, additional model validation and ethical review should precede these real-world applications.

Conclusions

We explored the use of 2 fine-tuned transformer-based models—SciBERT and Longformer—for classifying Cochrane PLS according to their level of conclusiveness. Both models demonstrated modest internal performance but poor generalization to newly published PLSs, particularly in distinguishing between inconclusive and unclear categories, likely due to their semantic overlap. An empirical analysis of different training and validation splits confirmed that larger training sets improve model performance, although the gains were modest. Most importantly, both models were outperformed by ChatGPT, which as a general-purpose language model, achieved higher accuracy (74%) and agreement with expert annotations, suggesting that state-of-the-art LLMs hold greater potential for health care information dissemination.

XSL•FO
RenderX

Funding

This study was funded by the Croatian Science Foundation “Professionalism in Health – Decision making in practice and research” (ProDeM) under grant IP-2019-04-4882. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets used and analyzed during the study are publicly available on the Open Science Framework website [46].

Authors' Contributions

Conceptualization: A Mijatović, IB, LU
Data curation: A Mijatović
Formal analysis: A Mijatović, BĆ, IB, LU, NB, RB
Funding acquisition: A Marušić
Investigation: A Mijatović
Methodology: A Mijatović
Project administration: A Marušić
Resources: IB
Software: A Mijatović
Supervision: A Marušić
Validation: A Mijatović, BĆ, NB, RB
Visualization: A Mijatović
Writing—original draft: A Mijatović
Writing—review and editing: A Marušić, BĆ, IB, LU, NB, RB
All authors approved the submitted version and take accountability for the work.

Conflicts of Interest

A Marušić and NB are active Cochrane members. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary material presenting detailed model training procedures, experimental setups, and full performance metrics for SciBERT and Longformer models.
[DOCX File, 764 KB - medinform_v14i1e72657_app1.docx]

References

1. Whiting P, Davenport C. Writing a plain language summary. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: John Wiley & Sons; 2023.
2. Higgins JP, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Collaboration; 2023. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-03-30]
3. Weiss BD. Health Literacy and Patient Safety: Help Patients Understand: Manual for Clinicians, 2nd edition: American Medical Association; 2007. URL: https://books.google.be/books?id=quJaYgEACAAJ [accessed 2026-03-30]
4. Schünemann HJ, Vist GE, Higgins JP, Santesso N, Deeks JJ, Glasziou P, et al. Chapter 15: Interpreting results and drawing conclusions. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane Collaboration; 2024.
5. Glenton C, Nilsen ES, Sporstøl Fønhus M, Goudie S, Noonan E. How to write a plain language summary of a Cochrane intervention review (version 10). Cochrane Norway. URL: https://www.cochrane.no/how-write-plain-language-summary [accessed 2026-03-22]
6. Lasserson T, Santesso N, Cumpston M, Marshall R, Ógáin ON. Incorporating GRADE in Cochrane reviews: feedback from the CEU screening programme. : Cochrane; 2016 URL: https://www.cochrane.org/sites/default/files/uploads/PDFs/MECIR/Incorporating%20GRADE%20in%20Cochrane%20Reviews.pdf [accessed 2026-03-22]
7. Santesso N, Glenton C, Dahm P, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. J Clin Epidemiol 2020 Mar;119:126-135. [doi: 10.1016/j.jclinepi.2019.10.014] [Medline: 31711912]
8. Changiz T, Yousefy A, Fakhari M. Research utilization process model: a cyclical, spiral, and developmental process to provide conclusive research knowledge in health professions education. Med J Islam Repub Iran 2020;34:79. [doi: 10.34171/mjiri.34.79] [Medline: 33306047]
9. Kurtzman ET, Greene J. Effective presentation of health care performance information for consumer decision making: a systematic review. Patient Educ Couns 2016 Jan;99(1):36-43. [doi: 10.1016/j.pec.2015.07.030] [Medline: 26277826]

XSL•FO
RenderX

10. Gray JA. Discussion and conclusion. AME Med J 2019;4:26. [doi: 10.21037/amj.2019.04.05]
11. Bujnowska-Fedak MM, Węgierek P. The impact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health 2020 Jan 31;17(3):880. [doi: 10.3390/ijerph17030880] [Medline: 32023828]
12. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9. [doi: 10.2196/jmir.5729] [Medline: 28104579]
13. Murray E, Lo B, Pollack L, et al. The impact of health information on the internet on the physician-patient relationship: patient perceptions. Arch Intern Med 2003 Jul 28;163(14):1727-1734. [doi: 10.1001/archinte.163.14.1727] [Medline: 12885689]
14. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? Can Fam Physician 2014 May;60(5):407-408. [Medline: 24828994]
15. Peng RX. How online searches fuel health anxiety: investigating the link between health-related searches, health anxiety, and future intention. Comput Human Behav 2022 Nov;136:107384. [doi: 10.1016/j.chb.2022.107384]
16. Müller A, Baumann E, Dierks ML. Cyberchondria - a new behavioral syndrome? [Article in German]. Psychother Psychosom Med Psychol 2021 Jun;71(6):243-255. [doi: 10.1055/a-1348-8059] [Medline: 34102694]
17. Doherty-Torstrick ER, Walton KE, Fallon BA. Cyberchondria: parsing health anxiety from online behavior. Psychosomatics 2016;57(4):390-400. [doi: 10.1016/j.psym.2016.02.002] [Medline: 27044514]
18. Šuto J, Marušić A, Buljan I. Linguistic analysis of plain language summaries and corresponding scientific summaries of Cochrane systematic reviews about oncology interventions. Cancer Med 2023 May;12(9):10950-10960. [doi: 10.1002/cam4.5825] [Medline: 36951519]
19. Mimouni M, Mimouni F, Segev F. Conclusiveness of the Cochrane Eye and Vision Group Reviews. BMC Res Notes 2015 Jun 16;8:242. [doi: 10.1186/s13104-015-1221-x] [Medline: 26076817]
20. Yin S, Chuai Y, Wang A, Zhang L. Conclusiveness of the Cochrane reviews in gynaecological cancer: a systematic analysis. J Int Med Res 2015 Jun;43(3):311-315. [doi: 10.1177/0300060515574922] [Medline: 25870179]
21. Mandel D, Littner Y, Mimouni FB, Lubetzky R. Conclusiveness of the Cochrane Neonatal Reviews: a systematic analysis. Acta Paediatr 2006 Oct;95(10):1209-1212. [doi: 10.1080/08035250600580537] [Medline: 16982491]
22. Cohen S, Lubetzky R, Mimouni FB, Marom R, Mandel D. Conclusiveness of the Cochrane Reviews in pediatric-gastroenterology: a systematic analysis. Eur J Gastroenterol Hepatol 2013 Feb;25(2):252-254. [doi: 10.1097/MEG.0b013e32835a1083] [Medline: 23044810]
23. Conway A, Conway Z, Soalheira K, Sutherland J. High quality of evidence is uncommon in Cochrane systematic reviews in Anaesthesia, Critical Care and Emergency Medicine. Eur J Anaesthesiol 2017 Dec;34(12):808-813. [doi: 10.1097/EJA.0000000000000691] [Medline: 29095726]
24. Banić A, Fidahić M, Šuto J, et al. Conclusiveness, linguistic characteristics and readability of Cochrane plain language summaries of intervention reviews: a cross-sectional study. BMC Med Res Methodol 2022 Sep 10;22(1):240. [doi: 10.1186/s12874-022-01721-7] [Medline: 36088293]
25. Bird S, Klein E, Loper E. Natural Language Processing with Python: O'Reilly Media; 2009.
26. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 2023;82(3):3713-3744. [doi: 10.1007/s11042-022-13428-4] [Medline: 35855771]
27. A complete guide to natural language processing. DeepLearning.AI. URL: https://www.deeplearning.ai/resources/natural-language-processing [accessed 2024-09-21]
28. Srinivasan A, Sitaram S, Ganu T, Dandapat S, Bali K, Choudhury M. Predicting the performance of multilingual NLP models. arXiv. Preprint posted online on Oct 21, 2021. [doi: 10.48550/arXiv.2110.08875]
29. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
30. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019. [doi: 10.18653/v1/D19-1371]
31. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on Apr 10, 2020. [doi: 10.48550/arXiv.2004.05150]
32. Bralić N, Buljan I. The association between research design and the perceived treatment effectiveness: a cross-sectional study. Front Med (Lausanne) 2023 Dec 22;10:1220999. [doi: 10.3389/fmed.2023.1220999] [Medline: 38196834]
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
34. Causevic S. Evolution of large language models—BERT, GPT3, MUM, and PaLM. Medium. 2022. URL: https://medium.com/data-science/self-supervised-transformer-models-bert-gpt3-mum-and-paml-2b5e29ea0c26 [accessed 2026-03-23]
35. Fine-tuning. Hugging Face. URL: https://huggingface.co/docs/transformers/training [accessed 2026-03-23]
36. Winastwan R. Text classification with BERT in PyTorch. Towards Data Science. 2021. URL: https://towardsdatascience.com/text-classification-with-bert-in-pytorch-887965e5820f [accessed 2024-09-21]

XSL•FO
RenderX

37. What is transfer learning? GeeksforGeeks. URL: https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning [accessed 2024-09-21]
38. Oztel I, Yolcu G, Oz C. Performance comparison of transfer learning and training from scratch approaches for deep facial expression recognition. Presented at: 2019 4th International Conference on Computer Science and Engineering (UBMK); Sep 11-15, 2019. [doi: 10.1109/UBMK.2019.8907203]
39. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alche'-Buc FD, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32: Curran Associates, Inc; 2019:8024-8035.
40. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas: IOS Press; 2016:87-90. [doi: 10.3233/978-1-61499-649-1-87]
41. Models. Hugging Face. URL: https://huggingface.co/models [accessed 2024-09-21]
42. AdamW. PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html#adamw [accessed 2024-09-21]
43. CrossEntropyLoss. PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html [accessed 2024-09-21]
44. Reed R, Marks RJ. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks: MIT Press; 1999.
45. scikit-learn: machine learning in Python. scikit-learn. URL: https://scikit-learn.org/stable [accessed 2024-09-21]
46. Readability, linguistic characteristics and conclusiveness of Cochrane plain language summaries of intervention reviews: a cross-sectional study. OSF. URL: https://osf.io/qvu3a/overview [accessed 2026-03-30]
47. Code of ethics of the Croatian Science Foundation. Hrvatska Zaklada za Znanost. 2020. URL: https://hrzz.hr/wp-content/uploads/2020/08/HRZZ-Code-of-Ethics.pdf [accessed 2026-03-30]
48. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol 2021 May;133:140-151. [doi: 10.1016/j.jclinepi.2020.11.003] [Medline: 33171275]
49. Povey N. ChatGPT: abstractive text summarization. Medium. 2022. URL: https://medium.com/@nadirapovey/chatgpt-text-summarization-44f768222a4c [accessed 2024-09-21]
50. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
51. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023 Aug 24;6(1):158. [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
52. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
53. Davidson T, Chae Y. Large language models for text classification: from zero-shot learning to instruction-tuning. Sociol Methods Res 2025. [doi: 10.1177/00491241251325243]

Abbreviations

AI: artificial intelligence
AUCROC: area under the curve of the receiver operating characteristic
BERT: Bidirectional Encoder Representations from Transformers
LLM: large language model
NLP: natural language processing
PLS: plain language summary
SciBERT: Scientific Bidirectional Encoder Representations from Transformers

Edited by A Castonguay, A Coristine; submitted 14.Feb.2025; peer-reviewed by FL Gaol, R Homayouni; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Apr.2026.

Please cite as:
Mijatović A, Ursić L, Bralić N, Bandić R, Ćaćić B, Buljan I, Marušić A
Classification of Cochrane Plain Language Summaries by Conclusiveness Using Transformer-Based Models and ChatGPT: Retrospective Observational Study
JMIR Med Inform 2026;14:e72657
URL: https://medinform.jmir.org/2026/1/e72657
doi:10.2196/72657

XSL•FO
RenderX

© Antonija Mijatović, Luka Ursić, Nensi Bralić, Ružica Bandić, Barbara Ćaćić, Ivan Buljan, Ana Marušić. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Rectal Cancer Radiotherapy Response Prediction: Retrospective Study of Development of a Deep Learning–Based Radiomics Model

Yiqun Li[1*], MS; Hengchang Liu[2*], PhD; Qiang Wei[1], MS; Zhenqi Wang[1], MS; Zhen Hu[1], MS

[1]Department of Colorectal and Anal Surgery, Shanxi Provincial People's Hospital, No. 29 Shuangtasi Street, Taiyuan, China

[2]Department of Colorectal Surgery, National Cancer Center/National Clinical Research Center of Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Zhen Hu, MS
Department of Colorectal and Anal Surgery, Shanxi Provincial People's Hospital, No. 29 Shuangtasi Street, Taiyuan, China

Abstract

Background: Radiotherapy (RT) is a cornerstone of multimodal treatment for rectal cancer (RC); yet, substantial interindividual variability in treatment response persists. Deep learning (DL)–based radiomics offers potential for pre-RT response prediction to support personalized decision-making. **Objective:** This study aimed to develop and compare multiple DL radiomics models for predicting RT response in RC, with emphasis on the performance and clinical utility of Transformer architectures. **Methods:** In this single-center retrospective study, 2000 pathologically confirmed patients with RC who received standard RT were included. Pretreatment computed tomographic and dynamic contrast-enhanced magnetic resonance images and clinical variables were collected. Treatment response was categorized according to RECIST (Response Evaluation Criteria in Solid Tumors) version 1.1 as good (complete or partial response) or poor (stable or progressive disease). The primary analysis used magnetic resonance imaging (MRI)–only input; computed tomography (CT) was used for registration and quality control and evaluated in a late-fusion CT + MRI sensitivity analysis. Data were randomly split into training, validation, and test sets (8:1:1), with 5-fold cross-validation within the training set. Test set tumor masks were manually delineated, whereas a U-Net assisted segmentation was performed only within training to prevent data leakage. Convolutional neural network, graph convolutional network, and Transformer classifiers were compared. Class imbalance (approximately 65% vs 35%) was addressed using class weighting. Performance was evaluated using area under the receiver operating characteristic curve (AUROC) and accuracy with 95% CIs obtained by bootstrapping. AUROC differences were assessed using the DeLong test. Clinical usefulness was evaluated using decision curve analysis. Segmentation performance was quantified by Dice coefficient and intersection over union. Model interpretability was assessed using Gradient-Weighted Class Activation Mapping. **Results:** In the MRI-only primary analysis, the Transformer achieved the best performance on the independent test set, with accuracy of 87.0% (95% CI 84.2%-89.5%) and AUROC of 0.921 (95% CI 0.901-0.945), significantly outperforming the convolutional neural network (AUROC 0.881; *P*=.02) and graph convolutional network (AUROC 0.894; *P*=.041). Sensitivity and specificity were 89.2% and 82.9%, respectively. Decision curve analysis demonstrated higher net benefit across threshold probabilities of 0.3-0.7. U-Net segmentation achieved a mean Dice coefficient of 0.892 and intersection over union of 0.814. In sensitivity analysis, CT + MRI late fusion yielded a comparable AUROC to MRI only (0.926 vs 0.921; *P*=.36), with modest incremental net benefit at higher thresholds. **Conclusions:** In this large pre-RT imaging cohort, an MRI-driven Transformer-based DL radiomics model outperformed conventional architectures in predicting RT response in RC and demonstrated superior clinical net benefit. Late fusion of CT and MRI did not significantly improve overall discrimination but may provide incremental benefit in specific decision contexts. Multicenter external validation is warranted.

(JMIR Med Inform 2026;14:e77313) doi:10.2196/77313

KEYWORDS

deep learning; radiomics; rectal cancer; radiotherapy; Transformer model

Introduction

Rectal cancer (RC) remains one of the leading causes of cancer-related incidence and mortality worldwide, with a particularly notable rise among middle-aged and older adult populations. Radiotherapy (RT) is a cornerstone of RC management; however, its therapeutic efficacy varies substantially among patients [1]. Current clinical indicators and conventional imaging approaches show limited ability to accurately predict individual responses to RT [2], often failing

XSL•FO
RenderX

to identify patients who are truly radiosensitive [3,4]. This variability poses a significant challenge in optimizing treatment strategies [2]. Consequently, there is a pressing need for robust predictive tools to guide personalized treatment and improve clinical outcomes.

The advent of radiomics has introduced the ability to extract high-dimensional quantitative features from medical images, uncovering latent biological information and offering novel insights into tumor behavior and treatment response. Unlike traditional imaging techniques, radiomics enables noninvasive prediction of therapeutic efficacy using imaging-derived biomarkers [5,6]. Nevertheless, conventional machine learning models often struggle to manage the complexity of radiomics data, particularly their high dimensionality and inherent nonlinearity [7,8]. This has led to a growing interest in the application of advanced deep learning (DL) algorithms, which offer superior capacity for feature representation and pattern recognition, to enhance the predictive power of radiomics analyses [9].

DL, particularly convolutional neural networks (CNNs), has achieved substantial advances in medical image analysis in recent years [10,11]. These models can automatically extract hierarchical features from imaging data and effectively recognize and classify complex patterns [12]. Compared with traditional approaches, DL offers notable advantages in handling high-dimensional data and performing intricate pattern recognition, making it especially well suited for analyzing radiomics data [13,14]. However, most existing studies have applied single DL algorithms in isolation, with few providing a systematic comparison of multiple models, particularly in the context of predicting RT response in patients with RC [15].

In recent years, the application of artificial intelligence (AI) in RT, cancer diagnosis, and clinical decision support has shifted from research exploration toward clinical deployment. Within the RT workflow, AI has been increasingly implemented in key steps such as organ autosegmentation, target delineation, (semi-)automated treatment planning, dose optimization, and image registration, yielding measurable time savings and improved plan consistency [16]. In the domain of cancer diagnosis, the integration of DL and radiomics has achieved substantial progress in RC lesion segmentation and feature quantification on magnetic resonance imaging (MRI). Such models can automatically localize tumor regions and characterize morphological and textural features of lesions, thereby facilitating earlier and more consistent assessment and diagnosis; this has been substantiated by both systematic reviews of RC lesion segmentation and specific studies on post–neoadjuvant chemoradiotherapy (nCRT) T2-weighted (T2W) MRI segmentation [17]. Moreover, AI models that integrate multimodal data have demonstrated practical value for personalized treatment strategies and risk prediction in clinical decision-making. For RC, a multicenter study showed that late-fusion models combining multiparametric MRI–derived DL or radiomics features with clinical variables outperformed unimodal approaches in predicting early recurrence risk and provided greater clinical net benefit as assessed by decision curve analysis (DCA) [18]. Against this background, this study integrates DL with radiomics to construct a predictive model aimed at accurately assessing the RT response in RC patients, thereby providing intelligent support for treatment optimization.

We systematically compared multiple state-of-the-art DL architectures, including CNNs, graph convolutional networks (GCNs), and Transformer models, for their ability to analyze and process radiomics imaging data in patients with RC. The Transformer architecture, in particular, exhibited superior capability in modeling complex image features through its self-attention mechanism, allowing it to effectively identify key regions relevant to RT response. By comparing these models in terms of feature extraction and predictive performance, this study investigated the potential of the Transformer model in predicting RT response in RC and evaluated its feasibility for clinical application. Unlike traditional applications of AI in tumor detection or initial diagnosis, this study focused on patients with pathologically confirmed RC. Our objective was to predict their responsiveness to RT based on pretreatment imaging, thereby enabling individualized treatment planning and minimizing unnecessary exposure to RT-related risks.

The key innovation of this study lay in the systematic application and comparison of multiple DL algorithms to enhance the accuracy of RT response prediction, with a particular emphasis on exploring the strengths of the Transformer model. By combining DL techniques with radiomics, we aimed to advance current predictive methodologies and develop a more precise model for RT response assessment. This work provides clinicians with a robust, data-driven tool to support personalized therapeutic strategies, ultimately contributing to the broader goal of precision medicine in oncology.

Methods

Ethical Considerations

This study was conducted in strict accordance with ethical standards. The use of imaging and clinical data was approved by the Medical Ethics Committee of Shanxi Provincial People's Hospital. As this was a retrospective analysis, all data were obtained from routine clinical records and were fully deidentified prior to analysis. The ethics committee confirmed that additional informed consent was not required. All data were anonymized before processing, and no identifiable patient images or information was included in the manuscript or supplementary materials. Data were stored on encrypted servers with access restricted to authorized investigators. Since this study did not involve prospective data collection or patient intervention, no monetary or nonmonetary compensation was provided to participants. For any future studies involving identifiable images, written informed consent will be obtained, and the relevant materials will be submitted with the manuscript.

Data Collection

This retrospective study included 2000 patients with RC who were treated at Shanxi Provincial People's Hospital between January 2015 and December 2023. All patients underwent standard RT and had complete paired pretreatment computed tomography (CT) and MRI data, along with follow-up assessments of RT response.

XSL•FO
RenderX

Inclusion criteria were (1) pathologically confirmed RC, (2) receipt of standard RT, (3) availability of complete pretreatment CT and dynamic contrast–enhanced MRI (DCE-MRI) data, and (4) documented RT outcomes with follow-up records. Exclusion criteria were (1) incomplete or poor-quality imaging data, (2) missing essential clinical information or outcome labels, (3) coexisting malignancies or severe systemic diseases, and (4) patients who did not receive standard RT protocols. Treatment regimens followed the Chinese Society of Clinical Oncology guidelines for RC and primarily consisted of long-course RT (50.4 Gy in 28 fractions). Some patients also received concurrent chemotherapy and subsequent surgical intervention, depending on clinical conditions.

To ensure transparency in patient selection, a CONSORT (Consolidated Standards of Reporting Trials)-style flowchart is provided in Figure S1 in Multimedia Appendix 1, detailing the inclusion process, exclusion criteria, final sample size, and classification of patients into RT-sensitive and RT-insensitive groups.

All imaging data were acquired from routine clinical scanning systems at a single institution (Shanxi Provincial People's Hospital), encompassing MRI and CT examinations performed across different years. Detailed information regarding scanner models and acquisition parameters is provided in Table S1 in Multimedia Appendix 2.

Data Preprocessing

Following data collection, the imaging data were preprocessed to ensure quality and consistency. The preprocessing steps included denoising, cropping, normalization, and resolution adjustment. All images were normalized to a range between 0 and 1 and resized to a resolution of 512×512 pixels. To enhance the model's generalization ability, data augmentation techniques such as rotation, flipping, and scaling were applied to generate more diverse training samples and effectively expand the dataset. In addition, clinical data were standardized to harmonize feature values with different units and ranges, ensuring compatibility with model input requirements. The complete data collection and preprocessing workflow is illustrated in Figure S1 in Multimedia Appendix 3.

To prevent data leakage, the statistical parameters required for image normalization (mean and SD) were computed exclusively from the training set and kept fixed for both the validation and test sets. The test set was not involved in estimating normalization parameters at any stage, thereby ensuring independence and rigor in both preprocessing and model evaluation.

All DL models used in this study were trained and inferred on 2D slices; thus, the input features were derived solely from spatial texture information within single images, without relying on voxel-level 3D volume information. During preprocessing, we did not perform volumetric resampling along the *z*-axis to reconcile slice thickness differences (3 mm vs 4 mm), in order to avoid interpolation-induced artifacts that could distort 2D texture features. Because model training and prediction were conducted at the single-slice level, and all patients' images were acquired using consistent scanning sequences and acquisition protocols, variations in slice thickness were expected to have limited impact on grayscale distribution or spatial structure of the input features. Nevertheless, we discuss the potential biases associated with differing slice thicknesses in the "Limitations" section, and we plan to explore full 3D volume resampling and the integration of 3D models in future work to further improve model robustness in cross-device settings.

Data Annotation and Quality Control

To ensure high-quality data for model training, strict annotation and quality control procedures were applied to all imaging data and RT response labels. For the imaging side, radiologists manually delineated tumor regions, with particular attention paid to tumor boundaries and clinically relevant areas. Semiautomated annotation tools were used to assist in the delineation process. All annotations were independently reviewed by 2 associate chief radiologists (ZW and ZH), each with more than 10 years of experience. All delineation results were further validated through random sampling to ensure the consistency and accuracy of boundary recognition.

RT response labels were assigned based on posttreatment MRI evaluations and classified according to the RECIST (Response Evaluation Criteria in Solid Tumors) version 1.1. Patients who achieved a complete response (CR) or a partial response (PR) were categorized as the "good response (GR) group," while those with stable disease or progressive disease were categorized as the "poor response group." All imaging evaluations were performed in a blinded manner, with evaluators unaware of treatment plans and clinical outcomes. Two independent reviewers (ZW and ZH) assessed the follow-up MRIs, and interobserver agreement was assessed using the Cohen κ coefficient, which reached 0.83, indicating good consistency. Final response classifications were determined through consensus and used as ground-truth labels for supervised model training.

In particular, all tumor masks in the test set were manually annotated by experts without reference to any DL model outputs, thereby avoiding error propagation and ensuring the independence and rigor of model evaluation. The U-Net model was used exclusively on the training set to generate tumor masks as auxiliary inputs for the classification models. Its outputs were not used in any segmentation or feature extraction tasks for the test set, minimizing the risk of information leakage.

DL Model Construction and Training

For the patient-level response prediction task, the model first performs independent inference on each 2D slice, producing a probability score indicating the likelihood of belonging to the "GR" category. Subsequently, the predicted probabilities of all slices from the same patient are averaged arithmetically to obtain an aggregated patient-level probability, and a threshold of 0.5 is applied to generate the final patient-level classification label. This aggregation strategy helps smooth slice-to-slice fluctuations caused by differences in imaging position, preventing any single atypical slice from disproportionately influencing the overall judgment. As a result, the stability and robustness of patient-level predictions are enhanced. This study used multiple DL algorithms to process radiomics imaging data and predict

XSL•FO
RenderX

patient responses to RT (Figure S1 in Multimedia Appendix 4). The models compared included CNNs, GCNs, and Transformer models. All models were trained and tested on the same dataset to ensure comparability.

The Transformer model served as the primary focus due to its self-attention mechanism, which enabled the capture of complex global features and provided unique advantages in tumor region recognition. In this model, input images were divided into 16×16 patches and processed with 8 multihead attention heads and 12 Transformer encoder layers. Standard sinusoidal positional encoding was incorporated to preserve spatial structure information. The model was trained using the Adam optimizer, with an initial learning rate of 1e-4 and a cosine annealing strategy for dynamic learning rate adjustment. Training was conducted over 100 epochs, with early stopping applied if validation loss did not decrease for 10 consecutive epochs.

As the comparative baseline, the CNN used a residual networks (ResNet-34) architecture, consisting of 34 residual convolutional layers organized into 4 residual stages with channel dimensions of 64, 128, 256, and 512, respectively. Feature extraction was performed on 2D patches of the input images. The output of the final global average pooling layer was fed into a 2-layer fully connected classification head (hidden layer size=256, activation function=ReLU) to generate predictions of RT response.

The GCN module adopted a 3-layer GCNConv structure, where nodes represented embedded features of image patches or regions of interest (ROIs). The adjacency matrix was constructed based on spatial proximity using a k-nearest neighbors approach (k=8). Each layer used a ReLU activation function, and dropout (P=.30) was applied between layers to mitigate overfitting. The aggregated node features after global pooling were passed to a classification layer. All models were trained under identical conditions, using the same loss function (weighted cross-entropy), optimizer (Adam, learning rate=1×10^{-4}), and training strategy (cosine annealing schedule with early stopping) to ensure a fair performance comparison. Hyperparameter tuning and L2 regularization were applied throughout training to reduce overfitting risks. The final model was independently evaluated on the test set and compared against other networks, with particular emphasis on its performance in feature extraction and classification accuracy.

Given the class imbalance in RT response labels (approximately 65% of patients in the GR group and 35% in the poor response group), a class weighting strategy was introduced during training. Specifically, inverse-frequency weights based on sample distribution were applied in the loss function to enhance the model's sensitivity to the minority class (poor response patients). This approach effectively mitigated the prediction bias caused by unequal class distribution and yielded higher sensitivity and overall stability in the validation set.

In the final implementation of this study, the Transformer architecture was used solely as an image-level classifier. The network consisted of multiple encoder layers and a classification head, without any decoder branch or explicit segmentation output (eg, a "Seg Head"). The tumor segmentation task was performed entirely by an independent U-Net model, and the 2 models operated independently throughout the training and evaluation processes.

Dataset Splitting and Cross-Validation

To evaluate the generalization ability of the models, the dataset was divided into training, validation, and test sets in a ratio of 8:1:1. Random sampling was used to ensure the representativeness of each subset. The 8:1:1 split was chosen based on the sufficiently large sample size (2000 patients), which supported the high-dimensional feature extraction needs of DL models. A larger training set enabled the model to better learn complex nonlinear patterns, while 200 patients each were retained for independent validation and testing. To further improve model stability and generalizability, 5-fold cross-validation was applied during training, with k set to 5. Stratified sampling was used to preserve consistent class distributions across all folds. All models were trained and evaluated on the same data splits, and the average performance metrics across the 5 folds were reported as the final results. This approach ensured robust evaluation, minimized bias, and enhanced the reproducibility of model performance across different subsets.

To evaluate model generalizability, the dataset was divided into training, validation, and testing subsets in an 8:1:1 ratio. All partitions were performed at the patient level, ensuring that all images, slices, and augmented samples from the same patient were assigned exclusively to a single subset to prevent information leakage. Stratified random sampling was applied to maintain consistent class distribution across subsets. To further enhance model robustness and generalization performance, 5-fold cross-validation (k=5) was conducted within the training cohort, also based on patient-level partitioning. This approach ensured the independence of the validation data and the reproducibility of the experimental results.

To avoid data leakage, data augmentation in this study was applied exclusively to the training set, while no augmentation was performed on the validation or test sets. All augmented samples were strictly linked to their original patients and existed only within the training subset, ensuring that none entered the independent test set. Consequently, each test-set patient corresponded to exactly 1 nonaugmented imaging input during inference, which explains why the total count in the confusion matrix matches the original number of patients (n=200).

Model Evaluation and Performance Analysis

Model performance was evaluated using multiple classification metrics, including accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curves, area under the curve (AUC) values, and confusion matrices, to comprehensively assess the predictive accuracy of RT response in RC. Ninety-five percent CIs were calculated for all metrics to strengthen the robustness and interpretability of results.

To statistically compare model performances and verify the relative advantage of the Transformer model, pairwise bootstrap resampling (10,000 iterations) and DeLong's test were performed for AUC comparisons, with P values reported. ROC curves and their confidence intervals were used to visually and

XSL•FO
RenderX

quantitatively compare the classification performance of different models.

Calibration was evaluated on the independent test set by plotting a calibration curve (reliability diagram) using 10 equal-frequency bins (deciles). For each bin, we reported the mean predicted probability and the observed event rate. Overall calibration was quantified with the Brier score (the mean squared error between predicted probabilities and true labels). All calibration analyses were performed once on the test set without any retraining or post hoc recalibration to avoid information leakage. To ensure fairness, the bin boundaries and the evaluation pipeline were kept identical across input settings (MRI-only vs CT + MRI).

During evaluation, particular emphasis was given to evaluating the Transformer model's performance under varying segmentation input conditions. Its prediction results were further analyzed through Gradient-Weighted Class Activation Mapping (Grad-CAM) visualization to interpret attention regions and their clinical relevance. Tumor segmentation performance using the U-Net model was evaluated on the independent test set, with the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU) used as primary metrics. Manually delineated tumor regions, confirmed by 2 experienced radiologists, served as the gold standard. Failure cases were manually analyzed to identify potential factors contributing to segmentation inaccuracies, such as blurred tumor boundaries, low tissue contrast, or irregular tumor morphology. Detailed segmentation outcomes are shown in the "Results" section, and representative case images are provided in multimedia appendices.

Image Data Analysis Method

The image data analysis pipeline in this study consisted of 2 independent stages: tumor segmentation and image-level classification. First, a U-Net segmentation network was used to delineate tumor regions on the raw MRI or CT images, generating binary masks (Figure S1 in Multimedia Appendix 5). This segmentation model served solely to provide ROIs and did not participate in subsequent classification learning. Next, image patches or ROI features extracted from the masked tumor regions were fed into 3 types of classification models, CNNs, GCNs, and a Transformer classifier (Figure S1 in Multimedia Appendix 6), to predict patients' RT response.

It is important to emphasize that the Transformer model used in this study was a pure classification architecture composed only of an encoder and a multilayer perceptron classification head. It did not include any decoder structure, upsampling pathway, or segmentation outputs. Accordingly, it was entirely independent of the U-Net segmentation model used to generate masks; the 2 models shared no parameters and had no information exchange during training or inference. The GCN was used to model spatial relationships within the tumor region, further enhancing classification performance. The overall model evaluation and performance analysis workflow is shown in Figure S1 in Multimedia Appendix 7.

Input Modalities and Multimodal Fusion Strategy

In the primary analysis, the classifier used MRI-only inputs—pre-RT DCE-MRI and T2W sequences—and extracted features or patches or embeddings strictly within expert-annotated tumor masks, thereby minimizing bias from cross-modal acquisition heterogeneity and registration error. CT was used solely for quality control and, when necessary, rigid registration and was not provided to the classifier. Per reviewer request, we additionally implemented late fusion of CT + MRI under the same data split, loss function, class weighting, and hyperparameter settings as the primary analysis. Specifically, patient-level representations were extracted from CT and MRI with identical backbone encoders; after feature alignment and normalization, the representations were concatenated immediately before the classification head and trained or evaluated jointly (all other procedures unchanged). To ensure a fair comparison, the independent test set was kept fixed (n=200; good responders, n=130 and poor responders, n=70); only the input modality configuration was varied.

Statistical Analysis

This study applied various DL techniques to analyze radiomics imaging data from patients with RC. Initially, CNNs were used to extract features from the image data, capturing fine-grained details of tumor tissues. During image segmentation, deep segmentation networks such as U-Net and Mask R-CNN were used to identify and annotate tumor regions, ensuring accurate localization of key lesions. The self-attention mechanism of the Transformer model was used to manage complex spatial relationships, thereby improving the model's pattern recognition capabilities. GCNs were also applied to extract global and local spatial features from the images, effectively modeling both tumor and peritumoral tissue characteristics. By integrating multiple DL models, the study aimed to achieve accurate predictions of RT response and support personalized treatment recommendations.

During model evaluation, various statistical methods were used to comprehensively assess predictive performance. Metrics, including sensitivity, specificity, precision, and recall, were calculated to evaluate classification outcomes across different categories. Confusion matrices were plotted to visualize model performance in distinguishing between positive and negative RT responses. To assess predictive ability across varying thresholds, ROC curves and AUC values were generated for each model. These statistical analyses facilitated a detailed comparison of the strengths and limitations of each DL approach in predicting RT response in patients with RC.

Results

Dataset Sample Distribution Balance and Generalization Ability

In this study, radiomic imaging data from 2000 patients with RC were collected and randomly divided into training (1600 cases), validation (200 cases), and test (200 cases) sets using an 8:1:1 ratio. Approximately 65% (1300/2000) of patients exhibited a favorable response to RT and were classified as the "sensitive" group, while the remaining 35% (700/2000) were categorized as "nonsensitive." As shown in Table 1, we further compared baseline demographic and clinical characteristics—including age, sex, T stage, N stage, tumor

XSL•FO
RenderX

location, and carcinoembryonic antigen levels—between the training and test sets. None of the variables showed statistically significant differences (P>.05). These results indicate that the random partitioning of the dataset was unbiased and that the training and test sets were highly comparable. This balanced distribution provides a reliable foundation for model training and independent validation, thereby enhancing the generalizability and robustness of the proposed approach.

Table . Baseline demographic and clinical characteristics of patients in the training and testing cohorts[a].

Variable	Training cohort (N=1600)	Testing cohort (N=200)	*P* value
Age (years), mean (SD)	58.7 (10.9)	59.1 (11.2)	.62
Sex, n (%)			.53
Male	980 (61.3)	118 (59.0)	
Female	620 (38.7)	82 (41.0)	
T stage, n (%)			.78
T2	430 (26.9)	50 (25.0)	
T3	930 (58.1)	120 (60.0)	
T4	240 (15.0)	30 (15.0)	
N stage, n (%)			.67
N0	610 (38.1)	80 (40.0)	
N1	710 (44.4)	86 (43.0)	
N2	280 (17.5)	34 (17.0)	
Tumor location, n (%)			.79
Upper rectum	310 (19.4)	36 (18.0)	
Midrectum	860 (53.8)	112 (56.0)	
Lower rectum	430 (26.9)	52 (26.0)	
CEA[b] level (ng/mL), median (IQR)	3.6 (2.1-6.8)	3.8 (2.2-7.0)	.73
Response category, n (%)			>.99
Good response	1040 (65.0)	130 (65.0)	
Poor response	560 (35.0)	70 (35.0)	

[a]This table shows the baseline differences between the training and test sets in demographic characteristics (age and sex), tumor clinical staging (T stage and N stage), tumor location, and CEA level. Statistical comparisons were performed using independent-samples *t* tests, chi-square tests, or Mann-Whitney *U* tests, as appropriate. All *P* values were greater than .05, indicating no significant differences between the training and test sets. This demonstrates that the random allocation process was unbiased and that the cohorts were well matched and comparable.

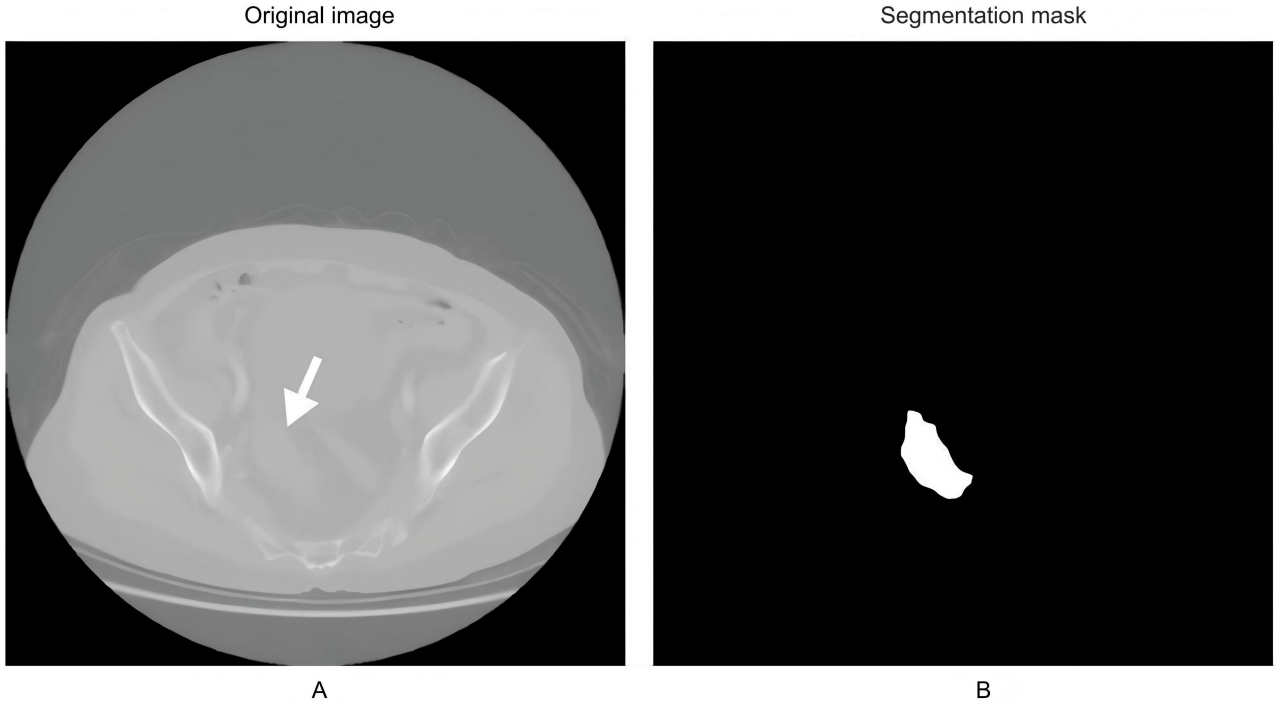
[b]CEA: carcinoembryonic antigen.

High Accuracy of U-Net Model in Tumor Region Segmentation

In this study, a U-Net segmentation network was applied to automatically segment tumor regions in the imaging data of patients with RC. The results demonstrated excellent performance of the model in tumor recognition and segmentation tasks. On the independent test set, the U-Net model achieved a mean DSC of 0.892 (95% CI 0.881 - 0.904) and a mean IoU of 0.814 (95% CI 0.800 - 0.829), indicating high segmentation consistency and robustness. The model precisely delineated tumor boundaries and clearly separated tumor regions from surrounding structures. The original image is shown in Figure 1A, and the corresponding segmentation mask generated by the U-Net model is shown in Figure 1B.

XSL•FO
RenderX

Figure 1. Tumor region segmentation results. (**A**) Original image. The white arrow indicates the rectal tumor lesion. (**B**) Model segmentation result image (tumor region highlighted).



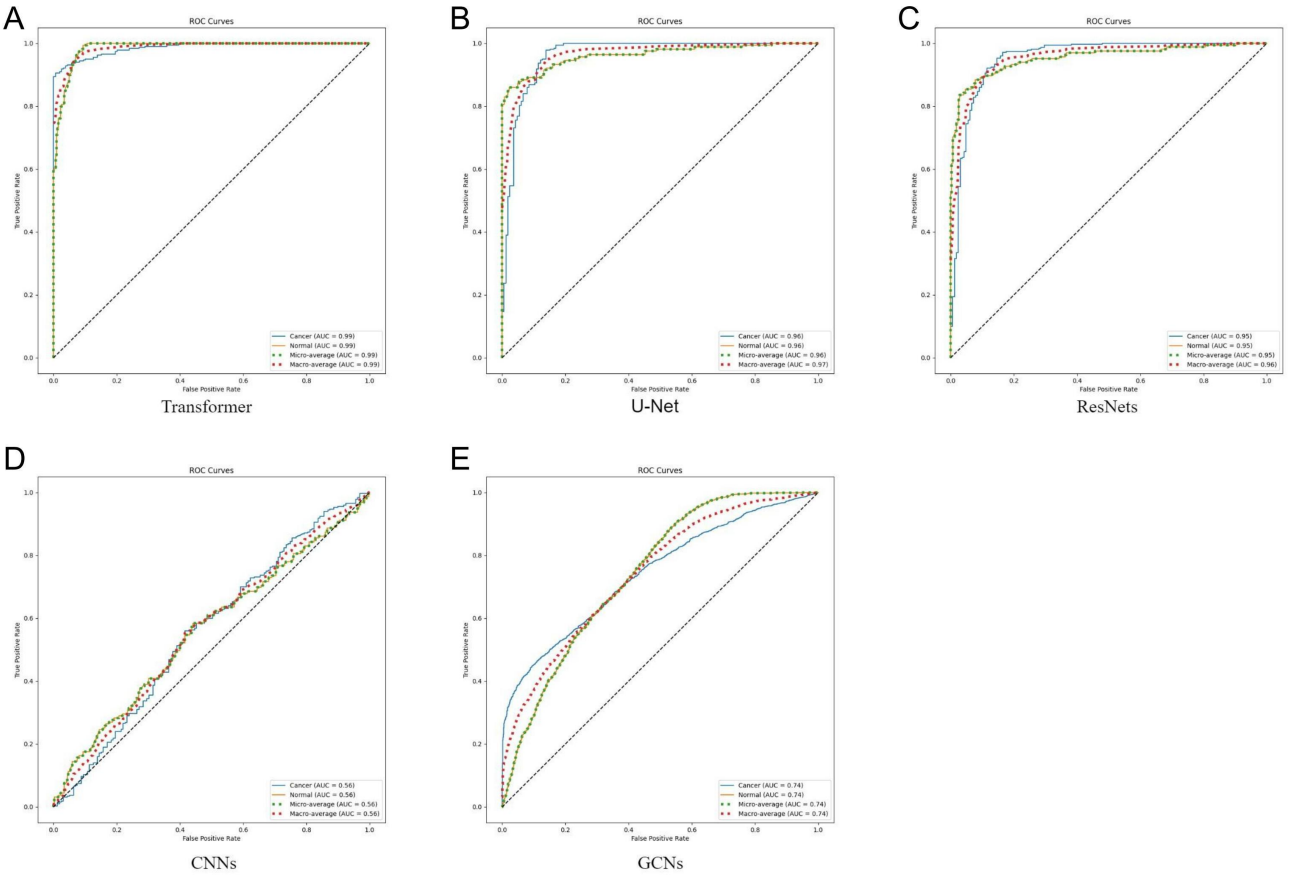
When compared with manually annotated gold-standard segmentations by expert radiologists, the model showed a high degree of agreement in both boundary definition and tumor morphology. Notably, the U-Net model maintained strong performance even when segmenting tumors with irregular shapes or blurred margins. However, a small number of inaccurate segmentations were observed, primarily in cases characterized by indistinct tumor margins, low contrast with adjacent tissues, or atypical tumor morphology. These challenging cases provide directions for future model optimization.

Outstanding Performance of the Transformer Model in Predicting RT Response

This study compared the performance of several DL models in predicting the RT response in patients with RC. The results showed that the Transformer model achieved the highest overall performance, with an accuracy of 87%. As illustrated in Figure 2A, the Transformer leveraged its self-attention mechanism to effectively process complex patterns and fine-grained features within the radiomics imaging data, offering a distinct advantage in capturing global contextual information. In comparison, the CNN (Figure 2D) and GCN (Figure 2E) models achieved accuracies of 82% and 84%, respectively, both lower than that of the Transformer model. Analysis of ROC curves (Figures 2A-2E) further highlighted the Transformer model's superior discriminative ability, as reflected by its AUC value of 0.921 (95% CI 0.901 - 0.945), which exceeded that of the CNN (Figure 2D) and GCN (Figure 2E).

XSL•FO
RenderX

Figure 2. ROC curve comparison of different models. (A) Transformer, (B) U-Net, (C) ResNet, (D) CNN, and (E) GCN. AUC: area under the curve; CNN: convolutional neural network; GCN: graph convolutional network; ROC: receiver operating characteristic.



To enhance the robustness of the results, 95% CIs were calculated for all classification metrics across the models (Table 2). On the independent test set, the Transformer model achieved an accuracy of 87.0% (95% CI 84.2% - 89.5%), sensitivity of 89.2% (95% CI 85.3% - 92.4%), specificity of 82.9% (95% CI 79.2% - 88.1%), and an AUC of 0.921 (95% CI 0.901 - 0.945). In contrast, the CNN model reached an AUC of 0.881 (95% CI 0.854 - 0.906), while the GCN model achieved an AUC of 0.894 (95% CI 0.869 - 0.918).

Table . Performance comparison of deep learning models in predicting radiotherapy response[a].

Model	Accuracy	Sensitivity	Specificity	AUC[b]	*P* value vs Transformer
Transformer	87.0% (84.2%-89.5%)	89.2% (85.3%-92.4%)	82.9% (79.2%-88.1%)	0.921% (0.901%-0.945%)	—[c]
CNN[d]	82.0% (78.9%-85.0%)	83.7% (79.4%-87.6%)	80.2% (75.0%-84.6%)	0.881% (0.854%-0.906%)	.02
GCN[e]	84.0% (80.4%-87.2%)	85.9% (81.7%-89.7%)	82.1% (77.0%-86.4%)	0.894% (0.869%-0.918%)	.041

[a]This table compares the performance of 3 deep learning models (transformer, CNN, and GCN) in predicting radiotherapy response in rectal cancer using an independent test set. All metrics are reported as mean values with 95% CI. The primary evaluation metrics include accuracy, sensitivity, specificity, and the area under the curve (AUC). Differences in AUC were assessed using the DeLong test. The results demonstrate that the Transformer model outperformed both CNN and GCN models across all evaluation metrics, with the differences in AUC reaching statistical significance (P<.05).

[b]AUC: area under the curve.

[c]Not available.

[d]CNN: convolutional neural network.

[e]GCN: graph convolutional network.

To evaluate the statistical significance of AUC differences between the Transformer and other models, pairwise bootstrap resampling (10,000 iterations) and DeLong's test were conducted. The results showed that the Transformer model significantly outperformed both the CNN model (P=.02) and the GCN model (P=.041) in terms of AUC. These findings indicated that the Transformer model provided a statistically significant advantage in identifying imaging features associated with RT response.

XSL•FO
RenderX

Confusion Matrix Reveals the Classification Capability of the Transformer Model

To further assess classification performance, a confusion matrix was constructed to evaluate the Transformer model's predictive ability. As shown in Table 3, among the 200 patients in the test set, the model correctly identified 89% of RT-sensitive patients. The false-positive rate, representing RT-insensitive patients incorrectly classified as sensitive, was only 8%. These results further validated the model's effectiveness across both classes, particularly in correctly identifying RT-insensitive patients. The Transformer model significantly reduced misclassification errors, a critical factor in clinical decision-making. Accurate identification of RT-insensitive patients was essential for enabling timely adjustments to treatment plans, avoiding ineffective RT, and ultimately improving patient prognosis and survival outcomes. On the independent test set, the Transformer's calibration curve closely followed the ideal 45° line, with only mild overconfidence observed in the high–predicted-probability region. The corresponding Brier scores were identical for both input settings: MRI-only, 0.145; and CT + MRI (late fusion), 0.145. Overall calibration was therefore comparable between modalities. Taken together with the AUCs (0.921 vs 0.927) and the DCA results, adding CT did not yield a meaningful improvement in global performance, but it produced a slight and consistent gain in net benefit within higher decision-threshold ranges (Figure 3 and Table 4).

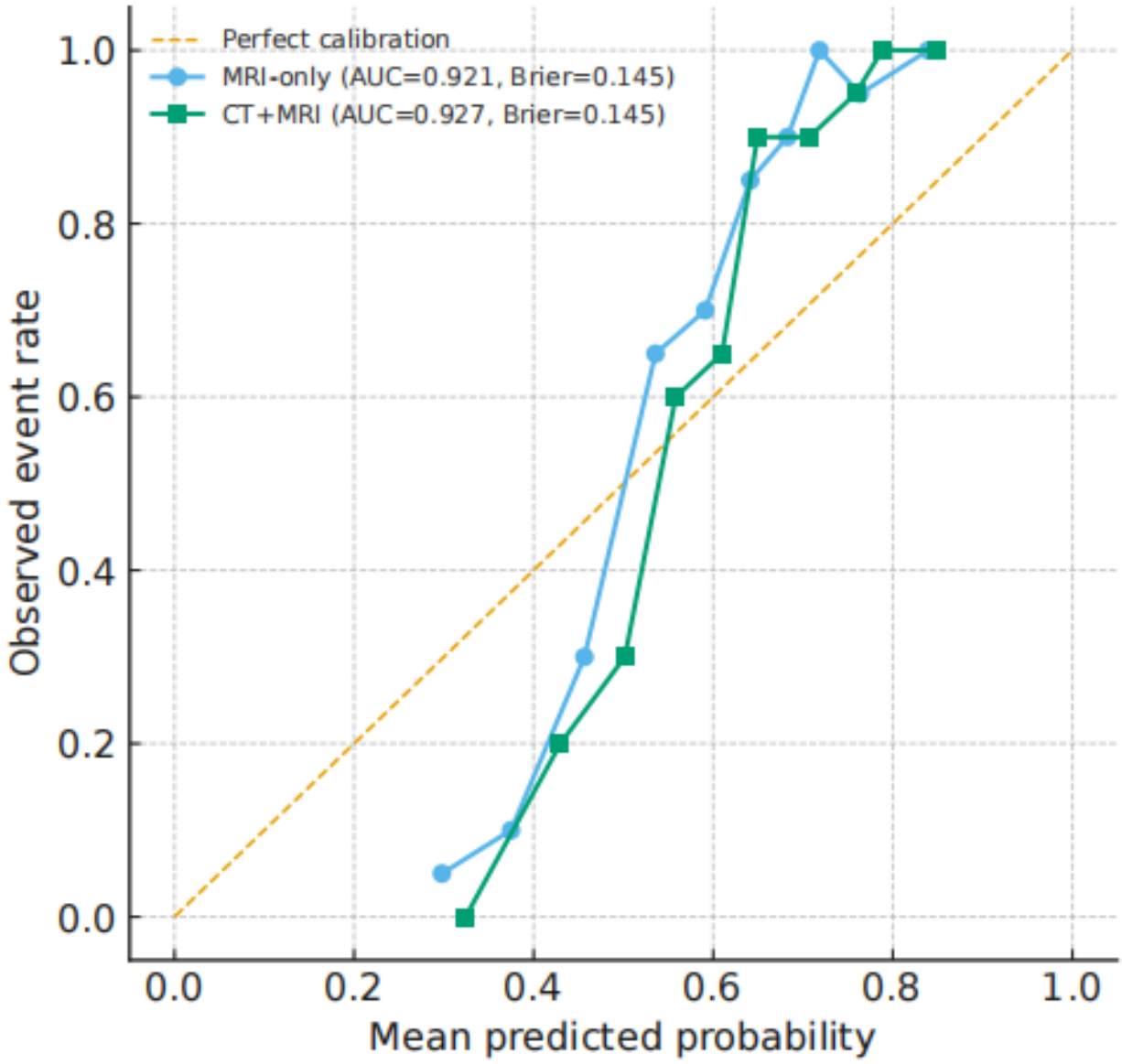
Table . Confusion matrix results for the Transformer model[a].

Category	Sensitive class (good response to radiotherapy)	Insensitive class (poor response to radiotherapy)
Predicted as sensitive class	116 (true positive)	12 (false positive)
Predicted as insensitive class	14 (false negative)	58 (true negative)

[a]This table shows the confusion matrix of the Transformer model on the independent test set (magnetic resonance imaging–only, n=200). True positives represent the number of patients correctly predicted as radiotherapy-sensitive (116 cases); false positives represent the number of patients incorrectly predicted as sensitive but actually nonsensitive (12 cases); false negatives represent the number of patients incorrectly predicted as nonsensitive but actually sensitive (14 cases); and true negatives represent the number of patients correctly predicted as radiotherapy nonsensitive (58 cases). Based on this matrix, the calculated performance metrics were as follows: accuracy=87.0%, sensitivity=89.2%, and specificity=82.9%.

XSL•FO
RenderX

Figure 3. Calibration of the Transformer classifier on the independent test set. Reliability curves of the Transformer classifier on the independent test set (n=200) using 10 equal-frequency bins. The dashed line indicates perfect calibration. Curves are shown for both MRI-only and CT + MRI input settings, with the corresponding AUC and Brier score reported within parentheses. AUC: area under the curve; CT: computed tomography; MRI: magnetic resonance imaging.



XSL•FO
RenderX

Table . Calibration results of the Transformer model under MRI[a]-only and CT[b] + MRI settings[c].

Setting	Bin	Count	pred_mean	obs_rate	pred_min	pred_max
MRI-only	1	20	0.297790855	0.05	0.182508038	0.338796612
MRI-only	2	20	0.374815574	0.1	0.350143687	0.407613655
MRI-only	3	20	0.456552744	0.3	0.411604379	0.488843116
MRI-only	4	20	0.535671487	0.65	0.489127784	0.56889754
MRI-only	5	20	0.591075985	0.7	0.573348236	0.619675996
MRI-only	6	20	0.641167999	0.85	0.620093517	0.662707158
MRI-only	7	20	0.682686695	0.9	0.663585095	0.697369603
MRI-only	8	20	0.718369677	1	0.697816083	0.742269577
MRI-only	9	20	0.763478434	0.95	0.742374832	0.788465779
MRI-only	10	20	0.839248874	1	0.789370839	0.919921412
CT + MRI	1	20	0.324232379	0	0.240873315	0.387211357
CT + MRI	2	20	0.429795288	0.2	0.389460186	0.473106637
CT + MRI	3	20	0.501537256	0.3	0.473698194	0.538277464
CT + MRI	4	20	0.558194525	0.6	0.538663984	0.585177755
CT + MRI	5	20	0.609770763	0.65	0.586338258	0.625575137
CT + MRI	6	20	0.649248771	0.9	0.626019055	0.677019798
CT + MRI	7	20	0.707037992	0.9	0.683165825	0.735004007
CT + MRI	8	20	0.758477804	0.95	0.737737455	0.775617046
CT + MRI	9	20	0.788180535	1	0.7760233	0.80785209
CT + MRI	10	20	0.847920043	1	0.808998069	0.928076363

[a]MRI: magnetic resonance imaging.

[b]CT: computed tomography.

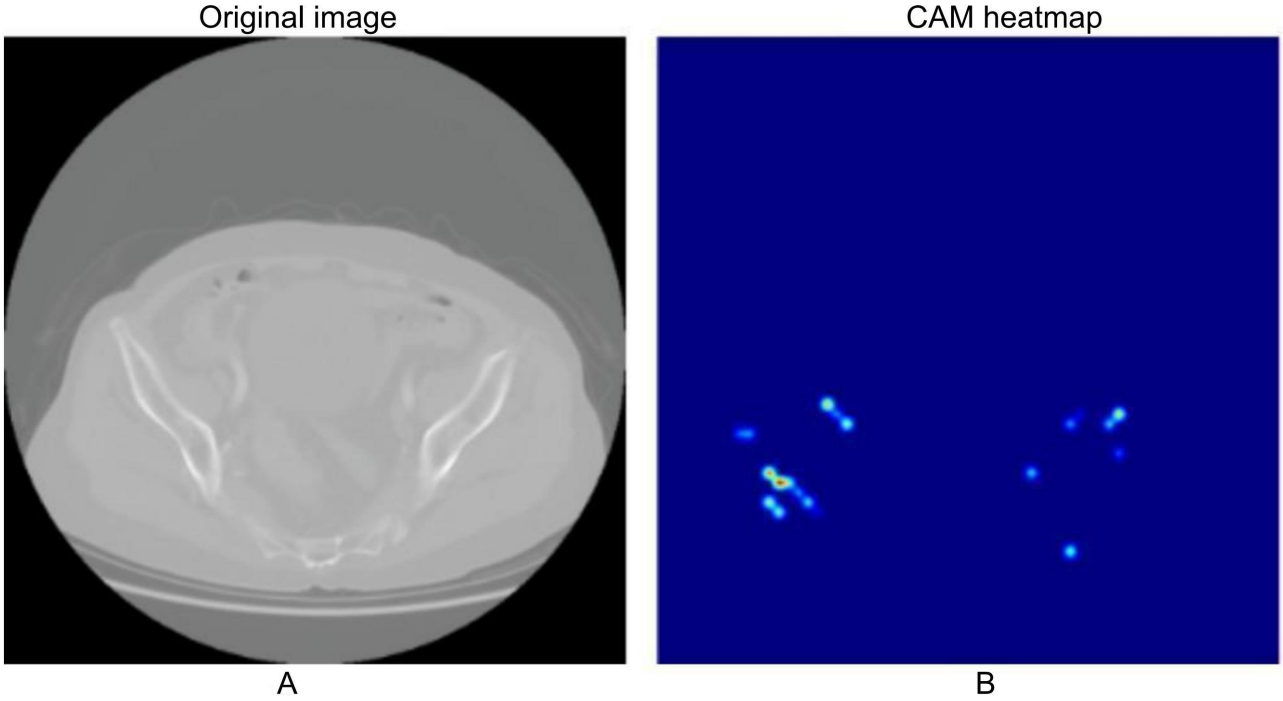
[c]This table shows the calibration binning results of the transformer model on the independent test set under 2 input settings: MRI-only and CT + MRI. The predicted probabilities (pred_mean) were divided into 10 equal-frequency bins (bin), each containing 20 samples. Obs_rate represents the observed positive rate (proportion of radiotherapy-sensitive patients), while pred_min and pred_max indicate the minimum and maximum predicted probabilities within each bin, respectively. The results show that both input settings exhibit calibration trends close to the ideal diagonal, indicating good agreement between predicted probabilities and observed rates. Notably, the CT + MRI setting demonstrated a slight improvement in net benefit within the higher threshold range.

Grad-CAM Visualization to Validate the Accuracy of Feature Extraction

The effectiveness of the Transformer model was reflected not only in its overall predictive performance but also in the depth and precision of its feature extraction. Grad-CAM was used to visualize the model's feature extraction process. The results revealed that the model primarily focused on tumor boundaries and high-density regions within the tumor. As shown in Figures 4A and 4B, the model was particularly adept at capturing complex structures in heterogeneous tumor areas and using them to predict RT response. The visualization results further demonstrated that the model effectively filtered out irrelevant background information and concentrated on diagnostically significant regions. This selective focus enhanced both the classification accuracy and the clinical interpretability of the model. The ability to highlight meaningful anatomical and pathological features supported the Transformer model as a reliable diagnostic aid for clinicians in treatment planning and response prediction.

XSL•FO
RenderX

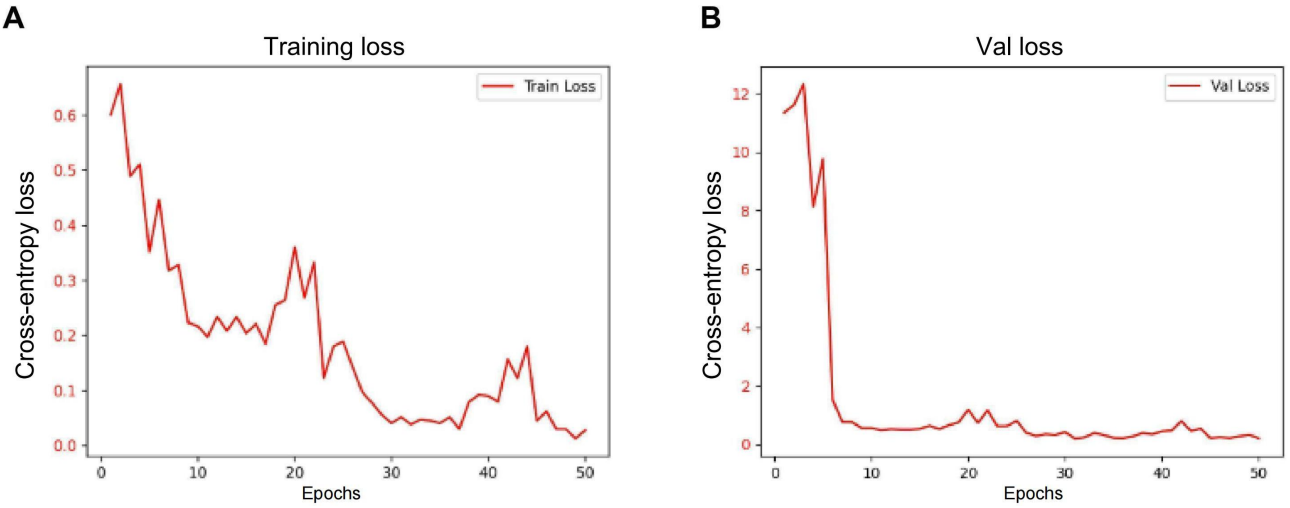
Figure 4. Gradient-Weighted Class Activation Mapping (Grad-CAM) visualization of feature extraction in the Transformer model. (A) Original image. (B) Grad-CAM visualization image: displays the areas where the model focuses, particularly on tumor edges and high-density regions. Red areas represent the regions of high attention by the model. CAM: Class Activation Mapping.



Hyperparameter Optimization and Model Iteration Analysis

During model training and optimization, adjustments to hyperparameters such as learning rate, batch size, and regularization substantially improved the performance of the Transformer model. After 50 training epochs, the validation loss steadily decreased to 0.01, indicating good model convergence and stable training dynamics. As training progressed, the loss curves of the training and validation sets gradually converged, with no apparent signs of overfitting, further confirming the critical role of hyperparameter optimization in enhancing model performance (see Figure 5A for training loss and Figure 5B for validation loss). The model maintained consistently low validation loss and favorable convergence characteristics, establishing a solid foundation for subsequent testing and external validation.

Figure 5. Changes in loss and accuracy for training and validation sets. (A) Loss curve: the loss values for both the training and validation sets decrease gradually over 50 epochs. (B) The validation loss curve (Val loss) indicates that the model stabilized during the later training phase without signs of overfitting.

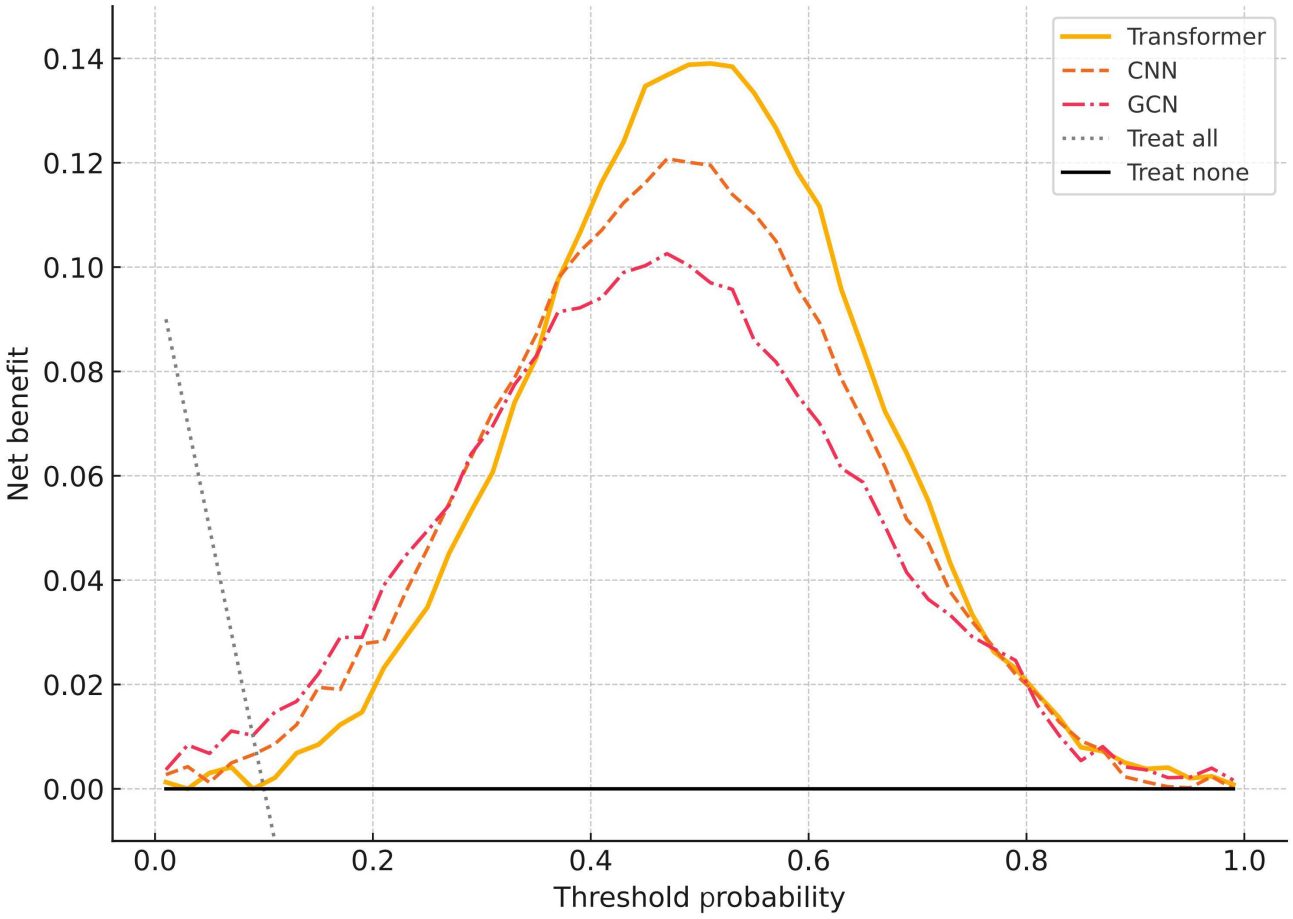


Clinical Benefit Analysis Based on DCA

To further evaluate the potential clinical use of the DL models in RT decision-making, DCA was performed to compare the net benefit of the Transformer, CNN, and GCN models across different probability thresholds (Figure 6). The results showed that across most threshold ranges—particularly between 0.3 and 0.7—the Transformer model achieved a higher net benefit than both the “treat all” and “treat none” strategies, as well as outperforming the CNN and GCN models. This suggests that the Transformer model provides greater decision support capacity in clinical practice.

XSL•FO
RenderX

Figure 6. Decision curve analysis. The figure shows the comparison of net benefit across different probability thresholds for the Transformer, CNN, and GCN models. The dashed lines represent the "treat all" and "treat none" strategies, respectively. The Transformer model demonstrated higher net benefits within the primary clinical decision-making range of 0.3 - 0.7, highlighting its clinical use in predicting radiotherapy sensitivity. CNN: convolutional neural network; GCN: graph convolutional network.



For example, when predicting RT-sensitive patients, the Transformer model maintained a high true-positive rate while significantly reducing the likelihood of unnecessary RT and associated side effects. These findings highlight the model's potential as a valuable tool for individualized RT decision-making.

Sensitivity Analysis of Imaging Input Settings (MRI-Only Versus CT + MRI)

To quantify the incremental value of CT, we compared MRI-only with late-fusion CT + MRI under identical training or validation procedures (same splits, loss, class weighting, and hyperparameters). Using the Transformer as an exemplar, overall discrimination on the independent test set was similar between modalities: Transformer (MRI-only): accuracy 87.0%, AUC 0.921 (95% CI 0.901 - 0.945), sensitivity 89.2% (true positive 116/130), specificity 82.9% (true negative 58/70), Brier score 0.145, and area under the precision–recall curve 0.900; Transformer (CT + MRI): accuracy 87.5%, AUC 0.926 (95% CI 0.905 - 0.948), sensitivity 90.0% (true positive 117/130), specificity 82.9% (true negative 58/70), Brier score 0.142, and area under the precision–recall curve 0.905.

DeLong's test indicated no significant difference in AUC between the 2 settings (P=.36). In DCA, CT + MRI yielded a slight and consistent gain in net benefit at higher threshold probabilities (>0.6), whereas curves largely overlapped within the primary clinical range (0.3 - 0.7). Overall, given a unified MRI acquisition protocol in this single-center cohort, MRI-only achieved robust performance; adding CT did not confer a statistically significant global improvement but may offer marginal clinical value at specific, higher decision thresholds. For completeness, test-set AUCs (95% CI) across all 3 architectures (CNN, GCN, and Transformer) under both input settings are summarized in Table 5, and full classification metrics for the Transformer are provided in Table 6.

XSL•FO
RenderX

Table . Comparison of the Transformer model performance under different input settings[a].

Input setting	Accuracy	AUC	95% CI (AUC[b])	Sensitivity	Specificity	Brier	AUCPR[c]	DeLong test *P* value
MRI[d]-only	87.00%	0.921	0.901 - 0.945	89.2% (116/130)	82.9% (58/70)	0.145	0.9	—[e]
CT + MRI (late fusion)	87.50%	0.926	0.905 - 0.948	90.0% (117/130)	82.9% (58/70)	0.142	0.905	.36

[a]This table compares the predictive performance of the Transformer model under 2 input settings: magnetic resonance imaging (MRI)–only and CT + MRI (late fusion). The main evaluation metrics include accuracy, area under the curve (AUC) with its 95% CI, sensitivity, specificity, Brier score, area under the precision–recall curve (AUCPR), and the DeLong test *P* value for AUC comparison. The results indicate that the overall discriminative performance of the 2 input settings was comparable (AUC difference, *P*=.36). The CT + MRI configuration showed slightly better Brier score and AUCPR at higher threshold ranges, suggesting a minor net benefit under certain clinical decision thresholds.

[b]AUC: area under the curve.

[c]AUCPR: area under the precision–recall curve.

[d]MRI: magnetic resonance imaging.

[e]Not available.

Table . Comparison of AUC[a] values across architectures under MRI[b]-only and CT[c] + MRI settings[d].

Architecture	MRI-only AUC (95% CI)	CT + MRI AUC (95% CI)	DeLong test *P* value
CNN[e]	0.881 (0.854 - 0.906)	0.888 (0.861 - 0.912)	.22
GCN[f]	0.894 (0.869 - 0.918)	0.899 (0.873 - 0.922)	.29
Transformer	0.921 (0.901 - 0.945)	0.926 (0.905 - 0.948)	.36

[a]AUC: area under the curve.

[b]MRI: magnetic resonance imaging.

[c]CT: computed tomography.

[d]This table compares the area under the curve (AUC) performance and corresponding 95% CI of 3 deep learning models—CNN, GCN, and Transformer—under 2 input conditions: MRI-only and CT + MRI. Differences in AUC were evaluated using the DeLong test. The results show that all 3 models exhibited slight improvements in AUC under the CT + MRI condition; however, none of the differences reached statistical significance (*P*>.05). Among the 3, the Transformer model achieved the highest AUC in both settings (MRI-only: 0.921; CT + MRI: 0.926).

[e]CNN: convolutional neural network.

[f]GCN: graph convolutional network.

Computational Efficiency and Inference Time Analysis

To quantitatively evaluate the computational overhead and deployment feasibility of different models in clinical scenarios, we further compared CNNs, GCNs, and Transformer-based models in terms of parameter count, theoretical computational complexity (Floating-Point Operations [FLOPs]), and inference time on an independent test set. Inference-time evaluation was conducted under a fixed hardware configuration (GPU: NVIDIA RTX 3090, 24 GB memory), using FP32 precision, with synchronized timing performed after model warm-up. The timing scope included model forward inference and patient-level probability aggregation (ie, arithmetic averaging of predicted probabilities across all 2D slices belonging to the same patient), while offline image preprocessing and data-loading overheads were excluded.

As summarized in Table 7, the Transformer model exhibits a larger parameter size (86.4 M) and higher FLOPs (17.8 G) than the CNN (ResNet-34) and GCN architectures, reflecting its increased model complexity. Nevertheless, on the test set, the Transformer model achieves an average per-slice inference time of 3.8 milliseconds. At the patient level (median number of slices: 20, IQR 16 - 24), the overall inference time is 78 (IQR 63-95) milliseconds per patient, which remains well within the acceptable range for real-time clinical decision support. By comparison, the patient-level inference times of the CNN and GCN models are 42 milliseconds and 55 milliseconds, respectively.

XSL•FO
RenderX

Table . Computational efficiency and inference latency of the evaluated models[a].

Model	Parameters, millions	FLOPs[b], billions	Inference time per slice (ms), mean (SD)	Average slices per patient, median (IQR)	Inference time per patient (ms)
CNN[c] (ResNet-34)	21.8	3.6	2.1 (0.3)	20 (16-24)	42 (34 - 50)
GCN[d]	28.5	5.2	2.7 (0.4)	20 (16-24)	55 (44 - 66)
Transformer	86.4	17.8	3.8 (0.5)	20 (16-24)	78 (63 - 95)

[a]Parameter counts and FLOPs were calculated based on a single region of interest input. Inference time was measured on an independent test set (n=200) and includes model forward inference as well as patient-level prediction probability aggregation, while offline image preprocessing and data-loading overheads were excluded. All results were obtained under identical hardware and software environments.

[b]FLOPs: Floating-Point Operations.

[c]CNN: convolutional neural network.

[d]GCN: graph convolutional network.

These results indicate that although the Transformer architecture incurs a higher computational cost than conventional convolutional models, its inference latency remains low. Moreover, when weighed against its substantial advantages in predictive performance and clinical net benefit, the increased computational overhead is acceptable for practical clinical applications.

Discussion

Principal Findings

In this study, a DL-based radiomics model was developed to predict RT response in patients with RC. A systematic comparison of multiple models, including CNNs, GCNs, and Transformers, revealed that the Transformer architecture demonstrated the best overall performance. By leveraging its self-attention mechanism, the Transformer effectively captured complex imaging features within key tumor regions, achieving an accuracy of 87% and an AUC of 0.921 on the test set. In addition to its predictive accuracy, the Transformer model demonstrated strong interpretability, as confirmed by Grad-CAM visualization, which highlighted its focus on clinically relevant tumor regions. These findings underscore the model's potential as a reliable and interpretable tool for supporting individualized RT planning in clinical settings.

During training, 5-fold cross-validation (k=5) was adopted based on the following considerations. First, it balanced statistical stability and sample efficiency. In this cohort (n=2000; class ratio ≈65%/35%), k=5 ensured approximately 200 validation cases per fold, providing sufficient positive and negative samples to stably estimate AUC, calibration, and DCA metrics. Larger k would yield smaller validation subsets, increasing variability in calibration and DCA estimates, whereas smaller k would raise estimation bias and reduce the resolution of hyperparameter selection. Second, it accounted for computational cost and practical reproducibility: compared with 10-fold cross-validation, 5 folds substantially reduced training time and resource requirements, making replication in a clinical research setting more feasible. To assess robustness, we conducted sensitivity analyses using k=3 and k=10 and repeated 5-fold cross-validation with different random seeds under the same patient-level stratification. Across all settings, the relative ranking of the Transformer and baseline models remained consistent, fold-to-fold AUC variability was comparable with the main analysis, and no systematic instability or overfitting was observed. Compared with 5 folds, 10 folds primarily increased training time and triggered more frequent early stopping without materially affecting final test performance. Three-folds showed slightly greater variance in validation metrics and reduced hyperparameter discriminability. In summary, k=5 provided an optimal balance among performance stability, computational feasibility, and methodological rigor. To avoid information leakage, cross-validation was performed exclusively within the training set using stratified, patient-level splits; all images, slices, and augmented samples from the same patient were confined to a single fold. The independent test set was reserved for a one-time final evaluation only.

Comparison With Prior Work

RT remains a cornerstone in RC treatment; however, predicting individual response to RT continues to be a major clinical challenge due to significant patient heterogeneity [19]. Radiomics, an emerging technology capable of extracting high-dimensional features from medical images, has opened new avenues for predicting treatment response [20]. However, traditional machine learning approaches have shown limited effectiveness in modeling the high-dimensional, heterogeneous nature of radiomics data [21]. The advent of DL, particularly CNNs, has markedly improved pattern recognition capabilities in this domain [7,9,22]. Unlike prior studies that applied single-model approaches, our work systematically compared 3 advanced DL architectures and identified the Transformer as the most effective model for complex image-based RT response prediction [23].

Several recent studies have attempted to integrate DL into tumor-imaging analysis and RT response prediction. For instance, Xu et al [24] applied CNNs for breast region segmentation using DCE-MRI in breast cancer, demonstrating strong performance in local feature extraction. However, their approach was limited to 2D analysis and failed to account for the complexity of 3D imaging data [24]. Similarly, Mzoughi et al [25] proposed a multiscale 3D CNN for glioma classification based on MRI, achieving approximately 85% accuracy, but still faced limitations in feature representation and generalizability.

XSL•FO
RenderX

Previous studies have systematically evaluated MRI-based radiomics and DL approaches for predicting response to nCRT in RC. A multicenter study published in *Radiology* demonstrated that MRI radiomics models could predict pathological complete response (pCR) with relatively high accuracy [26]. DL models based on post-nCRT T2W MRI also achieved discriminative performance for pCR or GR that was comparable to, or even surpassed, expert radiologist assessment [27]. Beyond single time point models, delta radiomics—which captures changes in features before and after treatment—showed superior performance in predicting nCRT response [28]. In terms of risk stratification and clinical translation, late-fusion multimodal models that integrated multiparametric MRI–derived DL or radiomics features with clinical variables outperformed unimodal models and provided greater clinical net benefit on DCA [18]. Furthermore, several systematic reviews and meta-analyses confirmed the robustness of imaging AI for predicting pCR, reporting pooled AUCs of approximately 0.90 - 0.91; models based on DL or clinical fusion generally outperformed conventional radiomics approaches [29]. These findings are consistent with the results of this study and support our design rationale of introducing more advanced architectures (eg, Transformers) and evaluating clinical use using tools such as DCA.

In contrast, our Transformer-based model achieved an accuracy of 87% and an AUC of 0.921, outperforming the CNN-based approaches in both precision and discriminative capacity. The multihead self-attention mechanism of the Transformer effectively captured global dependencies within tumor regions, demonstrating enhanced robustness and generalizability, especially in cases with complex morphology or blurred tumor boundaries. Compared with previously reported models, our approach offered distinct advantages in architectural innovation, predictive performance, and clinical interpretability, positioning it as a more reliable framework for RT response prediction.

Additionally, the U-Net model also demonstrated satisfactory performance in tumor segmentation, achieving a mean DSC of 0.892 and an IoU of 0.814 on the independent test set, indicating strong consistency with expert annotations. These results underscore the crucial role of accurate segmentation in radiomics workflows, as it ensures high-quality inputs for downstream feature extraction and classification, even in tumors with irregular edges.

Furthermore, Grad-CAM visualization revealed that the Transformer model concentrated on tumor edges and regions of heterogeneous density, areas known to correlate with RT sensitivity [20,21]. This interpretability enhanced the interpretability and clinical credibility of the model.

Limitations

Despite the promising results, this study had several limitations. First, the dataset was derived from a single center, with limited diversity in race, age, and pathological subtypes, which may have constrained the model's generalizability [20]. Although data augmentation and k-fold cross-validation were applied to mitigate this limitation, future studies should incorporate multicenter datasets to assess the model's robustness across broader populations. Second, the imaging data primarily consisted of DCE-MRI, without the integration of additional clinical variables such as genomic data, pathological classification, or treatment history. This single-modality approach may have limited the comprehensiveness of the predictive model [7]. Incorporating multimodal data in future models could improve both predictive accuracy and biological interpretability. Third, although Grad-CAM was used to enhance model interpretability, it functioned as a post hoc visualization method and did not directly reveal the underlying causal mechanisms driving model predictions. Future research that integrates molecular biomarkers and longitudinal follow-up data may help elucidate the biological basis of model outputs and strengthen their interpretability. Fourth, model performance evaluation was conducted on a single-center dataset and validated only on an internal holdout test set comprising 10% of the cohort. External multicenter validation was not performed. Although cross-validation and regularization techniques were used to reduce overfitting, radiomics models are known to be sensitive to cross-center variations—such as differences in imaging parameters, scanner models, and acquisition protocols—which may lead to substantial domain shift in real-world applications. Future studies should incorporate external datasets from multiple institutions to systematically assess robustness, generalizability, and clinical applicability. Fifth, expert-annotated tumor masks were used in the test set to avoid data leakage; however, this evaluation setup represents an "idealized" condition that does not fully reflect real-world clinical workflows. In practice, errors from automatic segmentation networks may propagate to the Transformer classifier, potentially leading to lower end-to-end performance than the manually masked results reported here. Future work will include validation of a fully automated pipeline and quantitative assessment of how segmentation inaccuracies affect downstream prediction tasks. Sixth, MRI slice thickness varied between 3 mm and 4 mm across patients. Although this study adopted a 2D slice–level modeling strategy, in which the inputs primarily rely on in-plane texture and morphological information, and volumetric resampling along the *z*-axis was intentionally avoided to prevent interpolation-induced artifacts, we acknowledge that differences in slice thickness and spatial resolution may still affect the statistical distributions of intensity and texture features, thereby constituting a potential confounding factor in radiomics-based analyses. We performed patient-level splitting for training, validation, and testing and used data augmentation and cross-validation to mitigate instability arising from acquisition-related heterogeneity; however, such effects cannot be completely eliminated. Future work will further explore standardized resampling strategies or explicit modeling of resolution differences (eg, stratification or covariate adjustment) and will incorporate 3D models and multicenter external validation to systematically evaluate the impact of slice thickness and scanner or protocol variability on model generalizability. Seventh, this study combined CR and PR patients into a single "GR" category to improve class balance, a strategy commonly used in statistical modeling. However, this grouping does not optimally reflect clinical practice: "watch-and-wait" strategies are typically suitable only for patients achieving clinical CR or pCR, whereas PR patients still require further treatment. While the binary classification

XSL•FO
RenderX

framework in this study aimed primarily to demonstrate the DL model's ability to predict overall RT response, future research should investigate multiclass models to enhance clinical interpretability and applicability. Eighth, although the Transformer model demonstrated superior predictive performance compared with CNN and GCN models, this improvement comes at the cost of increased computational burden. Transformers generally require larger parameter counts and higher floating-point computation, with the self-attention mechanism imposing substantial overhead when processing high-resolution image patches. To address this concern, we have supplemented the revised manuscript with a detailed comparison of model parameter counts, FLOPs, and patient-level inference time, thereby quantitatively assessing the computational efficiency of different architectures. The results demonstrate that despite its higher computational cost relative to ResNet-34 and the GCN model, the Transformer maintains a patient-level inference time within the acceptable range for real-time clinical decision support, indicating its practical feasibility for deployment under the current experimental settings. Nevertheless, in larger-scale datasets or resource-constrained clinical environments, model complexity and deployment cost may still pose limiting factors. Future work will therefore explore model compression strategies, such as knowledge distillation, structural pruning, and parameter sharing, to further reduce inference latency and computational resource consumption while preserving predictive performance, thereby enhancing the applicability of the model in real-world clinical workflows.

Future Directions

Future research should aim to extend the application of the Transformer-based radiomics model beyond RC, exploring its transferability and adaptability to other solid tumors such as cervical, lung, and esophageal cancers, in order to evaluate cross-cancer generalizability [15,30]. Another promising avenue involves integrating radiomics with additional data modalities, including genomic, transcriptomic, immunomic, and digital pathology features, to develop comprehensive multimodal fusion models. Such models have the potential to enhance predictive accuracy and provide deeper insights into the biological mechanisms underlying treatment response [21]. Furthermore, the development of lightweight and resource-efficient model versions suitable for deployment on edge devices will be critical for facilitating real-world clinical translation, enabling the routine use of DL-based radiomics models at the point of care to support individualized treatment planning and decision-making.

Conclusions

This study successfully constructed a Transformer-based DL radiomics model for predicting RT sensitivity in patients with RC. The results demonstrate that when using pre-RT MRI features as the primary input, the proposed model achieves superior performance in terms of predictive accuracy, interpretability, and model stability, thereby providing a feasible tool for pretreatment risk stratification and personalized RT decision-making. Supplementary analyses indicate that the multimodal integration of CT and MRI does not yield a statistically significant improvement in overall discriminative performance; however, it may offer a modest and consistent clinical net benefit at specific decision thresholds. Further validation using multicenter datasets is warranted, and future studies should more clearly delineate the potential value of multimodal information across different clinical application scenarios.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Disclaimer

No generative AI tools were used in the preparation, writing, analysis, or submission of this manuscript.

Data Availability

To enhance the reproducibility of this study, the complete implementation code, including scripts for data preprocessing, feature extraction, model training, and inference, will be made publicly available on GitHub upon publication. All data processing workflows have been automated and standardized to ensure that external users can replicate the model training process under identical data conditions. Considering patient privacy and institutional policy restrictions, the trained model weight files will be provided only upon reasonable academic request and limited strictly to noncommercial research purposes.

Authors' Contributions

YL and HL contributed equally to this work. YL, HL, and QW designed the study and performed data analysis. ZW and ZH were responsible for data collection and curation. YL developed the deep learning models and conducted experiments. ZH supervised the project and provided critical revisions. All authors contributed to the manuscript writing and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Patient inclusion flowchart for rectal cancer. The flowchart illustrates the inclusion and exclusion process. A total of 2450 patients were screened; 200 were excluded for missing imaging data, 120 for poor image quality, 80 for lack of radiotherapy outcome labels, and 50 for not receiving standard radiotherapy. Ultimately, 2000 patients were included for model analysis, with 1300 classified as radiotherapy-sensitive and 700 as radiotherapy-insensitive based on follow-up evaluations.
[PDF File, 345 KB - medinform_v14i1e77313_app1.pdf]

Multimedia Appendix 2
Scanner models and acquisition parameters used in the study.
[DOCX File, 12 KB - medinform_v14i1e77313_app2.docx]

Multimedia Appendix 3
Data collection and preprocessing flowchart. These figures illustrate the entire process from data collection, data anonymization, formatting, normalization, and data augmentation to the final dataset preparation. They depict data preprocessing steps such as denoising, cropping, normalization, and resolution adjustment.
[PDF File, 62 KB - medinform_v14i1e77313_app3.pdf]

Multimedia Appendix 4
Workflow of deep learning model training and validation. This figure illustrates the overall framework of the deep learning analysis pipeline used in this study, including data preprocessing, tumor region segmentation, feature extraction, model training with cross-validation, and performance evaluation. First, preradiotherapy images were cleaned, normalized, and augmented to generate standardized inputs. Subsequently, a U-Net model was used for automatic tumor segmentation, and the resulting masks were used to extract tumor region features from the magnetic resonance images. The Transformer model then performed classification training based on patch embeddings, leveraging multihead self-attention to capture global feature patterns. Five-fold cross-validation was adopted during training to enhance model robustness and generalizability. Finally, performance metrics, including accuracy, sensitivity, specificity, and area under the curve, were computed. This workflow provides a comprehensive overview of the deep learning analysis pathway from raw imaging data to model prediction.
[PNG File, 59 KB - medinform_v14i1e77313_app4.png]

Multimedia Appendix 5
U-shaped convolutional neural network (U-Net) model tumor segmentation process diagram. This diagram illustrates how the U-Net segmentation model extracts features layer by layer through convolution operations, ultimately segmenting the tumor region. It includes the original image input, convolution process, feature extraction, and segmentation results.
[PNG File, 70 KB - medinform_v14i1e77313_app5.png]

Multimedia Appendix 6
Schematic architecture of the transformer classification model. This figure illustrates the overall workflow of the Transformer-based classification model used to predict radiotherapy response in rectal cancer. First, pretreatment magnetic resonance images are processed using tumor-region masks (masked region of interest). The masked images are then divided into fixed-size 16×16 patches, which are linearly projected to obtain patch embeddings. The model consists of 12 Transformer encoder layers, where multihead self-attention mechanisms capture global contextual features. During classification, a dedicated classification token is passed through a multilayer perceptron (MLP head) to produce the final prediction of radiotherapy response (sensitive vs nonsensitive). This architecture operates exclusively at the image level and does not include any segmentation decoder, remaining fully independent from the U-Net segmentation model used elsewhere in the study.
[PNG File, 33 KB - medinform_v14i1e77313_app6.png]

Multimedia Appendix 7
Model evaluation and performance analysis flowchart. This figure illustrates the entire model evaluation process, including the calculation of accuracy, sensitivity, specificity, receiver operating characteristic curve, and area under the curve values, and how these metrics are used to assess the model's classification ability.
[PNG File, 36 KB - medinform_v14i1e77313_app7.png]

References

1. Barbieri P, Posa A, Lancellotta V, et al. Electrochemotherapy in the locoregional treatment of metastatic colorectal liver metastases: a systematic review. Curr Oncol 2024 Nov 20;31(11):7403-7413. [doi: 10.3390/curroncol31110546] [Medline: 39590176]
2. Ocanto A, Teja M, Amorelli F, et al. Landscape of biomarkers and pathologic response in rectal cancer: where we stand? Cancers (Basel) 2024 Dec 2;16(23):4047. [doi: 10.3390/cancers16234047] [Medline: 39682232]

XSL•FO
RenderX

3. Wojakowska A, Marczak L, Zeman M, et al. Proteomic and metabolomic signatures of rectal tumor discriminate patients with different responses to preoperative radiotherapy. Front Oncol 2024;14:1323961. [doi: 10.3389/fonc.2024.1323961] [Medline: 38410100]
4. Ellakwa DES, Mushtaq N, Khan S, et al. Molecular functions of microRNAs in colorectal cancer: recent roles in proliferation, angiogenesis, apoptosis, and chemoresistance. Naunyn-Schmiedeberg's Arch Pharmacol 2024 Aug;397(8):5617-5630. [doi: 10.1007/s00210-024-03076-w]
5. Mendes B, Domingues I, Santos J. Radiomic pipelines for prostate cancer in external beam radiation therapy: a review of methods and future directions. J Clin Med 2024 Jul 3;13(13):3907. [doi: 10.3390/jcm13133907] [Medline: 38999473]
6. Astaraki M, Häger W, Lazzeroni M, Toma-Dasu I. Radiomics and deep learning models for glioblastoma treatment outcome prediction based on tumor invasion modeling. Phys Med 2025 Jan;129:104881. [doi: 10.1016/j.ejmp.2024.104881] [Medline: 39724784]
7. Deasy JO. Data science opportunities to improve radiotherapy planning and clinical decision making. Semin Radiat Oncol 2024 Oct;34(4):379-394. [doi: 10.1016/j.semradonc.2024.07.012] [Medline: 39271273]
8. Wu M, Yan C, Sen G. Computer-aided diagnosis of hepatic cystic echinococcosis based on deep transfer learning features from ultrasound images. Sci Rep 2025;15(1). [doi: 10.1038/s41598-024-85004-9]
9. Zhang H, Dohopolski M, Stojadinovic S, et al. Multiomics-based outcome prediction in personalized ultra-fractionated stereotactic adaptive radiotherapy (PULSAR). Cancers (Basel) 2024 Oct 9;16(19):3425. [doi: 10.3390/cancers16193425] [Medline: 39410044]
10. Toba S, Mitani Y, Yodoya N, et al. Prediction of pulmonary to systemic flow ratio in patients with congenital heart disease using deep learning-based analysis of chest radiographs. JAMA Cardiol 2020 Apr 1;5(4):449-457. [doi: 10.1001/jamacardio.2019.5620] [Medline: 31968049]
11. Ran J, Li D. A faster privacy-preserving medical image diagnosis scheme with machine learning. J Digit Imaging Inform Med 2025;38(5):2716-2728. [doi: 10.1007/s10278-024-01384-4]
12. Alazwari S, Alsamri J, Alamgeer M, Alotaibi SS, Obayya M, Salama AS. Automated gall bladder cancer detection using artificial gorilla troops optimizer with transfer learning on ultrasound images. Sci Rep 2024 Sep 19;14(1):21845. [doi: 10.1038/s41598-024-72880-4] [Medline: 39300284]
13. Thompson HM, Kim JK, Jimenez-Rodriguez RM, Garcia-Aguilar J, Veeraraghavan H. Deep learning-based model for identifying tumors in endoscopic images from patients with locally advanced rectal cancer treated with total neoadjuvant therapy. Dis Colon Rectum 2023;66(3):383-391. [doi: 10.1097/DCR.0000000000002295]
14. Wichtmann BD, Albert S, Zhao W, et al. Are we there yet? The value of deep learning in a multicenter setting for response prediction of locally advanced rectal cancer to neoadjuvant chemoradiotherapy. Diagnostics (Basel) 2022 Jun 30;12(7):1601. [doi: 10.3390/diagnostics12071601] [Medline: 35885506]
15. Fu J, Zhong X, Li N, et al. Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer. Phys Med Biol 2020 Apr 2;65(7):075001. [doi: 10.1088/1361-6560/ab7970] [Medline: 32092710]
16. Kawamura M, Kamomae T, Yanagawa M, et al. Revolutionizing radiation therapy: the role of AI in clinical practice. J Radiat Res 2024 Jan 19;65(1):1-9. [doi: 10.1093/jrr/rrad090]
17. Yang M, Yang M, Yang L, et al. Deep learning for MRI lesion segmentation in rectal cancer. Front Med 2024;11. [doi: 10.3389/fmed.2024.1394262]
18. Li Z, Qin Y, Liao X, et al. Comparison of clinical, radiomics, deep learning, and fusion models for predicting early recurrence in locally advanced rectal cancer based on multiparametric MRI: a multicenter study. Eur J Radiol 2025 Aug;189:112173. [doi: 10.1016/j.ejrad.2025.112173] [Medline: 40403678]
19. Kaval G, Dagoglu Kartal MG, Azamat S, et al. Evaluating complete response prediction rates in locally advanced rectal cancer with different radiomics segmentation approaches. Pathol Oncol Res 2024;30:1611744. [doi: 10.3389/pore.2024.1611744] [Medline: 38694706]
20. Nardone V, Reginelli A, Rubini D, et al. Delta radiomics: an updated systematic review. Radiol Med 2024 Aug;129(8):1197-1214. [doi: 10.1007/s11547-024-01853-4] [Medline: 39017760]
21. Kwiatkowska-Miernik A, Wasilewski PG, Mruk B, Sklinda K, Bujko M, Walecki J. Estimating progression-free survival in patients with primary high-grade glioma using machine learning. J Clin Med 2024 Oct 16;13(20):6172. [doi: 10.3390/jcm13206172] [Medline: 39458122]
22. Ramireddy JK, Sathya A, Sasidharan BK, et al. Can pretreatment MRI and planning CT radiomics improve prediction of complete pathological response in locally advanced rectal cancer following neoadjuvant treatment? J Gastrointest Canc 2024 Sep;55(3):1199-1211. [doi: 10.1007/s12029-024-01073-z]
23. Zhu Y, Liu T, Chen J, Wen L, Zhang J, Zheng D. Prediction of therapeutic response to transarterial chemoembolization plus systemic therapy regimen in hepatocellular carcinoma using pretreatment contrast-enhanced MRI based habitat analysis and Crossformer model. Abdom Radiol 2024;50(6):2464-2475. [doi: 10.1007/s00261-024-04709-7]
24. Xu X, Fu L, Chen Y, et al. Breast region segmentation being convolutional neural network in dynamic contrast enhanced MRI. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:750-753. [doi: 10.1109/EMBC.2018.8512422] [Medline: 30440504]

XSL•FO
RenderX

25. Mzoughi H, Njeh I, Wali A, et al. Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. J Digit Imaging 2020 Aug;33(4):903-915. [doi: 10.1007/s10278-020-00347-9] [Medline: 32440926]
26. Shin J, Seo N, Baek SE, et al. MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy. Radiology 2022 May;303(2):351-358. [doi: 10.1148/radiol.211986] [Medline: 35133200]
27. Jang BS, Lim YJ, Song C, et al. Image-based deep learning model for predicting pathological response in rectal cancer using post-chemoradiotherapy magnetic resonance imaging. Radiother Oncol 2021 Aug;161:183-190. [doi: 10.1016/j.radonc.2021.06.019] [Medline: 34139211]
28. Wang L, Wu X, Tian R, et al. MRI-based pre-Radiomics and delta-Radiomics models accurately predict the post-treatment response of rectal adenocarcinoma to neoadjuvant chemoradiotherapy. Front Oncol 2023;13. [doi: 10.3389/fonc.2023.1133008]
29. Jia LL, Zheng QY, Tian JH, et al. Artificial intelligence with magnetic resonance imaging for prediction of pathological complete response to neoadjuvant chemoradiotherapy in rectal cancer: a systematic review and meta-analysis. Front Oncol 2022;12:1026216. [doi: 10.3389/fonc.2022.1026216] [Medline: 36313696]
30. Abbaspour S, Abdollahi H, Arabalibeik H, et al. Endorectal ultrasound radiomics in locally advanced rectal cancer patients: despeckling and radiotherapy response prediction using machine learning. Abdom Radiol 2022;47(11):3645-3659. [doi: 10.1007/s00261-022-03625-y]

Abbreviation

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
CONSORT: Consolidated Standards of Reporting Trials
CR: complete response
CT: computed tomography
DCA: decision curve analysis
DCE-MRI: dynamic contrast–enhanced magnetic resonance imaging
DL: deep learning
DSC: Dice Similarity Coefficient
FLOP: Floating-Point Operation
GCN: graph convolutional network
GR: good response
Grad-CAM: Gradient-Weighted Class Activation Mapping
IoU: Intersection over Union
MRI: magnetic resonance imaging
nCRT: neoadjuvant chemoradiotherapy
pCR: pathological complete response
PR: partial response
RC: rectal cancer
RECIST: Response Evaluation Criteria in Solid Tumors
ResNets: residual networks
ROC: receiver operating characteristic
ROI: region of interest
RT: radiotherapy
T2W: T2-weighted
U-Net: U-shaped convolutional neural network

Edited by A Coristine; submitted 11.May.2025; peer-reviewed by JCL Chow, P Sharma; revised version received 08.Jan.2026; accepted 12.Jan.2026; published 17.Mar.2026.

Please cite as:
Li Y, Liu H, Wei Q, Wang Z, Hu Z
Rectal Cancer Radiotherapy Response Prediction: Retrospective Study of Development of a Deep Learning–Based Radiomics Model
JMIR Med Inform 2026;14:e77313
URL: https://medinform.jmir.org/2026/1/e77313
doi:10.2196/77313

XSL•FO
RenderX

© Yiqun Li, Hengchang Liu, Qiang Wei, Zhenqi Wang, Zhen Hu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study

Jundan Chen*, PhD; Zhihe Mao*, BA

School of Physical Education, Hunan University of Arts and Science, 3150 Dongting Road, Changde, Hunan, China

*all authors contributed equally

Corresponding Author:
Zhihe Mao, BA
School of Physical Education, Hunan University of Arts and Science, 3150 Dongting Road, Changde, Hunan, China

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2026/1/e92491

Abstract

Background: Early life health risks can shape long-term morbidity trajectories, yet prevailing pediatric risk assessment paradigms are often fragmented and insufficiently capable of integrating heterogeneous data streams into actionable, individualized profiles.

Objective: This study aimed to design, implement, and validate an artificial intelligence–driven framework that fuses multimodal pediatric data and leverages advanced natural language processing and ensemble learning to improve early, accurate stratification of key pediatric health risks.

Methods: A retrospective dataset of over 40,000 pediatric participants aged 2 - 8 years was used to train and evaluate the framework. Data were split into training, validation, and test sets (70%, 15%, and 15%, respectively) with a temporally mindful partitioning strategy to approximate prospective evaluation. Baseline comparators included traditional statistical and machine learning models, and the statistical significance of area under the receiver operating characteristic curve (AUC-ROC) differences was assessed using the DeLong test.

Results: The proposed Bidirectional Encoder Representations From Transformers–based model achieved an AUC-ROC of 0.85 (95% CI 0.82 - 0.88), sensitivity of 0.78, specificity of 0.80, and F_1-score of 0.75 on the test set, outperforming multiple baseline models. In an additional manual comparison evaluation, automated and expert assessments aligned with 78% accuracy (78/100), and most discrepancies arose in “equivalent” cases.

Conclusions: This study provides a validated, artificial intelligence–driven, multimodal pediatric health risk stratification framework that translates heterogeneous child health data into clinically actionable risk profiles, demonstrating strong discriminative performance and meaningful agreement with expert assessment. The framework supports proactive, individualized pediatric care and offers a scalable foundation for further validation across broader populations and longitudinal follow-up.

(JMIR Med Inform 2026;14:e80163) doi:10.2196/80163

KEYWORDS

pediatric health; artificial intelligence; AI; risk stratification; ensemble learning; framework design; multisource data fusion; predictive modeling

Introduction

The assessment and early identification of health risks in children represent a critical area of research with profound implications for individual lifelong well-being and public health strategies. Establishing healthy developmental trajectories during childhood is fundamental, as health conditions and risk exposures in early life can significantly influence adult health outcomes and susceptibility to chronic diseases [1,2]. Traditional pediatric health surveillance often relies on periodic checkups, which may not fully capture the dynamic and multifaceted nature of health risks influenced by a complex interplay of genetic, environmental, and lifestyle factors [3,4]. Therefore, developing advanced, data-driven methodologies, such as those explored in recent research on artificial intelligence (AI) and machine learning (ML) for tasks like toxicity prediction and broader research applications, to proactively identify and stratify these risks is of paramount importance. Such systems, potentially incorporating guardian-interactive elements, can empower health

XSL•FO
RenderX

care providers and families with actionable insights for timely preventive interventions, ultimately contributing to a healthier future generation and optimizing health care resource allocation [5]. Global health initiatives increasingly emphasize the importance of leveraging innovative technologies, including advanced AI models, such as large language models adapted for specific domains [6], to enhance child health and well-being. The potential for deep learning to enhance diagnostic and prognostic capabilities in areas like child psychiatry further underscores this trend [7].

The rapid popularization of internet-based medical services, including telehealth, mobile health (mHealth) apps, and online consultation platforms, has significantly reshaped health care delivery and accessibility worldwide [8]. This digital transformation has not only provided patients and caregivers with more convenient access to medical advice and health information but also led to the generation of vast amounts of digital health data [9,10]. While much of the initial adoption has been observed in adult care, the principles of leveraging digital platforms for health monitoring, data collection, and remote consultation, often supported by automated systems for tasks like information synthesis [11], are increasingly recognized as valuable in pediatric care. The experience gained from the broader implementation of digital health solutions, including aspects of data security, interoperability, and user engagement, offers important lessons for developing effective and safe digital health tools specifically for children [12,13]. The expanding digital health infrastructure provides a rich ecosystem for collecting diverse health-related data that, if properly managed and ethically used with robust frameworks, can be invaluable for comprehensive risk assessment in pediatric populations [14,15]. Task-oriented dialogue systems, for instance, show potential for structured data gathering in such contexts. Despite progress in pediatric health surveillance, significant clinical voids persist, particularly in the settings of primary care and population health screening. Current risk assessment paradigms are often fragmented, struggling to effectively integrate heterogeneous data streams, such as structured electronic health record (EHR) data, unstructured clinical notes, parental reports, and real-time wearable sensor data, into a single, cohesive risk profile. This leads to a reactive, rather than proactive, approach to care, where risks are often identified only after they have manifested. This research addresses this gap by proposing a framework designed for the following specific end users: (1) pediatricians in primary care, who can use the system for point-of-care decision support; (2) public health officials, who can leverage population-level risk distributions for resource allocation and strategic planning; and (3) clinical care coordinators, who can manage and monitor high-risk pediatric cohorts more effectively.

The increasing availability of EHR data, administrative data, patient-generated health data, and other health-related information sources has fueled the application of health care big data analytics to improve service quality, operational efficiency, and patient outcomes [16-18]. Effective utilization of these data, supported by primers on data handling and the underlying principles of models like Bidirectional Encoder Representations From Transformers (BERT), can support evidence-based clinical decision-making, enhance population health management, facilitate epidemiological surveillance, and enable the development of personalized health care interventions [19]. To enhance clinical utility and facilitate shared decision-making, it is imperative that the reasons behind a child's risk stratification are transparent and interpretable. Shapley additive explanations (SHAP) are used due to their strong theoretical guarantee of consistency and local accuracy, which can provide a unified and reliable framework for model interpretation [20]. For a given prediction, SHAP assigns an importance value (SHAP value) to each input feature, and it is crucial for unlocking its full potential in transforming health care delivery, research, and policy.

AI algorithms can analyze complex, multidimensional datasets to identify subtle patterns, predict the likelihood of various health conditions, and stratify risk earlier and more accurately than traditional statistical methods [20,21]. Researchers have explored various AI-driven systems for health applications. For example, some systems focus on the early detection of developmental disorders using behavioral or imaging data, while others aim to predict adverse outcomes in neonatal intensive care units or identify children at risk for conditions like asthma, obesity, and mental health issues, sometimes using specific BERT-based approaches for analyzing medical records or predicting conditions like Alzheimer disease (AD-BERT), which, while not pediatric, demonstrates domain-specific adaptation [22]. Common approaches in developing such systems involve the curation of large, representative datasets; rigorous data preprocessing and feature engineering; the application of appropriate ML algorithms; and robust model validation using independent test sets and ideally prospective clinical evaluation [23]. For instance, deep learning models have shown promise in analyzing medical images for pediatric conditions, while ensemble methods and sophisticated scoring mechanisms like BERTScore for evaluating text generation quality [24] (analogous to assessing the quality of AI-generated health summaries) are often used to improve the robustness and accuracy of predictive models based on structured EHR data or textual information [25]. The primary contribution of our framework lies in its advanced natural language processing (NLP) capabilities, which distinguish it from prior AI systems in pediatric care that have largely relied on traditional NLP techniques. Methodologies, such as keyword matching, bag-of-words, and term frequency-inverse document frequency models, while useful, are fundamentally limited in their ability to interpret the complex and nuanced narratives found in clinical texts.

Therefore, this study aimed to design, implement, and validate a comprehensive, AI-driven framework for pediatric health risk stratification. We hypothesize that by integrating multimodal data, including EHRs, parental questionnaires, and wearable sensor data, and leveraging advanced NLP and ensemble learning models, our system can identify and stratify key pediatric health risks (eg, obesity and developmental delay) with greater accuracy and at an earlier stage than traditional assessment methods, thereby providing robust support for clinical decision-making and pre-emptive intervention. A

XSL•FO
RenderX

comparison of pediatric risk stratification frameworks is provided in Table 1.

Table . Comparison of the selection of pediatric risk stratification frameworks.

Study	Key method	Data modalities used	Clinical deployment status
Smith et al [26]	Logistic regression	EHR[a] (structured only)	Retrospective validation
Chen et al [27]	Random forest	EHR and parental surveys	Prototype on retrospective data
Our study	Fine-tuned BERT[b] + ensemble	EHR, surveys, and wearables	Prototype on retrospective data
Lee et al [28]	CNN[c] on images	Medical imaging	Conceptual framework

[a]EHR: electronic health record.

[b]BERT: Bidirectional Encoder Representations From Transformers.

[c]CNN: convolutional neural network.

Methods

Pediatric Health Risk Stratification Framework and Data Foundation

This section addresses the critical, complex, and multifaceted aspects of multimodal data acquisition, encompassing the collection of diverse data types from various sources and the requisite, often intricate, data preparation processes. These preparation stages, including but not limited to data cleaning, normalization, transformation, and feature engineering, are indispensable for effectively fueling the advanced AI models that form the analytical core of this research, ensuring they operate on high-quality, meaningful inputs to generate reliable and actionable insights [29,30]. The overarching goal is to establish robust data and conceptual groundwork upon which sophisticated risk prediction and stratification can be built, ultimately aiming to transform pediatric health care through proactive and personalized interventions. Conceptually, the framework aims to map a complex set of input variables Xdata (representing multimodal pediatric data) through a series of transformations and AI modeling of fAI to an interpretable risk stratification outcome Yrisk :

(1)Yrisk=fAI(Preprocessing(Xdata))

This high-level representation underscores the journey from raw data to actionable risk assessment.

Conceptual Framework for Pediatric Risk Stratification

The creation of a robust pediatric health risk stratification system requires a meticulously defined conceptual framework that delineates the core components, their interactions, and the overall data flow. As depicted in Figure 1, the proposed AI-driven framework is architected to be modular, scalable, and adaptable to diverse pediatric health contexts. Its primary objective is to convert raw, multisource pediatric data into actionable risk profiles, thereby facilitating proactive and personalized health care interventions.

Commencing with *multimodal data input*, the framework assimilates a wide array of data sources pertinent to child health. Subsequently, the *data preprocessing and harmonization* layer comes into play. Given the heterogeneity and potential quality issues inherent in real-world pediatric data, such as missing values, inconsistencies, and varying formats from different sources, this layer uses data cleaning, normalization, transformation, and integration techniques to create a unified and analysis-ready dataset.

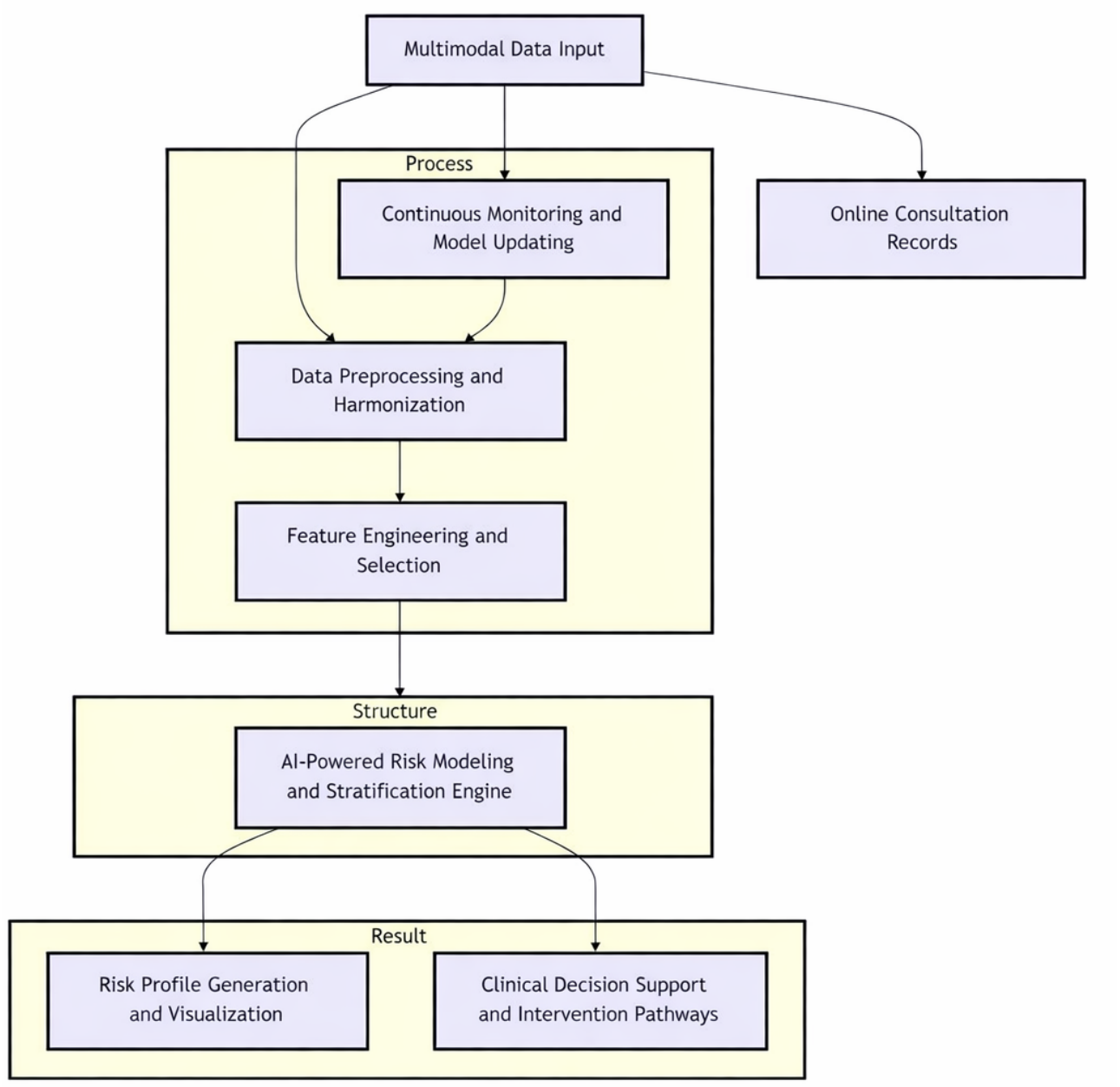
Next is the *feature engineering and selection* module, which is dedicated to extracting meaningful features from the preprocessed data that are highly indicative of pediatric health risks. This process may involve creating composite variables, transforming existing features, or applying dimensionality reduction techniques to optimize the input for AI models. Domain knowledge from pediatric medicine is crucial in guiding this process to ensure clinical relevance.

At the core of the framework lies the *AI-powered risk modeling and stratification engine*. This engine uses advanced ML algorithms, with a particular focus on ensemble learning techniques, as discussed in the Introduction, to construct predictive models for various pediatric health risks. Trained on historical data, these models learn complex patterns and relationships between input features and health outcomes or risk states. The output of this engine is not merely a binary prediction but a nuanced risk stratification, which may encompass risk scores, probability estimates, or the identification of distinct risk phenotypes. The determination of weights for each data modality is a critical aspect of our ensemble model. Rather than assigning predetermined, fixed weights, our *stacking* framework uses a data-driven approach.

The final layers of the framework are devoted to *risk profile generation and visualization* and *clinical decision support and intervention pathways*. The generated risk profiles are designed to be interpretable by health care professionals and understandable by caregivers, often incorporating visual aids and clear explanations. This output is intended to seamlessly integrate with clinical workflows, offering evidence-based insights to support shared decision-making regarding preventive strategies, further assessments, or targeted interventions. *Continuous monitoring* and *model updating* are also integral to the framework, ensuring that the AI models maintain their accuracy and relevance as new information emerges and clinical knowledge evolves [31].

XSL•FO
RenderX

Figure 1. Conceptual diagram of the AI-driven pediatric health risk stratification framework. The image visually represents the modules and their interconnections. AI: artificial intelligence.



Identification of Key Pediatric Health Domains and Risk Factors

The primary pediatric health domains identified for this framework, as detailed in Table 2, cover a broad spectrum of child well-being. *Patient safety* focuses on risks, such as adverse drug events and medical errors, typically informed by EHR data and incident reports. *Continuous health monitoring* is crucial, tracking growth parameters and developmental milestones using EHR data and wearable sensor inputs. For children with ongoing conditions, *chronic disease management* addresses factors like glycemic control or asthma exacerbation frequency, primarily using EHR data and home monitoring device information. *Mental health* is a significant domain, evaluating behavioral indicators and emotional well-being via parental, teacher, and direct child assessments. Furthermore, *preventive care* adherence is monitored through risk factors, such as vaccination status and well-child visit compliance, drawing from EHR data and parental questionnaires [32]. The interconnected domains of *nutrition* and *growth* consider risks associated with nutritional intake and growth percentiles, also informed by EHR and parental questionnaire data. Lastly, *physical activity* levels and sedentary behavior duration, identified as critical risk factors, are assessed using wearable sensor data and parental questionnaires.

For each identified health domain, a set of specific, measurable risk factors and associated potential data sources is cataloged. This systematic identification ensures that the subsequent data acquisition and AI modeling efforts are focused and aligned with the goal of comprehensive risk assessment. The dynamic nature of these risk factors across different developmental stages (infancy, early childhood, middle childhood, and adolescence)

XSL•FO
RenderX

is also a key consideration in the framework's design and application.

Table . Key pediatric health domains, associated risk factors, and potential data sources.

Key pediatric health domains	Associated risk factors	Potential data sources
Patient safety	Adverse drug events and medical errors	EHR[a] data and incident reports
Health monitoring	Growth parameters and developmental milestones	EHR data and wearable sensor data
Chronic disease management	Glycemic control and asthma exacerbation frequency	EHR data and home monitoring devices
Acute illness severity	Fever severity and respiratory distress indicators	EHR data and parental questionnaire responses
Mental health	Behavioral indicators and emotional well-being	Parental and teacher assessments, and direct child assessments
Preventive care	Vaccination status and well-child visit adherence	EHR data and parental questionnaire responses
Nutrition and growth	Nutritional intake and growth percentiles	EHR data and parental questionnaire responses
Physical activity	Activity levels and sedentary behavior duration	Wearable sensor data and parental questionnaire responses

[a]EHR: electronic health record.

Multimodal Data Acquisition and Preparation

The effectiveness of the AI-driven risk stratification framework is fundamentally dependent on the availability of comprehensive, high-quality, multimodal data that accurately reflect the various factors influencing child health. This section details the strategies for acquiring and preparing such data for input into the framework, emphasizing the critical importance of ethical considerations, including informed consent, data privacy, and security, in full compliance with relevant regulations, such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), adapted for pediatric populations.

The data sources for this framework are inherently multimodal, encompassing a variety of information types. EHRs serve as a primary source, offering longitudinal clinical data that include diagnoses (ICD codes), procedures, medications, laboratory results, growth chart data, and clinical notes. Access to both structured and unstructured EHR data is essential. *Parental and child questionnaires or surveys* are used to capture information often not systematically recorded in EHRs, such as detailed family history, socioeconomic status, lifestyle factors (diet, physical activity, and sleep), environmental exposures, and patient-reported outcomes or symptoms. *Wearable sensors and mHealth data* provide opportunities for continuous, real-world monitoring of physiological parameters (eg, heart rate, activity levels, and sleep patterns) and behavioral data. *School health records* can contribute information on immunizations, health screenings conducted at school, and potentially attendance or behavioral notes relevant to health [33]. *Public health and environmental databases* can offer community-level data on factors like air quality, neighborhood socioeconomic indicators, and access to health care or recreational facilities. *Genomic data*, when applicable and ethically sourced, can provide valuable genetic risk scores or specific genetic markers associated with pediatric conditions.

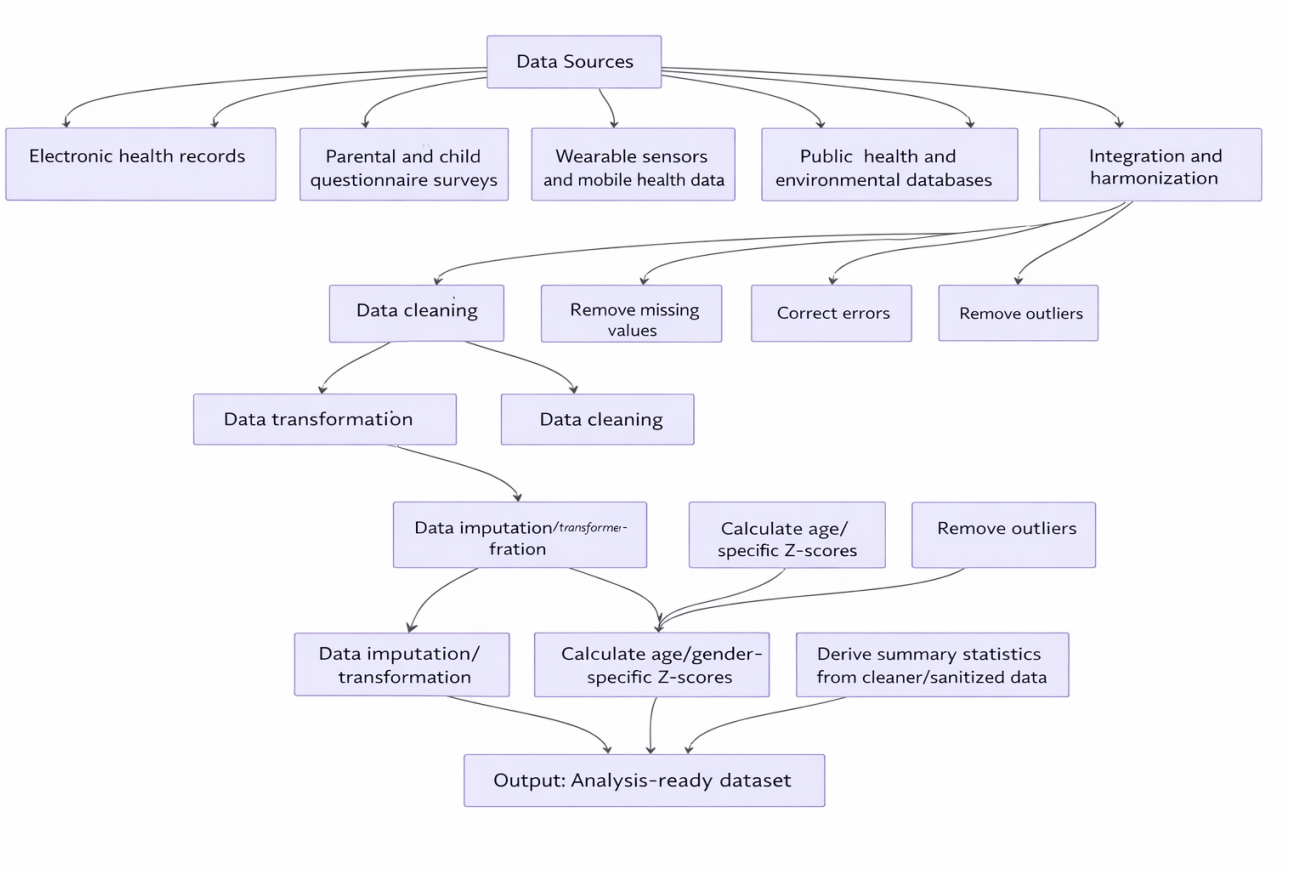
The data preparation pipeline, illustrated in Figure 2, involves several key steps to transform raw, multimodal data into an analysis-ready dataset. Initially, data from disparate sources must undergo integration and harmonization. This process includes mapping data elements to common terminologies, resolving inconsistencies, and creating a unified data schema. Data cleaning is then performed to address missing values, correct errors, and remove outliers. Data transformation may be necessary to convert raw data into formats suitable for analysis, such as calculating age-specific *z* scores for growth parameters or deriving summary statistics from time-series sensor data. For instance, if xraw is a raw measurement and μage,sex and σage,sex are the age- and sex-specific mean and SD, respectively, from a reference population, the *z* score can be calculated as follows:

$$z = x_{raw} - \mu_{age,sex} \sigma_{age,sex}$$

This normalization is crucial for enabling comparisons of measurements across different age groups and sexes. Finally, data anonymization or pseudonymization is rigorously applied to protect patient privacy before the data are used for model development. The resulting dataset provides the foundation for the feature engineering and AI modeling stages described in the subsequent part.(2)

XSL•FO
RenderX

Figure 2. Multimodal data acquisition and preparation pipeline.



Data Preprocessing and Feature Engineering

The integrity and utility of input data are paramount to the performance and clinical relevance of any AI model, particularly in the pediatric domain, where data can be sparse, longitudinal, and highly variable due to growth and development. Pediatric health data, often derived from multiple heterogeneous sources, such as EHRs, parental questionnaires, school health notes, and wearable sensors, typically present unique challenges, including systematically missing values (eg, developmental assessments not performed at certain ages), measurement noise (eg, variability in home-based measurements), and the presence of irrelevant or redundant information that can obscure true risk signals. Therefore, rigorous data preprocessing and thoughtful, domain-informed feature engineering are indispensable steps to prepare a high-quality dataset for subsequent AI modeling.

The initial phase of preprocessing involves comprehensive data cleaning. Missing data represent a pervasive issue in longitudinal pediatric health care datasets. For instance, growth parameters or developmental screening results might be missing for specific well-child visits. While simple imputation techniques like mean, median, and mode imputation can be applied for variables with a low percentage of missingness and random patterns, they often fail to capture the underlying data structure in pediatric cohorts. More sophisticated methods are typically required, such as k-nearest neighbors imputation, which identifies the k most similar pediatric cases (based on a suite of other observed features) and imputes the missing value for a feature Xj using a weighted average or majority vote from the neighbors [34]. Alternatively, model-based imputation, using algorithms like multivariate imputation by chained equations (MICE), can be used. MICE iteratively models each variable with missing values as a function of other variables in the dataset, cycling through variables until convergence is achieved. For a set of variables X1,…,Xp, MICE specifies a conditional model PXj ∣ X-j,φj for each Xj and iteratively samples from these conditional distributions.

Outlier detection and appropriate treatment are also critical in pediatric data, where extreme values might represent genuine clinical concern or measurement error. Statistical methods, such as the *z* score and IQR, are used. For a data point xi in a feature distribution, its *z* score is calculated as zi=(xi−μ)/σ, where μ is the mean and σ is the SD. Data points with |zi|>θoutlier may be flagged. The IQR method defines outliers as points falling below Q1-1.5×IQR or above Q3+1.5×IQR, where Q1 and Q3 are the first and third quartiles, respectively, and IQR=Q3-Q1. Decisions on handling outliers (removal, transformation, or winsorization) are made cautiously, considering the potential clinical significance of extreme values in child health.

Feature engineering in the pediatric context is a highly domain-driven process focused on creating new, informative features from raw data to capture developmental trajectories, critical exposure periods, and clinically relevant interactions. This enhances model performance and interpretability. In child health, this includes deriving age- and sex-adjusted *z* scores for growth parameters, such as height, weight, BMI, and head circumference, based on standardized pediatric growth charts like World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) growth standards. It also involves

XSL•FO
RenderX

creating features like developmental velocity, calculated as ΔD/Δt=Dt2-Dt1/t2-t1, reflecting the rate of change in developmental milestone scores between assessments. Additionally, it covers quantifying cumulative exposure to risk factors, such as days with poor air quality and screen time during early childhood, generating features for adherence to pediatric guidelines like vaccination completeness and well-child visit schedules, and creating interaction terms reflecting synergistic effects in child health, such as genetic predispositions interacting with environmental exposures.

To reduce dimensionality, mitigate the risk of overfitting (especially with potentially limited sample sizes in specific pediatric subpopulations), and improve model training efficiency and interpretability, various feature selection techniques are used. These are broadly categorized into filter methods, wrapper methods, and embedded methods. Filter methods evaluate features independently of the AI model, using statistical measures relevant to pediatric outcomes, such as the chi-square test for assessing associations between categorical risk factors (eg, maternal smoking during pregnancy) and a binary child health outcome, and the ANOVA *F* value for assessing the relationship between numerical predictors (eg, birth weight) and different risk groups. Wrapper methods, such as recursive feature elimination, use the performance of a specific AI model to iteratively build and evaluate models with different subsets of features. Embedded methods perform feature selection as an integral part of the model training process. For instance, tree-based algorithms like random forest and gradient boosting inherently provide feature importance scores based on how much each feature contributes to reducing impurity or error across the ensemble of trees. Regularization techniques, such as L1 regularization (Lasso), are particularly useful. The Lasso objective function for a linear model predicting a child's health outcome y based on features X and coefficients β is as follows:

$$(3) LLasso=\sum i=1n\ (yi-\beta 0-\sum j=1p\ xij\beta j)2+\lambda \sum j=1p\ |\beta j|$$

where λ is the regularization parameter that controls the penalty on the sum of absolute values of the coefficients, effectively shrinking less important feature coefficients to zero. The selection of λ is typically done via cross-validation.

To provide an overview of the preprocessing pipeline and feature engineering strategies across different data modalities, the key steps are summarized in Table 3.

Table . Summary of pediatric data preprocessing steps and feature engineering strategies.

Data type	Preprocessing methods	Feature engineering	Examples
Growth indicators	Missing data imputation (KNN[a] and MICE[b])	*z* score calculation and velocity features	Age- and sex-adjusted *z* scores
Early exposure	Outlier detection (*z* score and IQR)	Cumulative exposure and interaction terms	Poor air quality days and screen time
Questionnaires	Data cleaning and encoding	One-hot encoding and label encoding	Vaccination scores and visit adherence
School records	Data integration	Feature extraction	Immunization records and attendance
Wearables	Noise reduction and smoothing	Activity patterns and sleep metrics	Activity levels and sleep duration
Genomics	Data anonymization	Genetic risk scoring	Genetic predispositions

[a]KNN: k-nearest neighbors.

[b]MICE: multivariate imputation by chained equations.

Risk Stratification Algorithm

The process of risk stratification involves translating the continuous risk probabilities or composite risk scores generated by AI models into distinct, actionable risk categories that are meaningful in a clinical context. This is achieved by defining specific thresholds that convert model outputs into discrete risk strata. The thresholds are determined through a combination of clinical expertise, statistical methods, and decision theory principles, ensuring they are both clinically relevant and statistically valid. P(Risk / Xchild) represents the feature vector for a specific child. For a 4-tier stratification system, the risk categories can be defined as follows:

- Low risk: P(Risk /Xchild)<θ1
- Moderate risk: θ≤P(Risk / Xchild)<θ2
- High risk: θ2≤P(Risk / Xchild)<θ3
- Very high risk: P(Risk / Xchild)≥θ3

The selection of the thresholds θ1,θ2,θ3 is guided by clinical expertise, statistical methods, and decision theory. Pediatric specialists define thresholds based on established clinical guidelines, acceptable risk levels for specific age groups, and the availability and efficacy of preventive interventions. Statistical methods, such as using percentiles of the predicted risk distribution in a well-characterized pediatric reference population, can also be used. For example, thresholds might be set to correspond to the 50th, 80th, and 95th percentiles of this distribution. Decision theory principles can further refine threshold optimization by considering a utility function that balances the costs and benefits associated with true positives, false positives, true negatives, and false negatives. This might involve analyzing metrics like Youden J statistic (J=sensitivity+specificity−1) or finding points on the precision-recall curve that correspond to desired tradeoffs for pediatric screening or intervention programs.

XSL•FO
RenderX

When the system is designed to predict multiple distinct health risks for a child, such as the risk of obesity, developmental delay, or asthma exacerbation, the framework requires a mechanism to present these risks in a consolidated and understandable manner. This can be achieved by either displaying the stratified risk level for each condition independently or by developing a composite pediatric health vulnerability index (CPHVI). The CPHVI is calculated by assigning weights wi to each risk score *S*(Ri) based on factors, such as the clinical severity of risk Ri, its impact on long-term child development, its prevalence in the target pediatric population, and its responsiveness to an intervention. To translate the CPHVI into clinical practice, we propose a tiered intervention model based on predefined, clinically meaningful thresholds developed in collaboration with pediatric experts. For example, a CPHVI score exceeding a "high risk" threshold could automatically trigger an alert in the EHR system for the primary care physician, along with a recommendation for a direct referral to a relevant specialist. The formula for the CPHVI is as follows:

$$(4)\text{SCPHVI}=\sum i=1\text{N wi}\cdot\text{S(Ri)}\sum i=1\text{N wi}$$

The determination of the weights wi is a complex task, often requiring expert consensus or data-driven approaches. Another advanced approach involves using unsupervised learning techniques, such as clustering algorithms, applied to the vector of predicted individual risk probabilities for each child. This can help identify common co-occurring risk patterns or distinct pediatric subphenotypes that share similar multirisk profiles, guiding more tailored multifaceted interventions.

To enhance clinical utility and facilitate shared decision-making with families, the reasons behind a child's stratification into a particular risk category must be as transparent and interpretable as possible. Techniques, such as SHAP and Local Interpretable Model-Agnostic Explanations (LIME), are used. For a given prediction f(Xchild), SHAP assigns an importance value (SHAP value, φj) to each input feature j, representing its marginal contribution to pushing the prediction away from a baseline. The sum of SHAP values for all features plus the baseline prediction equals the model's output for that child:

$$(5)\text{fXchild} =\varphi 0+\sum j=1\text{M }\varphi j$$

This allows clinicians to understand which specific factors (eg, low birth weight, specific dietary patterns, and lack of physical activity) are most influential in determining a child's assessed risk level, thereby guiding targeted advice and interventions.

BERT-Based Knowledge Extraction and NLP Modeling

BERT-based NLP is used to extract informative representations from unstructured pediatric clinical text (eg, consultation records, clinical notes, and parental narratives), which are subsequently used as input features for pediatric health risk stratification models.

Pediatric medical journals serve as a vital source of pediatric disease knowledge, encompassing a wealth of information on pediatric diseases, including clinical symptoms, diagnostic methods, treatment approaches, and prognoses. These journals are written by pediatric experts and researchers and are based on extensive clinical practice and scientific research. They contain detailed case studies, research findings, and expert opinions, providing a solid foundation for the extraction and application of pediatric disease knowledge. The knowledge derived from these journals can help health care professionals better understand the characteristics and progression of pediatric diseases, thereby improving the quality of pediatric health care services [14].

The pediatric disease knowledge base is a structured repository that integrates information from pediatric medical journals and other authoritative sources. It includes various aspects of pediatric diseases, such as disease names, symptoms, signs, laboratory test results, imaging findings, treatment methods, and prognoses. This knowledge base is designed to provide comprehensive and accurate information on pediatric diseases, supporting clinical decision-making and research. By organizing and structuring the information from pediatric medical journals, the pediatric disease knowledge base enables efficient retrieval and utilization of pediatric disease knowledge, facilitating the application of this knowledge in clinical practice and research.

The process begins with the ingestion of pediatric medical journal data into our system. These textual data, rich in pediatric disease knowledge, undergo meticulous preprocessing to align with BERT's input requirements. The text is tokenized into a sequence of tokens [T1,T2,…,Tn], where each Ti represents a word or subword unit. Additionally, the pediatric disease knowledge base is incorporated to enhance the model's understanding of pediatric medical concepts. BERT's architecture consists of multiple transformer encoder layers, each equipped with self-attention mechanisms. The input tokens, enriched with special markers like [CLS] and [SEP], pass through these layers, generating a matrix of contextualized embeddings z={z1,z2,…,zn}. These embeddings capture intricate semantic relationships and contextual information within the text, forming the basis for subsequent knowledge extraction tasks.

To identify key pediatric disease entities and their relationships within the text, named entity recognition and relation extraction techniques are applied to the BERT-generated embeddings. For named entity recognition, a linear layer combined with a conditional random field is used. The conditional random field loss function is defined as follows:

$$(6)\text{LNER}=-1\text{N}\sum i=1\text{NlogP(yi} \mid \text{Xi)}$$

where N is the number of samples and Pyi ∣ Xi is the probability of the true label sequence yi given the input sequence Xi. For relation extraction between entities, a classifier is trained using a cross-entropy loss function as follows:

$$(7)\text{LRE}=-\sum i=1\text{M }\sum j=1\text{C yi,jlogy}\hat{}\text{i,j}$$

where M represents the number of entity pairs, C is the number of relation types, and yi,j and y^i,j are the true and predicted probabilities for relation *j* between the entity pair *i*, respectively. The extracted entities and relationships are then integrated into the pediatric disease knowledge base. This knowledge base, enriched with entities and relationships from pediatric medical literature, serves to enhance a medical chatbot's responses. When users input symptoms, the chatbot leverages this

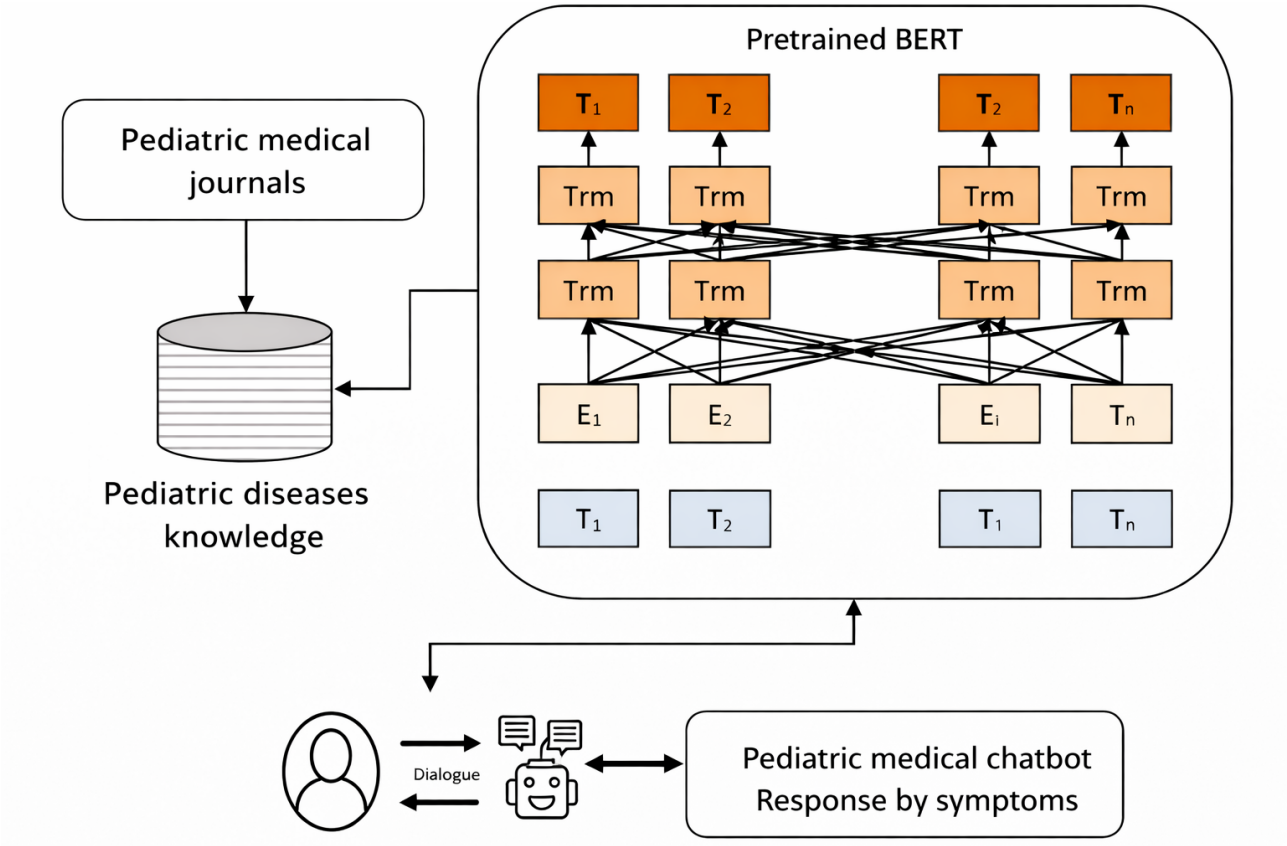
XSL•FO
RenderX

knowledge base to generate accurate and contextually relevant diagnostic suggestions and risk assessments.

Figure 3 illustrates a comprehensive workflow where BERT's robust language understanding capabilities are harnessed to extract valuable pediatric disease knowledge from medical texts. This knowledge is subsequently used to empower a medical chatbot, enabling it to deliver precise and informative responses to user inquiries regarding pediatric symptoms. The integration of BERT with pediatric medical journals and a pediatric disease knowledge base provides a powerful tool for advancing pediatric health care through improved diagnostic accuracy and personalized treatment recommendations.

Our knowledge extraction module is built upon BioBERT, a pretrained language model optimized for the biomedical domain. We chose BioBERT as our base model due to its demonstrated strong performance on various biomedical text mining tasks. For fine-tuning, we used a domain-specific corpus comprising over 50,000 articles from leading pediatric medical journals and over 100,000 deidentified clinical notes. The model was fine-tuned for 5 epochs with a learning rate of $2e^{-5}$. The performance of our fine-tuned model on key NLP tasks was rigorously evaluated on a manually annotated test set of 500 clinical notes, which were dual-annotated by pediatric domain experts to ensure quality.

Figure 3. BERT-driven pediatric disease knowledge extraction and medical dialogue response framework. T denotes the input token sequence obtained after tokenization of the pediatric clinical text, E represents the contextualized embedding vectors produced by the BERT encoder, and Trm refers to the Transformer encoder layers of the BERT model. BERT: Bidirectional Encoder Representations From Transformers.



System Architecture Design

The successful deployment, scalability, and usability of the AI-driven pediatric health risk stratification framework critically depend on a meticulously planned and well-engineered system architecture. This architecture must robustly support efficient multimodal data processing pipelines; reliable and timely execution of complex AI models; secure and compliant management of sensitive pediatric health data; and an intuitive, actionable interface for diverse end users, including pediatricians, specialist clinicians, public health officials, researchers, and potentially, with appropriate safeguards, parents or caregivers. A multitier, service-oriented architecture is proposed to ensure modularity, maintainability, and scalability. This typically comprises the following tiers:

1. Data tier: This foundational tier is responsible for the persistent storage and comprehensive management of all relevant pediatric health data. This includes raw input data from diverse sources (EHRs, wearables, and questionnaires), preprocessed and feature-engineered datasets, trained AI models (including their versions and metadata), generated risk profiles for individual children, and audit logs. This tier would likely involve a hybrid database strategy, combining relational databases (MySQL) for structured metadata, patient demographics, and well-defined clinical entities, with potentially NoSQL databases for handling large volumes of heterogeneous, unstructured, or streaming data (like continuous sensor readings). Robust data governance, backup, and recovery mechanisms are integral.
2. Application logic tier (backend services): This is the computational core of the system, housing the data

XSL•FO
RenderX

preprocessing pipelines, sophisticated feature engineering modules, the AI model inference engine, and the risk stratification algorithms. It handles all computational tasks, business logic for risk assessment, and interactions with the data tier. This tier would be developed using scalable and efficient programming languages and appropriate backend frameworks. It may be implemented as a set of microservices to enhance scalability and independent deployability of different functionalities.

3. Presentation tier (frontend interfaces): This tier provides the user interface and user experience for interacting with the system. This could manifest as a secure web-based application accessible via standard browsers, designed with responsive layouts for use on various devices. For specific user groups like parents or children (with age-appropriate design), a dedicated mobile app might be considered. The frontend allows authorized users to view individualized pediatric risk stratification results, explore interactive visualizations of risk factors and trends, access evidence-based decision support information or guideline recommendations, and potentially (for clinicians) trigger further diagnostic or intervention pathways.

Key functional system modules are designed to support the end-to-end pediatric risk stratification process (Figure 4).

1. Pediatric data ingestion and management module: It securely handles the import of data from diverse, often disparate, pediatric health information systems and sources. It performs initial data validation and schema mapping and manages data storage, versioning, and provenance tracking, which are crucial for longitudinal pediatric studies.
2. Preprocessing and pediatric feature engineering module: It implements complex algorithms described in the preceding section, tailored for pediatric data. This module needs to be highly configurable and extensible to accommodate new data types, evolving pediatric clinical knowledge, or updated feature definitions.
3. AI model execution and management engine: It loads versioned, trained AI models and performs inference on new or updated pediatric patient data to generate risk predictions. This engine must be optimized for performance and scalability, especially if real-time or near real-time risk assessment is required for acute pediatric conditions. It also manages the model lifecycle, including retraining triggers and performance monitoring.
4. Pediatric risk stratification and profiling module: It implements the sophisticated algorithms from the preceding section to convert raw model outputs into clinically interpretable risk strata and comprehensive, multidimensional pediatric risk profiles, potentially including age-adjusted interpretations.
5. Reporting, visualization, and alerting module: It generates customized reports, interactive dashboards (eg, showing population-level pediatric risk distributions and trends over time for specific age cohorts), and dynamic visualizations (eg, individual child risk timelines and feature importance charts for specific predictions). It may also include an alerting mechanism for clinicians when a child's risk profile crosses critical, predefined thresholds.
6. User authentication, authorization, and audit module: It ensures secure, role-based access to the system and its sensitive pediatric data, adhering strictly to privacy regulations (eg, HIPAA, GDPR, and Children's Online Privacy Protection Act [COPPA]). Comprehensive audit trails of data access and system actions are maintained.
7. Interoperability layer (application programming interfaces [APIs]): It provides well-defined APIs, possibly using standards like Fast Healthcare Interoperability Resources, to allow secure data exchange and interaction with other clinical systems (eg, pediatric EHRs, laboratory information systems, and clinical decision support tools embedded in existing workflows).

The data flow within the system begins with secure data ingestion from multiple sources into the *data tier*. The *application logic tier* then orchestrates the preprocessing and feature engineering pipelines. The processed features are subsequently fed into the *AI model execution engine* for risk prediction. These predictions are passed to the *risk stratification and profiling module*, and the final, interpretable pediatric risk profiles are stored and made available for secure access and visualization through the *presentation tier* or via the *API layer*. The technology stack will be carefully chosen based on criteria, such as scalability for large pediatric populations, real-time performance requirements, security mandates for child data, ease of development and maintenance, and compatibility with existing health care IT infrastructure. This might include Python with libraries like Pandas, NumPy, Scikit-learn, TensorFlow/Keras, and PyTorch for AI or ML development; web frameworks, such as Django/Flask (Python), Spring Boot (Java), and Node.js, for the backend services; modern JavaScript frameworks like React, Angular, and Vue.js for building responsive and interactive frontend interfaces; and a combination of SQL and NoSQL databases as described for the *data tier*. Containerization technologies like Docker and orchestration tools like Kubernetes may be used for deployment and scaling.

XSL•FO
RenderX

Figure 4. Multitier system architecture diagram for pediatric risk stratification. AI: artificial intelligence.

XSL•FO
RenderX

Prototype System Implementation

The pediatric health risk stratification prototype system integrates multisource data input with advanced AI models to deliver a comprehensive risk assessment platform. As illustrated in Figure 5, the system collects diverse data types, including patient demographics, clinical records, and parent-reported outcomes, which serve as the foundation for subsequent analysis and risk assessment. This integration ensures a holistic view of pediatric health, enabling more accurate and nuanced risk predictions.

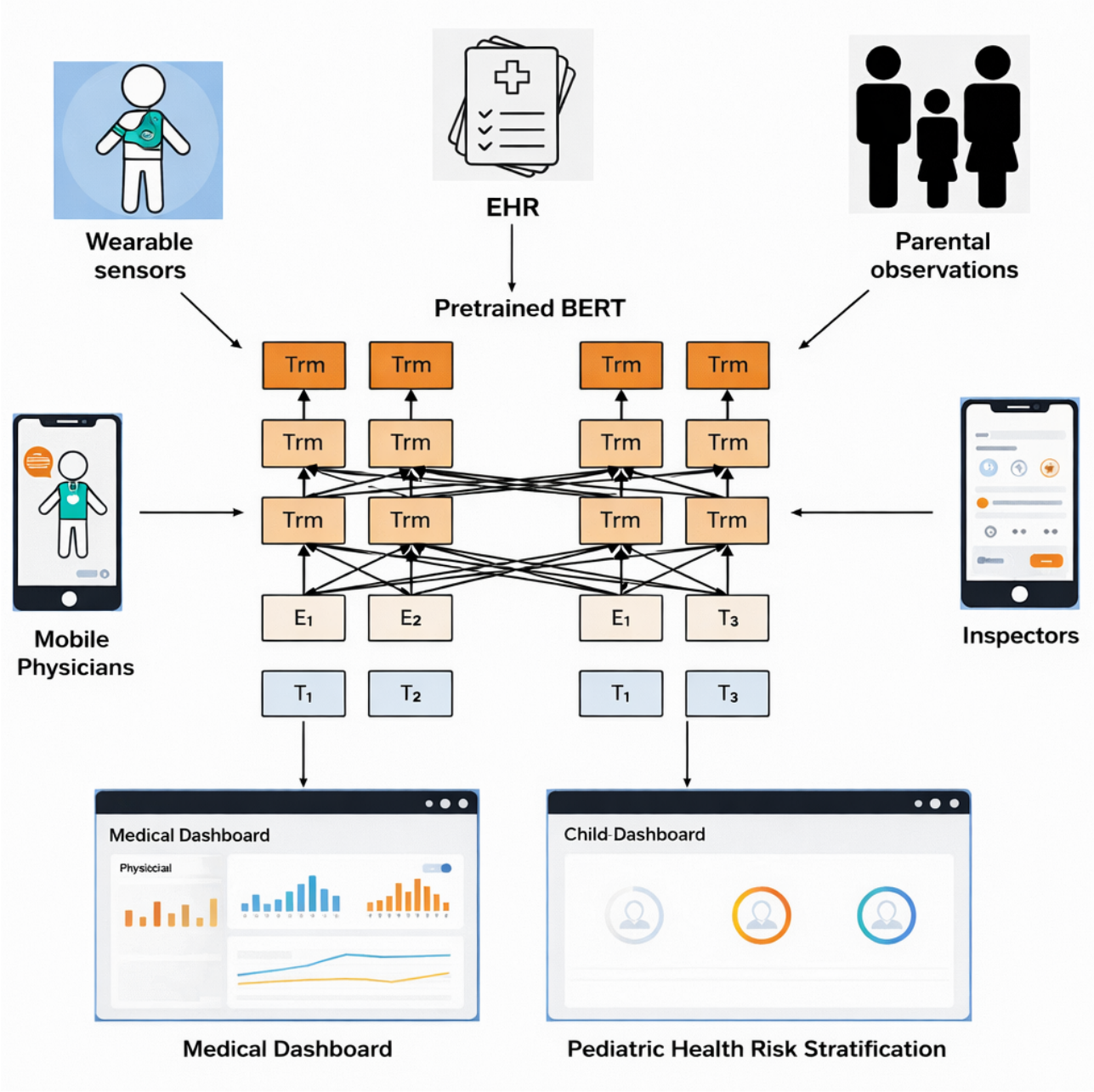
At the core of the system is a pretrained BERT model, specifically fine-tuned to handle pediatric health terminology and context. The model processes input text data through multiple Transformer modules, which encode the text to capture complex linguistic patterns and semantic relationships. Feature extraction layers then distil these encoded representations into risk-relevant features. These features are passed to risk assessment components where the actual risk stratification takes place, transforming raw data into actionable insights regarding children's health risks.

The system's implementation leverages a robust technology stack to ensure efficiency, scalability, and user-friendliness. The backend, developed using Python and its rich library ecosystem, handles data processing and model execution. It uses microservices, potentially orchestrated by lightweight frameworks like Flask, to manage distinct functionalities such as data ingestion, preprocessing, and model inference. Data storage follows a hybrid approach, combining PostgreSQL for structured data with MongoDB for unstructured or semistructured content. The frontend, built with modern JavaScript frameworks, offers tailored interfaces for different users, including physicians and parents. It includes features like risk profile visualizations and population-level risk trend dashboards. Docker is used for containerization to ensure smooth deployment and scalability of the application components.

To illustrate the practical utility of the framework, consider a hypothetical clinical scenario. Dr Smith, a pediatrician, begins her day by logging into the system's dashboard. The system flags one of her patients, a 5-year-old child, for a significant increase in the composite risk score for developing obesity. Dr Smith clicks on the patient's profile and is presented with an interactive, multimodal dashboard. An explainable AI feature, using SHAP values, highlights the primary contributing factors: a recent decrease in physical activity levels captured by wearable data and parent-reported dietary logs indicating high consumption of processed foods. Based on these insights, the system provides Dr Smith with evidence-based, actionable recommendations, including a referral to a pediatric nutritionist and a set of tailored educational materials for the parents. This scenario demonstrates how the framework can transform raw data into clinically actionable insights to facilitate timely and personalized preventive care.

XSL•FO
RenderX

Figure 5. Structure of the pediatric health risk stratification prototype system. T denotes the input token sequence obtained after tokenization of the pediatric clinical text, E represents the contextualized embedding vectors produced by the BERT encoder, and Trm refers to the Transformer encoder layers of the BERT model. BERT: Bidirectional Encoder Representations From Transformers; EHR: electronic health record.



Experimental Setup

The primary dataset used for training and validating the AI models included data from over 40,000 pediatric participants, with an age range of 2 to 8 years at the time of data collection or follow-up. The dataset was partitioned into 3 mutually exclusive sets: a training set (70% of the data), a validation set (15%), and an independent holdout test set (15%). To ensure temporal validity if longitudinal data were used, the split was performed such that data from earlier time periods were used for training and validation, while data from later periods were reserved for testing, mimicking a prospective evaluation.

To benchmark the performance of the proposed AI-driven framework, a selection of robust baseline models was implemented and rigorously evaluated on an identical dataset and predictive task. These comparators included established traditional statistical models frequently used in pediatric risk prediction, specifically logistic regression for binary risk outcomes, where the probability of risk $P(Y=1 \mid X)$ is modeled as $P(Y=1 \mid X)=(1+e^{-(\beta 0+\sum i=1p \beta iXi)})-1$, and Cox proportional hazards models in scenarios where time-to-event data for risk onset were available. Furthermore, simpler yet effective ML algorithms, namely support vector machine (SVM), configured with various kernels (linear and radial basis function), and standard single decision trees, were also incorporated as baselines. In instances where existing, validated pediatric risk scores or established rule-based systems were pertinent to the specific health outcomes under investigation and applicable to the dataset, these were also included in the comparative analysis to provide a comprehensive performance context. To rigorously

XSL•FO
RenderX

assess the statistical significance of our model's superior performance, we used the DeLong test to compare the area under the receiver operating characteristic curve (AUC-ROC) of our proposed model with that of each baseline model.

Ethical Considerations

This study involved research with human participants. The study protocol was reviewed and approved by the Institutional Review Board and Research Ethics Committee of Hunan University of Arts and Science (HUAS-20250504). All procedures were performed in accordance with the ethical standards of the responsible institutional committee and with the principles of the Declaration of Helsinki. Informed consent was obtained from all participants prior to their inclusion in the study. Participants were fully informed about the purpose of the study, the procedures involved, and their right to withdraw at any time without penalty. The privacy and confidentiality of all participants were strictly maintained throughout the study. All data were anonymized prior to analysis, and no personally identifiable information was collected, stored, or reported. No financial or material compensation was provided to participants for their participation in this study.

Results

This section reports the empirical performance of the proposed AI-driven pediatric health risk stratification framework evaluated on the independent test set, in comparison with multiple baseline models.

Table 4 summarizes the predictive performance of the proposed model and baseline methods across targeted pediatric health risks. The proposed model achieved an AUC-ROC of 0.85 (95% CI 0.82 - 0.88), an area under the precision-recall curve of 0.70 (95% CI 0.65 - 0.75), a sensitivity of 0.78, a specificity of 0.80, and an F_1-score of 0.75. In comparison, logistic regression, SVM, random forest, gradient boosting, and a conventional deep learning model yielded lower AUC-ROC and F_1-score values.

Table . Key performance metrics of the proposed artificial intelligence model and baseline models for targeted pediatric risks in the test set.

Model name	AUC-ROC[a], value (95% CI)	AUC-PR[b], value (95% CI)	Sensitivity	Specificity	F_1-score	Brier score
Proposed BERT[c] model	0.85 (0.82 - 0.88)	0.70 (0.65 - 0.75)	0.78	0.80	0.75	0.15
Logistic regression	0.72 (0.68 - 0.76)	0.60 (0.55 - 0.65)	0.65	0.70	0.62	0.20
SVM[d]	0.75 (0.71 - 0.79)	0.65 (0.60 - 0.70)	0.70	0.72	0.66	0.18
Random forest	0.78 (0.74 - 0.82)	0.67 (0.62 - 0.72)	0.72	0.75	0.69	0.17
Gradient boosting	0.80 (0.77 - 0.83)	0.69 (0.64 - 0.74)	0.75	0.76	0.72	0.16
Deep learning	0.83 (0.80 - 0.86)	0.72 (0.67 - 0.77)	0.76	0.79	0.74	0.14

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PR: area under the precision-recall curve.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]SVM: support vector machine.

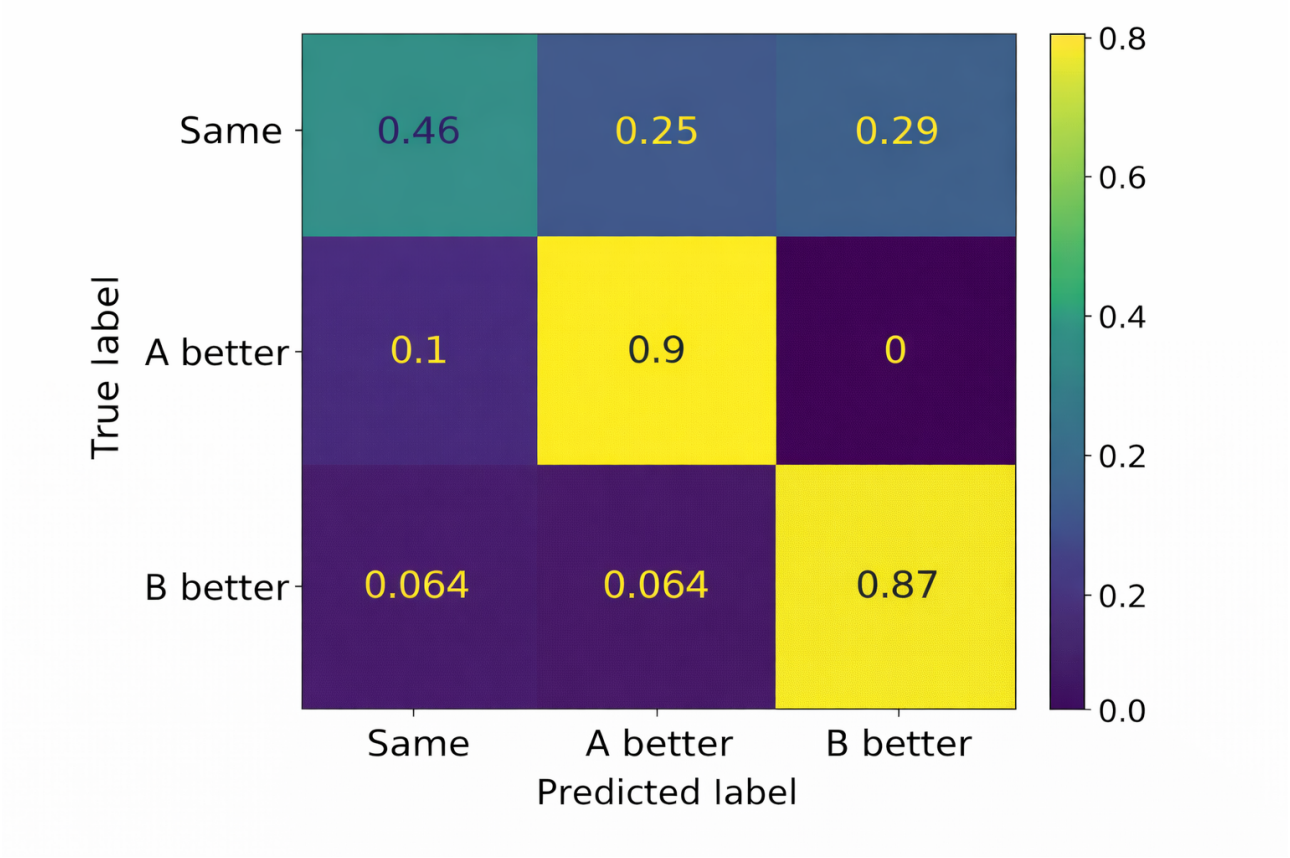
The DeLong test demonstrated that the AUC-ROC of the proposed model was significantly higher than that of each baseline model (all $P<.05$). Paired t tests on F_1-scores similarly indicated statistically significant improvements in predictive accuracy for the proposed model relative to all comparators.

To further assess alignment with expert judgment, a manual comparison evaluation was conducted using 100 randomly selected pairs of online consultation cases. Automated predictions and expert assessments were concordant in 78 cases, corresponding to an agreement rate of 78%.

A confusion matrix (Figure 6) shows that most discrepancies occurred in cases manually labeled as "equivalent," whereas predictions for clearly differentiated cases exhibited higher agreement with expert assessments. Misclassifications were predominantly observed within the equivalent category, while the majority of nonequivalent cases were correctly identified.

XSL•FO
RenderX

Figure 6. Artificial intelligence prediction confusion matrix.



Model performance was further examined across predefined pediatric subgroups, including age categories, sex, and socioeconomic status (Table 5). Comparable AUC-ROC and F_1-score values were observed across all subgroups. For age-based stratification, the AUC-ROC values were 0.84 for children younger than 2 years, 0.85 for those aged 2 - 5 years, and 0.86 for those older than 5 years. Similar performance consistency was observed across sex and socioeconomic strata.

Table . Model performance across different pediatric subgroups for early childhood obesity.

Group	AUC-ROC[a], value (95% CI)	F_1-score
Age		
<2 years	0.84 (0.80 - 0.88)	0.74
2 - 5 years	0.85 (0.82 - 0.89)	0.75
>5 years	0.86 (0.83 - 0.89)	0.76
Sex		
Male	0.85 (0.82 - 0.88)	0.75
Female	0.84 (0.81 - 0.87)	0.74
Status		
Low SES[b]	0.83 (0.79 - 0.87)	0.73
Medium SES	0.85 (0.82 - 0.88)	0.75
High SES	0.86 (0.83 - 0.89)	0.76

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]SES: socioeconomic status.

XSL•FO
RenderX

Discussion

Conclusions

This study successfully designed and validated a novel AI-driven framework for comprehensive pediatric health risk stratification. The primary findings were threefold. First, the framework demonstrates the capability to effectively integrate heterogeneous, multimodal data sources to create a holistic view of a child's health status. Second, our proposed predictive model, which combines a fine-tuned BERT architecture with an ensemble learning strategy, significantly outperformed established baseline models, such as logistic regression and SVM, in predicting key health risks, achieving an AUC-ROC of 0.85 for early childhood obesity. Third, the prototype system's risk assessments showed substantial agreement with manual expert evaluations (78% accuracy), confirming its potential clinical utility and feasibility. The successful implementation of the prototype system, featuring intuitive dashboards for both clinicians and parents, further illustrates the practical applicability of this approach in facilitating early and personalized interventions, thereby contributing a novel and robust technological foundation for proactive pediatric health care.

Limitations and Future Work

While the presented framework shows considerable promise, certain limitations and avenues for future research warrant discussion. The current validation, though rigorous, was based on a specific dataset of over 40,000 pediatric participants aged 2 - 8 years. Broader validation across more diverse pediatric populations and longitudinal follow-up are essential to ascertain long-term predictive accuracy and generalizability. One limitation of our study is the observed slight dip in model performance for the subgroup of children aged 3 - 5 years, which we attribute to the relative scarcity of rich, unstructured text data for this age cohort. To mitigate this issue in future work, we plan to incorporate alternative text sources. A promising approach, as suggested, is the inclusion of open-ended responses from parental questionnaires. These narratives, which capture detailed parental concerns and observations, can be processed by our fine-tuned BERT model to generate rich semantic features, thereby enriching the feature set for younger children and addressing the data sparsity issue. A crucial direction for future work is the implementation of advanced bias mitigation techniques. We plan to further explore methods, such as adversarial debiasing, which involves training the model to make predictions that are invariant to sensitive attributes, thereby proactively enhancing the fairness and equity of our risk stratification framework.

Funding

The research project was supported by the Education Department of Hunan Province: Multidimensional Risk Assessment and Sports-Health Integrated Intervention Model Development for Postural Health in Children in Northwestern Hunan (project number: 24B0629), and the Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province.

Conflicts of Interest

None declared.

References

1. Sales WB, Maranhão EF, Ramalho CST, Macêdo S, Souza GF, Maciel Á. Early life circumstances and their impact on health in adulthood and later life: a systematic review. BMC Geriatr 2024 Nov 28;24(1):978. [doi: 10.1186/s12877-024-05571-4] [Medline: 39609801]
2. Wadsworth ME, Kuh DJ. Childhood influences on adult health: a review of recent work from the British 1946 national birth cohort study, the MRC National Survey of Health and Development. Paediatr Perinat Epidemiol 1997 Jan;11(1):2-20. [doi: 10.1046/j.1365-3016.1997.d01-7.x] [Medline: 9018723]
3. Braveman P, Barclay C. Health disparities beginning in childhood: a life-course perspective. Pediatrics 2009 Nov;124 Suppl 3:S163-S175. [doi: 10.1542/peds.2009-1100D] [Medline: 19861467]
4. Noteboom A, Have MT, de Graaf R, Beekman ATF, Penninx B, Lamers F. The long-lasting impact of childhood trauma on adult chronic physical disorders. J Psychiatr Res 2021 Apr;136:87-94. [doi: 10.1016/j.jpsychires.2021.01.031] [Medline: 33581460]
5. Eadie P, Levickis P, McKean C, et al. Developing preschool language surveillance models - cumulative and clustering patterns of early life factors in the early language in Victoria study cohort. Front Pediatr 2022;10:826817. [doi: 10.3389/fped.2022.826817] [Medline: 35186809]
6. Olsen J, Pedersen LH. Reproductive health indicators and fetal medicine - many things will change. Scand J Work Environ Health 2016 Jun 1;42(6):561-562. [doi: 10.5271/sjweh.3596] [Medline: 27685865]
7. Young C, Grobelna A. Flash Glucose Monitoring Systems in Pediatric Populations With Diabetes: Canadian Agency for Drugs and Technologies in Health; 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK572013/ [accessed 2026-01-02]
8. Garbarino S, Bragazzi NL. Revolutionizing sleep health: the emergence and impact of personalized sleep medicine. J Pers Med 2024 Jun 4;14(6):598. [doi: 10.3390/jpm14060598] [Medline: 38929819]

XSL•FO
RenderX

9. Stidley CA, Shah VO, Narva AS, et al. A population-based, cross-sectional survey of the Zuni Pueblo: a collaborative approach to an epidemic of kidney disease. Am J Kidney Dis 2002 Feb;39(2):358-368. [doi: 10.1053/ajkd.2002.30557] [Medline: 11840378]
10. Vittoria Togo M, Mastrolorito F, Orfino A, et al. Where developmental toxicity meets explainable artificial intelligence: state-of-the-art and perspectives. Expert Opin Drug Metab Toxicol 2024 Jul 2;20(7):561-577. [doi: 10.1080/17425255.2023.2298827]
11. He O, Chen D, Li Y. A deep-learning approach to predict reproductive toxicity of chemicals using communicative message passing neural network. Front Toxicol 2025;7:1640612. [doi: 10.3389/ftox.2025.1640612] [Medline: 40765625]
12. Sambanis A, Osiecki K, Cailas M, Quinsey L, Jacobs DE. Using artificial intelligence to identify sources and pathways of lead exposure in children. J Public Health Manag Pract 2023;29(5):E208-E213. [doi: 10.1097/PHH.0000000000001759] [Medline: 37129378]
13. Huang YD, Zeng SL, Lin J, et al. Parents' understanding and attitudes toward the application of AI in pediatric healthcare: a cross-sectional survey study. Front Public Health 2025;13:1654482. [doi: 10.3389/fpubh.2025.1654482] [Medline: 40910043]
14. Sezgin E, Jackson DI, Kocaballi AB, et al. Can large language models aid caregivers of pediatric cancer patients in information seeking? A cross-sectional investigation. Cancer Med 2025 Jan;14(1):e70554. [doi: 10.1002/cam4.70554] [Medline: 39776222]
15. Shah U, Khan N, Alzubaidi M, Agus M, Househ M. ArtInsight: a multimodal AI framework for interpreting children's drawings and enhancing emotional understanding. Stud Health Technol Inform 2025 May 15;327(5):808-812. [doi: 10.3233/SHTI250471] [Medline: 40380579]
16. Till AC, Briganti G. Outcomes in machine learning models for child psychiatry: a systematic review of the literature. Psychiatr Danub 2025 Sep;37(Suppl 1):79-84. [Medline: 40982877]
17. Vandewouw MM, Syed B, Barnett N, et al. Using deep learning to predict internalizing problems from brain structure in youth. Transl Psychiatry 2025 Aug 29;15(1):326. [doi: 10.1038/s41398-025-03565-3] [Medline: 40883286]
18. Till AC, Briganti G. AI in child psychiatry: exploring future tools for the detection and management of mental disorders in children and adolescents. Psychiatr Danub 2023 Oct;35(Suppl 2):20-25. [Medline: 37800200]
19. Li Z, Lu F, Wu J, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26:e56556. [doi: 10.2196/56556] [Medline: 39589770]
20. Zhao H, Zhang Z, Tang J. Enhancing rural healthcare through internet-based remote collaborative outpatient services: a comprehensive evaluation in Changzhi, Shanxi Province. Medicine (Baltimore) 2024;103(36):e39614. [doi: 10.1097/MD.0000000000039614]
21. Liu F, Song T, Yu P, et al. Efficacy of an mHealth app to support patients' self-management of hypertension: randomized controlled trial. J Med Internet Res 2023 Dec 19;25:e43809. [doi: 10.2196/43809] [Medline: 38113071]
22. Hussain SA, Sezgin E, Krivchenia K, Luna J, Rust S, Huang Y. A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study. JAMIA Open 2021 Jul;4(3):ooab084. [doi: 10.1093/jamiaopen/ooab084] [Medline: 34604710]
23. Zayas-Cabán T, Haque SN, Kemper N. Identifying opportunities for workflow automation in health care: lessons learned from other industries. Appl Clin Inform 2021 May;12(3):686-697. [doi: 10.1055/s-0041-1731744] [Medline: 34320683]
24. Šuster S, Baldwin T, Lau JH, et al. Automating quality assessment of medical evidence in systematic reviews: model development and validation study. J Med Internet Res 2023 Mar 13;25:e35568. [doi: 10.2196/35568] [Medline: 36722350]
25. Abbasi M, Khera S, Dabravolskaj J, Aissiou A, Abbasi-Dezfouly R. Family physicians' feedback on the feature design of a digital health platform to streamline the care of older adults. Geriatrics (Basel) 2024 Nov 28;9(6):154. [doi: 10.3390/geriatrics9060154] [Medline: 39727813]
26. Smith J, Cheater F, Bekker H. Parents' experiences of living with a child with a long-term condition: a rapid structured review of the literature. Health Expect 2015 Aug;18(4):452-474. [doi: 10.1111/hex.12040] [Medline: 23311692]
27. Chen J, Lim CP, Tan KH, Govindan K, Kumar A. Artificial intelligence-based human-centric decision support framework: an application to predictive maintenance in asset management under pandemic environments. Ann Oper Res 2025 Jul;350(2):493-516. [doi: 10.1007/s10479-021-04373-w]
28. Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview. Korean J Radiol 2017;18(4):570-584. [doi: 10.3348/kjr.2017.18.4.570] [Medline: 28670152]
29. Kalla M, Bradford A, Schadewaldt V, et al. Co-designing a user-centered digital health tool for supportive care needs of patients with brain tumors and their caregivers: interview analysis. JMIR Cancer 2025 May 23;11:e53690. [doi: 10.2196/53690] [Medline: 40408266]
30. Kowalkowski M, Eaton T, Reeves KW, et al. Incorporating patient, caregiver, and provider perspectives in the co-design of an app to guide hospital at home admission decisions: a qualitative analysis. JAMIA Open 2024 Oct;7(3):ooae079. [doi: 10.1093/jamiaopen/ooae079] [Medline: 39156047]

XSL•FO
RenderX

31. Ibrahim ST, Patel J, Katapally TR. Digital citizen science for ethical monitoring of youth physical activity frequency: comparing mobile ecological prospective assessments and retrospective recall. PLOS Digit Health 2025 May;4(5):e0000840. [doi: 10.1371/journal.pdig.0000840] [Medline: 40315264]
32. Sumiya M, Nishimura T, Aizaki K, et al. Addressing the ethical, legal, and social issues of healthtech in education: insights from Japan. JMIR Form Res 2025 Jul 18;9:e72781. [doi: 10.2196/72781] [Medline: 40680270]
33. Pavarini G, Yosifova A, Wang K, et al. Data sharing in the age of predictive psychiatry: an adolescent perspective. Evid Based Ment Health 2022 May;25(2):69-76. [doi: 10.1136/ebmental-2021-300329] [Medline: 35346984]
34. Li E, Lounsbury O, Clarke J, Ashrafian H, Darzi A, Neves AL. Patient and caregiver perceptions of electronic health records interoperability in the NHS and its impact on care quality: a focus group study. BMC Med Inform Decis Mak 2024 Dec 3;24(1):370. [doi: 10.1186/s12911-024-02789-5] [Medline: 39627780]

Abbreviations

AI: artificial intelligence
API: application programming interface
AUC-ROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations From Transformers
CPHVI: composite pediatric health vulnerability index
EHR: electronic health record
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
MICE: multivariate imputation by chained equations
ML: machine learning
NLP: natural language processing
SHAP: Shapley additive explanations
SVM: support vector machine

Edited by A Benis; submitted 05.Jul.2025; peer-reviewed by B Zhang, K Li, MBR Mahmoud; revised version received 18.Oct.2025; accepted 23.Oct.2025; published 26.Jan.2026.

Please cite as:
Chen J, Mao Z
Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study
JMIR Med Inform 2026;14:e80163
URL: https://medinform.jmir.org/2026/1/e80163
doi:10.2196/80163

© Zhihe Mao, Jundan Chen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Scalable and Privacy-Conscious End-to-End Processing of Large-Scale Clinical Data for Precision Medicine: Empirical Evaluation Study

Jungwoo Lee[1], PhD; Sangwon Hwang[2], PhD; Kyu Hee Lee[2], PhD

[1]Artificial Intelligence Big Data Medical Center, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

[2]Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

Corresponding Author:
Kyu Hee Lee, PhD
Department of Precision Medicine
Yonsei University Wonju College of Medicine
20 Ilsan-ro, Wonju-si, Gangwon-do
Wonju, 26417
Republic of Korea
Phone: 82 82 741 0466
Email: powerpc@yonsei.ac.kr

Abstract

Background: In large-scale clinical data analysis, CSV and traditional relational database management system–based approaches are widely used but impose substantial storage and processing constraints that delay research preparation and hinder multicenter collaboration. Although column-oriented storage formats such as Apache Parquet have gained attention in data science, systematic end-to-end evaluations in clinical environments remain limited, particularly regarding efficiency and scalability.

Objective: This study aimed to empirically evaluate whether a Parquet-based end-to-end pipeline could improve computational efficiency and scalability in large-scale clinical data analysis while preserving predictive performance and protecting privacy.

Methods: Electronic health record data comprising 13.76 million rows from a large academic medical center in Korea were analyzed using Parquet, CSV, PostgreSQL, and DuckDB environments. Standardized SQL workloads and multilabel classification models—implemented using graphics processing unit–accelerated Extreme Gradient Boosting and classifier chain (CC) ensembles to address class imbalance—were applied to evaluate storage efficiency, time to analysis, and predictive performance. Statistical equivalence testing with prespecified clinical margins and bootstrap resampling ensured rigorous comparison, while privacy risks were assessed through advanced membership inference attacks (MIA), including shadow MIA and likelihood ratio attacks.

Results: Compared with CSV, Parquet demonstrated enhanced computational efficiency by lowering disk access from 940.2 to 44.2 seconds (95.3% reduction). End-to-end processing latency was substantially reduced across feature transformation (15.0 vs 9.3 s) and model training (8.1 vs 6.7 s). To address complex clinical correlations, we implemented CC and one-vs-rest architectures, which effectively captured interdependencies between disease labels. Classification performance remained statistically equivalent across area under the receiver operating characteristic curve, area under the precision-recall curve, accuracy, and F_1-score, with all differences falling within prespecified clinical equivalence margins (P<.001). Notably, the CC ensemble demonstrated high technical rigor, minimizing Hamming loss (2.2×10^{-4}) and ensuring robustness even in imbalanced cohorts. MIA performed at chance level (area under the curve=0.500), suggesting no measurable increase in privacy risk.

Conclusions: By significantly mitigating data processing bottlenecks, a Parquet-based pipeline enabled high-throughput, large-scale clinical evidence generation without compromising model integrity or patient privacy. This framework provides a scalable and robust infrastructure for precision medicine, facilitating agile multicenter collaborations and real-world data analysis in resource-constrained clinical environments.

(JMIR Med Inform 2026;14:e83487) doi:10.2196/83487

KEYWORDS

electronic health records; medical informatics; predictive performance; data privacy; precision medicine

XSL•FO
RenderX

Introduction

Clinical data, including diagnostic reports, prescriptions, laboratory results, and imaging data, hold substantial potential for precision medicine, predictive modeling, and collaborative research. The volume of electronic health records (EHRs) generated in clinical practice has grown exponentially with the expansion of digital health systems [1-4]. However, as datasets expand, bottlenecks, such as storage overhead, transmission delays, and slow query execution, become increasingly common. These challenges cannot be resolved simply by expanding storage capacity; instead, they necessitate more efficient frameworks for data storage and processing.

In medical informatics, data are most commonly stored in row-oriented formats, such as CSV files, spreadsheets, and relational databases. Although valued for their simplicity and interoperability, these approaches exhibit clear limitations in storage efficiency and analytic scalability. CSV files are uncompressed and large in size, restricting input and output throughput, while relational databases often underperform in column-level operations. In contrast, column-oriented formats offer high compression and selective input and output, making them well suited for high-volume and cloud-based environments. Among these, Apache Parquet is widely used as a standard solution owing to its scalability and interoperability [5,6]. However, few studies have systematically evaluated its end-to-end performance in clinical pipelines that incorporate deidentification procedures.

A further challenge in clinical data use is balancing privacy protection with data utility. Deidentification is mandatory in multicenter research and machine learning, but it may affect statistical distributions or compromise predictive performance. Understanding how storage architectures interact with deidentification to influence analytic outcomes is therefore of both academic and practical importance [7-13].

Unlike prior studies that focused solely on benchmarking computational efficiency, this work integrates privacy risk assessment and formal equivalence testing into a real-world end-to-end pipeline, thereby providing a clinically meaningful evaluation of Parquet as a foundation for precision medicine and collaborative research. Against this background, this study developed a scalable Parquet-based pipeline for high-volume clinical data and empirically assessed its storage efficiency, computational throughput, and preservation of predictive performance. This study also sought to establish a reliable infrastructure that enables cohort construction, secure data sharing, and clinical decision support in multicenter and precision medicine research [14-23].

Methods

Ethical Considerations

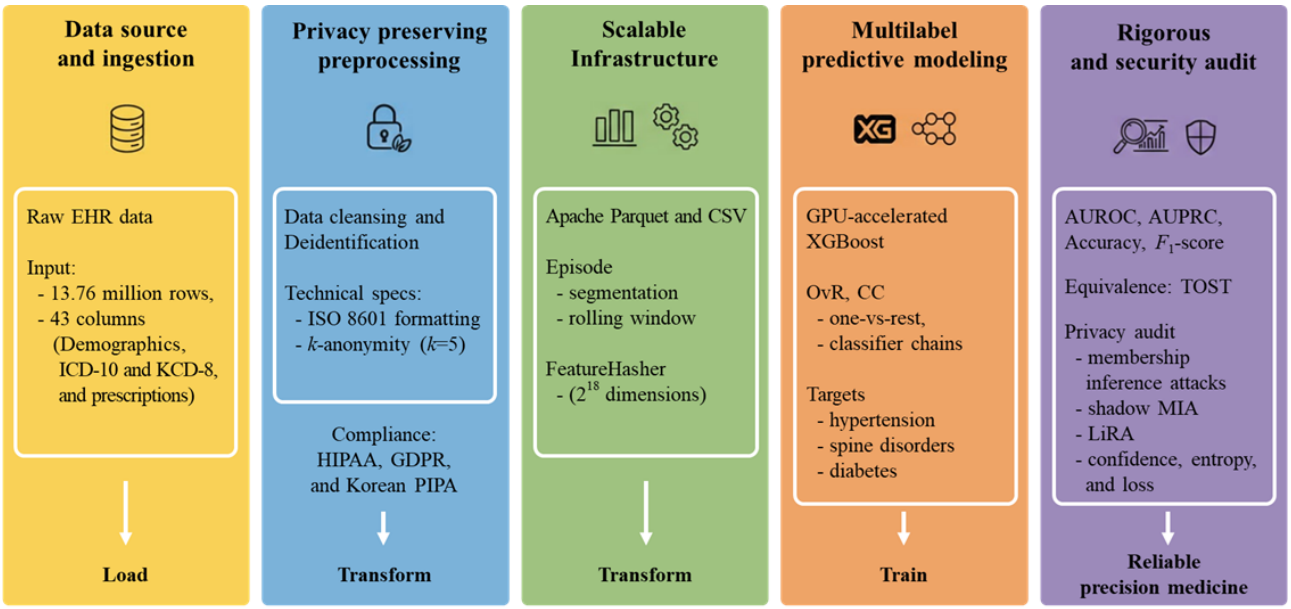
This retrospective observational study was conducted using real-world EHR data. The study protocol was approved by the institutional review board of Wonju Severance Christian Hospital (CR325065). The requirement for informed consent was waived by the institutional review board due to the retrospective nature of the study and the use of deidentified data. All data processing was performed within an independent honest broker system, and investigators had access only to deidentified data. No compensation was provided to the participants as there was no direct contact or intervention involved. The entire process complied with the Personal Information Protection Act of Korea, the Health Insurance Portability and Accountability Act Safe Harbor standard, and the General Data Protection Regulation [24-28].

Integrated End-to-End Clinical Data Processing Workflow

We developed an integrated pipeline connecting 5 core stages: raw EHR ingestion, privacy-preserving preprocessing (ISO 8601 and *k*-anonymity), scalable infrastructure using Apache Parquet, multilabel modeling via graphics processing unit (GPU)–accelerated Extreme Gradient Boosting (XGBoost), and rigorous statistical and privacy audits (Figure 1). To optimize analytic performance, we adopted a column-oriented storage format (Apache Parquet) over the traditional row-oriented structure (eg, CSV) [6]. For comparative evaluation, the same dataset was also maintained in CSV format. While row-oriented structures excel in transactional tasks, columnar storage significantly reduces input and output overhead and enhances compression efficiency for complex analytic queries. This architectural choice is specifically tailored for the high dimensionality and sparse distribution of clinical datasets, ensuring both system scalability and robust privacy protection across the entire workflow.

XSL•FO
RenderX

Figure 1. Integrated workflow for scalable and privacy-preserving clinical data processing. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; CC: classifier chains; EHR: electronic health record; GDPR: General Data Protection Regulation; GPU: graphics processing unit; HIPAA: Health Insurance Portability and Accountability Act; ICD-10: International Classification of Diseases, Tenth Revision; KCD: Korean Classification of Diseases, Eighth Revision; LiRA: Likelihood Ratio Attack; MIA: membership inference attack; OvR: one-vs-rest; PIPA: Personal Information Protection Act; TOST: two 1-sided tests; XGBoost: extreme gradient boosting.



Data Preparation and Transformation

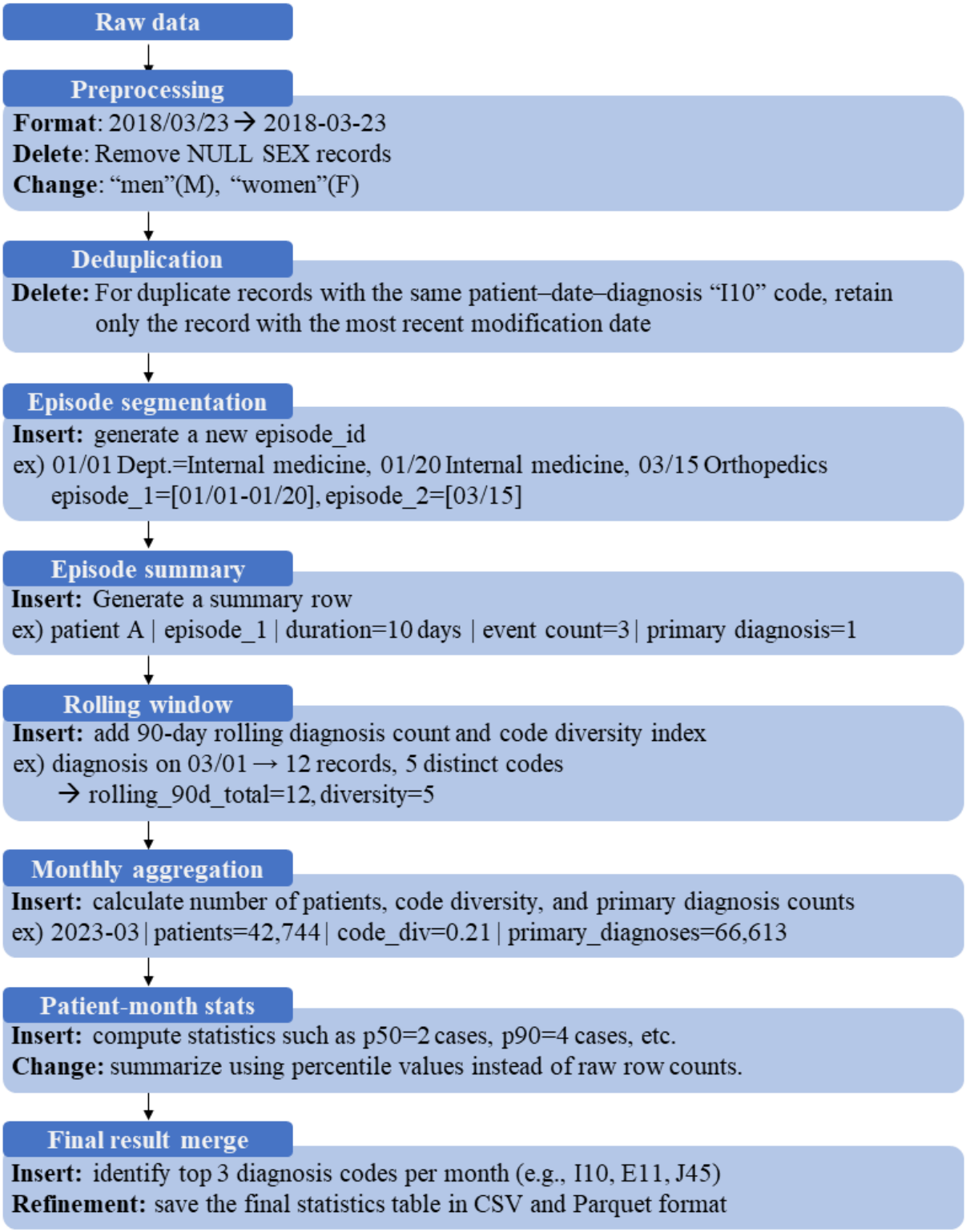
The dataset included patient demographics, diagnostic codes based on the International Classification of Diseases, Tenth Revision (ICD-10) [29-35] and the Korean Classification of Diseases, Eighth Revision (KCD-8), visit type, and prescription records. In total, the dataset comprised approximately 13.76 million rows (43 columns, approximately 3.2 GB), as detailed in Multimedia Appendix 1. After replacing patient identifiers with randomly generated surrogate IDs, quasi-identifiers were generalized to satisfy *k*-anonymity (*k*=5). Data cleansing further ensured quality control by standardizing date fields into the ISO-compliant format (YYYY-MM-DD).

Clinical Data Processing Pipeline

Clinical data processing followed a stepwise pipeline designed to reflect clinical context. In preprocessing, we standardized date formats, removed records with missing sex, and normalized sex codes (eg, “men” to M and “women” to F). Duplicate entries were resolved by retaining only the most recent record for identical combinations of patient, date, and diagnosis code. Visits were segmented into episodes by department changes or temporal gaps, and each episode was assigned an episode identifier. Episode-level summaries included episode duration, event count, and primary diagnosis status. To capture temporal continuity, a 90-day rolling window was applied to compute diagnosis counts and code diversity per patient. Monthly aggregates were then derived, including patient counts, code diversity, and primary diagnosis counts, with patient-month statistics summarized using quantiles (eg, p50 and p90). The diagnostic records for the targeted conditions—including I10.9, E10-E14, and M48.06—were identified and extracted to construct the final summary tables, which were then stored in both CSV and Parquet formats for comparative analysis. The stepwise process is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Stepwise scenario of the clinical data processing pipeline.



Database Workload Benchmarking

To evaluate the efficiency of storage formats, scenario-based workloads were executed in DuckDB (DuckDB Labs) [36] and PostgreSQL (PostgreSQL Global Development Group) [37] environments. Synthetic example data are provided in Multimedia Appendix 2. Identical datasets were loaded into Parquet-based DuckDB and row-store PostgreSQL, and complex SQL queries—adapted from TPC-DS patterns [38,39] (eg, window functions, multidimensional aggregation, and ROLLUP)—were applied to compare performance. Resource use was further examined by varying chunk sizes (10,000-50,000 records) with detailed results in Multimedia Appendix 3.

Predictive Modeling and Multilabel Strategy

Predictive modeling assessed the impact of storage formats on classification performance for hypertension (I10.9), spinal disorders (M48.06), and diabetes (E10-E14). Patient-level

XSL•FO
RenderX

diagnostic histories were transformed into 2^{18} dimensional sparse vectors (approximately 262,000 features) via FeatureHasher (scikit-learn; Python Software Foundation) [40-42]. The dataset was randomly partitioned into training (65%), validation (15%), and test (20%) subsets with a fixed seed of 42. Classification models were implemented using GPU-accelerated XGBoost with positive class weighting to address imbalance [43-45]. Hyperparameter details are provided in Multimedia Appendix 4.

For multilabel classification, we applied one-vs-rest (OvR) [46] and classifier chains (CC) [47]. To mitigate label-order dependence and evaluate interlabel dependencies, we constructed ensembles of 5 strategically ordered chains encompassing a prevalence gradient—from high-prevalence anchors (hypertension and diabetes) to the sparse target, spinal stenosis (M48.06, 0.50%; Multimedia Appendix 5). Prioritizing M48.06, a domain-independent and sparse condition, served to demonstrate model robustness even without the clustering benefits of related disease families (eg, I-series codes). To ensure the rigorous prevention of label leakage and validate the integrity of the performance gains, all defining ICD-10 codes were strictly excluded from the feature space. This setup ensured that the model's predictive capability relied solely on statistical associations across hundreds of thousands of nondiagnostic features. Model performance was evaluated using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, and F_1-score, calculated for each label as well as macroaverages and microaverages.

Statistical Evaluation and Privacy Assessment

Model performance was evaluated using AUROC, AUPRC, accuracy, and F_1-score. To assess model reliability, calibration was further evaluated using the Brier score and expected calibration error (ECE), with detailed results provided in Multimedia Appendix 6. CSV and Parquet results were compared using paired runs with shared identifiers. Normality was assessed via the Shapiro-Wilk test, followed by paired 2-tailed *t* tests or Wilcoxon signed-rank tests with Bonferroni correction (P<.05) [48-50].

Equivalence was assessed via two 1-sided tests (TOST) [51] and 95% bootstrap CIs. Equivalence margins (δ) were prespecified at 0.02 for AUROC or AUPRC and 0.01 for accuracy or F_1-score. These thresholds were selected to be narrower than the 0.05 margin commonly used in clinical equivalence trials [52] and are more conservative than performance variances typically observed in the external validation of clinical artificial intelligence (AI) systems [53]. Such strict margins were intended to constrain observed differences to ranges generally regarded as clinically negligible, particularly for high-performing models operating within an elevated accuracy range (0.98-1.00).

To mitigate stochastic variation, a fixed random seed was applied across all experiments, ensuring that observed differences reflected data handling effects rather than algorithmic randomness. Privacy was further audited using membership inference attacks (MIA)—including confidence, entropy, loss-based (Yeom), shadow, and Likelihood Ratio Attack (LiRA) strategies—under patient-level stratified splits (train: 240,596; test: 240,597) to verify model robustness.

Experimental Setup and Reproducibility

All experiments were conducted on a Linux Ubuntu 22.04 server with a 24-core central processing unit (CPU), 128-GB memory, and an NVIDIA RTX 5090 GPU (CUDA 12.2, cuDNN 8.9, driver 550.54). Data processing used Python (version 3.10; Python Software Foundation) with Pandas 1.5.3, PyArrow 12.0.1, scikit-learn 1.3.2, and XGBoost 2.0.3 (GPU-hist). Multiprocessing was implemented using the standard Python multiprocessing module.

As specified in the modeling strategy, XGBoost was configured with a learning rate of 0.08, maximum depth of 6, and 1000 estimators. To handle class imbalance, task-specific *scale_pos_weight* and an ensemble of 5 CCs were used. All random operations used a fixed seed of 42 for reproducibility. For robust performance evaluation, we conducted 20 independent experimental runs and used 5000-iteration bootstrap resampling to compute 95% CIs. Detailed hyperparameters and the statistical framework for equivalence testing are provided in Multimedia Appendix 4.

Results

Storage and Pipeline Efficiency

Conversion to Parquet substantially improved storage efficiency, reducing file size by nearly 6.9-fold relative to CSV. As summarized in Table 1, Parquet also consistently outperformed CSV across the end-to-end pipeline. File loading (input and output) time decreased by more than 95%, and both feature transformation (FeatureHasher) and model training (XGBoost) were faster under Parquet. Although peak CPU memory was about 1.3 GB higher, GPU memory consumption decreased by 326 MB, resulting in stable overall resource use. Resource consumption remained stable at 9 to 10 GB across chunk sizes, reflecting consistent scalability, with efficiency varying by chunk size. The values in Table 1 represent peak use in end-to-end runs, with minor variations arising from measurement conditions, whereas CPU, memory, and disk input and output patterns remained consistent across chunk sizes (Multimedia Appendix 3).

XSL•FO
RenderX

Table 1. System performance comparison between CSV and Parquet formats.

Metric	CSV, mean (SD)	Parquet, mean (SD)	Difference[a] (95% CI)
Time (s)			
Input and output duration	940.2 (53.3)	44.2 (1.6)	−896.0 (−930.0 to −860.0)
Feature transformation	15.0 (0.9)	9.3 (0.7)	−5.7 (−7.0 to −4.5)
Model training	8.1 (1.2)	6.7 (0.9)	−1.4 (−2.5 to −0.3)
Resource use			
Memory maximum (GB)	21.2 (0.8)	22.5 (0.6)	1.3 (0.1 to 2.0)
GPU[b] memory maximum (MB)	4873.0 (582.6)	4547.0 (596.9)	−326.0 (−450.0 to −200.0)

[a]Calculated as the Parquet mean minus CSV mean.

[b]GPU: graphics processing unit.

Classification Performance and Calibration

Single-label results are presented in Table 2. All 3 tasks achieved AUROC values above 0.920, with AUPRC ranging from 0.488 to 0.580. Hypertension (I10.9) yielded the highest F_1-score (0.599), followed by spinal disorders (M48.06, 0.536) and diabetes (E10-E14, 0.521). Macroaveraged performance reached an AUROC of 0.920, an AUPRC of 0.545, an accuracy of 0.934, and an F_1-score of 0.552. Calibration analyses [54,55], reported in Multimedia Appendix 6, demonstrated excellent reliability for hypertension (Brier score 2.1×10^{-5}; ECE 6.0×10^{-6}), while spinal disorders (ECE=0.016) and diabetes (ECE=0.018) showed mild miscalibration but remained within recommended thresholds (ECE<0.020).

Table 2. Per-label classification performance using extreme gradient boosting.

Labels	AUROC[a]	AUPRC[b]	Accuracy	F_1-score
Hypertension	0.920	0.580	0.906	0.599
Spine disorder	0.941	0.488	0.982	0.536
Diabetes[c]	0.899	0.567	0.915	0.521
Macroaverage	0.920	0.545	0.934	0.552

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]Diabetes mellitus (International Classification of Diseases, Tenth Revision codes: E10-E14).

Table 2 evaluates label-specific discrimination under substantial class imbalance, whereas Table 3 summarizes overall multilabel prediction performance using microaveraged aggregation across all label–instance pairs. Specifically, Table 2 focuses on the predictive performance of individual sparse labels, while Table 3 reflects aggregated performance across all prediction events. Multilabel results in Table 3 confirmed high-performance consistency for both OvR and CC, with AUROC >0.999 and AUPRC >0.998. Subset accuracy was identical (0.999), but CC showed slight improvements in Hamming loss and Jaccard index [56], suggesting modest benefits from capturing label dependencies. A numerical gap was observed between per-label AUPRC (approximately 0.5; Table 2) and multilabel AUPRC (near 1.0; Table 3). This reflects the technical difference in aggregation: multilabel metrics were calculated using microaveraging to evaluate the overall model efficacy.

XSL•FO
RenderX

Table 3. Multilabel classification performance.

Models	AUROC[a]	AUPRC[b]	Hamming loss	Subset accuracy	Jaccard index
One-vs-rest					
Microaverage	0.999	0.998	2.4×10^{-4}	0.999	0.997
Macroaverage	0.999	0.998	—[c]	—	—
Classifier chains					
Microaverage	0.999	0.998	2.2×10^{-4}	0.999	0.997
Macroaverage	0.999	0.998	—	—	—

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]Not applicable.

Although discrimination remained high across all tasks, calibration analyses revealed modest overconfidence for diabetes and spinal disorder models. These deviations, while within accepted thresholds, emphasize the importance of calibration-aware adjustments when deploying predictive models in clinical workflows.

Database Benchmarking

As summarized in Table 4, DuckDB executed analytic workloads with an approximately 28-fold reduction in wall-clock time compared with PostgreSQL, which required 2546.4 seconds. DuckDB also exhibited higher CPU use (72.1% vs 39.6%), while memory consumption was comparable across systems.

Table 4. Performance comparison of DuckDB and PostgreSQL in scenario-based SQL workloads.

Metrics	DuckDB[a]	PostgreSQL[b]
Wall-clock time (s)	90.2	2546.4
CPU[c] use (%)	72.1	39.6
Memory use (GB)	13.7	14.8
Read throughput (MB/s)	554.7	365.2
Write throughput (MB/s)	1116.0	998.0

[a]DuckDB: column-oriented database optimized for analytical workloads.

[b]PostgreSQL: row-oriented relational database primarily designed for transactional workloads, with support for analytical queries.

[c]CPU: central processing unit.

Equivalence Analysis

Paired experiments (20 runs per condition) were conducted using shared run identifiers, with median and IQR values provided in Table S1 in Multimedia Appendix 7. Several metrics converged across all repetitions, producing zero IQR values and demonstrating high algorithmic stability. Overall, both formats achieved robust numerical consistency across all experimental iterations; for instance, mean AUROC was 1.000 for CSV and 0.984 for Parquet ($\Delta=-0.015$). While other metrics (AUPRC, accuracy, and F_1-score) were slightly higher under the CSV format, paired *t* tests and Wilcoxon signed-rank tests confirmed statistical significance after Bonferroni correction ($P<.001$; Table S2 in Multimedia Appendix 7). However, these numerical variations remained minimal.

Notably, all observed differences fell within the prespecified equivalence margins ($\delta=0.02$ for AUROC and AUPRC and $\delta=0.01$ for accuracy and F_1-score). TOST confirmed statistical equivalence ($P<.05$) across all evaluated metrics, and bootstrap 95% CIs for AUROC Δ (−0.016 to −0.015) further reflected near-zero variability consistent with asymptotic performance stability (Table S3 in Multimedia Appendix 7).

Privacy Assessment

The MIA evaluations, performed based on established methodologies [57-59] and summarized in Table 5, produced area under the curve (AUC) values of 0.501, 0.502, and 0.499 for confidence, entropy, and loss-based strategies, respectively. The maximum advantage did not exceed 0.007, and the gain over random guessing was ≤0.003. Attack performance was indistinguishable from random guessing, indicating no additional privacy risk. Additional shadow and LiRA [60,61] results are available in Multimedia Appendix 8, corroborating these findings.

XSL•FO
RenderX

Table 5. Membership inference attack results across different attack strategies.

Attack strategy	Attack AUC[a] (95% CI)	Advantage	Best accuracy	Gain vs baseline
Confidence	0.501 (0.501-0.503)	0.006	0.503	0.003
Entropy	0.502 (0.502-0.503)	0.007	0.503	0.003
Loss (Yeom)	0.499 (0.498-0.499)	0.002	0.500	0.001

[a]AUC: area under the curve.

Discussion

Principal Findings

This study successfully implemented an integrated end-to-end pipeline (Figure 1) that ensures a seamless transition from raw EHR data ingestion to predictive modeling. By leveraging Apache Parquet, the framework effectively addressed storage and processing bottlenecks—specifically, input and output latency—inherent in traditional row-oriented workflows, substantially enhancing data throughput and accelerating the analytic preparatory phase while maintaining high predictive fidelity. Furthermore, our findings indicate that the minor performance discrepancies between formats are attributable to systematic data handling effects under a strictly controlled environment. Beyond these computational efficiencies, privacy evaluations demonstrated that the optimized workflow maintains rigorous data protection standards, providing a secure foundation for multicenter research operations without introducing additional vulnerabilities.

Although paired statistical tests yielded highly significant *P* values ($P<.001$), this result reflects the sensitivity of paired testing under large-scale data and repeated bootstrap evaluation rather than clinically meaningful divergence. From an equivalence perspective, all performance differences remained well within the prespecified margins, confirming that the Parquet-based pipeline achieves practical performance equivalence to the traditional CSV workflow.

The substantial improvement in multilabel AUPRC compared with per-label performance highlights the system-level predictive advantage gained by explicitly modeling interlabel dependencies. While individual disease prediction remains challenging under substantial class imbalance, the multilabel formulation enables the model to capture clinically meaningful comorbidity structures, thereby reframing the observed performance gap as a reflection of holistic predictive efficacy rather than a metric artifact. This gap was also driven by the extreme class imbalance inherent in clinical datasets (eg, 0.50% prevalence for spinal disorders), where high AUROCs can coexist with lower F_1-scores. By integrating these dependencies within the CC architecture, the model leveraged frequent clinical co-occurrence patterns—such as hypertension and diabetes—to stabilize predictions for sparse conditions, including spinal stenosis. Importantly, the consistent performance observed despite the strict exclusion of all defining ICD-10 codes provided strong evidence against label leakage and underscored the model's reliance on nondiagnostic clinical associations.

Clinical Implications and Research Operations

The clinical implications were considerable. By substantially shortening preparatory phases and enabling analyses to be completed within hours rather than days, Parquet-based pipelines allowed investigators to progress more rapidly from raw data to actionable evidence. Such improvements not only accelerated study initiation but also facilitated the integration of predictive models into research and clinical workflows. In multicenter settings, scalability supported the efficient integration of heterogeneous patient populations without excessive computational cost, thereby enhancing reproducibility and advancing precision medicine initiatives. These efficiencies were directly relevant to clinical workflows and regulatory-grade analyses. They enabled interim analyses and timely regulatory submissions, supporting more responsive and efficient research operations.

Privacy and Security Safeguards

Privacy evaluation reinforced the suitability of this approach. Deidentification procedures satisfied k-anonymity, and MIA yielded chance-level performance. The chance-level performance of multiple MIA methods confirmed that computational gains do not compromise patient confidentiality, a prerequisite for secure collaboration across institutions. Patient-level stratification and consistent pipeline design acted as effective safeguards. While residual risks remain inherent to secondary data use, our findings suggested that Parquet does not introduce additional vulnerabilities. For collaborative research networks—where privacy assurances are essential for sharing—these results provide practical reassurance.

Interpretation of Predictive Performance and Label Dependencies

Classification analyses revealed stable performance for common conditions such as hypertension, whereas imbalanced disorders such as diabetes and spinal disease yielded lower scores. This pattern highlights the need to account for dataset structure and interlabel dependencies when interpreting predictive outputs. The elevated AUROC values likely reflect the combination of high-prevalence conditions and extensive diagnostic feature sets. Consistent with the modeling strategy, defining codes were excluded from the feature space; however, correlated diagnostic features may still have influenced performance estimates.

While performance convergence at elevated levels constrained the sensitivity to subtle comparative differences, it underscored the robustness and reproducibility of the pipeline, supporting its suitability for large-scale clinical research, where stability and reliability are critical. Predictive reliability was not compromised by storage structure, supporting Parquet as a

XSL•FO
RenderX

dependable option for secondary use of EHR data in diverse research settings.

Model Reliability and Calibration

Calibration analyses further showed that hypertension models were well aligned with observed risks, whereas diabetes and spinal disorder tasks exhibited modest miscalibration. These results underscore that even high-performing classifiers may generate overconfident probability estimates, a limitation directly relevant to clinical decision support. Future studies should evaluate calibration-aware techniques, such as temperature scaling or isotonic regression, to improve the reliability of probability estimates (Multimedia Appendix 6).

Beyond technical validation, the findings demonstrated practical value for clinical research operations. Large-scale trials and observational studies often face delays in cohort assembly and iterative validation; Parquet-based pipelines shorten these steps, enabling more adaptive study designs. In multicenter collaborations, they promote secure and efficient data exchange, reducing delays that commonly impede cooperative studies.

Limitations

Several limitations should be acknowledged. First, the datasets analyzed were derived from specific institutional cohorts, which may constrain generalizability to other health care systems. Second, the elevated performance baseline of the predictive models limited the sensitivity to subtle differences between CSV and Parquet formats. Third, while MIA offered a widely accepted proxy for privacy risk, broader adversarial evaluations were not included. Fourth, the evaluation was primarily conducted in on-premise settings, so the scalability of findings in distributed or cloud-based environments remains to be established. Finally, despite the clear efficiency advantages of Parquet, its adoption may introduce a technical learning curve for clinical researchers who are less familiar with column-oriented storage formats or big-data frameworks. Transitioning from traditional row-oriented workflows to optimized columnar architectures requires initial investments in technical training and infrastructure adjustment. However, these upfront efforts are likely to be offset by the long-term gains in scalability and processing speed demonstrated in our results, ultimately supporting more efficient and reproducible clinical evidence generation.

Future Work

Future research should extend the evaluation of Parquet-based pipelines to distributed and cloud-based environments, focusing on both scalability and cost-effectiveness. In addition, validation across hybrid and federated architectures will be important to confirm applicability in multicenter collaborations where data cannot be centralized. Beyond structured EHR data, future studies should also investigate applications to unstructured modalities, such as clinical narratives, imaging metadata, and time-series signals. Such efforts would test the versatility of Parquet-based pipelines in supporting multimodal learning and integrative analyses. Privacy assessments should likewise be expanded to encompass attribute inference, linkage, and other adversarial scenarios, ideally within federated or privacy-preserving analytic frameworks. Moreover, interoperability across international coding standards should be validated using external multicenter datasets to ensure global applicability. Pursuing these directions will enhance the technical robustness of Parquet-based infrastructures while reinforcing their clinical relevance, ultimately enabling secure collaboration and more efficient integration of evidence into patient care.

Conclusions

This study demonstrated that Parquet-based end-to-end processing effectively alleviated the storage and computational bottlenecks inherent in traditional row-oriented and relational database workflows while maintaining predictive fidelity and privacy safeguards. By improving computational efficiency and enabling the rapid preparation of analysis-ready cohorts, Parquet reduces the resources required for large-scale clinical data handling. These efficiencies translate into tangible research benefits, enabling earlier study initiation and more adaptive study designs across diverse clinical research settings.

The integration of efficiency with privacy safeguards further ensures that accelerated workflows remain suitable for secure data sharing across institutions. In parallel, these capabilities support regulatory-grade analyses and health technology assessments, where timely and reproducible evidence is critical. For clinicians and researchers, the central implication is that Parquet-based pipelines reduce the manual effort and computational overhead required to transform raw EHR data into analysis-ready cohorts. This capability not only strengthens multicenter collaboration but also facilitates the timely translation of evidence into clinical decision support.

Ultimately, by aligning computational efficiency with privacy-conscious infrastructure, Parquet-based pipelines reinforce the evidence base that informs clinical decision-making, positioning this approach as a robust foundation for sustainable precision medicine.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (grant RS-2025-24683718). This research was supported by the Ministry of Trade, Industry and Energy, Republic of Korea, under the project titled “Research and Development on AI Standardization in the Healthcare Industry Based on Human Big Data and Artificial Intelligence” (grant RS-2025-13882968), funded through the Korea Evaluation Institute of Industrial Technology. This research was supported by a grant from the medical data-driven hospital support project through the Korea Health Information Service, funded by the Ministry of Health and Welfare, Republic of Korea. This research was supported by the Regional Innovation

XSL•FO
RenderX

System and Education program through the Gangwon Regional Innovation System and Education Center, funded by the Ministry of Education and the Gangwon State, Republic of Korea (grant 2025-RISE-10-006).

Funding

This study was funded by the following grants: (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (grant RS-2025-24683718); (2) a grant from the medical data-driven hospital support project through the Korea Health Information Service, funded by the Ministry of Health and Welfare, Republic of Korea; and (3) the Regional Innovation System and Education program through the Gangwon Regional Innovation System and Education Center, funded by the Ministry of Education and the Gangwon State, Republic of Korea (grant 2025-RISE-10-006). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KHL
Data curation: JL
Formal analysis: JL, KHL
Methodology: SH
Project administration: KHL
Resources: JL
Writing – original draft: JL, SH
Writing – review and editing: KHL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Column definitions for the structured clinical data table.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v14i1e83487_app1.xlsx]

Multimedia Appendix 2
Synthetic example data used for scenario-based workload evaluation.
[XLSX File (Microsoft Excel File), 24 KB - medinform_v14i1e83487_app2.xlsx]

Multimedia Appendix 3
Resource use profiles during Parquet-based data processing across different chunk sizes.
[DOCX File , 252 KB - medinform_v14i1e83487_app3.docx]

Multimedia Appendix 4
Hyperparameter configurations and the formal statistical framework for equivalence testing.
[DOCX File , 16 KB - medinform_v14i1e83487_app4.docx]

Multimedia Appendix 5
Detailed demographic characteristics and visit statistics of the study cohort.
[PNG File , 235 KB - medinform_v14i1e83487_app5.png]

Multimedia Appendix 6
Calibration metrics (Brier score and expected calibration error [ECE] and ECE) for each binary classification task.
[DOCX File , 14 KB - medinform_v14i1e83487_app6.docx]

Multimedia Appendix 7
Integrated statistical performance and equivalence analysis.
[DOCX File , 18 KB - medinform_v14i1e83487_app7.docx]

XSL•FO
RenderX

Multimedia Appendix 8
Membership inference results with shadow and Likelihood Ratio Attack attackers.
[DOCX File , 15 KB - medinform_v14i1e83487_app8.docx]

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
2. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
3. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
4. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
5. Vohra D. Apache Parquet. In: Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. Berkeley, CA: Apress; 2016:325-335.
6. Apache parquet documentation. Apache Software Foundation. URL: https://parquet.apache.org/ [accessed 2025-06-01]
7. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010 Aug 02;10(1):70 [FREE Full text] [doi: 10.1186/1471-2288-10-70] [Medline: 20678228]
8. Negash B, Katz A, Neilson CJ, Moni M, Nesca M, Singer A, et al. De-identification of free text data containing personal health information: a scoping review of reviews. Int J Popul Data Sci 2023 Dec 12;8(1):2153 [FREE Full text] [doi: 10.23889/ijpds.v8i1.2153] [Medline: 38414537]
9. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 2012 Jul;50 Suppl(Suppl):S82-101 [FREE Full text] [doi: 10.1097/MLR.0b013e3182585355] [Medline: 22692265]
10. El-Hayek C, Barzegar S, Faux N, Doyle K, Pillai P, Mutch SJ, et al. An evaluation of existing text de-identification tools for use with patient progress notes from Australian general practice. Int J Med Inform 2023 May;173:105021. [doi: 10.1016/j.ijmedinf.2023.105021] [Medline: 36870249]
11. Liu J, Gupta S, Chen A, Wang CK, Mishra P, Dai HJ, et al. OpenDeID pipeline for unstructured electronic health record text notes based on rules and transformers: deidentification algorithm development and validation study. J Med Internet Res 2023 Dec 06;25:e48145 [FREE Full text] [doi: 10.2196/48145] [Medline: 38055317]
12. Jabet A, Bérot V, Chiarabini T, Dellière S, Bosshard PP, Siguier M, et al. Trichophyton mentagrophytes ITS genotype VII infections among men who have sex with men in France: an ongoing phenomenon. J Eur Acad Dermatol Venereol 2025 Feb;39(2):407-415. [doi: 10.1111/jdv.20439] [Medline: 39587983]
13. Crane HM, Nance RM, Ruderman SA, Drumright LN, Mixson LS, Heckbert SR, et al. Smoking and type 1 versus type 2 myocardial infarction among people with HIV in the United States: results from the center for AIDS research network integrated clinical systems cohort. J Assoc Nurses AIDS Care 2024 Sep 6;35(6):507-518. [doi: 10.1097/JNC.0000000000000492] [Medline: 39241219]
14. Solomon J, Dauber-Decker K, Richardson S, Levy S, Khan S, Coleman B, et al. Integrating clinical decision support into electronic health record systems using a novel platform (EvidencePoint): developmental study. JMIR Form Res 2023 Oct 19;7:e44065 [FREE Full text] [doi: 10.2196/44065] [Medline: 37856193]
15. Trinkley KE, Kroehl ME, Kahn MG, Allen LA, Bennett TD, Hale G, et al. Applying clinical decision support design best practices with the practical robust implementation and sustainability model versus reliance on commercially available clinical decision support tools: randomized controlled trial. JMIR Med Inform 2021 Mar 22;9(3):e24359 [FREE Full text] [doi: 10.2196/24359] [Medline: 33749610]
16. Yoo J, Lee J, Rhee PL, Chang DK, Kang M, Choi JS, et al. Alert override patterns with a medication clinical decision support system in an academic emergency department: retrospective descriptive study. JMIR Med Inform 2020 Nov 04;8(11):e23351 [FREE Full text] [doi: 10.2196/23351] [Medline: 33146626]
17. Jing X, Min H, Gong Y, Biondich P, Robinson D, Law T, et al. Ontologies applied in clinical decision support system rules: systematic review. JMIR Med Inform 2023 Jan 19;11:e43053 [FREE Full text] [doi: 10.2196/43053] [Medline: 36534739]
18. Sommers SW, Tolle HJ, Trinkley KE, Johnston CG, Dietsche CL, Eldred SV, et al. Clinical decision support to increase emergency department Naloxone Coprescribing: implementation report. JMIR Med Inform 2024 Nov 06;12:e58276 [FREE Full text] [doi: 10.2196/58276] [Medline: 39504560]
19. Barton HJ, Maru A, Leaf MA, Hekman DJ, Wiegmann DA, Shah MN, et al. Academic detailing as a health information technology implementation method: supporting the design and implementation of an emergency department-based clinical decision support tool to prevent future falls. JMIR Hum Factors 2024 Apr 18;11:e52592 [FREE Full text] [doi: 10.2196/52592] [Medline: 38635318]

XSL•FO
RenderX

20. Dong J, Jin Z, Li C, Yang J, Jiang Y, Li Z, et al. Machine learning models with prognostic implications for predicting gastrointestinal bleeding after coronary artery bypass grafting and guiding personalized medicine: multicenter cohort study. J Med Internet Res 2025 Mar 06;27:e68509 [FREE Full text] [doi: 10.2196/68509] [Medline: 40053791]
21. Jung C, Mamandipoor B, Fjølner J, Bruno RR, Wernly B, Artigas A, et al. Disease-course adapting machine learning prognostication models in elderly patients critically ill with COVID-19: multicenter cohort study with external validation. JMIR Med Inform 2022 Mar 31;10(3):e32949 [FREE Full text] [doi: 10.2196/32949] [Medline: 35099394]
22. Alvarez-Romero C, Martinez-Garcia A, Ternero Vega J, Díaz-Jiménez P, Jiménez-Juan C, Nieto-Martín MD, et al. Predicting 30-day readmission risk for patients with chronic obstructive pulmonary disease through a federated machine learning architecture on Findable, Accessible, Interoperable, and Reusable (FAIR) data: development and validation study. JMIR Med Inform 2022 Jun 02;10(6):e35307 [FREE Full text] [doi: 10.2196/35307] [Medline: 35653170]
23. Blasi L, Bordonaro R, Serretta V, Piazza D, Firenze A, Gebbia V. Virtual clinical and precision medicine tumor boards-cloud-based platform-mediated implementation of multidisciplinary reviews among oncology centers in the COVID-19 era: protocol for an observational study. JMIR Res Protoc 2021 Sep 10;10(9):e26220 [FREE Full text] [doi: 10.2196/26220] [Medline: 34387553]
24. Ahmed A, Shahzad A, Naseem A, Ali S, Ahmad I. Evaluating the effectiveness of data governance frameworks in ensuring security and privacy of healthcare data: a quantitative analysis of ISO standards, GDPR, and HIPAA in blockchain technology. PLoS One 2025 May 23;20(5):e0324285 [FREE Full text] [doi: 10.1371/journal.pone.0324285] [Medline: 40408373]
25. Bradford L, Aboy M, Liddell K. International transfers of health data between the EU and USA: a sector-specific approach for the USA to ensure an 'adequate' level of protection. J Law Biosci 2020;7(1):lsaa055 [FREE Full text] [doi: 10.1093/jlb/lsaa055] [Medline: 34221424]
26. Gomase VS, Ghatule AP, Sharma R, Sardana S. Cybersecurity and compliance in clinical trials: the role of artificial intelligence in secure healthcare management. Rev Recent Clin Trials 2025 Apr 25;20(4):332-351. [doi: 10.2174/0115748871366467250413070850] [Medline: 40277117]
27. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res 2021 Feb 25;23(2):e25120 [FREE Full text] [doi: 10.2196/25120] [Medline: 33629963]
28. Broen K, Trangucci R, Zelner J. Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. Int J Health Geogr 2021 Jan 07;20(1):3 [FREE Full text] [doi: 10.1186/s12942-020-00256-8] [Medline: 33413390]
29. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. JMIR Med Inform 2021 Aug 31;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
30. Chen PF, Chen KC, Liao WC, Lai F, He TL, Lin SC, et al. Automatic international classification of diseases coding system: deep contextualized language model with rule-based approaches. JMIR Med Inform 2022 Jun 29;10(6):e37557 [FREE Full text] [doi: 10.2196/37557] [Medline: 35767353]
31. Chen PF, He TL, Lin SC, Chu YC, Kuo CT, Lai F, et al. Training a deep contextualized language model for international classification of diseases, 10th revision classification via federated learning: model development and validation study. JMIR Med Inform 2022 Nov 10;10(11):e41342 [FREE Full text] [doi: 10.2196/41342] [Medline: 36355417]
32. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform 2019 Nov 29;7(4):e14325 [FREE Full text] [doi: 10.2196/14325] [Medline: 31553307]
33. Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G. Neural translation and automated recognition of ICD-10 medical entities from natural language: model development and performance assessment. JMIR Med Inform 2022 Apr 11;10(4):e26353 [FREE Full text] [doi: 10.2196/26353] [Medline: 35404262]
34. Ghasemi P, Lee J. Unsupervised feature selection to identify important ICD-10 and ATC codes for machine learning on a cohort of patients with coronary heart disease: retrospective study. JMIR Med Inform 2024 Jul 26;12:e52896 [FREE Full text] [doi: 10.2196/52896] [Medline: 39087585]
35. Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, et al. Evaluating a natural language processing-driven, AI-assisted international classification of diseases, 10th revision, clinical modification, coding system for diagnosis related groups in a real hospital environment: algorithm development and validation study. J Med Internet Res 2024 Sep 20;26:e58278 [FREE Full text] [doi: 10.2196/58278] [Medline: 39302714]
36. DuckDB documentation. DuckDB Foundation. URL: https://duckdb.org/ [accessed 2025-05-25]
37. PostgreSQL documentation. PostgreSQL Global Development Group. URL: https://www.postgresql.org/docs/ [accessed 2025-06-10]
38. Tardío R, Maté A, Trujillo J. Beyond TPC-DS, a benchmark for Big Data OLAP systems (BDOLAP-Bench). Futur Gener Comp Syst 2022 Jul;132:136-151. [doi: 10.1016/j.future.2022.02.015]
39. Ding B, Chaudhuri S, Gehrke J, Narasayya V. DSB: a decision support benchmark for workload-driven and traditional database systems. Proc VLDB Endow 2021 Oct 28;14(13):3376-3388. [doi: 10.14778/3484224.3484234]

XSL•FO
RenderX

40. Pedragosa M, Riera G, Casella V, Esteve-Codina A, Steuerman Y, Seth C, et al. Linking cell dynamics with gene coexpression networks to characterize key events in chronic virus infections. Front Immunol 2019 May 3;10:1002 [FREE Full text] [doi: 10.3389/fimmu.2019.01002] [Medline: 31130969]
41. Pang Y, Zhang X, Gao R, Xu L, Shen M, Shi H, et al. Efficacy of web-based self-management interventions for depressive symptoms: a meta-analysis of randomized controlled trials. BMC Psychiatry 2021 Aug 11;21(1):398 [FREE Full text] [doi: 10.1186/s12888-021-03396-8] [Medline: 34380440]
42. Bugiardini R, Cenko E. Sex differences in myocardial infarction deaths. Lancet 2020 Jul 11;396(10244):72-73. [doi: 10.1016/S0140-6736(20)31049-7] [Medline: 32445694]
43. Kaul V, Gross S, Corbett FS, Malik Z, Smith M, Tofani C, et al. Clinical utility of wide-area transepithelial sampling with three-dimensional computer-assisted analysis (WATS3D) in identifying Barrett's esophagus and associated neoplasia. Dis Esophagus 2020 Dec 07;33(12):32607543. [doi: 10.1093/dote/doaa069] [Medline: 32607543]
44. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
45. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010 Jun;48(6 Suppl):S106-S113. [doi: 10.1097/MLR.0b013e3181de9e17] [Medline: 20473190]
46. Rifkin R, Klautau A. In defense of one-vs-all classification. J Mach Learn Res 2004;5:41. [doi: 10.5555/1005332.1005336]
47. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Mach Learn 2011 Jun 30;85(3):333-359. [doi: 10.1007/S10994-011-5256-5]
48. Dunn OJ. Multiple comparisons among Means. J Am Stat Assoc 1961 Mar;56(293):52-64. [doi: 10.1080/01621459.1961.10482090]
49. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan 21;310(6973):170 [FREE Full text] [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]
50. Choi JE, Kang HW, Hong YM, Sohn S. C-reactive protein and N-terminal pro-brain natriuretic peptide discrepancy: a differentiation of adenoviral pharyngoconjunctival fever from Kawasaki disease. Korean J Pediatr 2018 Jan;61(1):12-16 [FREE Full text] [doi: 10.3345/kjp.2018.61.1.12] [Medline: 29441107]
51. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 1987 Dec 1;15(6):657-680. [doi: 10.1007/BF01068419] [Medline: 3450848]
52. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med 2011 Feb 21;26(2):192-196 [FREE Full text] [doi: 10.1007/s11606-010-1513-8] [Medline: 20857339]
53. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2025 Jan 06;25(1):5 [FREE Full text] [doi: 10.1186/s12911-024-02830-7] [Medline: 39762808]
54. Glenn WB. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2]
55. Dawood T, Chen C, Sidhu BS, Ruijsink B, Gould J, Porter B, et al. Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images. Med Image Anal 2023 Aug;88:102861 [FREE Full text] [doi: 10.1016/j.media.2023.102861] [Medline: 37327613]
56. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. IEEE Trans Med Imaging 2020 Nov;39(11):3679-3690. [doi: 10.1109/tmi.2020.3002417]
57. Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. J Biomed Inform 2022 Jan;125:103977 [FREE Full text] [doi: 10.1016/j.jbi.2021.103977] [Medline: 34920126]
58. Cobilean V, Mavikumbure HS, Drake D, Stuart M, Manic M. Investigating membership inference attacks against CNN models for BCI systems. IEEE J Biomed Health Inform 2025 Nov;29(11):8164-8174. [doi: 10.1109/jbhi.2025.3593443]
59. Chen J, Wang WH, Shi X. Differential privacy protection against membership inference attack on machine learning for genomic data. Pac Symp Biocomput 2021;26:26-37 [FREE Full text] [Medline: 33691001]
60. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham VV, et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet 2019 Apr 18;51(4):749-754 [FREE Full text] [doi: 10.1038/s41588-019-0366-2] [Medline: 30886424]
61. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Zhao Y, Jin H, et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat Genet 2022 Oct 26;54(10):1564-1571 [FREE Full text] [doi: 10.1038/s41588-022-01180-2] [Medline: 36163278]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUPRC: area under the precision-recall curve

XSL•FO
RenderX

AUROC: area under the receiver operating characteristic curve
CC: classifier chains
CPU: central processing unit
ECE: expected calibration error
EHR: electronic health record
GPU: graphics processing unit
ICD-10: International Classification of Diseases, Tenth Revision
KCD-8: Korean Classification of Diseases, Eighth Revision
LiRA: Likelihood Ratio Attack
MIA: membership inference attack
OvR: one-vs-rest
TOST: two 1-sided tests
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 03.Sep.2025; peer-reviewed by A Kalluchi, Y He; comments to author 19.Jan.2026; accepted 05.Feb.2026; published 04.Mar.2026.

Please cite as:
Lee J, Hwang S, Lee KH
Scalable and Privacy-Conscious End-to-End Processing of Large-Scale Clinical Data for Precision Medicine: Empirical Evaluation Study
JMIR Med Inform 2026;14:e83487
URL: https://medinform.jmir.org/2026/1/e83487
doi:10.2196/83487
PMID:

©Jungwoo Lee, Sangwon Hwang, Kyu Hee Lee. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Bilayer Feature Fusion Framework for Pan-Cancer Survival Prediction Based on Multihead Attention and Adaptive Differential Privacy: Model Development and Validation Study

Yun Chen[1], BE; Zhifang Deng[1], BE; Lili Wang[1], ME; Huanhuan Wang[1], PhD; Xiang Wu[1,2,3], PhD

[1]School of Medical Information and Engineering, Xuzhou Medical University, Xuzhou, China

[2]Yunlong Lake Laboratory of Deep Underground Science and Engineering, Xuzhou, China

[3]The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China

Corresponding Author:
Xiang Wu, PhD
School of Medical Information and Engineering
Xuzhou Medical University
No. 209 Tongshan Road
Xuzhou, 221004
China
Phone: 86 18752167676
Email: wuxiang@xzhmu.edu.cn

Abstract

Background: In the field of precision medicine, pan-cancer survival prediction is crucial for individualized oncology diagnosis and treatment. Although multimodal data fusion techniques have significantly improved prediction accuracy, existing studies generally overlook the sensitivity of medical data and the need for privacy protection.

Objective: To address the aforementioned problem, this study aims to propose a bilayer feature fusion framework based on the multihead attention mechanism and adaptive differential privacy, which balances precise feature extraction and sensitive data protection.

Methods: Specifically, the multihead attention mechanism was integrated into bilayer feature extraction and fusion. Layer-wise relevance analysis was used to calculate the correlation between features and outcomes, and Laplacian noise was adaptively added based on the calculation results to achieve collaborative optimization of precise feature extraction and sensitive data protection. Additionally, the concordance index (C-index) and 5-fold cross-validation were used to compare the proposed method with state-of-the-art approaches.

Results: The proposed model achieved superior performance in both pan-cancer and single-cancer survival prediction, validated via the C-index and 5-fold cross-validation. In pan-cancer scenarios, the trimodal combination of clinical data, messenger RNA data, and microRNA expression data achieved the highest C-index of 0.799, outperforming multiple existing multimodal survival prediction approaches. After adaptive Laplacian noise injection for privacy protection, the model's accuracy decreased by only 0.01-0.03 while satisfying -differential privacy. For single-cancer prediction, the proposed method achieved higher C-index values in 18 out of 20 cancer types compared with representative deep learning–based survival models, with 7 types showing significant improvements (C-index difference >0.1) and more stable performance distributions. Furthermore, pan-cancer–trained models generally outperformed single-cancer–trained counterparts in most cancer types, highlighting the value of shared predictive features across cancers.

Conclusions: This study provides a solution that balances prediction accuracy and privacy security for pan-cancer survival prediction, laying the foundation for the efficient use of medical data under privacy protection. Future work may further integrate pathological images and proteomics data to expand the model's applications in cancer subtype classification and biomarker discovery.

(JMIR Med Inform 2026;14:e83743) doi:10.2196/83743

KEYWORDS

adaptive differential privacy; multihead attention; pan-cancer survival prediction

XSL•FO
RenderX

Introduction

In the era of precision medicine, pan-cancer survival prediction refers to the task of modeling and predicting patient survival outcomes across multiple cancer types simultaneously by leveraging shared and cancer-specific prognostic patterns. As a core component of individualized oncology diagnosis and treatment, pan-cancer survival prediction has been widely validated for its clinical value. Accurately estimating survival risk for patients with different cancer types not only enables clinicians to optimize treatment strategies and personalize follow-up schedules, but also facilitates risk stratification and more efficient allocation of medical resources [1].

Early survival prediction efforts were predominantly anchored in unimodal data paradigms, relying on either genomic features or clinical variables in isolation. Statistical models laid the foundational groundwork for this field: the random survival forest (RSF) [2], for instance, addressed key challenges in survival analysis (eg, right-censored data and missing values) through log-rank–based tree splitting and adaptive imputation, establishing a robust nonparametric framework. Meanwhile, the classic Cox proportional hazards model (CPH) [3] became a clinical staple for its interpretability via linear combinations of clinical indicators, though its inability to capture complex nonlinear biological mechanisms limited its predictive power in heterogeneous cancer cohorts. The advent of deep learning catalyzed a paradigm shift in unimodal modeling: approaches like DeepSurv [4] integrated neural networks with Cox regression to automatically learn nonlinear covariate–treatment relationships, outperforming traditional models on genomic datasets. Subsequent studies further validated deep learning's superiority in capturing complex variable interactions; for example, in clinical data for oral squamous cell carcinoma [5] and histopathological images for colorectal cancer [6]. Weakly supervised learning also emerged as a promising direction for leveraging unannotated whole-slide images across multiple cancer types [7], while transfer learning facilitated cross-dataset generalization in disease-specific prognostic tasks [8]. Despite these advances, unimodal models inherently suffer from incomplete information capture: genomic data alone overlooks the spatial heterogeneity of tumor microenvironments (eg, angiogenesis patterns in pathology slides), while clinical variables fail to reveal molecular-level prognostic markers, creating an inherent ceiling for predictive performance.

In response to these limitations, multimodal data fusion has emerged as a transformative research focus in pan-cancer survival prediction, driven by the recognition that complementary data modalities (omics, imaging, and clinical) can synergistically enhance prognostic accuracy. Early multimodal frameworks integrated clinical, transcriptomic (messenger RNA [mRNA]/microRNA [miRNA]), and pathological data to achieve cross-cancer prediction [9], while recent innovations have incorporated advanced techniques such as unsupervised learning and attention mechanisms to dynamically weight modality contributions [10]. A key trend in this space is the development of specialized fusion architectures: the Multimodal Affinity Fusion Network [11] and MultiCoFusion [12]; for example, use attention mechanisms and graph convolutional networks to model intermodal dependencies and gene-gene interactions, respectively. Furthermore, bilinear fusion approaches, such as the attention-based multimodal bilinear fusion [13] and hierarchical factorized models [14], have addressed the curse of dimensionality to enable deeper cross-modal interactions. Notably, methods like Dynamic-DeepHit [15] have pioneered the use of recurrent neural network–based attention to model longitudinal dependencies for dynamic risk prediction, while cross-modal translation and alignment (CMTA) [16] explicitly learn and aligns intra- and cross-modal representations between pathology and genomics via encoder-decoder structures. End-to-end multimodal models such as MultiSurv [17], which integrate up to 6 modalities (whole-slide images, clinical data, and multiomics), have demonstrated state-of-the-art performance (concordance index [C-index]=0.779) across 33 cancer types, validating the value of comprehensive data integration. Other frameworks have incorporated multitask learning [18] or alternating training [19] to jointly optimize survival prediction and auxiliary tasks (eg, cancer grading), further improving model robustness. Recent work by Flack et al [20] introduced a robust multimodal survival prediction framework that effectively handles data heterogeneity and missing modalities through a regularization-based fusion strategy, showing competitive performance across multiple cancer types. Despite these advancements, critical challenges persist: inefficient cross-modal information alignment, inadequate intramodal feature refinement, and unresolved issues of intermodal heterogeneity and noise propagation remain major barriers to achieving consistent performance across diverse cancer types at the pan-cancer scale. More critically, the majority of these state-of-the-art multimodal survival prediction models, including MultiSurv [17], HFBSurv [14], attention-based multimodal bilinear fusion [13], and MBFusion [18], operate under the implicit assumption of centralized, nonsensitive data. They fundamentally lack integrated mechanisms to protect the highly sensitive genetic and clinical information inherent in medical datasets, creating a significant gap between methodological advancement and real-world clinical applicability where privacy is paramount.

Although multimodal data fusion has significantly improved pan-cancer survival prediction accuracy, existing research generally ignores the sensitivity of medical data and the need for privacy protection. Medical data contain sensitive information such as patients' genomic data, pathological images, and clinical records, whose leakage may lead to ethical risks such as identity recognition and insurance discrimination. For instance, empirical studies have demonstrated that even aggregated genomic data remain vulnerable to linkage attacks using auxiliary public databases, potentially inferring an individual's identity with high confidence [21]. Additionally, during the training and sharing of deep learning models, gradients or intermediate features may inadvertently leak sensitive patterns of the original training data [22]. While Chen et al [23] proposed an optimized logistic regression model (batch gradient descent logistic regression and balanced differential privacy logistic regression based on hybrid feature selection (Pearson correlation test + random forest out-of-bag algorithm) and differential privacy (DP) protection to address the problems

XSL•FO
RenderX

of insufficient prediction accuracy and privacy leakage in existing machine learning models for breast cancer prediction, and Chai et al [24] proposed the decentralized federated learning framework AdFed, combining regularization methods to train models without sharing raw data while achieving feature selection and privacy protection. To protect DNA data, Wu et al [25] proposed the DP-Motif Finding algorithm based on -DP, which uses closed frequent patterns to reduce redundant motifs, allocates privacy budget by constructing a perturbed extension tree, and uses best linear unbiased estimation postprocessing to optimize the noisy support. Wu et al [26] also proposed an adaptive federated learning scheme integrating DP, which adjusts the gradient descent process through an adaptive learning rate algorithm to avoid model overfitting and fluctuations, while introducing a DP mechanism to resist various background knowledge attacks, providing quantifiable privacy protection for the federated learning process. Wang et al [27] proposed a blockchain-based access control framework for genome-wide association studies in federated learning to protect the security of genetic data—this framework implements automated quality control to ensure training data quality, designs a blockchain-based authentication mechanism to filter malicious attackers, and adopts a periodic aggregation method combined with DP to accelerate cloud model training and resist multiple attacks [28]. Wang et al [29] proposed a method based on the ant colony optimization algorithm to detect gene interactions for genome-wide association studies—an intelligent privacy-preserving scheme. Current studies still suffer from problems such as single datasets and heavy reliance on data anonymization or centralized storage, failing to embed dynamic privacy protection mechanisms.

To solve the above problems, this study proposes a bilayer feature fusion framework based on the multihead attention (MHA) mechanism and adaptive DP, achieving collaborative optimization of precise feature extraction and sensitive data protection through technological innovation. Specifically, for feature fusion, traditional methods typically rely on fixed weights or simple concatenation to handle heterogeneous data. Simple concatenation, however, merely "physically stacks" feature vectors from different modalities, posing notable limitations. Multisource data (clinical, mRNA, miRNA, and gene copy number variation [CNV]) vary drastically in feature dimensions, semantics, and distributions; forcing them into a single-vector space postconcatenation overlooks inherent intermodal correlations; for example, links between specific mRNA expression and clinical survival time, or regulatory relationships between miRNA and CNV. Additionally, concatenation assumes "all features are equally important," failing to differentiate their contributions to survival prediction. This often buries key signals (eg, driver gene mutation sites) under redundancy, and may trigger the curse of dimensionality via rapid feature expansion, increasing computational complexity and overfitting risk. Given the aforementioned limitations of simple concatenation, our first contribution is the design of a structured bilayer feature extraction module. The first layer performs modality-specific feature extraction on clinical, mRNA, miRNA, and CNV data through fully connected (FC) networks and embedding layers. The second layer innovatively uses an MHA mechanism not merely as a fusion tool, but as a structured cross-modal interaction model that explicitly quantifies feature-level contributions. Unlike single-head attention, which models only single correlations, MHA captures differentiated correlations between features from different subspaces through parallelized independent attention heads and focuses on key survival-related features through dynamic weight allocation. Our second and primary contribution lies in the novel integration of adaptive DP directly into the multimodal feature extraction pipeline. At the privacy protection level, this study proposes adaptive DP based on layer-wise relevance analysis, balancing prediction accuracy and privacy security through a 2-step strategy. First, the layer-wise relevance propagation algorithm is used to quantify the association strength between neurons at each layer and survival prediction outcomes, identifying highly sensitive features (eg, driver gene mutation sites) in mRNA, miRNA, and CNV data; then privacy budgets are dynamically allocated based on correlation scores to inject Laplacian noise into the gradient update process—features with higher correlations receive less noise, preserving key prediction information while suppressing privacy leakage risks.

In summary, the novelty of our work is not merely the combination of MHA and DP, but the creation of a synergistic framework where (1) the bilayer MHA structure provides the granular feature importance signals necessary for adaptive noise allocation and (2) the adaptive DP mechanism, guided by layer-wise relevance analysis, protects privacy in a targeted manner that minimizes damage to the model's core predictive capability. This integrated solution addresses the specific limitations of prior multimodal models (which neglect privacy) and prior DP applications (which impair use in complex models), paving the way for clinically viable, privacy-preserving pan-cancer prediction tools.

Methods

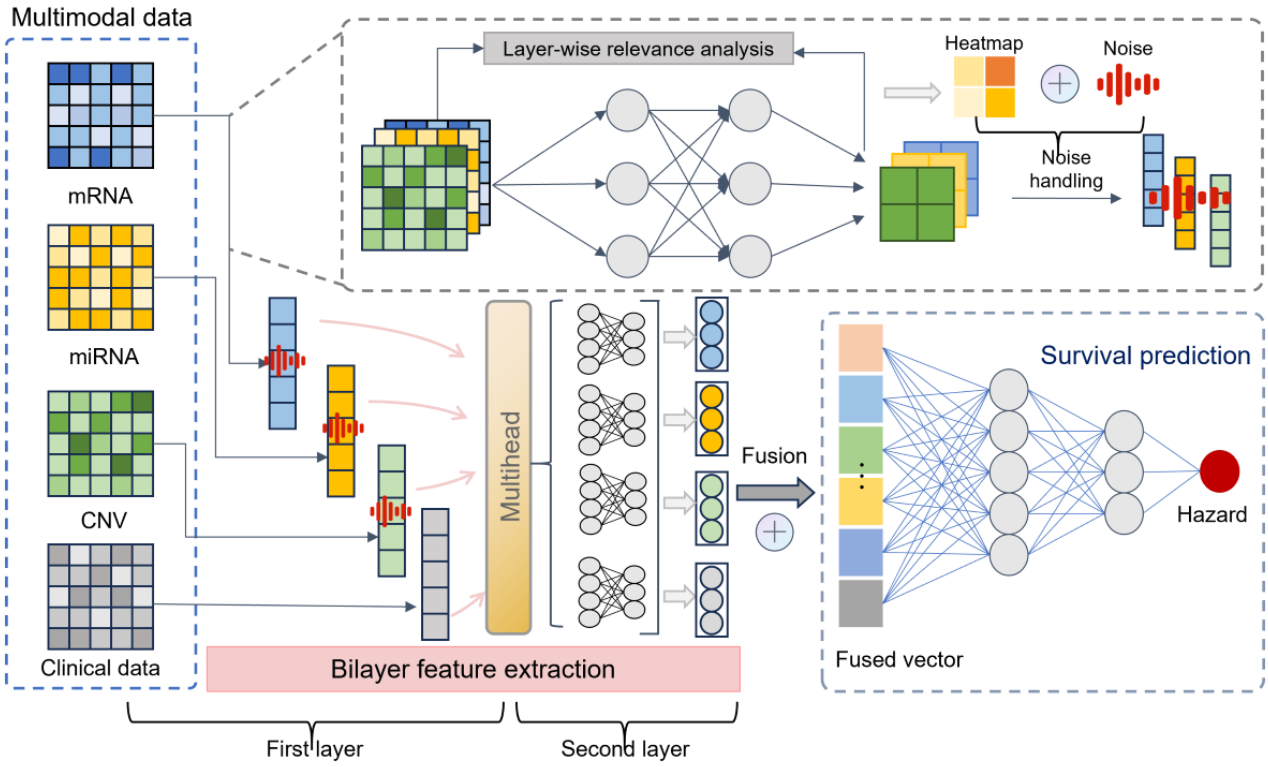
Overview

As shown in Figure 1, the proposed model architecture consists of 2 parts: a bilayer feature extraction module based on MHA and an adaptive DP protection module based on layer-wise relevance analysis. The bilayer feature extraction part, based on MHA, first performs first-layer feature extraction through FC layers and embedding layers, and then uses the MHA mechanism in the second layer for further feature extraction and multimodal data fusion. By integrating the MHA mechanism, the bilayer feature extraction module fully uses the advantage of parallel heads to capture different correlations, comprehensively and deeply mining key information from multimodal data. Additionally, a privacy protection component is added during the first-layer feature extraction, integrating layer-wise relevance analysis into the FC layers to deeply analyze the correlation between features and survival prediction outcomes through backpropagation. Noise is adaptively added based on the obtained correlation results. This study introduces a bilayer feature extraction module into the multimodal survival prediction model to better mine data features and provides privacy protection, effectively preserving the privacy information in the training data. The data preprocessing and

XSL•FO
RenderX

detailed method descriptions are presented in the following sections.

Figure 1. The architecture of the proposed model. CNV: gene copy number variation; miRNA: microRNA; mRNA: messenger RNA.



Data Preprocessing

In this study, 4 types of data were used: clinical data, gene expression (mRNA) data, miRNA data, and CNV data.

This section performs data screening for different data features. For clinical features, patient data with missing values and/or missing follow-up times were excluded. For gene expression data, a variance threshold method was used to select features with variances greater than a given threshold calculated from all patients [18,30]. In this study, the same thresholds as Fan et al [10] were used for mRNA and CNV modalities: 7 and 0.2, respectively. This processing retained 1579 mRNA features and 2711 CNV features. Subsequently, all continuous variables were normalized to the interval (0, 1) using minimum-maximum normalization [31].

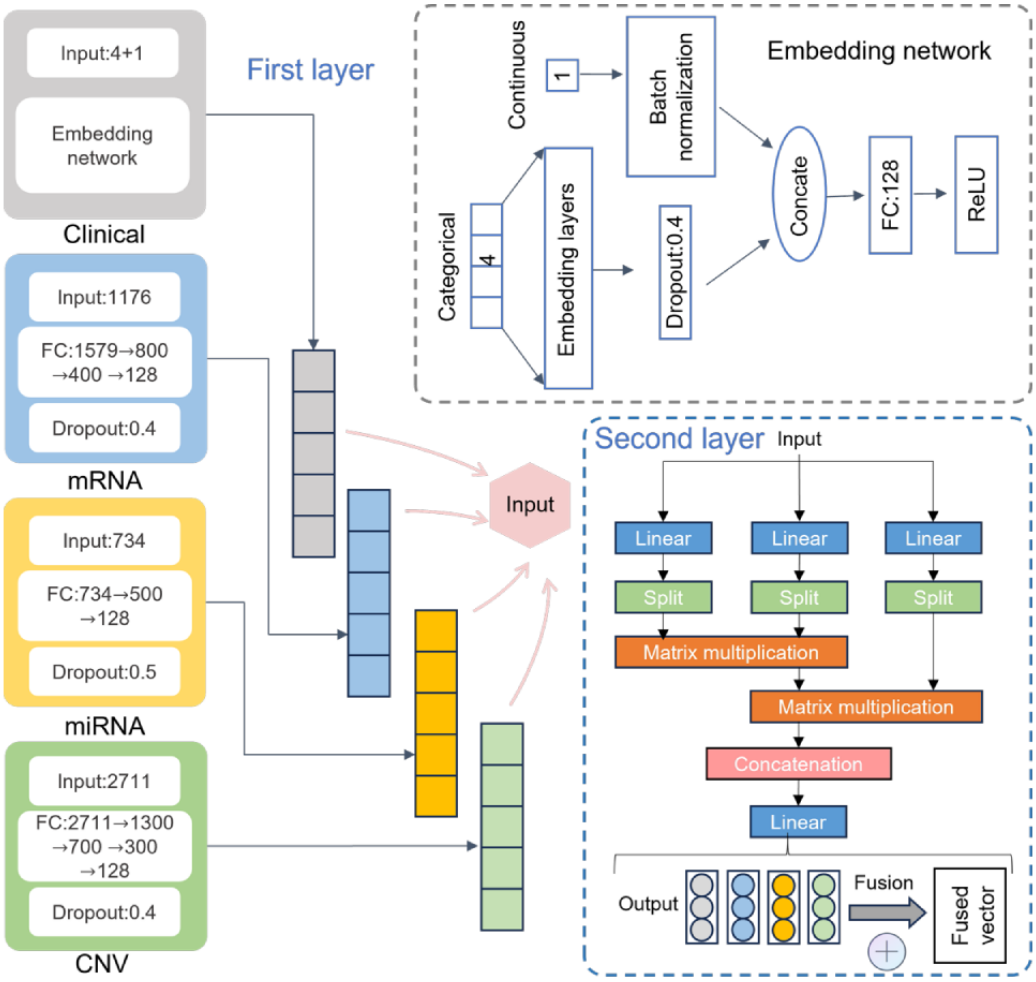
Considering the variations in diagnosis and treatment methods across different cancer types, samples in the dataset may lack some required modalities. This study adopts zero-vector imputation for missing modalities, with its rationale mainly reflected in 2 aspects: first, zero vectors can explicitly indicate "no measured data for the modality," avoiding false biological signals introduced by statistical imputation methods. Second, zero-vector imputation ensures the consistency of feature dimensions across all samples, simplifies the model's input structure, and avoids additional computational overhead from complex missing-value processing modules, providing stability guarantees for gradient updates in the integration of adaptive DP mechanisms.

Bilayer Feature Extraction Based on MHA

To improve the accuracy and efficiency of feature extraction, this study designs a bilayer feature extraction method based on MHA, which can more comprehensively and deeply mine key information from multimodal data, as shown in Figure 2.

XSL•FO
RenderX

Figure 2. The architecture of bilayer feature extraction based on multihead attention. CNV: gene copy number variation; FC: fully connected; miRNA: microRNA; mRNA: messenger RNA; ReLU: rectified linear unit.



In recent years, neural networks with FC layers have been widely used for representation extraction [15]. In this study, different neural network settings with FC layers were used as the first-layer feature extraction to obtain representations from the raw numerical data of different modalities. Batch normalization was adopted to normalize the layer inputs and accelerate neural network training [30]. For the clinical data, we used categorical embedding layers for the 4 categorical variables, namely, cancer type, gender, race, and histological type, with dropout to encode the categorical variables into a numerical vector. The normalized continuous variables (ie, age) were then concatenated with these numerical vectors and fed into an FC layer with a fixed representation length. For miRNA, mRNA, and CNV modalities, 2-4 FC layers with rectified linear unit activation and batch normalization were used to extract fixed-length representations. The representation vectors were obtained from modality-specific configurations to form the representation matrix where m is the sample size and d is the feature dimension.

To effectively capture the complex interaction relationships between multimodal data and enhance feature representation capabilities, the MHA mechanism was introduced in the second-layer feature extraction stage.

The proposed MHA mechanism consists of 3 components similar to the self-attention in the transformer model: query (Q), key (K), and value (V). The matrix B is projected to generate keys K and values V. Specifically, V is processed using an identity function because the features extracted in the first layer already have sufficient representational power. For K, it is generated by:

Where (d represents the feature dimension) is a trainable matrix randomly initialized from a standard Gaussian distribution and updated during training; B is the representation matrix obtained from the first-layer feature extraction. The dropout layer and rectified linear unit activation function are used to prevent overfitting and introduce nonlinearity. The query is a learnable vector initialized randomly, with elements sampled from a uniform distribution .

The attention aggregation feature is calculated as:

Where A is the attention weight. After softmax normalization along the sample dimension, the sum of the attention weights

XSL•FO
RenderX

A for *N* samples is 1, enabling dynamic focusing on key survival-related features.

To capture multidimensional survival associations, the attention layer is extended to *H* heads. Specifically, *Q*, *K*, and *V* are split into blocks of size *d*/*H* along the embedding dimension. The attention for each head is calculated as:

These heads are then concatenated into a multihead output:

This process allows different heads to capture interactions between different types of data separately.

Finally, the fused features are obtained by:

Where is the weight of the last layer. In evaluation mode, the dropout layer is disabled. Thus, the fused feature *Y* can be written as:

The attention mechanism focuses on key information through dynamic weight allocation, efficiently handling long-range dependencies in sequence data and suppressing noise interference. However, its single-head structure has limitations in modeling multidimensional correlations. Therefore, based on the features extracted from the 4 modalities in the first layer, this study introduces the MHA mechanism. By parallelizing multiple independent attention heads, it captures differentiated correlations between features from different subspaces, providing more comprehensive feature representations for pan-cancer survival prediction. The above method achieves deep fusion of multimodal survival features, offering high-precision feature representations for pan-cancer survival prediction.

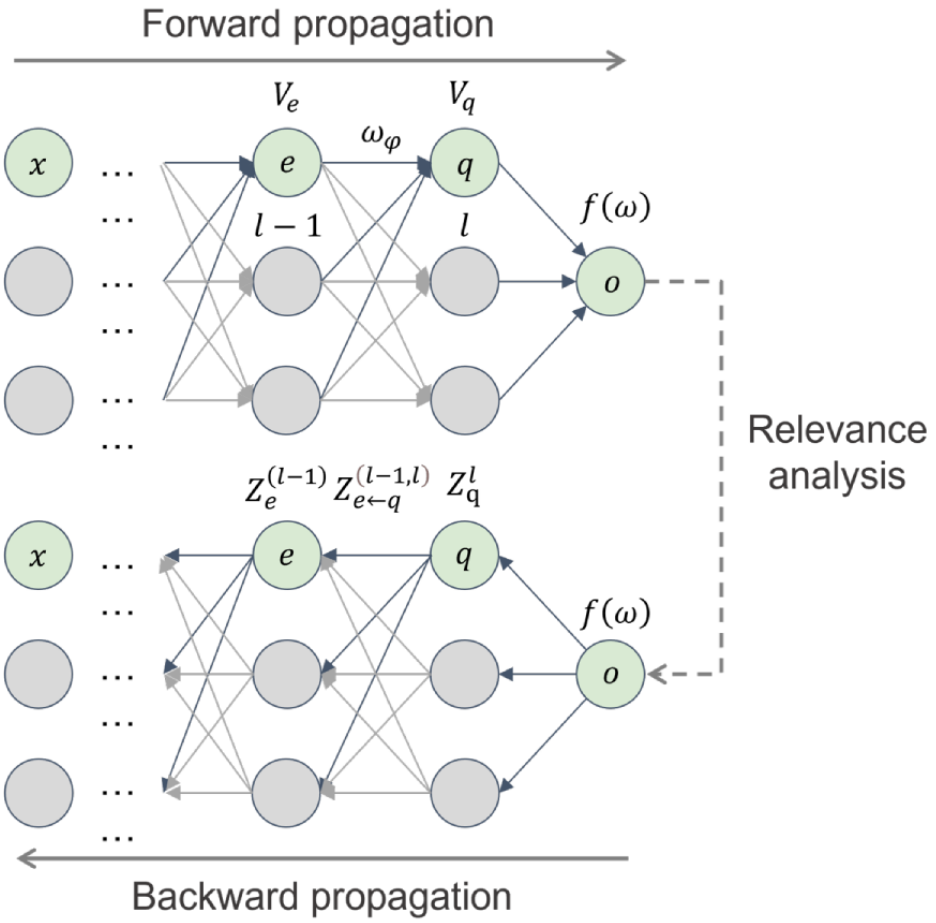
Adaptive DP Based on Layer-Wise Relevance Analysis

In the field of cancer survival prediction, privacy protection of medical data is of utmost importance. Genetic data, such as mRNA, miRNA, and CNV data, contain highly sensitive personal information, posing a risk of privacy leakage. Therefore, during the feature extraction process, this study designs a method based on layer-wise relevance analysis to construct a noise addition mechanism by adding Laplacian noise for privacy protection, ensuring data security.

First, during the extraction of mRNA, miRNA, and CNV features in the first layer, this paper adds appropriate Laplace noise to the training gradient of neurons based on analyzing the correlation between each neuron layer and the output layer. The core theoretical operation process is shown in Figure 3. Relevance analysis begins after forward and backward propagation, calculating the correlation results of each feature in a layer of neurons. Importantly, the correlation analysis results are adjusted with the forward and backward propagation processes until the propagation ends.

XSL•FO
RenderX

Figure 3. The process of the layer-wise relevance analysis method of the proposed model based on the backpropagation algorithm.



The correlation between each input feature x_{ij} and the output $Hx_i(0)$ is calculated by decomposing the neurons of the previous layer. Given $Z_e^{(l)}(x_i)$ as the correlation result between x_i and neuron e in layer l, define as the process of neuron e sending information to q. The neuron correlation is:

The decomposition rule for layer-wise relevance analysis is:

Where is a predefined stabilizer to address the unboundedness of $Z_e^{(l)}(x_i)$. R_q is the affine transformation of neuron e, defined as:

Where v_e is the value of neuron e,$_{eq}$ is the weight between neuron e and q,u_q is the bias term.

In the last hidden layer, for the output variable o, the correlation is calculated as:

This study directly calculates the correlation score $Z_j(x_{ij})$ between each feature x_{ij} inlayer j and the survival prediction result for mRNA, miRNA, and CNV features.

Next, a noise addition rule is defined. Through the above layer-wise relevance analysis, the value range of these correlation scores is divided into n nonoverlapping threshold intervals , and T_1 corresponds to the features with the highest relevance scores.

A privacy-budget mapping function is then defined, which assigns a specific privacy parameter $_k$ to each interval T_k. The function is designed to be strictly monotonic, satisfying the following condition: for any $g<h$,$\eta(T_g)<\eta(T_h)$, thereby ensuring that features with stronger predictive correlations are allocated a higher privacy budget (less noise), while features with weaker correlations receive a lower privacy budget (more noise).

Formally, for an interval index $k\in\{1,2,...,n\}$,the privacy parameter $_k$ is given by:

Where $_{max}$=0.5 and $_{min}$=0.1 define the permissible range of the privacy budget; $\delta_k \sim U(-\Delta,\Delta)$ is a bounded random perturbation term that preserves the strict monotonicity $_1 > _2 > ... > _n$.

XSL•FO
RenderX

Privacy budgets are adaptively allocated according to the function :

Where denotes the global privacy budget.

Laplacian noise is added to the training gradients of neurons, where is the total privacy budget for protecting mRNA, miRNA, and CNV features.

At the start of training, the gradient update objective function of the general optimization method is defined. In each training step, using a set of random training samples L from the genetic data feature set X (where $X = \{x_{11}, x_{12},..., x_{ij}\}$), starting from the initial point $_0$, the parameters are updated at step t as:

where is the learning rate at step t, is the regularization parameter, and is the loss function.

Subsequently, the training gradients are perturbed to ensure the security of genetic data during training and sharing:

Where Y_t is the Laplacian noise.

After detailing the process of adding Laplacian noise to features based on layer-wise relevance analysis, this section proceeds to prove the implementation of -DP. The following proof shows how the proposed method satisfies -DP requirements:

Proof: assume L and L' are 2 adjacent batches, and $_{t+1(L)}$ and $_{t+1(L')}$ are the parameters for L and L', respectively. The formula is expressed as follows:

Then, the inequality for the difference between the 2 output results is:

From the above formula and DP, $_t$ is the sensitivity of the neural network **Inline graphic 27**. To protect the privacy information of the neural network, the gradient is perturbed based on relevance analysis, and the noise can be written as:

We have:

This proves that the method satisfies -DP, ie,

, where M is the model after adding Laplacian noise to features. It ensures that the model's noise addition mechanism is consistent with the definition of DP while protecting the privacy of mRNA, miRNA, and CNV features.

This method not only effectively reduces the risk of data leakage but also maximizes data usability, providing a more secure and reliable solution for cancer survival prediction and promoting the efficient use and sharing of medical data under privacy protection.

Ethical Considerations

All data used in this study were obtained from The Cancer Genome Atlas (TCGA), a publicly available database. The TCGA project obtained informed consent from all participants and received ethical approval from the appropriate institutional review boards. As this study involved only the use of deidentified, publicly accessible data, no additional ethical approval was required.

Results

Data

The clinical, mRNA, and miRNA data were downloaded from the Pan-Cancer Atlas of TCGA project (publicly accessible at the Genomic Data Commons), where more than 11,000 patient samples across 33 tumor types [32,33] were collected. The clinical dataset provides annotations for 11,160 patients, using 5 variables consistent with Silva and Rohr [17]: cancer type, gender, race, histological type, and age. The miRNA data set contains 743 miRNA feature records, and the mRNA dataset contains RNA sequencing counts for 20,531 mRNAs. The CNV data were downloaded from the University of California Santa Cruz Xena [34].

Table 1 describes the data information after preprocessing in detail. In this study, the pan-cancer datasets were partitioned into training (6656/11094, 60%), validation (2219/11094, 20%), and testing (2219/11094, 20%) subsets, with model performance evaluated using 5-fold cross-validation.

XSL•FO
RenderX

Table 1. Summary information of the different modalities after preprocessing.

Modality	Patients, n	Features, n		All zero vector, n/N (%)
		Continuous	Categorical	
Clinical	11,094	1	4	—[a]
miRNA[b]	11,094	743	—	72/743 (9.7)
mRNA[c]	11,094	1579	—	49/1579 (3.1)
CNV[d]	11,094	2711	—	225/2711 (8.3)

[a]Not available.

[b]miRNA: microRNA.

[c]mRNA: messenger RNA.

[d]CNV: gene copy number variation.

Experimental Settings

We conduct all experiments on a simulation environment equipped with a 64-bit Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz processor, 32 GB RAM, and an NVIDIA GeForce RTX 3080 GPU for accelerated computation. The experiments are implemented on the Windows 11 operating system, with Python (version 3.10; Python Software Foundation) as the primary programming language and PyTorch (version 2.6.0; Meta AI) as the deep learning framework.

For model training, the Adam optimizer is adopted with a fixed learning rate of 0.001. Key hyperparameters are configured as follows: sequence length is set to 128, batch size is 256, and total training epochs are 100.

Evaluation Metrics

In this study, we assessed the performance of our model by using the C-index, a widely adopted metric for evaluating survival predictions in censored survival data [2,35]. The C-index measures the proportion of concordant pairs among all possible evaluation pairs, defined as:

where 1 is the indicator function of whether the expression in parentheses is true or false. The C-index ranges from 0 to 1. The closer the C-index is to 1, the closer the prediction order is to the real one; the closer the C-index is to 0.5, the closer the model's prediction is to a random prediction. Notably, the C-index focuses on the ordinal relationship of predictions rather than the accuracy of individual sample forecasts, making it particularly suitable for evaluating proportional hazards models.

Accuracy was also used for performance evaluation. The metric is evaluated as follows:

Layer-Wise Relevance Analysis

To evaluate the model's performance in privacy preservation, this study determines the privacy budget through hierarchical correlation analysis and subsequently adds Laplacian noise based on the results. Specifically, this study analyzed the strength of correlations between features and different FC layers using heatmaps for the 3 modalities: mRNA, miRNA, and CNV. The computed correlation scores are sorted in descending order, and the feature scores at the 25th, 50th, 75th, and 100th percentiles are visualized. The heatmaps show that higher correlation scores indicate stronger associations between the corresponding features and survival prediction outcomes.

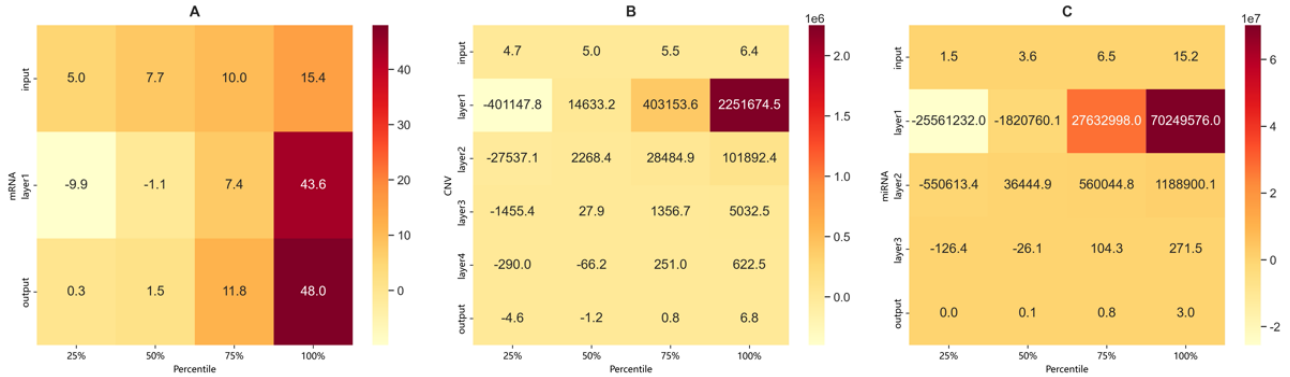
A more detailed analysis of each modality's heatmap reveals that, as shown in Figure 4A, the output layer and mRNA layer 1 exhibit significantly higher correlation scores at the 100th percentile (48.0 and 43.6, respectively), with notable increases compared to the 25th, 50th, and 75th percentiles. This suggests that these 2 layers are particularly effective in capturing high-correlation features, which are strongly associated with survival outcomes. These findings indicate that privacy-preserving noise injection should focus on these layers to avoid the leakage of critical predictive information.

For the CNV and miRNA modalities, as illustrated in Figures 4B and 4C, layer 1 demonstrates markedly higher correlation scores at the 100th percentile compared to subsequent layers and the output layer. This suggests that the model primarily extracts key features from layer 1, with diminishing correlation strength in deeper layers. Therefore, noise injection strategies should prioritize high-correlation features in layer 1, enhancing privacy protection while maintaining the model's feature extraction performance.

Overall, the heatmaps of layer-feature correlations across the 3 modalities not only identify the core layers and key percentile-based features involved in the model's internal processing, but also provide empirical support for refining noise injection strategies. This contributes to achieving a better balance between privacy protection and model performance.

XSL•FO
RenderX

Figure 4. Heatmaps of correlation scores between layers and features in (A) messenger RNA (mRNA), (B) gene copy number variation (CNV), and (C) microRNA (miRNA) modalities.



Privacy Protection Effect Comparison of Multimodal Methods

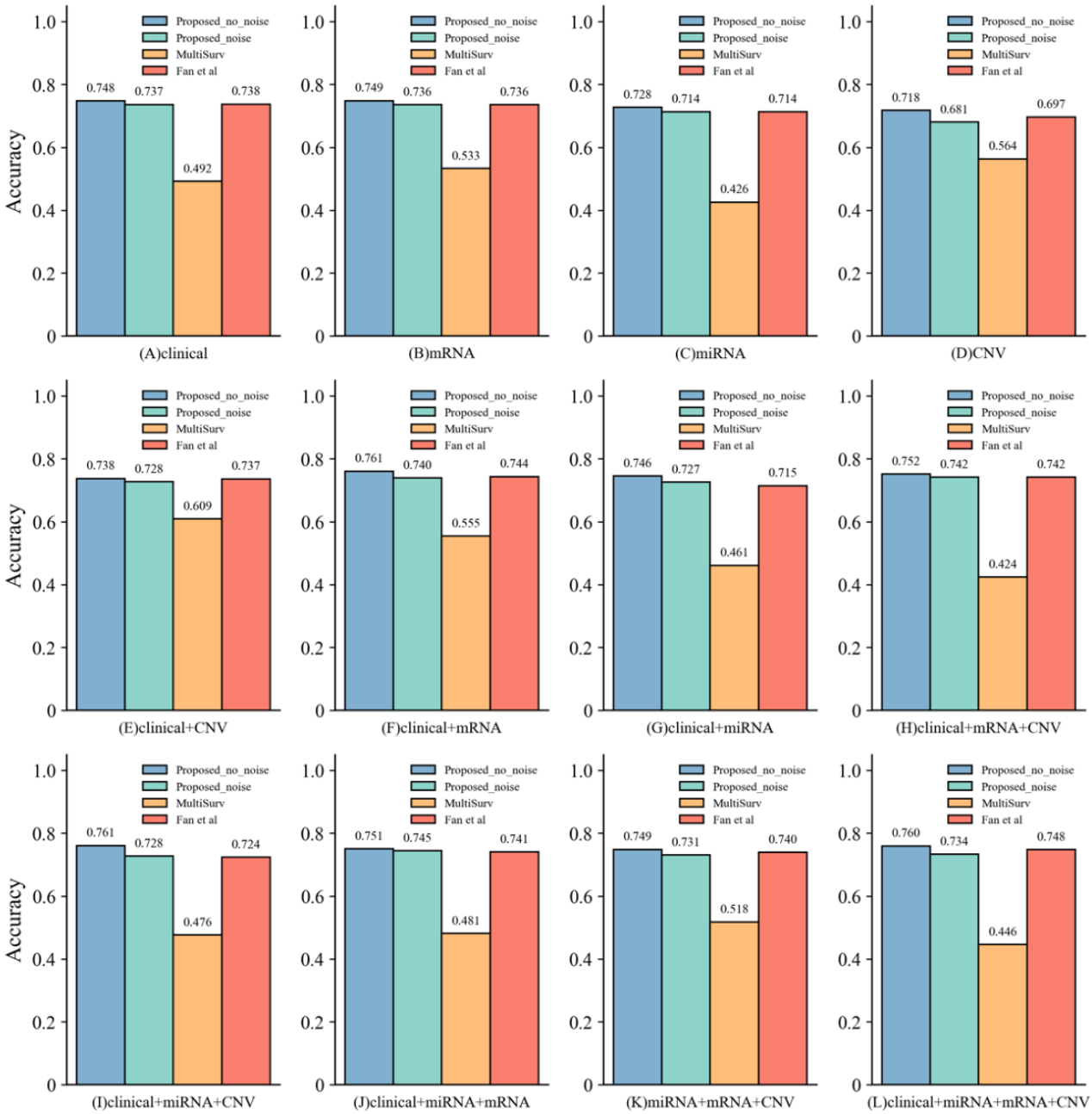
To further evaluate the impact of noise addition on model performance, this study uses accuracy as an auxiliary metric, with patients' survival status as the classification criterion (where deceased patients are labeled as 1 and surviving patients as 0), and implements the classification and prediction of survival status through a neural network with 3 FC layers. Leveraging the pan-cancer dataset, a systematic analysis is conducted on the model's survival classification performance across 4 data combinations: clinical data, mRNA, miRNA, and CNV. The proposed model is compared with 2 state-of-the-art attention-based models for pan-cancer survival prediction: MultiSurv [17] and Fan et al [10].

In unimodal settings, as shown in Figures 5A-5D, the proposed model demonstrates strong robustness. For instance, in the miRNA modality, accuracy slightly decreases from 0.728 before noise injection to 0.714 after; in the mRNA modality, accuracy drops modestly from 0.749 to 0.736.

Experiments involving multimodal combinations were conducted in dual-, tri-, and 4-modality settings. As shown in Figures 5E-5L, the model also maintains stable performance under multimodal conditions. For example, the clinical + miRNA + mRNA combination achieves an accuracy of 0.751 before noise addition and 0.746 afterward; for the clinical + mRNA combination, accuracy declines from 0.761 to 0.740 after noise injection.

XSL•FO
RenderX

Figure 5. Accuracy comparison results between "Proposed_no_noise" (blue bars), "Proposed method" (green bars), "MultiSurv" (orange bars), and Fan et al [10] (red bars) across (A) clinical, (B) messenger RNA (mRNA), (C) microRNA (miRNA), (D) gene copy number variation (CNV), (E) clinical + CNV, (F) clinical + mRNA, (G) clinical + miRNA, (H) clinical + mRNA + CNV, (I) clinical + miRNA + CNV, (J) clinical + miRNA + mRNA, (K) miRNA + mRNA + CNV, and (L) miRNA + mRNA + CNV modalities.



Survival Prediction Using Pan-Cancer Training Dataset

To evaluate the survival prediction advantages of the proposed model across different modalities, this study adopted the C-index as the performance metric and compared it against 6 state-of-the-art models: the CPH [3], RSF [2], MultiSurv, DeepHit [15], CMTA [16], and Fan et al [10]. The experiments are conducted using clinical data, mRNA, miRNA, and CNV data in both unimodal and multimodal settings. All results were obtained via 5-fold cross-validation and are summarized in Table 2.

XSL•FO
RenderX

Table 2. Model performance on mixed cancer types using single-modal and multimodal training datasets by the measurements of concordance index (C-index).

Modality	C-index						
	CPH[a] [3]	RSF[b] [2]	MultiSurv [17]	DeepHit [15]	CMTA[c] [16]	Fan et al [10]	Proposed model
Clinical	0.633 (±0.010)	0.745 (±0.008)	0.749 (±0.021)	0.742 (±0.013)	0.758 (±0.009)	0.756 (±0.013)	0.775 (±0.008)[d]
mRNA[e]	0.715 (±0.013)	0.739 (±0.013)	0.747 (±0.012)	0.744 (±0.008)	0.760 (±0.010)	0.764 (±0.013)	0.779 (±0.007)[d]
miRNA[f]	0.715 (±0.010)	0.735 (±0.005)	0.725 (±0.012)	0.701 (±0.008)	0.712 (±0.007)	0.738 (±0.011)	0.760 (±0.011)[d]
CNV[g]	0.589 (±0.008)	0.626 (±0.007)	0.649 (±0.006)	0.579 (±0.006)	0.618 (±0.0011)	0.646 (±0.009)	0.659 (±0.009)[d]
Clinical + CNV	—[h]	—	0.752 (±0.008)	0.747 (±0.006)	0.758 (±0.012)	0.753 (±0.005)	0.763 (±0.006)[d]
Clinical + mRNA	—	—	0.773 (±0.008)	0.764 (±0.011)	0.785 (±0.010)	0.784 (±0.007)	0.792 (±0.006)[d]
Clinical + miRNA	—	—	0.758 (±0.011)	0.760 (±0.008)	0.763 (±0.007)	0.766 (±0.010)	0.783 (±0.007)[d]
Clinical + mRNA + CNV	—	—	0.766 (±0.012)	0.768 (±0.013)	0.774 (±0.013)	0.779 (±0.007)	0.785 (±0.004)[d]
Clinical + miRNA + CNV	—	—	0.760 (±0.009)	0.761 (±0.010)	0.762 (±0.009)	0.764 (±0.007)	0.775 (±0.008)[d]
Clinical + miRNA + mRNA	—	—	0.771 (±0.014)	0.773 (±0.014)	0.779 (±0.013)	0.777 (±0.013)	0.799 (±0.012)[i]
miRNA + mRNA + CNV	—	—	0.747 (±0.013)	0.752 (±0.012)	0.770 (±0.007)	0.767 (±0.008)	0.778 (±0.010)[d]
Clinical + miRNA + mRNA + CNV	—	—	0.768 (±0.013)	0.770 (±0.010)	0.772 (±0.012)	0.779 (±0.010)	0.787 (±0.007)[d]

[a]CPH: Cox proportional hazards model.

[b]RSF: random survival forest.

[c]CMTA: cross-modal translation and alignment.

[d]$P<.05$ vs MultiSurv, DeepHit, CMTA, and Fan et al [10] (paired *t* test with Bonferroni correction, n=1000 bootstrap samples).

[e]mRNA: messenger RNA.

[f]miRNA: microRNA.

[g]CNV: gene copy number variation.

[h]Not available.

[i]$P<.01$ vs CMTA (top competitor; paired *t* test with Bonferroni correction, n=1000 bootstrap samples).

To substantiate claims of superiority despite modest C-index differences, statistical significance testing was performed using bootstrapped resampling and 2-tailed paired *t* tests. For each model and modality combination, 1000 bootstrap samples were generated from the test set. Paired *t* tests were conducted on the bootstrapped C-index values (n=1000) to test the null hypothesis that the mean performance difference is 0, with Bonferroni correction applied for multiple comparisons. Additionally, 95% CIs were derived from the bootstrap distributions. Nonoverlapping CIs are interpreted as indicating statistically significant differences ($P<.05$).

In unimodal prediction scenarios, the proposed model achieves C-index scores of 0.775, 0.779, 0.760, and 0.659 for clinical, mRNA, miRNA, and CNV data, respectively—significantly outperforming the CPH and RSF, which are limited to unimodal survival prediction. Notably, the proposed model yields the highest performance on mRNA data, with a C-index of 0.779, compared to 0.715 for CPH and 0.739 for RSF.

Furthermore, in the multimodal setting, this study compared the performance of MultiSurv, DeepHit, CMTA, Fan et al [10], and the proposed model across various combinations of data modalities. As shown in Table 2, the proposed model consistently achieves C-index values around 0.78 across different combinations. The highest performance is observed in the trimodal configuration (clinical + mRNA + miRNA), which yields a C-index of 0.799—surpassing MultiSurv (0.771),

XSL•FO
RenderX

DeepHit (0.773), CMTA (0.779), and Fan et al [10] (0.777). To statistically validate these performance improvements, paired significance testing with multiple-comparison adjustment was conducted. For each model and modality combination, 1000 bootstrap samples were generated from the test set. Paired *t* tests were performed on the bootstrapped C-index values (n=1000) with Bonferroni correction applied for multiple comparisons. The proposed model demonstrated statistically significant improvements over all competitors in the optimal trimodal configuration (clinical + mRNA + miRNA) after multiplicity adjustment: vs CMTA (*P*=.008), Fan et al [10] (*P*=.006), MultiSurv (*P*=.003), and DeepHit (*P*=.004). These adjusted *P* values (<.05) confirm that the observed C-index improvements (0.020, 0.022, 0.028, and 0.026 relative to CMTA, Fan et al [10], MultiSurv, and DeepHit, respectively) are statistically robust and not attributable to random variation.

This superior performance is attributed to the model's dual-layer feature extraction architecture, which, leveraging multiview parallel processing, enables finer-grained automatic weight adjustment. Notably, our proposed model further integrates an adaptive DP protection mechanism to safeguard sensitive medical data, while still achieving a higher C-index compared to existing state-of-the-art multimodal models that lack privacy-preserving designs. These findings reinforce the advantage of multimodal inputs over unimodal inputs, as multimodal data provide more comprehensive and informative cues for survival prediction.

Comparison With Standard Privacy Protection Methods

To comprehensively validate the superiority of the proposed adaptive DP framework in balancing privacy protection and predictive utility, this study adopted 2 core metrics—C-index for survival outcome ranking accuracy and classification accuracy for binary survival status prediction (deceased=1 and surviving=0)—and systematically compared it against 3 standard DP-aware baseline models: DP-Stochastic Gradient Descent (SGD) [36], DP-Adam [37], and Private Aggregation of Teacher Ensembles (PATE) [38]. The experiments were conducted under the optimal multimodal setting (clinical + mRNA + miRNA), with a fixed global privacy budget of = 0.8 (δ = 1e-5 for DP-SGD and DP-Adam, as per their standard implementations) to ensure a fair comparison of use under identical formal privacy guarantees. Consistent data preprocessing (zero-vector imputation for missing modalities, minimum-maximum normalization) and model training configurations (Adam optimizer, learning rate=0.001, batch size=256, and 100 epochs) were used across all methods to eliminate confounding variables. All results were derived from 5-fold cross-validation on the pan-cancer test set and supplemented with 1000 bootstrap resamples for statistical significance testing (paired *t* test with Bonferroni correction, *P*<.05). Detailed performance metrics and comparative analysis are summarized in Table 3.

Table 3 clearly shows that the proposed adaptive DP framework outperforms all standard DP baselines in maintaining predictive utility under the same formal privacy guarantee (=0.8). Regarding the C-index—the gold standard for survival prediction that assesses the ordinal consistency between predicted risk scores and actual survival times—the proposed model achieves 0.799 (±0.012), which is 4.3, 3.7, and 6.5 percentage points higher than DP-SGD (0.756 ± 0.015), DP-Adam (0.762 ± 0.014), and PATE (0.734 ± 0.016), respectively. Its utility loss compared to the non-DP counterpart (0.821 ± 0.008) is only 0.022, which is less than half of DP-SGD's loss (0.065) and about 1/3 of PATE's loss (0.087). This confirms that the layer-wise relevance propagation–guided targeted noise injection strategy more efficiently uses the privacy budget, preserving high-relevance prognostic features while satisfying -DP.

In terms of binary survival status classification accuracy, the proposed model maintains a high level of 0.745 (±0.013), outperforming the aforementioned baselines by 0.8, 1.0, and 1.4 percentage points, respectively. Its accuracy drop (0.006) is also significantly smaller than that of the baselines (0.014-0.020), further validating the efficiency of adaptive noise allocation in minimizing utility degradation for a given . Statistical analysis (paired *t* test with Bonferroni correction, n=1000 bootstrap samples) reveals that the proposed model's C-index and accuracy are significantly different from all baselines (*P*<.05), with nonoverlapping 95% CIs, confirming the robustness of the observed improvements.

Table 3. Use comparison of proposed model vs standard differential privacy (DP) baselines (multimodal: clinical + messenger RNA + microRNA; =0.8).

Modality	C-index[a]	Accuracy
Proposed_no_noise	0.821 (±0.008)	0.751 (±0.008)
Proposed model	0.799 (±0.012)	0.745 (±0.013)
DP-SGD[b]	0.756 (±0.015)	0.737 (±0.010)
DP-Adam	0.762 (±0.014)	0.735 (±0.008)
PATE[c]	0.734 (±0.016)	0.731 (±0.009)

[a]C-index: concordance index.

[b]SGD: Stochastic Gradient Descent

[c]PATE: Private Aggregation of Teacher Ensembles.

XSL•FO
RenderX

Analysis of Privacy Protection Effectiveness

The proposed adaptive DP mechanism is designed to satisfy -DP (=0.8) as formally proven in equation 20. Compared with traditional uniform noise injection methods (DP-SGD and DP-Adam) under the same global privacy budget, our method significantly improves the accuracy of survival prediction (with the C-index increased by 3.7-6.5 percentage points; Table 3). This result indicates that, for the same formal privacy guarantee ( ), our adaptive allocation strategy achieves higher predictive utility. The adaptive mechanism identifies features with high relevance to prediction and allocates a larger portion of the privacy budget to them (injecting less noise), thereby preserving critical prognostic information. Conversely, features with low predictive contribution receive a smaller budget (more noise), which theoretically enhances privacy protection for those elements without harming overall model utility. Therefore, the proposed method provides a stronger formal privacy guarantee for low-relevance features and achieves superior utility for high-relevance features under the same constraint, representing a more efficient use of the privacy budget.

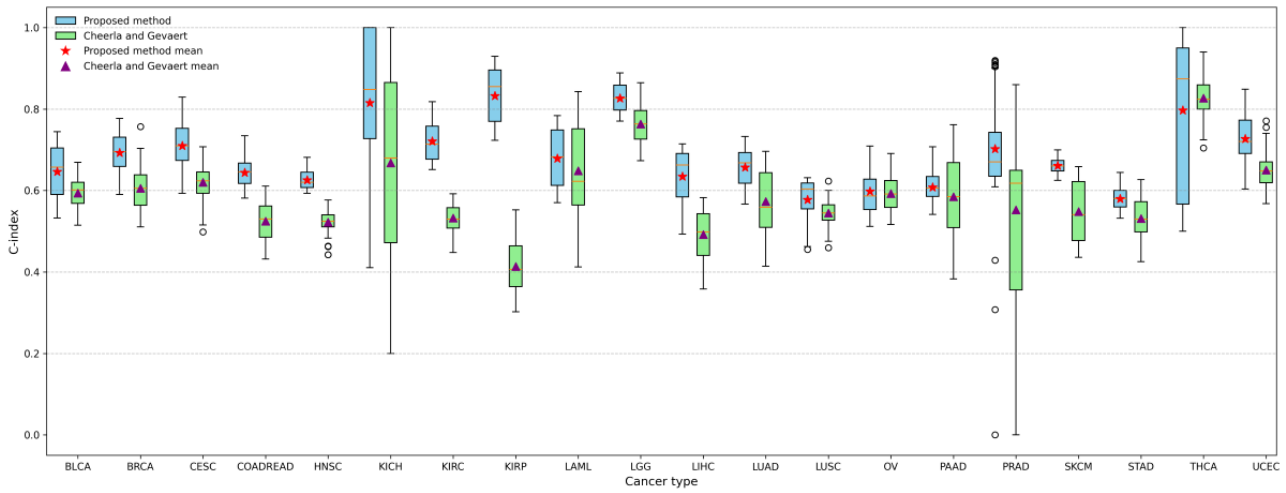
Comparison of Survival Prediction Performance for Individual Cancer Types

To investigate the effectiveness of the proposed model in predicting survival outcomes for individual cancer types, this study conducted a comparison between our proposed model and the state-of-the-art deep learning–based model (Cheerla and Gevaert [9]) for single-cancer prediction using pan-cancer datasets. For consistency, this study adopted a multimodal input comprising clinical, miRNA, and mRNA data. The results are presented in Figure 6.

Compared with Cheerla and Gevaert [9], our proposed method achieved a higher C-index in the majority of cancer types—18 out of 20. Notably, in 7 of these 18 cancer types, the performance improvement was substantial, with C-index differences exceeding 0.1.

Furthermore, as illustrated in Figure 6, our model exhibits higher median C-index values for several cancer types and demonstrates more concentrated distributions in the upper range. This indicates both superior and more stable predictive performance. In contrast, Cheerla and Gevaert [9] show relatively lower C-index scores in some cancer types, with several boxplots revealing greater data dispersion and reduced stability.

Figure 6. Concordance index (C-index) of the proposed model (blue bars) and the previous work (Cheerla and Gevaert [9]; green bars) on the 20 cancer types using the modality combination of clinical, microRNA, and messenger RNA. BLCA: bladder urothelial carcinoma; BRCA: breast invasive carcinoma; CESC: cervical squamous cell carcinoma; COADREAD: colon adenocarcinoma/rectum adenocarcinoma; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: low-grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.



Comparison Between Using Single-Cancer and Pan-Cancer Training Datasets

To further investigate the predictive performance of the proposed model across different types of datasets, this study conducted experiments using both pan-cancer and single-cancer datasets. The first objective was to examine whether training on pan-cancer data improves survival prediction accuracy for individual cancer types. In this study, we selected the same 20 cancer types as those used by Cheerla and Gevaert [9], where patients have significantly different survival patterns.

This study first compared the performance of models trained solely on single-cancer data with those trained on all pan-cancer samples, using consistent test sets for each cancer type. For single-cancer experiments, this study selected patients with the same cancer type in pan-cancer training-validation-test sets to form the single-cancer training-validation-test sets. The model was trained using integrated clinical, mRNA, and miRNA modalities, and the results are shown in Figure 7.

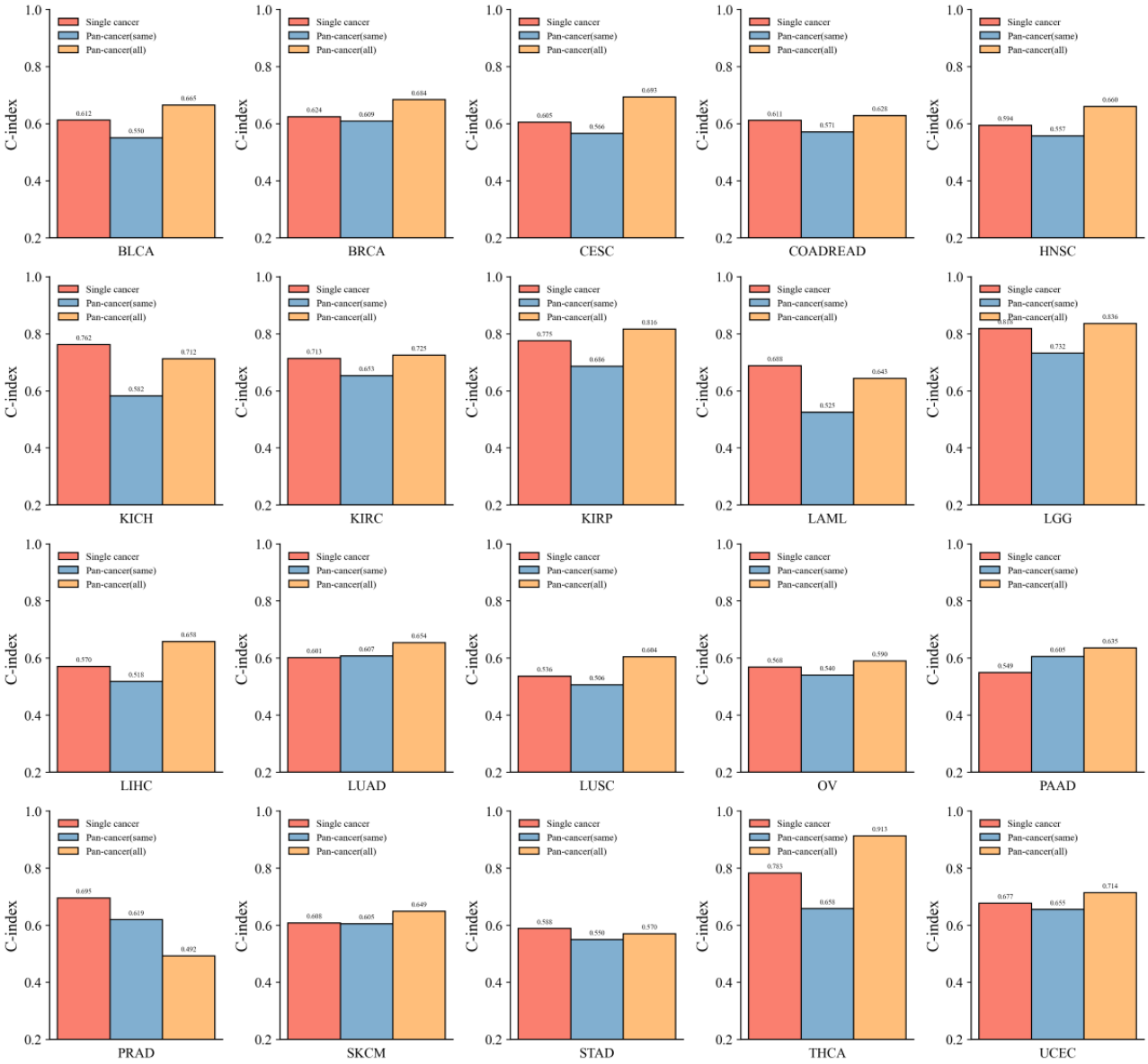
As shown by the bar plots labeled “Single cancer” and “Pan-cancer (all)” in Figure 7, the model trained on pan-cancer data generally outperformed the single-cancer–trained model

XSL•FO
RenderX

on the same test sets, with the exception of 4 cancer types: kidney chromophobe, acute myeloid leukemia, prostate adenocarcinoma, and stomach adenocarcinoma. For example, in the case of bladder urothelial carcinoma the pan-cancer–trained model achieved a C-index of 0.665, significantly higher than the 0.612 achieved by the model trained only on bladder urothelial carcinoma data.

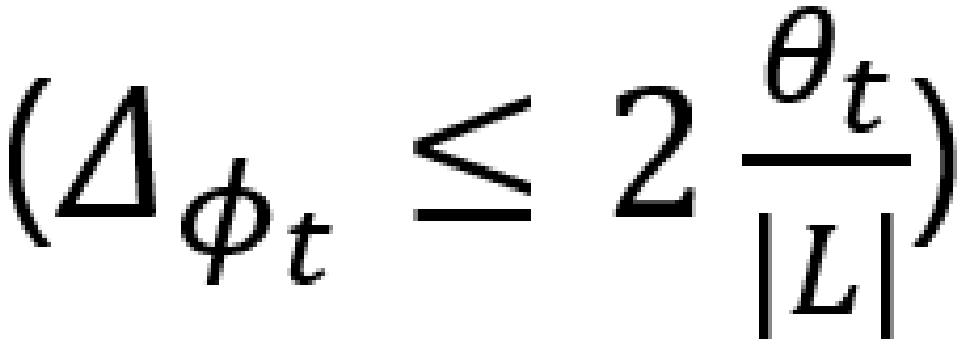
To control for potential biases due to varying training sample sizes, this study further evaluated model performance when trained on an equal number of samples from the pan-cancer and single-cancer datasets. Specifically, we selected subsets of pan-cancer data matching the sample size of each individual cancer dataset. As illustrated by the bar plots labeled "Single cancer" and "Pan-cancer (same)" in Figure 7, the single-cancer–trained models generally outperformed the pan-cancer–trained counterparts under equal training size conditions.

Figure 7. Concordance index (C-index) scores of the proposed model trained on single-cancer and pan-cancer datasets on 20 cancer types using clinical, messenger RNA, and microRNA modalities. Red bars represent models trained on single-cancer datasets, where only patients with the same cancer type were used. Blue bars indicate models trained on pan-cancer datasets with sample sizes matched to each single-cancer dataset to control for sample size bias. Orange bars correspond to models trained on the full pan-cancer dataset. BLCA: bladder urothelial carcinoma; BRCA: breast invasive carcinoma; CESC: cervical squamous cell carcinoma; COADREAD: colon adenocarcinoma/rectum adenocarcinoma; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: low-grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.



XSL•FO
RenderX

Figure 34. Inline graphic 27.



External Validation on Independent Cohorts

To address the critical question of generalizability beyond the TCGA dataset, we conducted an external validation on 2 independent cancer cohorts from the International Cancer Genome Consortium (ICGC) database [39]. Specifically, we selected the hepatocellular carcinoma (HCC; ICGC-HCC) and clear cell renal cell carcinoma (ccRCC; ICGC-ccRCC) projects, which provide clinical survival information and mRNA expression data compatible with our model's input requirements. To maintain consistency with our multimodal input configuration (clinical + mRNA + miRNA), the same zero-vector imputation strategy described in the Methods section ("Data Preprocessing") was applied to handle the missing miRNA modality and any mRNA features absent in the external data. The ICGC-HCC cohort comprised 242 patients with overall survival data, and the ICGC-ccRCC cohort comprised 530 patients. For each cohort, we preprocessed the mRNA expression data by aligning gene features to the 1579 mRNA features used in our TCGA training set. All continuous features were normalized using the minimum and maximum values derived from the TCGA training set to ensure consistent scaling.

This study applied our proposed model (trained on the pan-cancer TCGA dataset with clinical + mRNA + miRNA modalities) directly to the preprocessed ICGC cohorts without any retraining or fine-tuning. The model's survival prediction performance was evaluated using the C-index.

Our proposed model, trained on the pan-cancer TCGA dataset, achieved a C-index of 0.721 on the independent ICGC-HCC cohort and 0.708 on the ICGC-ccRCC cohort. This demonstrates that the model maintains its generalizable prognostic discriminatory power on unseen, external patient populations.

Discussion

Principal Results

In this study, we evaluated the model's classification performance in survival prediction across 4 data combinations—clinical data, mRNA, miRNA, and CNV—using the pan-cancer dataset. Compared with MultiSurv [17], the proposed model exhibits an approximately 0.3 improvement in prediction accuracy across all modality combinations after noise injection. When compared with Fan et al [10], the proposed model achieves higher accuracy in all settings before noise addition, and maintains comparable accuracy after noise injection, despite the latter model not incorporating any privacy-preserving mechanisms.

The improvement in survival prediction across different modalities stems from the model's ability to capture key information from different subspaces in parallel, thereby overcoming the representational limitations of single-attention mechanisms. By effectively integrating core features from each modality, the model enhances the flexibility and expressiveness of the fusion process. Notably, the inclusion of the CNV modality generally leads to a decline in predictive performance, suggesting that CNV data contain relatively fewer informative features relevant to patients' survival. This further underscores the value of multimodal data fusion in enhancing the generalizability of survival prediction models.

Comparisons in single-cancer survival prediction further demonstrate that the model achieves higher and more stable C-index values in 18 out of 20 cancer types compared with existing deep learning methods, highlighting its robustness in individualized survival prediction. Our study not only confirms the effectiveness and superiority of the proposed model in individualized cancer survival prediction but also demonstrates its ability to preserve patients' privacy. Collectively, these results provide a more reliable and privacy-conscious solution for precision oncology applications.

XSL•FO
RenderX

In addition, experiments conducted on both pan-cancer and single-cancer datasets show that the proposed model maintains strong predictive capability across diverse data settings, demonstrating robust generalization ability.

Challenges

While the model demonstrates strong privacy-preserving capabilities through DP, the integration of DP mechanisms into clinical artificial intelligence models introduces certain challenges. First, DP may impact the interpretability of the model, as the noise added to the gradients or model parameters can obscure the relationships between input features and predicted outcomes. This could reduce the transparency of the model, making it more difficult for clinicians to trust and understand the model's decisions, which is critical in medical applications. In addition, DP mechanisms may complicate auditability, as it becomes harder to trace the influence of individual data points on the model's predictions due to the noise injected for privacy protection.

Second, the integration of privacy mechanisms such as DP into clinical workflows or federated research environments can be practically challenging. In clinical settings, where data are highly sensitive, ensuring that privacy is preserved while maintaining the usability and accuracy of predictive models requires careful consideration of computational resources and model deployment strategies. In federated research environments, where data are distributed across multiple institutions, ensuring that DP mechanisms are applied consistently across all sites without compromising the model's predictive performance can be technically difficult. Moreover, adapting these privacy-preserving mechanisms to existing clinical systems and ensuring compliance with data protection regulations (eg, General Data Protection Regulation requires close collaboration with regulatory bodies and careful planning of data-sharing protocols.

These considerations highlight the need for further research into improving the balance between privacy protection, model interpretability, and practical integration in clinical and research settings.

Limitations

The proposed model has certain limitations. First, while we have demonstrated the model's effectiveness on the TCGA pan-cancer dataset and provided initial evidence of generalizability through external validation on ICGC cohorts, the performance on independent datasets shows a nonnegligible decline. This underscores the inherent challenge of translating models trained on 1 cohort (even a large and diverse one like TCGA) to others due to population, treatment, and technical variability. More extensive validation across multiple, prospectively collected cohorts is needed to fully establish clinical readiness. For example, training efficiency requires improvement. Analysis suggests that one possible reason is the introduction of the bilayer feature extraction module based on MHA and the adaptive DP protection mechanism, which increases algorithmic computational complexity. Additionally, the model exhibits increased inference time compared with simpler architectures such as CPH or RSF, due to the MHA mechanism and the privacy-preserving noise injection process. This may limit its deployment in real-time clinical settings where rapid predictions are required. Meanwhile, comparative experiments with other advanced models on larger-scale multimodal datasets are necessary. Indeed, in the era of big data and multimodal data fusion, new challenges lie ahead.

Future Work

Future work will expand in several directions. On the one hand, integrating pathological images and proteomics data will enhance multimodal feature fusion, extracting and fusing more information to further improve prediction performance. On the other hand, the model will be applied to additional tasks such as cancer subtype classification and biomarker discovery. Finally, research on other noise mechanisms based on DP is needed to protect sensitive information from multiple perspectives and ensure the security of model training and sharing. Moreover, efforts will be made to optimize the computational efficiency and reduce inference time through techniques such as model pruning, quantization, or lightweight attention mechanisms, thereby enhancing the model's practical applicability in clinical environments. To address potential limitations of zero-vector imputation, future work will also explore advanced missing-modality handling strategies, including masking strategies and missing-modality–aware attention mechanisms.

Conclusions

To address the key challenges of inefficient multimodal integration, insufficient privacy protection, and limited generalizability in pan-cancer survival prediction, this study proposes a bilayer feature fusion model integrating the MHA mechanism and adaptive DP. Innovatively, the framework uses modality-specific FC networks in the first layer to extract features from clinical, mRNA, miRNA, and CNV data. The second layer uses MHA to model cross-modal interactions through parallel attention heads, dynamically allocating weights to focus on survival-related features and overcoming the limitations of static fusion. Additionally, layer-wise relevance analysis is used during first-layer feature extraction to quantify the correlation between features and outcomes, enabling adaptive DP—injecting less Laplacian noise into the gradients of high-correlation features and more noise into low-correlation features to balance privacy and utility. Experimental results on pan-cancer datasets from TCGA demonstrate the feasibility and superiority of our proposed method. Furthermore, external validation on independent ICGC cohorts provided preliminary evidence of the model's generalizability, showing retained predictive power superior to clinical-only models despite expected performance attenuation due to cohort heterogeneity.

XSL•FO
RenderX

Acknowledgments

The authors sincerely thank all colleagues and collaborators who provided helpful discussions and valuable suggestions during the preparation of this study. In addition, the authors note that XW and HW contributed equally as co-corresponding authors.

Funding

This work was supported by the Yunlong Lake Laboratory of Deep Underground Science and Engineering Project under grant 104024005, Jiangsu Provincial Innovation Capacity Building Program under grants BM2022009 and BM2023017, the Integrated Traditional Chinese and Western Medicine Chronic Disease Management Research Project under grant CXZH2024088, and the Serving the “343” Industrial Development Project for Universities in Xuzhou under grant gx2024017.

Data Availability

The data that support the findings of this study are openly available in the Xena database [40] and the TCGA database [41].

Authors' Contributions

Conceptualization: YC
Data curation: ZD
Formal analysis: YC (lead), LW (supporting)
Funding acquisition: XW
Investigation: ZD (lead), LW (supporting)
Methodology: XW
Project administration: XW
Resources: XW
Supervision: LW (lead), HW (supporting)
Validation: YC (lead), HW (supporting)
Visualization: YC (lead), ZD (supporting)
Writing—original draft: YC (lead), ZD (supporting)
Writing—review & editing: XW (lead), HW (supporting)

Conflicts of Interest

None declared.

References

1. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep 2021;11(1):13505 [FREE Full text] [doi: 10.1038/s41598-021-92799-4] [Medline: 34188098]
2. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann. Appl. Stat 2008;2(3):841-860. [doi: 10.1214/08-AOAS169]
3. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 2018;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
4. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018;18(1):24 [FREE Full text] [doi: 10.1186/s12874-018-0482-1] [Medline: 29482517]
5. Kim DW, Lee S, Kwon S, Nam W, Cha I, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep 2019;9(1):6994 [FREE Full text] [doi: 10.1038/s41598-019-43372-7] [Medline: 31061433]
6. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med 2021;4(1):71 [FREE Full text] [doi: 10.1038/s41746-021-00427-2] [Medline: 33875798]
7. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One 2020;15(6):e0233678 [FREE Full text] [doi: 10.1371/journal.pone.0233678] [Medline: 32555646]
8. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Wu X, Li P. Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative. Front Bioeng Biotechnol 2023;11:1164655 [FREE Full text] [doi: 10.3389/fbioe.2023.1164655] [Medline: 37122858]
9. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 2019;35(14):i446-i454 [FREE Full text] [doi: 10.1093/bioinformatics/btz342] [Medline: 31510656]
10. Fan Z, Jiang Z, Liang H, Han C. Pancancer survival prediction using a deep learning architecture with multimodal representation and integration. Bioinform Adv 2023;3(1):vbad006 [FREE Full text] [doi: 10.1093/bioadv/vbad006] [Medline: 36845202]

XSL•FO
RenderX

11. Guo W, Liang W, Deng Q, Zou X. A multimodal affinity fusion network for predicting the survival of breast cancer patients. Front Genet 2021;12:709027 [FREE Full text] [doi: 10.3389/fgene.2021.709027] [Medline: 34490038]
12. Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. Artif Intell Med 2022;126:102260. [doi: 10.1016/j.artmed.2022.102260] [Medline: 35346442]
13. Na H, Wang L, Zhuang X. Attention-based multimodal bilinear feature fusion for lung cancer survival analysis. : IEEE; 2023 Presented at: IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE); 2023 December 04-06; Dayton, OH, USA p. 219-225. [doi: 10.1109/bibe60311.2023.00042]
14. Li R, Wu X, Li A, Wang M. HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics 2022;38(9):2587-2594 [FREE Full text] [doi: 10.1093/bioinformatics/btac113] [Medline: 35188177]
15. Lee C, Yoon J, Schaar MVD. Dynamic-deepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng 2020;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]
16. Kim T, Cho H, Yoon KJ. Cmta: Cross-modal temporal alignment for event-guided video deblurring. Cham: Springer Nature Switzerland; 2024 Presented at: European Conference on Computer Vision; 2024 September 29–October 4; Milan, Italy p. 1-19. [doi: 10.1007/978-3-031-72943-0_1]
17. Silva LAV, Rohr K. Pan-cancer prognosis prediction using multimodal deep learning. 2020 Presented at: IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020 April 3-7; Iowa City, IA, USA p. 568-571. [doi: 10.1109/isbi45749.2020.9098665]
18. Zhang Z, Yin W, Wang S, Zheng X, Dong S. MBFusion: Multi-modal balanced fusion and multi-task learning for cancer diagnosis and prognosis. Comput Biol Med 2024;181:109042. [doi: 10.1016/j.compbiomed.2024.109042] [Medline: 39180856]
19. Wang C, Guo J, Zhao N, Liu Y, Liu X, Liu G, et al. A cancer survival prediction method based on graph convolutional network. IEEE Trans Nanobioscience 2020;19(1):117-126. [doi: 10.1109/TNB.2019.2936398] [Medline: 31443039]
20. Flack D, Tripathi A, Waqas A, Rasool G, Dera D. Robust multimodal fusion for survival prediction in cancer patients. Cancer Inform 2025;24:11769351251376192 [FREE Full text] [doi: 10.1177/11769351251376192] [Medline: 41024938]
21. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. Sci Adv 2021;7(47):eabg3296 [FREE Full text] [doi: 10.1126/sciadv.abg3296] [Medline: 34788101]
22. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2(6):305-311. [doi: 10.1038/s42256-020-0186-1]
23. Chen H, Wang N, Zhou Y, Mei K, Tang M, Cai G. Breast cancer prediction based on differential privacy and logistic regression optimization model. Applied Sciences 2023;13(19):10755. [doi: 10.3390/app131910755]
24. Chai H, Huang Y, Xu L, Song X, He M, Wang Q. A decentralized federated learning-based cancer survival prediction method with privacy protection. Heliyon 2024;10(11):e31873 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e31873] [Medline: 38845954]
25. Wu X, Wei Y, Mao Y, Wang L. A differential privacy DNA motif finding method based on closed frequent patterns. Cluster Comput 2018;22(S2):2907-2919. [doi: 10.1007/s10586-017-1691-9]
26. Wu X, Zhang Y, Shi M, Li P, Li R, Xiong NN. An adaptive federated learning scheme with differential privacy preserving. Future Generation Computer Systems 2022;127:362-372. [doi: 10.1016/j.future.2021.09.015]
27. Wang H, Zhang X, Xia Y, Wu X. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. Computer Standards & Interfaces 2023;84:103694. [doi: 10.1016/j.csi.2022.103694]
28. Wang H, Zhang X, Xia Y, Wu X. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. Computer Standards & Interfaces 2023;84:103694. [doi: 10.1016/j.csi.2022.103694]
29. Wang H, Wu X. IPP: An intelligent privacy-preserving scheme for detecting interactions in genome association studies. IEEE/ACM Trans Comput Biol Bioinform 2023;20(1):455-464. [doi: 10.1109/TCBB.2022.3155774] [Medline: 35239492]
30. Al Fatih Abil Fida M, Ahmad T, Ntahobari M. Variance threshold as early screening to boruta feature selection for intrusion detection system. 2021 Presented at: 13th International Conference on Information & Communication Technology and System (ICTS); 2021 October 19-21; Surabaya, Indonesia p. 46-50. [doi: 10.1109/icts52701.2021.9608852]
31. Gajera V, Gupta R, Jana PK. An effective multi-objective task scheduling algorithm using min-max normalization in cloud computing. 2016 Presented at: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT); 2016 July 21-23; SJB Institute of Technology, Bengaluru, Karnataka, India p. 812-816. [doi: 10.1109/icatcct.2016.7912111]
32. Hutter C, Zenklusen JC. The cancer genome atlas: Creating lasting value beyond its data. Cell 2018;173(2):283-285 [FREE Full text] [doi: 10.1016/j.cell.2018.03.042] [Medline: 29625045]
33. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Cancer Genome Atlas Research Network, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell 2018;173(2):338-354.e15 [FREE Full text] [doi: 10.1016/j.cell.2018.03.034] [Medline: 29625051]
34. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 2015. [doi: 10.5860/choice.189890]

XSL•FO
RenderX

35. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247(18):2543-2546. [Medline: 7069920]
36. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K. Deep learning with differential privacy. 2016 Presented at: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security; 2016 October 24 - 28; Vienna Austria p. 308-318. [doi: 10.1145/2976749.2978318]
37. Tang Q, Lécuyer M. DP-Adam: Correcting DP bias in adam's second moment estimation. arXiv:2304.11208 2023. [doi: 10.48550/arXiv.2304.11208]
38. Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv:1610.05755 2016. [doi: 10.48550/arXiv.1610.05755]
39. ICGC ARGO. URL: https://www.icgc-argo.org/ [accessed 2026-03-06]
40. UCSC Xena. URL: https://xenabrowser.net/ [accessed 2026-03-06]
41. GDC Data Portal. URL: https://portal.gdc.cancer.gov/ [accessed 2026-03-06]

Abbreviations

ccRCC: clear cell renal cell carcinoma
C-index: concordance index
CMTA: cross-modal translation and alignment
CNV: gene copy number variation
CPH: Cox proportional hazards model
DP: differential privacy
FC: fully connected
HCC: hepatocellular carcinoma
ICGC: International Cancer Genome Consortium
MHA: multihead attention
miRNA: microRNA
mRNA: messenger RNA
PATE: Private Aggregation of Teacher Ensembles
RSF: random survival forest
SGD: Stochastic Gradient Descent
TCGA: The Cancer Genome Atlas

Edited by J Klann; submitted 08.Sep.2025; peer-reviewed by Y Yang, F Amakye, D Dera, J Zhu; comments to author 11.Dec.2025; accepted 09.Feb.2026; published 30.Mar.2026.

Please cite as:
Chen Y, Deng Z, Wang L, Wang H, Wu X
A Bilayer Feature Fusion Framework for Pan-Cancer Survival Prediction Based on Multihead Attention and Adaptive Differential Privacy: Model Development and Validation Study
JMIR Med Inform 2026;14:e83743
URL: https://medinform.jmir.org/2026/1/e83743
doi:10.2196/83743
PMID:

©Yun Chen, Zhifang Deng, Lili Wang, Huanhuan Wang, Xiang Wu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trust in AI-Supported Screening in General Practice Among Urban and Rural Citizens: Cross-Sectional Study

Larisa Wewetzer, Dr Rer Hum Biol; Katja Goetz, Prof Dr Phil, Dipl-Soz; Soenke Freischmidt, Dr Med; Jost Steinhauser, Prof. Dr Med

Institute of Family Medicine, University Medical Center Schleswig-Holstein, Ratzeburger Allee 160, Luebeck, Germany

Corresponding Author:
Larisa Wewetzer, Dr Rer Hum Biol
Institute of Family Medicine, University Medical Center Schleswig-Holstein, Ratzeburger Allee 160, Luebeck, Germany

Abstract

Background: The early detection of diseases is one of the tasks of general practice. Artificial intelligence (AI)–based technologies could be useful for identifying diseases at an early stage in general practices. As approximately 90% of the population regularly consults a general practitioner during one year, this could increase the percentage of citizens who take part in meaningful screening measures.

Objective: This study aimed to evaluate the level of trust among citizens in rural and urban areas in AI-supported early detection measures in general practice.

Methods: This cross-sectional study was conducted in the federal state of Schleswig-Holstein, Germany, from November 2023 to December 2023, on the topic of early detection measures with AI in general practice care, among other things. For this purpose, 5000 adult residents of rural areas (Ostholstein, Pinneberg, and Nordfriesland) and urban areas (the city of Kiel) were invited to take part in the survey. Data analysis was carried out using descriptive statistics, subgroup analysis, and linear and stepwise regression analysis to identify the factors that influenced trust in AI-based diagnoses.

Results: Most respondents (787/1790, 44.0%) considered the introduction of an AI-based screening measure to be a sign of modern medicine. Moreover, 21.7% (n=388) of respondents feared that the introduction of such services could lead to a deterioration in the physician-patient relationship. The role of AI in future care was rated as very important by 35.4% (n=634) of respondents. The stepwise regression analysis showed that a positive attitude toward AI in medicine was the strongest predictor (ß=0.420) of trust in AI-based diagnoses. In contrast, trust in physician diagnoses was associated with lower age (ß=–0.111) and shorter waiting times for test results (ß=0.077).

Conclusions: Trust in general practitioner–based diagnoses was approximately 6 times greater than trust in AI applications. Despite concerns about their impact on the physician-patient relationship, approximately one-third of participants believed that the role of AI in health care will grow.

(JMIR Med Inform 2026;14:e69777) doi:10.2196/69777

KEYWORDS

artificial intelligence; AI; citizens; cross-sectional study; diabetic retinopathy; general practice; trust

Introduction

The basic care of all patients in emergency, acute, and long-term care, as well as in prevention and rehabilitation, constitutes a central responsibility of general medicine [1]. Therefore, the early detection of diseases represents an important aspect of secondary prevention in a general practitioner (GP) practice. Appropriate measures can be initiated through early detection, which can considerably improve the course of an individual illness. GPs are often the first to diagnose diseases such as diabetes, hypertension, and cardiovascular diseases [2].

More than 90% of the population regularly consults their GP per year [1], which offers a significant opportunity to increase the proportion of citizens who take part in appropriate screening measures.

Screening for eye diseases is a promising extension of these early detection measures. Diabetic retinopathy (DR), in particular, can lead to severe visual impairment or even blindness if it is not detected and treated early. The number of people in Germany who become blind as a result of DR is approximately 15,000 per year [3]. DR is therefore the second most common cause of blindness [4]. According to a European study, 48% of patients with diabetes do not undergo the recommended screening examination [5].

The current diagnosis of these diseases requires a referral to an ophthalmologist, which can lead to long waiting times, especially in rural areas with limited access to specialists [6].

XSL•FO
RenderX

In this context, artificial intelligence (AI)–supported technologies could play an important role by complementing and extending traditional screening approaches. For example, in general practice, these technologies may include AI-assisted analysis of skin lesions for cancer screening or algorithms that identify abnormal heart rhythms in long-term electrocardiograms. In this study, we focused on DR as an illustrative example.

AI-assisted diagnostic tools could be used to analyze retinal images without requiring the presence of an eye specialist [7,8]. At the same time, recent studies indicate good diagnostic performance of deep learning algorithms. For example, in a study conducted in 2019, the detection of DR was analyzed on 25,326 retinal images of diabetes patients from Thailand. The algorithm showed a sensitivity of 97%, higher than the 74% observed for human experts, and a specificity of 96%, compared with 98% of specialists [9].

The implementation of AI-supported technologies could enable screening for DR directly in the GP practice as part of disease management programs. In this way, suspicious findings could be detected early, and patients could be referred to ophthalmologists accordingly.

It is still unknown whether trust in AI-supported technologies for early disease detection varies across population groups. However, significant variations in attitudes toward eHealth were observed between urban and rural respondents, highlighting a potential disparity in the future implementation of these technologies in general practice across demographic areas [10,11]. Therefore, the aim of the study was to evaluate the level of trust citizens in rural and urban areas have in AI-supported early detection measures in general practice.

Methods

This cross-sectional study was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [12].

Study Design and Participants

A cross-sectional study on early detection measures using AI in GP care was conducted in the federal state of Schleswig-Holstein, Germany, from November 2023 to December 2023. To this end, 5000 adult citizens from Schleswig-Holstein were invited to take part in the survey. The citizens were randomly selected from the databases of the local registration office.

Following the first invitation, 2 reminders to participate in the study were sent out at 3-week intervals. A stamped and preaddressed return envelope was included in all mailings to facilitate participation. A short version of the questionnaire was enclosed with the third reminder letter. This approach, which is consistent with the Tailored Design Method in survey research, was chosen to maximize the response rate and reduce the participation barrier for citizens who had not yet responded, by minimizing the time required.

Data Collection and Questionnaire

Data were collected using a self-constructed questionnaire in paper form. The development of the questionnaire was rigorously based on the findings of a qualitative study that served as a formal pilot phase and informed the generation of items [13]. This preceding qualitative study explored the determinants of AI implementation for DR screening in primary care. In addition, a subset of items—especially those assessing trust in AI vs physician diagnosis and general attitudes toward AI—was pilot-tested in a small qualitative pretest with GPs (1 female GP and 3 male GPs) to ensure clarity and face validity.

To account for different levels of respondent burden, 2 versions of the questionnaire were used:

- The long version consisted of 35 items, covering all conceptual domains in depth (telemedicine use, attitudes toward AI, perceived impact of AI-supported screening, importance ratings for screening attributes, expectations about examination duration, trust in diagnoses, and sociodemographic variables).
- The short version was reduced to 5 core study-related items, capturing only the core constructs (general attitude toward telemedicine, general attitude toward AI in medicine, trust in AI diagnosis, trust in physician diagnosis, age, gender, and region).

The use of long and short questionnaire versions was chosen because shorter instruments in postal surveys can increase response rates and reduce dropout, especially when sent as reminders [14].

The long questionnaire consisted of five sections:

1. Telemedicine experience and attitudes
2. Attitudes toward AI in health care
3. Perceived impact of AI-based screening in general practice
4. Importance of specific aspects of AI-assisted eye disease screening
5. Trust in AI-based diagnosis vs diagnosis by physicians

The participants' general attitude toward AI applications in general practice was assessed using a 6-point Likert scale ranging from "very positive" to "very negative." In addition, questions were asked about the expected future role of AI in their own health care and the potential impact of integrating AI into general practice.

Participants were asked to rate from a list of options what impact the use of AI-based screening measures would have on their everyday life and care, such as improving or worsening patient care and physician-patient relationships. In addition, specific aspects were asked about the importance of performing AI-based screening tests. The importance of these aspects was rated on a 6-point scale, ranging from "very important" to "very unimportant." Finally, respondents were asked about their trust in diagnoses made by AI compared with diagnoses made by GPs.

The questionnaire also included sociodemographic parameters such as age, gender, and regional origin. Sociodemographic information (age, gender, and regional origin) was collected at the end of the instrument. Regional origin was dichotomized

XSL•FO
RenderX

into rural (Ostholstein, Pinneberg, and Nordfriesland) and urban (Kiel).

Statistical Data Analysis

The analyses were carried out using the SPSS software (version 29.0; IBM Corp). The sociodemographic characteristics of the sample (full and short versions of the questionnaire) were analyzed descriptively. Variables with continuous values were summarized using mean values and SDs. Variables with categorical values were represented by absolute (n) and relative (%) frequencies. The regional origin was dichotomized: the regions Ostholstein, Pinneberg, and Nordfriesland were assigned to the rural area, and the Kiel region to the urban area.

The sociodemographic characteristics of the sample (full and short versions of the questionnaire) were analyzed descriptively. Variables with continuous values were summarized using mean values and SDs. Variables with categorical values were represented by absolute (n) and relative (%) frequencies. The regional origin was dichotomized: the regions Ostholstein, Pinneberg, and Nordfriesland were assigned to the rural area, and the Kiel region to the urban area.

The nonparametric Mann-Whitney *U* test was used to compare the full and short versions of the questionnaire regarding the general preferences of telemedicine and AI. Furthermore, a subgroup analysis was performed by using the nonparametric Mann-Whitney *U* test to assess whether the regional origin (0=rural region and 1=urban region) differed concerning the preferences for telemedicine and AI.

Moreover, linear correlations between trust in diagnosis with AI and trust in diagnosis through physicians, the sociodemographic variables, and various influencing factors were calculated using Spearman correlation coefficient. Testing for connections using a correlation is a common procedure to reduce the number of variables for subsequent model calculations. In a further step, significant correlations identified in this way were taken into account in the linear regression models.

Afterward, 2 stepwise regression analyses were performed to evaluate which factors were strongly associated with the outcome variables. Trust in diagnosis with AI and trust in diagnosis through physicians were the outcome variables, while other aspects, such as preferences for telemedicine and AI, attitudes to early detection measures, and some characteristics of participants (such as gender, age, regional origin), were considered potential predictors and handled as covariates. Additionally, the possibility for multicollinearity was considered in both models. The variance inflation factor (VIF) and the tolerance value were reported for the final step of the regression models. Values for VIF should not exceed 5.0, and tolerance values should not be lower than 0.25 [15]. Statistical significance was set at $P<.05$.

Ethical Considerations

The ethics application for conducting the study was approved by the ethics committee of the University of Lübeck in March 2023 (2023-258_1). All participants provided informed consent prior to participation and were informed that their participation was voluntary and that they could withdraw from the study at any time without consequences. The study involved only the analysis of data collected within the scope of the approved protocol; no additional data were evaluated beyond those covered by the original ethical approval. All data were anonymized prior to analysis, and no personally identifiable information was accessible to the researchers. Appropriate technical and organizational measures were implemented to ensure data protection and confidentiality in accordance with applicable data protection regulations. No incentives were offered for participation.

Results

Description of the Sample

Of the 5000 citizens contacted, 1790 (35.8%) took part in the survey, corresponding to a response rate of almost 36%. Of the participants, 785 (43.9%) were male and 972 (54.3%) were female. A total of 10 participants (0.6%) identified themselves as diverse, while 23 (1.3%) did not specify their gender. Of these participants, 352 (19.7%) chose the short version of the questionnaire. Of these, 160 (44.1%) were male, 189 (52.1%) were female, 3 (0.8%) were diverse, and 7 (1.9%) had missing data for this variable. The average age of all participants was 56 years (range 18-96 years).

Of the 1790 participants, 764 (42.7%) lived in the city and 979 (54.7%) lived in a rural area. Most participants lived in Kiel (764/1790, 42.7%), followed by 520 (29.1%) in Ostholstein, 335 (18.7%) in Pinneberg, and 124 (6.9%) in Nordfriesland districts. Table 1 provides further details on the comparison between the long version and the short version of the questionnaire.

XSL•FO
RenderX

Table . Sociodemographic characteristics of participants (n=1790).

Characteristics[a]	Long version questionnaire (n=1427)	Short version questionnaire (n=363)	*P* value
Gender, n (%)			.64
Female	783 (54.9)	189 (52.1)	
Male	625 (43.8)	160 (44.1)	
Diverse	7 (0.5)	3 (0.8)	
Age (y), mean (SD; range)	55.8 (16.8; 18 - 96)	57 (16.4; 18 - 91)	.11
Regional origin, n (%)			.09
Kiel (urban)	606 (42.5)	158 (43.5)	
Ostholstein (rural)	406 (28.5)	114 (31.4)	
Pinneberg (rural)	275 (19.3)	60 (16.5)	
Nordfriesland (rural)	109 (7.6)	15 (4.1)	
Assessment of local medical care[b], mean (SD)	2.97 (1.43)	2.93 (1.32)	.88

[a]The n values vary due to missing data.

[b]Ranging from 1 (very good) to 6 (very bad).

General Attitude, Future Role of AI, and Trust in Diagnosis Through AI and Physicians

A descriptive analysis of general preferences regarding telemedicine and AI in health care was carried out (Table 2). The variables analyzed included general attitudes toward telemedicine and AI, as well as specific aspects such as trust in diagnoses by AI and the future role of AI in one's own health care.

Table . General preferences for telemedicine and artificial intelligence (AI)—descriptive analysis.

Variables	Long version, mean (SD)	Short version, mean (SD)	*P* value
General attitude to the topic of telemedicine[a]	3.44 (1.50)	3.58 (1.37)	.06
General attitude to applications of AI in medicine[a]	3.35 (1.49)	3.47 (1.47)	.19
Future role of AI for own health care	2.91 (1.48)	—[b]	—
Trust in diagnosis through AI[c]	3.53 (1.43)	—	—
Trust in diagnosis through physician[c]	1.96 (0.82)	—	—

[a]Ranging from 1 (very positive) to 6 (very negative).

[b]Not available.

[c]Ranging from 1 (very small) to 6 (very big). These items were not asked in the short version.

For the general attitude toward applications of AI in medicine, a mean value of 3.35 (SD 1.49) was observed for the long version of the questionnaire, and a mean value of 3.47 (SD 1.47) was observed for the short version.

The assessment of the future role of AI in personal health care had a mean value of 2.91 (SD 1.48). The SD of 1.48 suggests a wide distribution of responses. It was rated important by 35.4% (n=634) of respondents.

Trust in diagnoses by AI is rated with a mean value of 3.53 (SD 1.43). In comparison, trust in diagnoses by physicians had a mean value of 1.96 (SD 0.82).

Subgroup Analysis Concerning Regional Origin and General Preferences to Telemedicine and AI

Participants from urban regions were more positive toward telemedicine ($mean_{urban}$ 3.26, SD 1.5; $mean_{rural}$ 3.56, SD 1.5; $P<.001$) and toward applications of AI in medicine ($mean_{urban}$ 3.11, SD 1.4; $mean_{rural}$ 3.52, SD 1.5; $P<.001$) than participants from rural regions. Moreover, for respondents from urban regions, AI was perceived to play a bigger future role for own health care ($mean_{urban}$ 2.65, SD 1.4; $mean_{rural}$ 3.09, SD 1.5; $P<.001$), and they reported greater trust in diagnoses made by AI ($mean_{urban}$ 3.38, SD 1.4; $mean_{rural}$ 3.63, SD 1.5; $P<.001$) than participants from rural regions. No significant difference was

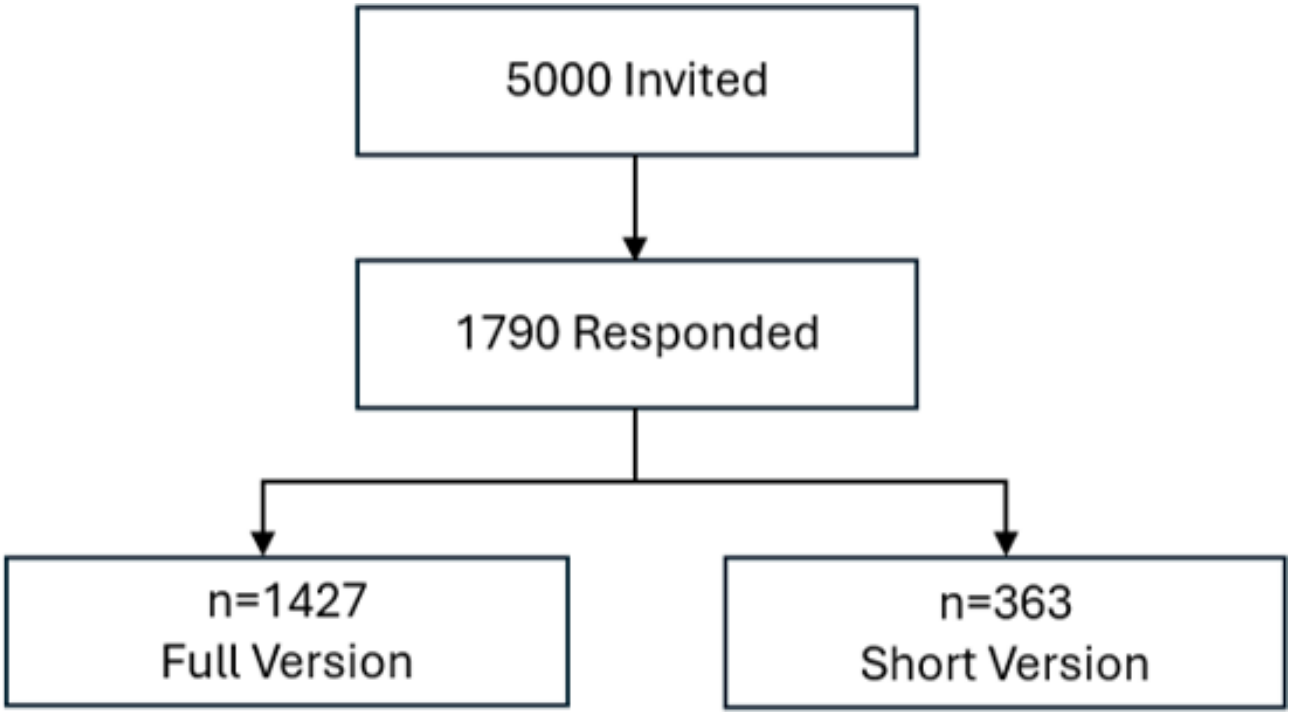
XSL•FO
RenderX

observed for trust in diagnosis through physicians ($mean_{urban}$ 1.99, SD 0.80; $mean_{rural}$ 1.93, SD 0.81; *P*=.14).

Relevance of AI-Based Screening Measures in the GP Practice

Participants were asked to rate the use of AI-based screening measures in GP practices (Figure 1). Respondents had the opportunity to select several statements that applied.

Figure 1. Response rates for the long version vs short version of the questionnaire.



The analysis of the survey results showed a differentiated perception of the AI-based screening measures, which reflected both potential benefits and concerns among the citizens surveyed.

Out of the 1790 respondents, the most frequently mentioned statement was “signs of modern medicine,” selected by 787 (44.0%) participants. The statement “improved patient care” was selected by 618 participants (34.5%), followed by “easing everyday life” with 445 (24.9%) responses. In contrast, 388 (21.7%) participants stated that they would expect a “worsened doctor-patient relationship,” and 270 (15.1%) respondents described the measure as “intensifying uncertainty.”

The statements “improved doctor-patient relationship” (n=167, 9.3%), “signs of unnecessary technologization” (n=162, 9.1%), “deterioration in patient care” (155, 8.7%), and “difficulty in everyday life” (n=37, 2.1%) were selected less frequently (Multimedia Appendix 1).

Aspects of the Implementation of Screening Examinations for Eye Diseases Using AI in GP Practices

In the data received from the long version of the questionnaire, the most important factors from the citizens’ perspective were trust in the GP (mean 1.78, SD 1.24) and the reliability of the examination (mean 1.43, SD 1.01), followed by data security (mean 2.19, SD 1.52).

With regard to organizational factors, the accessibility of the practice (mean 2.14, SD 1.31) and the waiting time for an ophthalmologist appointment (mean 2.13, SD 1.26) were also rated as relevant. The protection against additional examination costs (eg, Individual Health Services; mean 2.30, SD 1.42) was rated as moderately important.

Factors considered less important from the citizens’ perspectives included not having their pupils dilated (mean 3.24, SD 1.65), the duration of the examination (mean 2.93, SD 1.45), and the distance to the nearest ophthalmologist (mean 2.50, SD 1.39).

The results showed that trust in the GP conducting the examination, the reliability of the examination, and data security were the most important aspects considered by the participants when implementing such technologies in GP practices (Table 3).

XSL•FO
RenderX

Table . Evaluation of the aspects of artificial intelligence–supported screening for eye diseases in general practitioner practices.

Variables[a]	Factors, mean (SD; 95% CI)
Trust in general practitioner	1.78 (1.24; 1.71 - 1.86)
Duration of the examination	2.93 (1.45; 2.84 - 3.02)
Waiting time until the test results are available	2.44 (1.29; 2.37 - 2.52)
Data security	2.19 (1.52; 2.10 - 2.29)
Reliability of the examination	1.43 (1.01; 1.37 - 1.49)
Not having the pupils dilated	3.24 (1.65; 3.14 - 3.35)
Accessibility of the practice	2.14 (1.31; 2.06 - 2.22)
Distance to the nearest ophthalmologist	2.50 (1.39; 2.41 - 2.58)
Waiting time for an eye specialist appointment	2.13 (1.26; 2.05 - 2.21)
Waiting time in an ophthalmologist's practice	2.49 (1.31; 2.41 - 2.57)
Protection against additional examination costs	2.30 (1.42; 2.22 - 2.39)
Performance of the examination by medical assistants	2.12 (1.30; 2.04 - 2.20)

[a]Scale from 1=very important to 6=very unimportant.

Factors Associated With Trust in Diagnosis With AI

The stepwise regression analysis is presented in Table 4 and reports only coefficients with statistical significance at the $P<.05$ level. A 9-step model was fitted and explained more than 60% ($R^2=0.602$) of the variance in the dependent variable, "trust in diagnosis through AI." In the first step of the stepwise regression analysis, the variable "general attitude to applications of AI in medicine" showed the highest explained variance ($R^2=0.509$). A higher positive level of attitude toward applications of AI in medicine was strongly associated with higher level of trust in diagnosis through AI. Individual characteristics were not included in the regression model, as $P>.05$. Collinearity statistics ranged between 2.112 (VIF value) and 0.473 (tolerance value) for "general attitude to applications of AI in medicine" and 1.140 (VIF value) and 0.877 (tolerance value) for "improved patient-physician relationship."

Table . Associations of individual characteristics, general preferences for telemedicine and artificial intelligence (AI), and attitudes toward early detection measures in general practitioner care with general trust in AI (results of stepwise linear regression analysis).

Variables[a]	β	*P* value
General attitude to applications of AI in medicine	0.420	<.001
Intensifying uncertainty	0.136	<.001
Future role of AI for own health care	0.164	<.001
Signs of unnecessary technologization	0.120	<.001
Improved physician-patient relationship	−0.063	.003
Providing security	−0.091	<.001
Duration of the examination	0.063	.004
Worsened physician-patient relationship	0.066	.004
Reliability of the examination	0.053	.02

[a]Last step was reported.

Factors Associated With Trust in Diagnosis Through Physicians

The stepwise regression analysis is presented in Table 5 and reports only coefficients with statistically significances at the $P<.05$ level. A 5-step model was fitted and explained more than 4.8% ($R^2\approx.048$) of the variance in the dependent variable, "trust in diagnosis through physicians." For example, a higher positive level of trust in diagnosis through physicians was strongly associated with lower age and greater importance attributed to the waiting time until the test results were available. Collinearity statistics ranged between 1.308 (VIF value) and 0.765 (tolerance value) for "distance to the nearest ophthalmologist" and 1.007 (VIF value) and 0.993 (tolerance value) for "providing security."

XSL•FO
RenderX

Table . Associations of individual characteristics, general preferences for telemedicine and artificial intelligence, and attitudes toward early detection measures in general practice care on general trust in diagnosis through physicians (results of stepwise linear regression analysis).

Variables	β	*P* value
Age of the participants	−0.111	<.001
Waiting time until the test results are available	0.077	.03
Providing security	0.093	.003
Local general medical care	0.089	.005
Distance to the nearest ophthalmologist	0.090	.01

[a]Last step was reported.

Discussion

This study examines citizens' attitudes toward the use of AI in general practice, particularly in the context of early detection measures. The study showed that the introduction of AI-supported technologies is met with mixed reactions.

Differences Between Rural and Urban Areas

Significant differences emerged between urban and rural participants. People from urban areas were significantly more positive about the use of AI in medicine than those from rural areas.

Differences in access to medical services between urban and rural areas are deeply rooted and reflect a complex mix of infrastructural, socioeconomic, and cultural factors [16,17]. While urban areas often provide a wide range of medical services and specialists, some rural areas experience a shortage of physicians, greater distances to health facilities, and limited access to specialized treatment [17].

The introduction of AI and telemedicine is often promoted as a solution to these inequalities by extending access to care. However, emerging research suggests that these technologies are not always met with acceptance [10,11]. The reasons for this are varied and include technological barriers, a lack of trust in digital diagnostic methods, and a preference for in-person physician-patient interactions [12,18,19].

To ensure equal access to care, a deeper understanding of the specific needs and concerns of the rural population is needed. This means not only providing technology but also taking into account the cultural and psychological aspects that influence attitudes toward technology.

Dynamics of Trust in General Practice Care and Technology

This study pointed to a trust deficit in AI-supported technologies. Trust plays an essential role both in interpersonal interactions and in the use of technology. The patient-physician relationship is built on trust, particularly in the physician's commitment to voluntary responsibility and compassionate judgment [20]. This dimension of trust is crucial, as patients in vulnerable health situations depend heavily on the expertise and care of their treating physicians.

This trust is primarily shaped by the quality of the physician-patient relationship, in particular effective communication, interpersonal care, and the physician's knowledge of the patient's individual needs. The study emphasized that these factors play an important role in strengthening patient trust, while other aspects, such as the duration of registration at a practice or the frequency of visits, had less influence on trust [21].

Interpersonal interaction and trust in the treating physician are essential for patient satisfaction and adherence to treatment [20]. Research also highlighted that perceived empathy and individualized care are particularly important to patients, reinforcing the importance of these qualities in maintaining trust in medical professionals [21].

Central to this trust is the concept of continuity of care, particularly relational continuity, in which patients consistently see the same physician over time.

The association between continuity of primary care and patient mortality has been described by categorizing continuity into 3 types: relational continuity, informational continuity, and management continuity. A review of 13 quantitative studies found that greater continuity of care is significantly associated with lower all-cause mortality, particularly among older patients and those with chronic conditions. This protective effect was explained by improved physician knowledge, increased patient confidence, and better adherence to medical advice [22].

Therefore, in an increasingly digitalized world, it is essential that new digital health applications are designed from a patient-centered perspective.

Trust in technologies is predominantly based on functional aspects, such as reliability, security, and predictability [23]. From the perspective of the citizens in our study, the reliability of the examination was the most important factor.

Trust in technology is a construct that is strengthened by a combination of experiences with the technology, trust in users, and trust in regulators. Each of these components plays a critical role in maintaining and fostering trust over time. Past experiences, especially the reliability of the technology, had a significant impact on user trust [24]. In addition, trust is strengthened by the integrity and competence of those who use the technology, especially in the medical field [25], as well as by trust in regulatory authorities responsible for the safety and effectiveness of the technology [26].

The ways in which AI technologies work are currently incomprehensible to the public. This technological opacity requires a certain degree of explainability and interpretability

XSL•FO
RenderX

[27]. This lack of transparency means that damage caused by interaction with an AI system cannot be easily proven or clearly attributed [28]. This poses a major challenge and underlines the urgency of developing mechanisms that clearly regulate responsibility and liability in connection with AI decisions.

AI is often seen as a potential solution to current challenges in the health care system, such as relieving the burden on physicians and promoting a person-centered physician-patient relationship [29,30]. In a systematic analysis of 45 studies, the impact of AI on the person-centered physician-patient relationship was investigated. The researchers identified various ways in which the use of AI tools could promote person-centered care. One of these ways was the use of AI tools in a supportive role [31].

The rising interest in AI for medical applications should not overshadow the need for proper scientific review and evaluation. It is crucial that the medical field maintains its high standards of evidence and ethical responsibility to ensure both the credibility and effectiveness of patient care in an increasingly digital world.

While a large body of research demonstrates the benefits of primary care, there is comparatively less evidence for the use of AI technologies in health care. This discrepancy has substantial implications for evidence-based medicine, where clinical decisions should be based on solid scientific research [32].

A lack of uniform evaluation protocols introduces several challenges, such as doubts concerning the long-term effectiveness, safety of patients, security of data, and ethical dilemmas. To integrate telemedicine and AI into health care systems responsibly and effectively, we need flexible and comprehensive evaluation methods that conform to clinical objectives and regulatory guidelines [33].

The literature highlights that the factors influencing health are complex and influenced by a variety of factors [34,35]. Future studies should continue to explore how aspects beyond technology influence health to ensure that the design of new health systems considers a balanced integration of digital and human resources.

Future Role of AI in Medical Care

While concerns remain about the integration of AI in health care, a significant proportion of participants believed that AI-based early detection measures in general practice could enhance patient care. The application of patient-specific medical data combined with AI assistance systems has the potential to improve treatment outcomes by facilitating prevention, early diagnosis, and individualized therapies. Such advancements could lead to the identification of new medical correlations and innovative approaches to prevention [36].

However, despite these potential benefits, significant challenges and concerns remain. The use of AI necessitates access to substantial amounts of sensitive medical data, raising concerns about data privacy and security. Robust security protocols are essential to safeguard patient privacy and prevent data misuse [37].

The development and integration of AI in health care also bring significant ethical considerations, particularly around accountability for poor decisions or unexpected outcomes. Establishing clear guidelines and defining responsibilities are essential to address these ethical concerns and build trust with patients [38].

In this context, the European Union's "AI Act" represents a critical step toward creating uniform regulations for AI use within the European Union to ensure safety and trustworthiness. This regulation uses a risk-based approach, categorizing AI applications into different risk levels, with high-risk applications, such as those often found in the medical field, subjected to strict requirements. These requirements include transparency, security, and the avoidance of discrimination, all aimed at ensuring that AI systems in the health care sector adhere to high-quality standards and guarantee patient safety [39].

Strengths and Limitations

With a response rate of approximately 36%, the study provides a robust insight into the attitudes and opinions of the population of Schleswig-Holstein on the use of AI in GP practices.

The study was limited to Schleswig-Holstein, which could limit the generalizability of the results to the whole of Germany. Furthermore, the sample could be biased by self-selection of participants who may have had a greater interest in or familiarity with digital technologies. Future studies should therefore aim for a broader geographical spread to improve the representativeness and applicability of the results to the entire German population.

Conclusions

These findings suggest that despite the growing importance of AI in some areas of health care, the traditional physician-patient relationship is seen as absolutely essential. This emphasizes the enduring value of the human elements in health care—personal interaction, empathy, and understanding—features that AI cannot replicate.

Nevertheless, the introduction of AI is still recognized as a sign of modern medicine. People in urban areas are somewhat more open to AI technologies in health care. The integration of AI into health care raises crucial considerations regarding the principles of evidence-based medicine. Balancing technological innovation with the fundamentals of patient-centered care will be crucial in navigating the future landscape of health care.

Funding

This work was supported by grants from the Joachim Herz Foundation as part of the PASBADIA (smartphone-based point-of-care diagnostics) Project.

XSL•FO
RenderX

Authors' Contributions

LW and JS initiated and designed the study. LW and SF coordinated the study. KG carried out data analysis. LW wrote the first draft of the manuscript. All authors (LW, KG, SF, and JS) read earlier versions of the manuscript, provided critical comments, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Evaluation of an artificial intelligence–based screening measure in general practice.
[DOCX File, 16 KB - medinform_v14i1e69777_app1.docx]

References

1. Fachdefinition allgemeinmedizin [Web page in German]. Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin. 2024. URL: https://www.degam.de/fachdefinition [accessed 2025-05-25]
2. Prävention: Früherkennungsuntersuchungen und Impfungen [Web page in German]. Kassenärztliche Bundesvereinigung. 2024. URL: https://www.kbv.de/praxis/patientenversorgung/praevention [accessed 2025-05-25]
3. Hörle S, Grüner F, Kroll P. Epidemiologie diabetischer Erblindungen - eine Übersicht [Article in German]. Klin Monatsbl Augenheilkd 2002 Nov;219(11):777-784. [doi: 10.1055/s-2002-36318]
4. Cavan D, Makaroff L, da Rocha Fernandes J, et al. The Diabetic Retinopathy Barometer Study: global perspectives on access to and experiences of diabetic retinopathy screening and treatment. Diabetes Res Clin Pract 2017 Jul;129:16-24. [doi: 10.1016/j.diabres.2017.03.023] [Medline: 28499163]
5. Diabetic retinopathy barometer report. International Diabetes Federation Europe. 2017. URL: https://drbarometer.com/wp-content/uploads/2022/08/drbarometer_global_report-1.pdf [accessed 2025-05-25]
6. Krogh M, Germund Nielsen M, Byskov Petersen G, et al. Patient acceptance of AI-assisted diabetic retinopathy screening in primary care: findings from a questionnaire-based feasibility study. Front Med 2025 Sep 10;12:1610114. [doi: 10.3389/fmed.2025.1610114]
7. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1(1):39. [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
8. Ting DS, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017 Dec 12;318(22):2211-2223. [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
9. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Erratum: author correction: deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. NPJ Digit Med 2019;2:68. [doi: 10.1038/s41746-019-0146-5] [Medline: 31341955]
10. von Solodkoff M, Strumann C, Steinhäuser J. Akzeptanz von Versorgungsangeboten zur ausschließlichen Fernbehandlung am Beispiel des telemedizinischen Modellprojekts „docdirekt“: ein mixed-methods design [Article in German]. Gesundheitswesen 2021 Mar;83(3):186-194. [doi: 10.1055/a-1173-9903]
11. Strumann C, Steinhäuser J. Evaluation of a direct-to-patient telehealth service in Germany (Docdirekt) based on routine data. Dtsch Arztebl Int 2021 Dec 17;118(50):872-873. [doi: 10.3238/arztebl.m2021.0378] [Medline: 35300786]
12. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg 2014 Dec;12(12):1500-1524. [doi: 10.1016/j.ijsu.2014.07.014] [Medline: 25046751]
13. Held LA, Wewetzer L, Steinhäuser J. Determinants of the implementation of an artificial intelligence-supported device for the screening of diabetic retinopathy in primary care - a qualitative study. Health Informatics J 2022;28(3):14604582221112816. [doi: 10.1177/14604582221112816] [Medline: 35921547]
14. Edwards P, Roberts I, Clarke M, et al. Increasing response rates to postal questionnaires: systematic review. BMJ 2002 May 18;324(7347):1183. [doi: 10.1136/bmj.324.7347.1183] [Medline: 12016181]
15. Field A. Discovering Statistics Using IBM SPSS Statistics: Sage Publications; 2011.
16. Hartley D. Rural health disparities, population health, and rural culture. Am J Public Health 2004 Oct;94(10):1675-1678. [doi: 10.2105/ajph.94.10.1675] [Medline: 15451729]
17. Loftus J, Allen EM, Call KT, Everson-Rose SA. Rural-urban differences in access to preventive health care among publicly insured Minnesotans. J Rural Health 2018 Feb;34 Suppl 1(Suppl 1):s48-s55. [doi: 10.1111/jrh.12235] [Medline: 28295584]
18. Haggstrom DA, Lee JL, Dickinson SL, et al. Rural and urban differences in the adoption of new health information and medical technologies. J Rural Health 2019 Mar;35(2):144-154. [doi: 10.1111/jrh.12358] [Medline: 30830983]

XSL•FO
RenderX

19. LeBlanc M, Petrie S, Paskaran S, Carson DB, Peters PA. Patient and provider perspectives on eHealth interventions in Canada and Australia: a scoping review. Rural Remote Health 2020 Sep;20(3):5754. [doi: 10.22605/RRH5754] [Medline: 32949485]
20. DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. Lancet Digit Health 2019 Dec;1(8):e390. [doi: 10.1016/S2589-7500(19)30197-9] [Medline: 33323220]
21. Tarrant C, Stokes T, Baker R. Factors associated with patients' trust in their general practitioner: a cross-sectional survey. Br J Gen Pract 2003 Oct;53(495):798-800. [Medline: 14601357]
22. Baker R, Freeman GK, Haggerty JL, Bankart MJ, Nockels KH. Primary medical care continuity and patient mortality: a systematic review. Br J Gen Pract 2020 Sep;70(698):e600-e611. [doi: 10.3399/bjgp20X712289] [Medline: 32784220]
23. Lankton N, McKnight DH, Tripp J, Marshall University, Michigan State University, Baylor University. Technology, humanness, and trust: rethinking trust in technology. JAIS 2015;16(10):880-918. [doi: 10.17705/1jais.00411]
24. Hafizoğlu FM, Sen S. Understanding the influences of past experience on trust in human-agent teamwork. ACM Trans Internet Technol 2019 Nov 30;19(4):1-22. [doi: 10.1145/3324300]
25. Wewetzer L, Held LA, Goetz K, Steinhäuser J. Determinants of the implementation of artificial intelligence-based screening for diabetic retinopathy-a cross-sectional study with general practitioners in Germany. Digit Health 2023;9:20552076231176644. [doi: 10.1177/20552076231176644] [Medline: 37274367]
26. Li X, Hess TJ, Valacich JS. Why do we trust new technology? A study of initial trust formation with organizational information systems. The Journal of Strategic Information Systems 2008 Mar;17(1):39-71. [doi: 10.1016/j.jsis.2008.01.001]
27. Herzog C. Technological opacity of machine learning in healthcare. Presented at: 2nd Weizenbaum Conference: Challenges of Digital Inequality; May 16-17, 2019 URL: https://www.ssoar.info/ssoar/bitstream/handle/document/62580/ssoar-2019-herzog-Technological_Opacity_of_Machine_Learning.pdf;jsessionid=2A48C886E963221F3FD376851A02F14A?sequence=1 [accessed 2025-05-25]
28. Beck S, Faber M, Gerndt S. Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege [Article in German]. Ethik Med 2023 Jun;35(2):247-263. [doi: 10.1007/s00481-023-00763-9]
29. Aminololama-Shakeri S, López JE. The doctor-patient relationship with artificial intelligence. AJR Am J Roentgenol 2019 Feb;212(2):308-310. [doi: 10.2214/AJR.18.20509] [Medline: 30540210]
30. Printz C. Artificial intelligence platform for oncology could assist in treatment decisions. Cancer 2017 May 15;123(6):905-905. [doi: 10.1002/cncr.30655] [Medline: 28263386]
31. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak 2023 Apr 20;23(1):73. [doi: 10.1186/s12911-023-02162-y] [Medline: 37081503]
32. Evidence summary: the benefits of relational continuity. Alberta Medical Association. 2017. URL: https://www.albertadoctors.org/resource-centre/relational-continuity-resources/evidence-summary-the-benefits-of-relational-continuity/ [accessed 2025-05-25]
33. Burrell DN. Dynamic evaluation approaches to telehealth technologies and artificial intelligence (AI) telemedicine applications in healthcare and biotechnology organizations. Merits 2023;3(4):700-721. [doi: 10.3390/merits3040042]
34. Stronks K, Kunst AE. The complex interrelationship between ethnic and socio-economic inequalities in health. J Public Health (Oxf) 2009 Sep;31(3):324-325. [doi: 10.1093/pubmed/fdp070] [Medline: 19589801]
35. Deutekom M, Vansenne F, McCaffery K, Essink-Bot ML, Stronks K, Bossuyt PMM. The effects of screening on health behaviour: a summary of the results of randomized controlled trials. J Public Health (Oxf) 2011 Mar;33(1):71-79. [doi: 10.1093/pubmed/fdq050] [Medline: 20667898]
36. Budde K, Dasch T, Kirchner E, et al. Künstliche Intelligenz: Patienten im Fokus [Article in German]. Dtsch Arztebl 2020;117(49):A-2407 [FREE Full text]
37. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
38. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019 Feb 22;363(6429):810-812. [doi: 10.1126/science.aaw0029] [Medline: 30792287]
39. On artificial intelligence - a European approach to excellence and trust. European Commission. 2020. URL: https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf [accessed 2025-05-25]

Abbreviations

AI: artificial intelligence
DR: diabetic retinopathy
GP: general practitioner
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factor

XSL•FO
RenderX

Edited by A Benis; submitted 08.Dec.2024; peer-reviewed by A Wahbeh, L Buhr, M Bleckwenn; revised version received 15.Dec.2025; accepted 17.Dec.2025; published 12.Feb.2026.

Please cite as:
Wewetzer L, Goetz K, Freischmidt S, Steinhauser J
Trust in AI-Supported Screening in General Practice Among Urban and Rural Citizens: Cross-Sectional Study
JMIR Med Inform 2026;14:e69777
URL: https://medinform.jmir.org/2026/1/e69777
doi:10.2196/69777

© Larisa Wewetzer, Katja Goetz, Soenke Freischmidt, Jost Steinhauser. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Stakeholder Perspectives on Affinity Domains in Digital Health Interoperability: Qualitative Study

Jan Bruthans[1,2], Prof Dr; Petra Hospodková[1], MBA, PhD; Ilona Fejfarová[1], ING

[1]Department of Biomedical Technology, Faculty of Biomedical Engineering, Czech Technical University in Prague, Sítná Squere 3105, Kladno, Czech Republic

[2]Faculty of Health Studies, Technical University of Liberec, Liberec, Czech Republic

Corresponding Author:
Petra Hospodková, MBA, PhD
Department of Biomedical Technology, Faculty of Biomedical Engineering, Czech Technical University in Prague, Sítná Squere 3105, Kladno, Czech Republic

Abstract

Background: Operational health data interoperability in post-transition health systems requires not only technical standards but also enforceable governance arrangements. Affinity domains, as defined in the IHE XDS (Integrating the Healthcare Enterprise Cross-Enterprise Document Sharing) framework, represent a structured organizational-technical model for cross-enterprise document sharing. However, evidence from Central and Eastern Europe on their governance feasibility and implementation readiness remains limited, particularly in systems characterized by institutional fragmentation and evolving regulatory mandates.

Objective: This study aimed to examine stakeholder perceptions of the prerequisites, risks, benefits, and governance conditions for implementing affinity domains in the Czech health care system. The study further sought to identify system-level readiness factors relevant to national interoperability initiatives in the context of the European Health Data Space regulation.

Methods: We conducted 18 semistructured interviews between January 2025 and April 2025 with policymakers, regional health authorities, health care providers, health insurance funds, health IT vendors, and independent experts. Participants were selected using purposive and snowball sampling to ensure institutional diversity. Interviews explored governance roles, legal accountability for shared data, enforcement of standards, financing models, and technical readiness. Data were analyzed inductively using thematic analysis in MAXQDA 24 (VERBI Software GmbH), supported by dual coding, consensus discussions, and reflexive memoing. Inductively derived themes were subsequently interpreted against core governance dimensions of affinity domains (rule-making authority, membership and participation, accountability, enforcement, and trust).

Results: Five major thematic categories emerged: (1) ambiguous roles and responsibilities marked by fragmented mandates and vendor influence; (2) perceived risks, including institutional distrust, legal uncertainty regarding liability, technical fragmentation, and vendor lock-in; (3) system-level prerequisites such as the need for an empowered coordinating authority, binding interoperability standards (IHE XDS and HL7 FHIR [Health Level 7 Fast Healthcare Interoperability Resources]), sustainable financing, and human resource capacity; (4) perceived benefits, including improved continuity of care, reduced duplication, and enhanced transparency; and (5) structural implementation barriers including political discontinuity, weak enforcement mechanisms, and limited stakeholder engagement. Across stakeholder groups, feasibility was consistently linked to 3 actionable readiness conditions: clearly assigned responsibility and liability for shared data, enforceable technical requirements supported by certification or audit mechanisms rather than voluntary compliance, and financing models addressing both IT infrastructure and organizational change.

Conclusions: Affinity domains are recognized as a viable interoperability model in the Czech Republic, but their implementation is constrained primarily by governance deficits rather than technical immaturity. Establishing an empowered coordinating authority, piloting enforceable domain structures, and aligning national governance with European Health Data Space requirements represent necessary steps toward operational interoperability. These findings provide empirically grounded insights into the governance conditions required for scalable health information exchange in post-transition health systems.

(JMIR Med Inform 2026;14:e83894) doi:10.2196/83894

KEYWORDS

electronic health records; interoperability; health information exchange; qualitative research; stakeholder participation; European Health Data Space; Integrating the Healthcare Enterprise Cross-Enterprise Document Sharing; IHE XDS; Czech Republic

XSL•FO
RenderX

Introduction

The digital transformation of health care systems is widely recognized as a strategic priority across Europe, aiming to improve care quality, efficiency, and patient safety while enabling sustainable system governance [1]. A key enabler of this transformation is the establishment of interoperability infrastructure that allows secure and standardized exchange of medical information among health care providers, patients, and administrative entities [2]. Within this context, the concept of affinity domains, as defined within the IHE XDS (Integrating the Healthcare Enterprise Cross-Enterprise Document Sharing) architecture, constitutes a sociotechnical construct that combines a technical framework for cross-enterprise document sharing with an explicit governance arrangement. Beyond interoperability per se, affinity domains are characterized by shared rule-making, defined membership, accountability for data stewardship, and enforcement mechanisms governing participation and compliance among involved organizations [3,4]. While health information exchange (HIE) or electronic health record (EHR) interoperability primarily refers to the technical ability of systems to exchange data, an affinity domain additionally defines shared governance rules, membership criteria, accountability, and enforcement mechanisms for participating organizations. In this paper, we use the term "interoperability infrastructure" to refer broadly to the technical and governance ecosystem enabling cross-enterprise data exchange, "HIE" to denote operational exchange systems, and "affinity domains" to describe a governance-anchored interoperability model as conceptualized in the IHE XDS framework.

While the implementation of affinity domains has been successfully institutionalized in several Western and Northern European countries—such as Austria (Elektronische Gesundheitsakte [ELGA]), Switzerland (Elektronisches Patientendossier [EPD]), the Netherlands (Landelijk Schakelpunt), and Sweden (Národní plán obnovy)—their adoption across Central and Eastern Europe remains limited and fragmented. Although conceptual frameworks and technical standards such as IHE XDS have been available for over a decade [4], countries in this region continue to face systemic, organizational, and legislative barriers to real-world deployment [5]. The Czech Republic, despite recent improvements in its eHealth strategy and digital infrastructure, continues to lack the operational governance and institutional capacity required to implement affinity domains at scale. In practical terms, the Czech Republic currently lacks a national HIE system comparable to ELGA or EPD, and interoperability initiatives remain primarily project-based or regionally bounded. Although Act No. 325/2021 Collection of Laws established a legal framework for digital health governance, implementation mechanisms, certification procedures, and clearly mandated coordinating authorities remain only partially operational. As a result, interoperability is shaped more by institutional negotiation and vendor configurations than by enforceable national governance structures.

Since the early 2000s, the Czech health care system has initiated multiple digitalization efforts—most prominently the Internet Health Information Portal (Internetový zdravotní informační portál [IZIP]) and ZDRAVeL platforms—but these have largely failed to achieve widespread adoption due to fragmented governance, insufficient standardization, institutional resistance, and lack of sustained stakeholder engagement. The IZIP system, launched in 2004, was a pioneering national web-based EHR that demonstrated clear potential but achieved limited integration and impact in routine care [6]. Its successor, ZDRAVeL, introduced improved functionality but similarly failed to secure institutional alignment and embed itself into clinical workflows [7]. At present, the only digital health applications with nationwide coverage and consistent daily use are the Outpatient Electronic Prescription System [8] and the Digital Imaging and Communications in Medicine (DICOM) image exchange platform [9]. Other systems exist but remain fragmented, region-specific, or limited to pilot projects, which restrict large-scale interoperability. A notable example of a successful local secure health data exchange platform is the eMeDOcS (Exchange Medical Documents System) project [10]. The eMeDOcS system is an electronic platform designed to enable structured data exchange across affinity domains, supporting interoperability and standardized clinical communication. It is mainly used by hospitals in the Vysočina Region. The eMeDOcS system is based on the ICZ ISAC (Integration, Share and Communication System) platform [11], which enables easy integration between EHRs and other specific systems, including telemedicine solutions [12].

More recent evaluations note that although the COVID-19 pandemic accelerated telemedicine adoption, this progress remains fragmented and contingent, hampered by legal uncertainty and the absence of a functional interoperability infrastructure [13]. Comparable challenges have been documented in other post-transition systems, where weak enforcement mechanisms, unclear institutional mandates, and limited technical standardization obstruct effective interoperability [14,15]

Despite these conceptual and technical advances, affinity domains have not progressed beyond policy discussions in the Czech Republic, and no operational implementation currently exists. Several domain experts note that while the IHE-affinity domain concept is acknowledged and debated at the policy level in the Czech Republic, actual system-level implementations are deferred due to government ambivalence about central steering versus federated emergence, and the absence of pilot registries or unified patient identity mechanisms. Similarly, an interoperability pilot (Semantic Interoperability Platform) launched in the mid-2000s demonstrated the technical readiness to adopt international standards but failed to scale due to institutional fragmentation and a lack of sustained governance frameworks [16].

As of 2024, electronic medical record systems are widely adopted at the institutional level in the Czech Republic, with nearly all hospitals and most outpatient providers operating some form of digital documentation. However, these systems remain predominantly vendor-specific and lack cross-enterprise interoperability. No nationwide HIE system comparable to ELGA or EPD is operational. Regional exchange initiatives, such as eMeDOcS in the Vysočina Region, demonstrate

XSL•FO
RenderX

technical feasibility using IHE XDS–based architectures, but their implementation remains geographically limited and institutionally fragmented. National digital infrastructure is currently centered around selected vertical services (ePrescription and DICOM image exchange), while unified patient identity management, master patient indexing, and certification mechanisms for interoperability remain underdeveloped. Although the legal framework for digital health governance was established by Act No. 325/2021 Collection of Laws, operational governance structures and mandatory interoperability enforcement are still evolving.

Despite the availability of technical standards and formal policy documents, existing literature [17] provides limited insight into how the governance attributes of affinity domains, such as rule-making authority, accountability, participation, enforcement, and trust, are interpreted and operationalized by institutional actors in practice. While affinity domains are technically defined within the IHE XDS framework, these governance dimensions are only partially reflected in legal texts or technical specifications and tend to emerge through organizational interactions, negotiation processes, and implementation experiences. Existing research has primarily focused on legal or technical aspects rather than governance practices. To address this gap, we adopted a qualitative design based on stakeholder interviews, which enables analysis of how abstract interoperability models translate into real-world decision-making contexts and how governance responsibilities are perceived and enacted by key system actors [18].

Therefore, in this study, we aim to contribute to the evolving discourse on digital health governance in post-transition health care systems by (1) presenting key findings from stakeholder perceptions regarding the implementation of affinity domains in Czech health care, (2) analyzing these findings in comparison with selected European experiences, and (3) offering evidence-informed recommendations for advancing interoperability through affinity domains as part of the national eHealth policy.

Methods

Study Design and Objective

This study used a qualitative, exploratory design using semistructured interviews and stakeholder analysis to examine perceptions, expectations, and barriers related to the implementation of affinity domains in the Czech health care system. The qualitative approach was selected to capture governance-related attributes of affinity domains—such as the allocation of roles, accountability for shared data, enforcement mechanisms, and trust relationships—that are central to their feasibility but are not directly observable through policy documents or technical specifications. The research objective was to identify enabling and constraining factors from the perspective of key stakeholder groups and derive actionable insights to inform policy and strategy. At the start of each interview, participants were provided with a standardized, brief explanation of an “affinity domain” based on IHE XDS, emphasizing both its governance components (shared policies, membership, accountability, and enforcement) and its technical scope (cross-enterprise document sharing under agreed profiles). Participants were asked to reflect on prerequisites, risks, benefits, and barriers specifically in relation to this defined model. Where participants referred to interoperability challenges more broadly, such statements were retained in the analysis only insofar as they directly constrained or enabled the feasibility of implementing affinity domains in practice. The study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) 32-item checklist to ensure methodological transparency (Checklist 1) [19].

Research Team and Reflexivity

Interviews were conducted by PH and IF, both of whom have formal training in qualitative health research and prior experience conducting semistructured interviews in digital health policy projects. JB co-led the analytic phase. A brief relationship with some participants existed through professional networks used for initial identification; however, no direct supervisory or contractual dependency was present. Participants were informed about the researchers’ institutional affiliation, study aims, and interest in interoperability governance. To mitigate potential bias, we used an a priori protocol, a piloted interview guide, maintained an audit trail, and implemented dual independent coding with consensus resolution; reflexive memos documented assumptions and analytic decisions throughout. We did not compute formal intercoder reliability coefficients, opting instead for iterative consensus and peer debriefing.

Stakeholder Framework and Selection

The selection of stakeholder groups was based on a combined analytical and policy-relevant approach, aligned with established digital health governance frameworks, including multilevel governance approaches and actor-based interoperability models [17,18]. Stakeholders were defined as individuals or institutions that directly influence or are affected by the implementation of national eHealth infrastructure, specifically interoperability components such as affinity domains. Stakeholder groups were selected to reflect core governance roles within an affinity domain, including regulatory authority, payers, providers, vendors, and independent technical experts. These roles correspond to the key governance dimensions of affinity domains, namely rule-setting authority (Ministry of Health [MoH]), implementation coordination (regional health authority [RHA]), operational participation (health care providers), financing and regulatory oversight (health insurance fund [HIF]), technical standard implementation (health IT vendors), and advisory or normative alignment (independent consultants and expert advisors). Inclusion criteria were (1) active involvement in drafting or implementing national or regional eHealth strategies, (2) responsibility for management or governance of hospital information system or interoperability solutions, or (3) technical or advisory roles in pilots or standard-setting activities. Exclusion criteria included the absence of formal responsibility for digital health governance or purely end-user roles without decision authority. Stakeholders were identified through a purposive sampling strategy supplemented by snowball sampling [20]. First, 6 stakeholder categories were defined a priori based on their formal roles in national and regional digital health governance. Within these predefined categories, individuals

XSL•FO
RenderX

were primarily identified through purposive selection based on their formal responsibilities, documented involvement in national or regional eHealth initiatives, and participation in legislative or technical working groups. Snowball sampling was subsequently used as a supplementary strategy to identify additional relevant informants, particularly within the independent expert group and in cases where formal role-based identification was insufficient. The recruitment process continued until adequate coverage of all predefined categories was achieved. This resulted in a final pool of 25 identified candidates, of whom 22 were formally invited; 4 declined due to time constraints, resulting in 18 completed interviews.

To ensure both diversity of perspectives and relevance to the Czech health care context, the following 6 stakeholder groups were identified.

Table 1 shows the stakeholder groups and corresponding inclusion criteria applied in a qualitative semistructured interview study examining the implementation of IHE XDS–based affinity domains in the Czech health care system. Eighteen interviews were conducted between January 2025 and April 2025. Participants were recruited using purposive and snowball sampling to represent regulatory, financing, service provision, health IT vendor, and advisory roles. Detailed participant characteristics are provided in Multimedia Appendix 1.

Table . Stakeholder selection criteria (N=18).

Stakeholder group	Role in eHealth system	Inclusion criteria	Participants, n (%)
Ministry of Health[a], Czech Republic	Policy development, legislation, and national coordination	Direct involvement in drafting or implementing Act No. 325/2021 Collection of Laws or national eHealth strategies	3 (16.6)
RHAs[h]	Regional eHealth implementation and funding oversight	Responsibility for managing or co-funding regional eHealth projects, or an active role in local Hospital Information System governance	3 (16.6)
Health care providers	Operational users of interoperability infrastructure	Senior managers or digital transformation leads in hospitals participating in national eHealth projects or data-sharing pilots	4 (22)
HIF[c]	Data consumers, payers, and system regulators	Participation in interoperability-related initiatives; responsibility for data security or claims data infrastructure	2 (11)
Health IT vendors	System developers and integrators of interoperability solutions	Experience with IHE[d] XDS[e] or HL7[f] FHIR[g] implementations; vendors of Hospital Information Systems used in the Czech or European Union interoperability frameworks	4 (22)
Independent consultants and expert advisors	Strategic or technical advising, legislation support	Recent collaboration with MoH or regions on legislative drafting, standard setting, or coordination of digital health pilot programs	2 (11)

[a]MoH: Ministry of Health.

[b]RHA: regional health authority.

[c]HIF: health insurance fund.

[d]IHE: Integrating the Healthcare Enterprise.

[e]XDS: Cross-Enterprise Document Sharing.

[f]HL7: Health Level 7.

[g]FHIR: Fast Healthcare Interoperability Resources (HL7).

The purposive strategy targeted participants with substantial experience or formal responsibility in health information systems, eHealth projects, or regulatory implementation. Recruitment prioritized individuals actively engaged in national-level working groups, projects funded through the Integrated Regional Operational Program or National Recovery Plan (Národní plán obnovy) schemes, and the development of Czech national digital health legislation. This ensured that all selected informants had first-hand knowledge of operational, regulatory, or strategic aspects of digital health implementation.

As noted above, snowball sampling complemented the primary purposive strategy by enabling the inclusion of additional consultants and advisors involved in recent pilot initiatives or standard-setting activities. Recruitment followed a structured, multiphase logic: initial identification through professional networks, follow-up invitations by email, and confirmation of eligibility against the inclusion criteria.

XSL•FO
RenderX

The final sample reflected both policymaking and operational perspectives, as well as representation from public and private sector stakeholders. While not statistically representative, the composition was sufficient to achieve thematic saturation across the predefined analytical categories, thereby increasing the credibility and contextual depth of the findings. Group-level distributions (MoH, RHAs, health care providers, HIFs, health IT vendors, and independent consultants and expert advisors) are detailed in Checklist 1 to support COREQ-aligned reporting.

Data Collection

Data were collected through semistructured interviews, providing a flexible yet systematic approach to exploring stakeholder perspectives. Interviews were conducted between January 2025 and April 2025. An interview guide was developed based on a review of relevant literature and pilot-tested with 2 external experts, which led to minor refinements in question wording and flow. Interviews were scheduled for approximately 50 minutes and were conducted online via secure video conferencing tools to accommodate participants' geographical distribution.

All interviews were audio-recorded with participants' consent, and detailed field notes were authored immediately after each session. No repeat interviews were conducted. Only the interviewer and participant were present during each interview; no third parties attended. Interview transcripts were returned to all participants for member checking. Of the 18 participants, 12 provided feedback or explicit confirmation, and minor clarifications were incorporated where relevant. The remaining participants did not request modifications. The full semistructured interview guide, including main questions and illustrative probes, is provided in Multimedia Appendix 2 to enhance transparency regarding question framing and potential prompting effects.

The interview guide focused on 4 main domains: stakeholders' perceptions of opportunities and barriers in digital health implementation, experiences with project governance and interinstitutional collaboration, anticipated impacts of national digital health strategies and legislation, and suggestions for improving effectiveness, efficiency, and trust in digital health projects.

Data saturation was monitored throughout the process and was judged to have been achieved after the 12th interview; subsequent interviews confirmed existing themes without generating substantial new insights. A saturation grid documenting the last occurrence of novel codes is included in Multimedia Appendix 3. A summary of preliminary themes was also shared with participants, who provided confirmation or minor clarifications.

Data Analysis

To support construct validity, the concept of affinity domains was analytically specified through a set of governance and technical dimensions derived from the IHE XDS framework and relevant digital health governance literature [3,4,17].

Table 2 shows the conceptual specification of governance and technical dimensions constitutive of affinity domains, which were used as an interpretive reference in this qualitative interview study conducted within the Czech health care system (18 interviews; January 2025-April 2025). The dimensions were derived from the IHE XDS framework and relevant digital health governance literature, serving as an analytical lens rather than a deductive coding framework.

Table . Conceptual specification of affinity domain dimensions used as an interpretive reference.

Affinity domain dimension	Conceptual basis	Relevance for analysis
Shared rule-making	IHE[a] XDS[b] governance model; digital health governance literature	Framing how rules and standards are defined and legitimized
Membership and participation	Affinity domain concept; institutional governance	Understanding inclusion or exclusion of actors
Accountability and liability	Governance and health law literature	Interpreting responsibility for shared data
Enforcement mechanisms	EHDS[c] and interoperability governance debates	Assessing compliance beyond voluntary adoption
Trust relationships	Trust and legitimacy literature in digital health	Explaining willingness to participate
Technical scope (IHE XDS)	IHE XDS architecture	Anchoring findings in document-based exchange

[a]IHE: Integrating the Healthcare Enterprise.

[b]XDS: Cross-Enterprise Document Sharing.

[c]EHDS: European Health Data Space.

These dimensions were not applied as a deductive coding scheme but served as an interpretive reference for examining how inductively derived themes relate to the core attributes of affinity domains. This approach preserved the inductive nature of the thematic analysis while ensuring conceptual alignment with the affinity domain construct.

Interview transcripts were subjected to a thematic analysis using MAXQDA 24 (VERBI Software GmbH, Berlin, Germany, version 24.1.0) to support the organization and coding of qualitative data. An initial inductive codebook was developed and iteratively refined; the final coding tree with major themes and subthemes is provided in Multimedia Appendix 4. Minor and divergent views were also reported, ensuring that less frequent but important perspectives were not overlooked. A worked example (quote → in-vivo code → category → theme) is provided in Multimedia Appendix 5, along with a table of representative quotes by theme and subtheme. The higher-order thematic categories were subsequently interpreted in relation to the governance dimensions constitutive of affinity domains.

XSL•FO
RenderX

The analysis followed a multistep, reflexive process consistent with best practices in qualitative health research:

- Familiarization: researchers conducted repeated readings of transcripts and field notes to identify initial patterns.
- Coding: inductive, in vivo codes were applied independently by 2 coders (PH and IF); discrepancies were discussed until consensus was reached, with an audit trail documenting decisions. We did not compute a formal intercoder reliability coefficient; instead, we used iterative consensus meetings and peer debriefing to ensure dependability.
- Theme development: codes were clustered into broader thematic categories capturing recurrent stakeholder perspectives. Some empirical phenomena spanned multiple domains and were analytically differentiated according to their primary governance implications. To enhance readability in the "Results" section, conceptually related phenomena are discussed under their primary governance function, with cross-references where appropriate. For instance, vendor dominance was interpreted under "roles and responsibilities" as an issue of rule-making authority, whereas vendor lock-in was classified under "perceived risks" as a structural constraint affecting enforceability and participation. This approach ensured conceptual clarity while maintaining alignment with governance attributes central to affinity domains.
- Verification: a third qualitative expert reviewed the coding scheme and thematic structure to ensure rigor and validity. Conceptually related codes (eg, vendor dominance and vendor lock-in, and legal uncertainty) were critically examined during iterative refinement. Where retained in separate higher-order categories, distinctions were based on their primary governance function (eg, structural power asymmetry vs enforceability constraints), thereby avoiding post hoc theme allocation. The use of MAXQDA facilitated systematic data management and the creation of thematic maps. The relative prominence of each domain was visualized through frequency-weighted displays, supporting the interpretation and presentation of the findings.

Quotations in the "Results" section are labeled using a dual system: a unique respondent code (R1-R18) and the corresponding stakeholder group (MoH, RHAs, health care providers, HIFs, health IT vendors, and independent consultants and expert advisors). Each code was assigned to a single participant and used consistently throughout the analysis. The distribution of respondent codes mirrors the sample composition described in Table 1: MoH (n=3), RHAs (n=3), health care providers (n=4), HIFs (n=2), health IT vendors (n=4), and independent consultants and expert advisors (n=2). This labeling strategy enhances the transparency of reporting while preserving participant anonymity. Where appropriate, we report countervailing or minority views to avoid over-generalization.

To contextualize the Czech findings, we conducted a focused desk-based review of selected European countries with operational HIE systems comparable to IHE XDS–based affinity domains (Austria, Switzerland, the Netherlands, Sweden, and Belgium). Countries were selected purposefully to reflect different governance architectures (centralized, federated, and regional) and based on the availability of official documentation and peer-reviewed analyses.

Data were extracted from national eHealth portals, legislative documents, and relevant academic literature. Information was synthesized descriptively, focusing on governance structures, legal mandates, implementation scopes, and coordination mechanisms. This component served as a contextual comparison and did not constitute a separate systematic review.

Ethical Considerations

The study was approved by the ethics committee of the Faculty of Biomedical Engineering under protocol C59/2025. Research procedures were conducted in compliance with local legislation and institutional requirements [21]. The ethical approval covered the recruitment of participants from external public and private institutions in their professional capacity. No additional institutional approvals were required beyond individual informed consent.

All participants received an information sheet explaining the study's purpose, procedures, and their rights, including confidentiality and the voluntary nature of participation. Written informed consent was obtained before data collection. Participant identities were anonymized and institutional affiliations were masked to protect confidentiality and comply with data protection standards. Audio files and transcripts were stored on secure institutional servers with restricted access, and deidentified transcripts were used for analysis. There was no compensation awarded.

Results

Overview

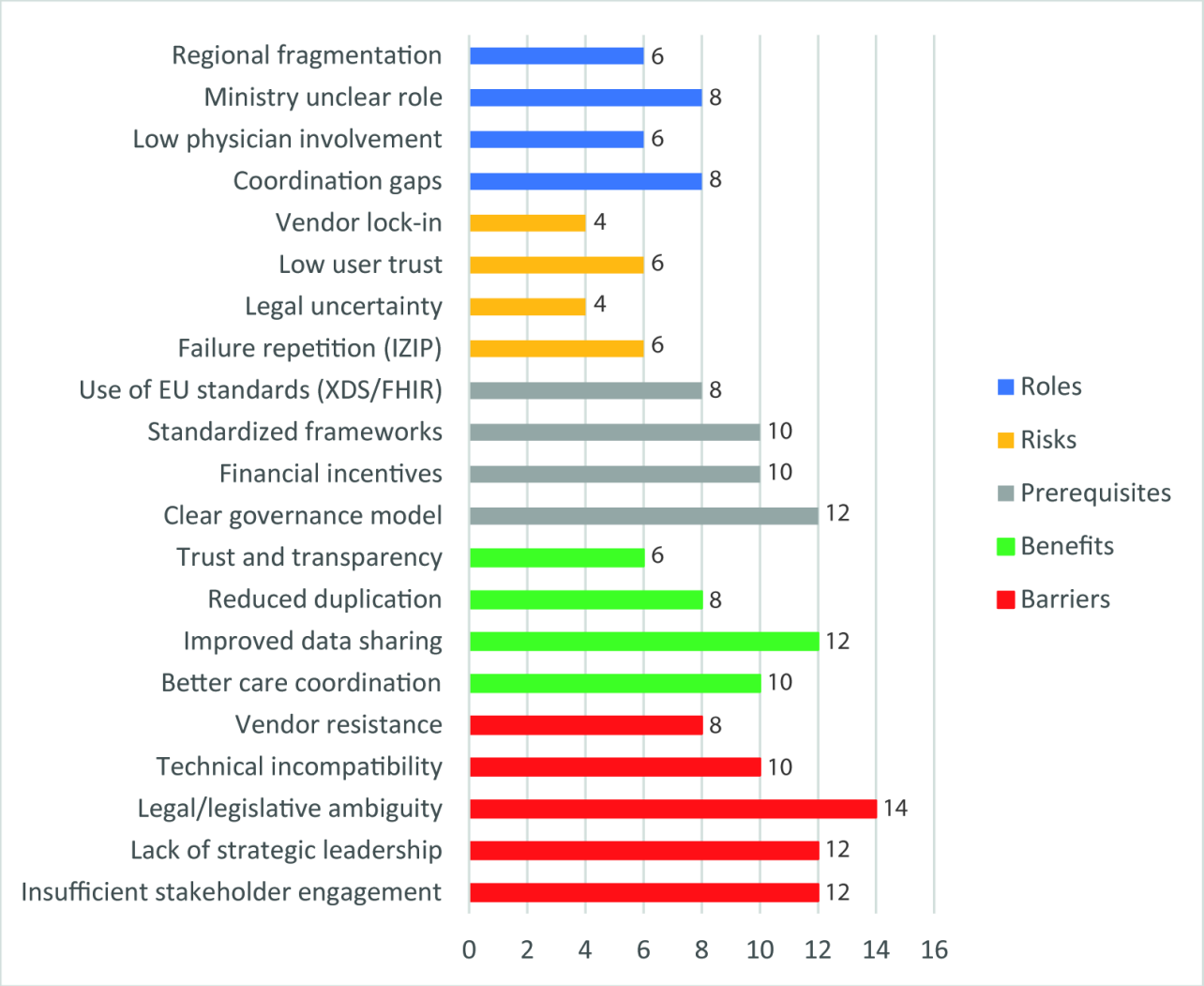
The thematic analysis yielded five major categories concerning the implementation of affinity domains in the Czech health care system: (1) roles and responsibilities, (2) perceived risks, (3) system-level prerequisites, (4) perceived benefits and opportunities, and (5) implementation barriers. These categories capture institutional expectations, governance concerns, and perceived feasibility conditions articulated across stakeholder groups.

The categories are presented thematically according to the dominant governance implications identified across stakeholder groups. Where phenomena intersected multiple governance dimensions (eg, vendor influence or legal uncertainty), they are discussed under their primary governance function.

Figure 1 illustrates the distribution of major thematic categories derived from 18 semistructured stakeholder interviews conducted within the Czech health care system between January and April 2025. The bars represent the frequency of coded text segments assigned to each theme across the dataset. The values reflect relative thematic prominence and do not indicate the number of participants endorsing a theme or statistical significance.

XSL•FO
RenderX

Figure 1. Distribution of major thematic categories derived from stakeholder interviews (N=18 interviews). FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7; IZIP: Internetový zdravotní informační portál; MoH: Ministry of Health; XDS: Cross-Enterprise Document Sharing.



Thematic Categories

Roles and Responsibilities

The analysis of stakeholder interviews revealed a systemic ambiguity in the definition and operationalization of roles within the governance of affinity domains in the Czech health care system. While the MoH formally holds responsibility for strategic direction in digital health, respondents consistently emphasized its limited engagement in operational implementation. As 1 participant noted:

> *The Ministry sets strategies but does not enforce them. We're left guessing what they really want.* [R18, independent consultant and expert advisor]

Several participants highlighted the absence of an empowered intermediary body with the technical and legal authority to coordinate interoperability initiatives—a role that, in other countries, is fulfilled by institutions such as Gematik (Germany).

At the regional level, fragmentation emerged as a critical barrier. Regional IT departments reportedly operate with minimal alignment, resulting in uncoordinated and heterogeneous solutions. This decentralization exacerbates interoperability issues and complicates accountability. According to another respondent:

> *Each region develops its own system, with no national integration in sight. We're building isolated islands.* [R4, RHA]

The lack of nationally endorsed technical and procedural standards reinforces these inconsistencies.

Stakeholders also identified a pronounced disconnect between digital governance structures and clinical end-users. Physicians, despite being primary users of EHRs, are reportedly excluded from planning and decision-making processes, as 1 respondent explained:

> *We receive new systems and tools, but no one asks whether they match how we work. That's why so many doctors ignore them.* [R7, health care provider]

This exclusion leads to suboptimal system design and limited adoption.

Finally, vendor dominance was perceived as a significant factor undermining coherent governance. In many facilities, technology suppliers exert considerable influence, effectively establishing

XSL•FO
RenderX

de facto standards and resisting external integration. As 1 respondent put it:

> *In some hospitals, the vendor decides more than the IT department. They lock the system and then dictate what's possible.* [R13, health IT vendor]

The absence of regulatory mechanisms to counterbalance these dynamics contributes to role diffusion and undermines long-term governance sustainability. This asymmetry also underpins later concerns regarding vendor lock-in and resistance during implementation.

In governance terms, this ambiguity signals the absence of clearly defined rule-making authority, membership criteria, accountability allocation, and enforceable coordination mechanisms that are constitutive of affinity domains as structured interoperability arrangements.

Perce ived Risks

Interviewees articulated a range of interrelated risks that threaten the feasibility and sustainability of affinity domains. The most frequently cited concern was reputational damage and institutional distrust rooted in the failed IZIP initiative. As 1 respondent recalled:

> *IZIP left a scar. We've all seen what happens when things go wrong—and no one wants to be blamed again.* [R17, independent consultant and expert advisor]

Respondents noted that the legacy of IZIP continues to generate reluctance among stakeholders to participate in digital health projects, particularly those lacking transparent governance and legal protection.

Legal uncertainty regarding liability for shared data was another dominant theme. Many stakeholders expressed concern over the absence of clear accountability frameworks for data integrity and clinical liability once information is shared across domains. As 1 respondent emphasized:

> *If a patient is harmed due to outdated or missing data, who will take responsibility? No one wants to be exposed.* [R8, health care provider]

This regulatory vacuum exacerbates institutional risk aversion, particularly in environments where errors could directly affect patient safety.

Technical risks were linked to the heterogeneous landscape of health IT systems and the lack of mandatory, enforceable interoperability standards. Interviewees reported variability in the implementation of HL7 FHIR and IHE profiles, often contingent upon vendor willingness rather than policy mandates. As 1 respondent observed:

> *Everyone does their own thing. Some use FHIR; some don't. There's no technical backbone that guarantees compatibility.* [R14, health IT vendor]

The absence of a national HIE system or compliance framework further limits reliable data exchange.

Building on earlier concerns about vendor dominance in rule-setting processes, stakeholders described vendor lock-in as a structural risk limiting enforceable participation. One respondent remarked, "Why would a vendor enable data sharing if it means losing business to a competitor?" (R15, health IT vendor). Additionally, low trust among clinical professionals—linked to past failures and limited inclusion in governance—was reported as a persistent obstacle to stakeholder engagement. As 1 respondent stated:

> *It feels like another bureaucratic project we're supposed to accept without discussion. That erodes trust, not builds it.* [R9, HP]

These risks indicate fragile institutional preconditions for affinity domains, particularly with respect to trust, liability clarity, and enforceable participation rules.

System-Level Prerequisites

Participants outlined several prerequisites critical for the successful implementation of affinity domains. Chief among these was the establishment of a functionally independent governance structure capable of coordinating the technical, legal, and strategic aspects of interoperability deployment. Respondents advocated for a body with a defined mandate, cross-sectoral legitimacy, and operational capacity. As 1 respondent pointed out:

> *You can't just publish a strategy and hope that hospitals will fall in line. Someone has to own it, drive it, and manage it.* [R2, MoH]

Funding was highlighted as a key enabling factor, given the substantial costs of system upgrades, training, and organizational redesign. According to 1 respondent:

> *We're not against interoperability. But who's going to pay for all the technical and organizational changes it requires?* [R11, HIF]

Incentive structures, including European Union–linked funding or performance-based reimbursement, were proposed as potential levers to foster participation.

Technical prerequisites included the uniform adoption of recognized standards such as HL7 FHIR and IHE XDS, as well as the creation of centralized compliance frameworks, reference implementations, and certification schemes. Interviewees also emphasized the need for inclusive interoperability beyond hospitals, incorporating social services, emergency care, and public health actors. As 1 respondent highlighted:

> *You can't build a regional domain if the social care system is left out. It's half of the equation.* [R5, RHA]

Such integration would necessitate coordinated policymaking across ministries.

Human resource capacity was also noted as a critical constraint. Respondents cited a shortage of professionals with the requisite skills to implement and manage interoperability solutions at both the institutional and regional levels. As 1 respondent explained:

> *The biggest bottleneck isn't technology. It's people who understand the standards and can actually implement them.* [R18, independent consultant and expert advisor]

XSL•FO
RenderX

These prerequisites correspond to the foundational governance conditions of affinity domains, including the need for an empowered coordinating authority, formally recognized participation rules, technical compliance enforcement, and sustainable operational capacity.

Perceived Benefits and Opportunities

The perceived benefits of affinity domains were principally associated with improved clinical coordination, operational efficiency, and patient safety. The most prominent advantage cited by stakeholders was seamless access to structured patient data across institutional boundaries. Respondents described scenarios in which inadequate information sharing resulted in redundant diagnostics, delayed care, or clinical uncertainty. As 1 respondent described:

> *If a patient comes from another hospital, I usually start from scratch. Even allergy history is often missing.* [R10, health care provider]

Affinity domains were also perceived as a mechanism to reduce redundant diagnostic procedures, particularly in oncology and chronic care, where patients often traverse multiple providers. One respondent reported:

> *We see the same CT scans being done twice just because the systems don't talk. That's both costly and unethical.* [R8, health care provider]

In addition to financial savings, stakeholders emphasized the ethical imperative to avoid unnecessary patient burdens.

Improved continuity of care and support for integrated service delivery was also recognized. Interoperable access to discharge summaries, medication histories, and treatment plans was seen as essential for primary care physicians managing complex patients. As 1 respondent mentioned, "A GP who sees a patient after hospital discharge should not have to guess what happened during hospitalization" (R7, health care provider).

Some respondents identified the potential for interoperability to facilitate system-level accountability through audit trails and standardized documentation. According to 1 respondent:

> *If we can see who accessed what, when, and why, it becomes easier to ensure compliance and trace responsibility.* [R1, MoH]

Secondary uses of data, including research, surveillance, and planning, were identified as long-term opportunities, albeit contingent on the establishment of trust and robust legal and technical safeguards. As 1 respondent suggested:

> *Affinity domains could eventually feed anonymised data into registries or support AI models. But that's the next step.* [R17, ICEA]

These anticipated benefits reflect the governance potential of affinity domains to structure accountability, enable traceable exchanges, and foster coordinated cross-institutional participation.

Implementation Barriers

Stakeholders identified multiple barriers hindering the adoption of affinity domains, spanning legal, organizational, technical, and cultural domains. The lack of a comprehensive legal framework was cited as the most significant impediment. Beyond the liability-related uncertainty described under the "Perceived Risks" section, respondents also identified broader legal uncertainty—particularly regarding consent procedures and data stewardship—as an operational barrier. As 1 respondent stated:

> *We operate in legal gray zones. Everyone is afraid of making a mistake, so they'd rather do nothing.* [R6, RHA]

Political instability and weak leadership continuity were also seen as undermining implementation. Respondents described repeated strategic resets with governmental turnover, leading to policy fatigue and limited stakeholder confidence. One respondent observed:

> *There's no continuity. One government introduces a concept, the next one cancels it. That's how we fail by design.* [R18, independent consultant and expert advisor]

Technical fragmentation, particularly the coexistence of proprietary, noninteroperable systems, was noted across both hospitals and outpatient facilities. The absence of national technical infrastructure (eg, a HIE system) was seen as a structural constraint on scalability. As 1 respondent described:

> *Every vendor builds their own castle. And no one checks if the drawbridges align.* [R16, health IT vendor]

In operational terms, stakeholders also described vendor resistance as a practical barrier during implementation, including commercial disincentives such as excessive application programming interface fees and limited technical transparency. As 1 respondent explained:

> [If interoperability isn't part of the contract, don't expect vendors to care.] [R15, HV]

Finally, limited stakeholder engagement—especially among frontline providers—was identified as a barrier to implementation. The top-down nature of digital health policy development was seen as excluding key actors, undermining system ownership, and contributing to resistance at the operational level. According to 1 respondent:

> *Doctors, nurses, even IT departments—most of them hear about new projects for the first time at the rollout stage. [R9, health care provider]*

Taken together, these barriers reflect a structural misalignment between current institutional arrangements and the governance logic required for affinity domains, particularly in terms of enforceability, defined domain membership, leadership continuity, and coordinated policy execution.

Discussion

Summary of Main Findings

Our findings indicate that the feasibility of affinity domains in the Czech health care system is primarily shaped by governance-related constraints rather than by the absence of

XSL•FO
RenderX

technical standards. Across stakeholder groups, barriers clustered around unclear institutional mandates, weak enforcement mechanisms, fragmented implementation structures, and limited stakeholder inclusion—factors that collectively undermine trust and coordination required for cross-enterprise data sharing.

Synthesis of Themes and Policy Implications

To enhance the policy relevance and readability of the findings, we synthesized the 5 empirical themes into a structured overview linking governance gaps to corresponding policy implications. While the themes emerged inductively from stakeholder interviews, their interpretation reveals actionable dimensions relevant to affinity domain implementation.

Table 3 shows the synthesis of empirically derived thematic categories and their corresponding governance implications for affinity domain implementation in the Czech health care system. The table is based on findings from 18 semistructured stakeholder interviews conducted between January 2025 and April 2025 and reflects author-derived implications grounded in qualitative analysis rather than direct participant recommendations.

Table . From thematic findings to policy implications for affinity domain implementation.

Empirical theme	Governance gap identified	Policy implication
Roles and responsibilities	• Ambiguous mandates • Absence of empowered coordinating body • Vendor dominance	• Establish an independent national interoperability authority with clear rule-making and enforcement powers
Perceived risks	• Legal uncertainty regarding liability • Institutional distrust • Vendor lock-in	• Clarify liability frameworks • Introduce binding interoperability requirements • Create contractual enforcement mechanisms
System-level prerequisites	• Lack of certification frameworks • Inconsistent standard adoption • Limited human capacity	• Implement mandatory technical standards (IHE[a] XDS[b] and HL7[c] FHIR[d]) supported by certification, audit, and national reference implementations
Perceived benefits	• Recognized value for continuity of care and accountability, but unrealized due to governance gaps	• Use pilot affinity domains to demonstrate measurable clinical and operational benefits
Implementation barriers	• Political discontinuity • Fragmented regional approaches • Weak stakeholder engagement	• Introduce phased pilots with co-design mechanisms and multilevel governance coordination aligned with EHDS[e].

[a]IHE: Integrating the Healthcare Enterprise.

[b]XDS: Cross-Enterprise Document Sharing.

[c]HL7: Health Level 7.

[d]FHIR: Fast Healthcare Interoperability Resources.

[e]EHDS: European Health Data Space.

Comparison With International Literature

The study identifies several systemic, structural, and contextual obstacles that complicate the implementation of affinity domains within the Czech health care system. While some of these challenges correspond to broader trends observed across European countries, the Czech case is distinguished by shortcomings in governance architecture, legal frameworks, and stakeholder involvement.

Although affinity domains are not always explicitly addressed in major European policy and governance studies, the structural conditions underpinning their creation and sustainability are frequently analyzed. Affinity domains, as defined by the IHE XDS framework, require not only interoperable technical architectures but also clearly defined governance structures, implementation standards, and legal-operational certainty. Recent literature in digital health governance thus offers critical insights into the enabling—or inhibiting—factors for their deployment.

To contextualize the Czech findings, we purposively selected European countries representing distinct governance architectures of affinity domain–like models (centralized, federated, semicentralized, and regional). These cases were chosen to reflect variation in institutional authority, legal mandates, and implementation maturity, thereby enabling analytical comparison with the governance gaps identified in the Czech interviews.

Table 4 shows a comparative overview of affinity domain–like interoperability governance models in selected European countries, based on desk-based document review conducted as part of this qualitative study. The comparison contextualizes findings from 18 stakeholder interviews (Czech Republic, January 2025-April 2025) against operational governance

XSL•FO
RenderX

structures in Austria, Switzerland, the Netherlands, Sweden, and Belgium.

Table . Comparative overview of affinity domain implementation in selected European countries.

Country	Affinity domain model	Governance structure	National coverage	Key characteristics
Austria	ELGA[a]	Centralized, state-owned	National	Mandatory for providers; opt-out for patients; based on IHE[b] XDS[c] [22]
Switzerland	EPD[d]	Federated; regulated at the federal level	Regional or national	Voluntary for patients; multiple affinity domains operated by certified communities [23]
The Netherlands	LSP[e]	Semicentralized	National	Opt-in system national broker connects providers to shared exchange services [24]
Sweden	NPÖ[f]	Regional (county-level)	Regional	Based on regional initiatives, integrated into national strategy and opt-out model [25]
Belgium	Vitalink	Federated (per region)	Regional	Strong focus on authentication via national electronic identity and platform-based document exchange [26]
Czech Republic	Not operational (conceptual phase)	Central coordination planned via MoH[g] and ÚZIS[h]	None	Legal framework in place, no piloting or operational deployment, and unclear governance

[a]ELGA: Elektronische Gesundheitsakte

[b]IHE: Integrating the Healthcare Enterprise.

[c]XDS: Cross-Enterprise Document Sharing.

[d]EPD: Elektronisches Patientendossier.

[e]LSP: Landelijk Schakelpunt.

[f]NPÖ: Národní plán obnovy.

[g]MoH: Ministry of Health.

[h]ÚZIS: Institute of Health Information and Statistics of the Czech Republic.

Comparative analysis illustrates that successful affinity domain models, from Austria's centralized ELGA to Switzerland's federated EPD, share common elements: empowered coordinating authorities, integrated legal mandates, robust access infrastructures, and sustained stakeholder engagement. In alignment with interviewees' accounts of legal uncertainty, fragmented coordination, and the absence of empowered governance bodies, the Czech Republic currently appears to remain at a largely conceptual stage, with limited piloting, incomplete governance clarification, and no operational affinity domain deployment.

The comparison is not intended as an exhaustive review but as an analytical reference illustrating governance configurations that directly contrast with constraints identified in the Czech case. These comparative patterns reinforce the governance gaps identified in the Czech interviews. Our analysis demonstrates that without clearly defined responsibilities, legal clarity, and enforceable standards, practical progress in interoperability remains unattainable. A comparable pattern is observed in a recent World Health Organization (WHO) or European study, which shows that although 83% of countries in the region have adopted a formal digital health strategy, only 79% have fully developed implementation and performance monitoring mechanisms—a relatively high figure, yet with substantial variation in maturity, funding, and institutional authority [27]. A multi-country qualitative study by Papadopoulos et al [18] further underscores that public trust in national EHR systems hinges on data security, privacy, and the perceived legitimacy of actors—key governance issues also relevant for affinity domains. In parallel, a scoping review by Gazzarata et al [14] confirms that although adoption of HL7 FHIR is growing, particularly in chronic disease management, the lack of binding implementation guides, certification frameworks, and coordinating authorities continues to hamper interoperability and data reusability. In affinity domain terms, such structural vendor dependencies undermine enforceable participation rules and coordinated document exchange across institutional boundaries.

Vendor lock-in and proprietary architecture remain structural barriers. As Kapitan et al [28] emphasize, the adoption of open standards such as HL7 FHIR is not sufficient unless accompanied by open-source implementations and integration toolkits tailored to local constraints. A scoping review of European chronic care implementations found that only 20%

XSL•FO
RenderX

of projects reference formal implementation guides, and certification or central coordination mechanisms are rare, resulting in persistent heterogeneity that undermines interoperability [14]. A recent review on telemonitoring [29] likewise highlights that despite using FHIR and ISO/IEEE (International Organization for Standardization and the Institute of Electrical and Electronics Engineers) 11073, inconsistent governance models and fragmented implementations remain widespread.

Clinician and stakeholder engagement also remains minimal. This finding directly resonates with interviewees' descriptions of the limited inclusion of clinical end users in governance processes. According to the Organisation for Economic Co-operation and Development [30], the average quality score for stakeholder engagement in primary lawmaking rose modestly from 2.00 in 2014 to 2.26 in 2024 (on a 0 - 4 scale). Engagement in subordinate regulation remains weaker (1.95-2.16), with oversight and quality control scoring particularly low (0.37 out of 1). Despite rapid growth in remote care and patient access metrics, the sustainability of interoperability gains remains uncertain. The Czech Republic's maturity score increased by 26 points in 2024—the largest annual improvement in the European Union—yet full interoperability remains unachieved, as key structural mechanisms are still absent or only partially implemented.

Building public trust in digital infrastructures is essential. A structured review of semantic interoperability methods shows that meaningful trust-building relies on transparent data standards supported by policy frameworks [31,32]. A cross-country analysis of personal EHR access further validates the importance of standards, usability, and stakeholder inclusion in fostering acceptance and trust [33]. In the absence of such frameworks, Czech narratives around health data governance often neglect data security, privacy, and stakeholder legitimacy—undermining the sustainable implementation of affinity domains.

Implications for Policy and Practice

The following implications represent author-derived interpretations informed by the empirical findings, relevant literature, and EHDS regulatory requirements, rather than direct requests expressed by the interviewed stakeholders.

Given stakeholders' repeated emphasis on legal uncertainty, fragmented governance structures, weak enforcement mechanisms, and institutional risk aversion, phased pilot deployments may represent a pragmatic policy instrument for reducing uncertainty and strengthening trust in domain-based interoperability arrangements.

International evidence [22,23,26] suggests that pilots anchored in regional settings and supported by co-design can facilitate stakeholder engagement and accelerate scaling. In the Czech context, and consistent with the governance constraints identified in the interviews, the current absence of pilot deployments, test registries, or structured sandbox environments limits opportunities for iterative validation of both technical and governance arrangements under real-world conditions.

Building on the governance gaps identified in this study, we recommend that the Czech Republic adopt maturity assessment frameworks such as those used in the WHO or European survey [27] or embedded in the EHDS implementation model [34] to evaluate readiness across legal, organizational, and technical dimensions. These tools may guide prioritization, funding allocation, and benchmarking toward operational interoperability.

In alignment with the EHDS regulation, national interoperability governance frameworks—including potential affinity domain structures—should be coordinated with national metadata catalogs and designated health data access bodies to ensure transparency, lawful data access, and structured secondary use.

Limitations

This study has several limitations. First, purposive and snowball sampling may have over-represented stakeholders who were more accessible or willing to participate, which could limit the transferability of the findings. Second, interviews were conducted with senior policymakers and institutional leaders, including representatives of the MoH and regional authorities. The hierarchical and professional status of these participants may have influenced the framing of responses, potentially introducing elements of social desirability or correctness bias. Although anonymity was guaranteed and reflexive strategies were applied, responses may still reflect institutionally aligned perspectives rather than fully critical accounts. The sample predominantly consisted of highly experienced professionals in senior governance roles. While appropriate for analyzing institutional decision-making and regulatory structures, this composition may under-represent early-career or operational-level perspectives and thus limit generational diversity in the findings. In addition, representatives of primary care professional associations, organized medical chambers, and patient advocacy groups were not directly included. Although senior provider representatives participated, frontline primary care clinicians and organized patient stakeholders were under-represented, which may privilege governance-level perspectives over operational or user-level experiences. Third, while dual independent coding and consensus resolution were implemented, no formal intercoder reliability coefficients were calculated. Fourth, the perspectives analyzed are context-specific to the Czech governance and regulatory environment and may not be generalizable to countries with more mature certification frameworks or long-standing national HIE systems. In addition, the study did not include patients or the general public, despite recurring interview themes related to trust, legitimacy, and institutional credibility. The absence of these perspectives represents a substantive limitation, as public trust is a foundational condition for sustainable affinity domain governance and cross-enterprise data sharing. Furthermore, data collection (January 2025-April 2025) coincided with the finalization and publication of the EHDS regulation in February 2025. The findings may therefore reflect stakeholder perceptions during an early transition phase marked by regulatory uncertainty rather than a stabilized postimplementation environment. Finally, the research did not evaluate pilot implementations or real-world affinity domain deployments, which limits the ability to draw conclusions about operational

XSL•FO
RenderX

feasibility. These limitations should be considered when interpreting the results and designing future studies.

Future Directions

Future research should build on these findings in several directions. First, empirical evaluation of pilot affinity domains is needed to examine how the identified governance attributes—such as role allocation, accountability, and enforcement—operate in practice once interoperability infrastructure moves beyond the conceptual stage. Longitudinal studies following pilot implementations could assess how trust, compliance, and institutional coordination evolve over time.

Second, future research should prioritize the inclusion of patients and public representatives, particularly to examine how trust, legitimacy, and perceived data governance affect the willingness to participate in affinity domain–based information exchange. Incorporating these perspectives would allow for the triangulation of governance assumptions identified in this study.

Third, comparative research across post-transition health systems could help distinguish context-specific barriers from more general governance patterns affecting affinity domain implementation. Such work would also support benchmarking national readiness in relation to EHDS requirements and inform transferable policy lessons.

Conclusions

This study examined stakeholder perceptions of affinity domains as a governance and interoperability construct within the Czech health care system using qualitative interviews. The findings show that while affinity domains are recognized as a promising model for cross-enterprise data sharing, their feasibility is constrained by unclear governance mandates, limited enforcement mechanisms, legal uncertainty, and fragmented implementation capacity. By explicitly linking these challenges to the governance attributes of affinity domains, the study clarifies why interoperability barriers persist despite the availability of technical standards. These findings contribute empirical insight into the governance conditions required for operational interoperability in post-transition health systems and provide a grounded basis for further policy and implementation research.

Acknowledgments

The authors thank all stakeholders for their participation and time, as well as colleagues from the Czech Technical University and the Technical University of Liberec for their support. The authors also acknowledge administrative support from the Faculty of Biomedical Engineering. Generative artificial intelligence tools were used for language clarification and editorial support in cases of ambiguity. No generative artificial intelligence was used for data generation, analysis, or interpretation.

Funding

This study was supported by a grant from the Czech Technical University in Prague (SGS23/197/OHK5/3T/17).

Data Availability

Due to confidentiality and General Data Protection Regulation restrictions, full interview transcripts cannot be publicly shared. Deidentified coding frameworks and illustrative quotes are available in the multimedia appendices. Additional anonymized materials may be made available upon reasonable request to the corresponding author. This statement complies with JMIR Publications' data availability policy.

Authors' Contributions

JB and PH conceived the study design and co-led the analysis. PH and IF conducted the interviews and initial coding. All authors contributed to the interpretation of findings, drafting, and revising the manuscript, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participant characteristics.
[DOCX File, 26 KB - medinform_v14i1e83894_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 20 KB - medinform_v14i1e83894_app2.docx]

Multimedia Appendix 3
Saturation grid.
[DOCX File, 26 KB - medinform_v14i1e83894_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Coding tree and exemplar codes.
[DOCX File, 26 KB - medinform_v14i1e83894_app4.docx]

Multimedia Appendix 5
Extended codebook with illustrative quotes.
[DOCX File, 25 KB - medinform_v14i1e83894_app5.docx]

Checklist 1
COREQ checklist.
[DOCX File, 30 KB - medinform_v14i1e83894_app6.docx]

References

1. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending directive 2011/24/EU and regulation (EU) 2024/2847 (text with EEA relevance). EUR-Lex: Access to European Union law. 2025. URL: https://eur-lex.europa.eu/eli/reg/2025/327/oj [accessed 2026-03-12]
2. Readiness of electronic health record systems to contribute to national health information and research. : OECD Publishing; 2017 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/12/readiness-of-electronic-health-record-systems-to-contribute-to-national-health-information-and-research_ed3beaed/9e296bf3-en.pdf [accessed 2026-03-12]
3. IHE IT infrastructure (ITI) technical framework. : IHE International, Inc; 2020 URL: https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Rev17-0_Vol1_FT_2020-07-20.pdf [accessed 2026-03-12]
4. Dogac A, Laleci GB, Aden T, Eichelberg M. Enhancing IHE XDS for federated clinical affinity domain support. IEEE Trans Inf Technol Biomed 2007 Mar;11(2):213-221. [doi: 10.1109/titb.2006.874928] [Medline: 17390991]
5. Tornero Costa R, Adib K, Salama N, et al. Electronic health records and data exchange in the WHO European region: A subregional analysis of achievements, challenges, and prospects. Int J Med Inform 2025 Feb;194:105687. [doi: 10.1016/j.ijmedinf.2024.105687] [Medline: 39556969]
6. Zvárová J, Lhotská L, Seidl L, Zvára K. Health data collecting and sharing: case studies of Czech e-health applications. Stud Health Technol Inform 2012;180:672-676. [Medline: 22874276]
7. Těšinová JK, Dobiášová K, Dušek Z, Tobiášová A. Development of telemedicine in the Czech Republic from patients' and other key stakeholders' perspective. Front Public Health 2023;11:1202182. [doi: 10.3389/fpubh.2023.1202182] [Medline: 37937075]
8. Bruthans J, Duftschmid G, Hammar T, et al. Comparison of electronic prescription systems in the European Union: benchmarking development, use, and future trends. IEEE J Biomed Health Inform 2025 May;29(5):3712-3722. [doi: 10.1109/JBHI.2025.3531317] [Medline: 40036418]
9. Bruthans J. The successful usage of the DICOM images exchange system (ePACS) in the Czech Republic. Appl Clin Inform 2020 Jan;11(1):104-111. [doi: 10.1055/s-0040-1701252] [Medline: 32023639]
10. eMeDOcS. URL: https://www.emedocs.cz/ [accessed 2025-09-10]
11. Exchange Medical Documents System—National healthshare system. ICZ Group. URL: https://www.iczgroup.com/en/pa-reference/exchange-medical-documents-system-national-healthshare-system/ [accessed 2025-09-10]
12. Vlasakova M, Muzik J, Holubová A, et al. A telemedicine system intervention for patients with type 1 diabetes: pilot feasibility crossover intervention study. JMIR Form Res 2023 Apr 28;7:e35064. [doi: 10.2196/35064] [Medline: 37115601]
13. Angelovská O, Dobiášová K, Těšinová JK. Pandemic COVID-19 as a challenge for telemedicine in the Czech Republic. Int J Health Plann Manage 2025 Jan;40(1):271-286. [doi: 10.1002/hpm.3863] [Medline: 39497023]
14. Gazzarata R, Almeida J, Lindsköld L, et al. HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) In digital healthcare ecosystems for chronic disease management: scoping review. Int J Med Inform 2024 Sep;189:105507. [doi: 10.1016/j.ijmedinf.2024.105507] [Medline: 38870885]
15. Digital decade eHealth indicators: progress throughout 2023. European Health Telematics Association. URL: https://ehtel.eu/media-room/latest-news/291-digital-decade-ehealth-indicators-progress-throughout-2023.html [accessed 2025-07-25]
16. Nagy M, Hanzlícek P, Precková P, et al. Semantic interoperability in Czech healthcare environment supported by HL7 version 3. Methods Inf Med 2010;49(2):186-195. [doi: 10.3414/ME09-02-0018] [Medline: 19936441]
17. Belrhiti Z, Bigdeli M, Lakhal A, Kaoutar D, Zbiri S, Belabbes S. Unravelling collaborative governance dynamics within healthcare networks: a scoping review. Health Policy Plan 2024 Apr 10;39(4):412-428. [doi: 10.1093/heapol/czae005] [Medline: 38300250]
18. Papadopoulos K, Ammenwerth E, Lame G, et al. Understanding public trust in national electronic health record systems: a multi-national qualitative research study. Digit Health 2025;11:20552076251333576. [doi: 10.1177/20552076251333576] [Medline: 40190337]

XSL•FO
RenderX

19. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
20. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
21. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). : Publications Office of the European Union; 2016 URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-03-21]
22. Elektronische Gesundheitsakte. URL: https://www.elga.gv.at/ [accessed 2025-07-25]
23. Bundesamt für Gesundheit BAG. URL: https://www.bag.admin.ch/de [accessed 2025-07-25]
24. De Vereniging van Zorgaanbieders voor Zorgcommunicatie [Website in Dutch]. URL: https://www.vzvz.nl/ [accessed 2025-07-25]
25. E-hälsomyndigheten. URL: https://www.ehalsomyndigheten.se/ [accessed 2025-07-05]
26. Vitalink. 2025. URL: https://www.vitalink.be/ [accessed 2025-07-25]
27. Fernandes FA, Chaltikyan G, Adib K, Caton-Peters H, Novillo-Ortiz D. The role of governance in the digital transformation of healthcare: results of a survey in the WHO Europe Region. Int J Med Inform 2024 Sep;189:105510. [doi: 10.1016/j.ijmedinf.2024.105510] [Medline: 38901269]
28. Kapitan D, Heddema F, Dekker A, Sieswerda M, Verhoeff BJ, Berg M. Data interoperability in context: the importance of open-source implementations when choosing open standards. J Med Internet Res 2025 Apr 15;27:e66616. [doi: 10.2196/66616] [Medline: 40232773]
29. Martins D, Lewerenz S, Carmo A, Martins H. Interoperability of telemonitoring data in digital health solutions: a scoping review. Front Digit Health 2025;7:1502260. [doi: 10.3389/fdgth.2025.1502260] [Medline: 40330872]
30. Government at a glance 2025. : Organisation for Economic Co-operation and Development; 2025 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/government-at-a-glance-2025_70e14c6c/0efd0bcd-en.pdf [accessed 2026-03-12]
31. Palojoki S, Lehtonen L, Vuokko R. Semantic interoperability of electronic health records: systematic review of alternative approaches for enhancing patient information availability. JMIR Med Inform 2024 Apr 25;12:e53535. [doi: 10.2196/53535] [Medline: 38686541]
32. Scheibner J, Sleigh J, Ienca M, Vayena E. Benefits, challenges, and contributors to success for national ehealth systems implementation: a scoping review. J Am Med Inform Assoc 2021 Aug 13;28(9):2039-2049. [doi: 10.1093/jamia/ocab096] [Medline: 34151990]
33. Moll J, Scandurra I, Bärkås A, et al. Sociotechnical cross-country analysis of contextual factors that impact patients' access to electronic health records in 4 European countries: framework evaluation study. J Med Internet Res 2024 Aug 26;26:e55752. [doi: 10.2196/55752] [Medline: 39186760]
34. European Health Data Space Regulation (EHDS). European Commission. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en [accessed 2025-07-25]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DICOM: Digital Imaging and Communications in Medicine
EHDS: European Health Data Space
EHR: electronic health record
ELGA: Elektronische Gesundheitsakte
eMeDOcS: Exchange Medical Documents System
EPD: Elektronisches Patientendossier
FHIR: Fast Healthcare Interoperability Resources
HIE: health information exchange
HIF: health insurance fund
HL7: Health Level 7
IEEE: Institute of Electrical and Electronics Engineers
IHE: Integrating the Healthcare Enterprise
ISAC: Integration, Share and Communication System
ISO: International Organization for Standardization
IZIP: Internetový zdravotní informační portál
MoH: Ministry of Health
RHA: regional health authority
WHO: World Health Organization

XSL•FO
RenderX

XDS: Cross-Enterprise Document Sharing

Edited by A Coristine; submitted 10.Sep.2025; peer-reviewed by C Vorisek, GLP Bodagala, M Nakayama; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 02.Apr.2026.

Please cite as:
Bruthans J, Hospodková P, Fejfarová I
Stakeholder Perspectives on Affinity Domains in Digital Health Interoperability: Qualitative Study
JMIR Med Inform 2026;14:e83894
URL: https://medinform.jmir.org/2026/1/e83894
doi:10.2196/83894

©Jan Bruthans, Petra Hospodková, Ilona Fejfarová. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Current Status, Challenges, and Opportunities Associated With Implementing Clinical Data Interchange Standards Consortium Standards in Japanese Academic Medical Centers: Cross-Sectional Survey

Kaori Nagai[1], MSN; Noriko Ito[1], MEng; Yuya Ikeda[2], BEng; Toshiki I Saito[1,3], MD, PhD

[1]Clinical Research Center, NHO Nagoya Medical Center, 4-1-1 Sannomaru, Naka-ku, Nagoya, Aichi, Japan

[2]Clinical Research Support Office, National Cancer Center Hospital East, Kashiwa, Chiba, Japan

[3]Clinical Research Center, NHO Headquarters, Meguro-ku, Tokyo, Japan

Corresponding Author:
Kaori Nagai, MSN
Clinical Research Center, NHO Nagoya Medical Center, 4-1-1 Sannomaru, Naka-ku, Nagoya, Aichi, Japan

Abstract

Background: The implementation of Clinical Data Interchange Standards Consortium (CDISC) standards is essential for accelerating clinical research and is mandated for new drug applications in Japan. However, the current status of their implementation and associated challenges in Japanese academic medical centers has not been comprehensively investigated.

Objectives: This study aimed to comprehensively clarify the current status, challenges, and opportunities associated with implementing CDISC standards in Japanese academic medical centers.

Methods: We conducted a questionnaire survey targeting 91 academic medical centers (83 university hospitals, 7 national centers for advanced and specialized medical care, and 1 national hospital organization) from June to August 2024. The survey assessed experience with contracted CDISC data creation, implementation status, barriers, and available resources. We performed Mann-Whitney *U* tests to assess the association between the number of data managers or biostatisticians and CDISC implementation status.

Results: Responses were received from 84 institutions (response rate: 92%). While 29% (26/91) had experience with contracted CDISC standards data creation, 22% (20/91) had implemented CDISC standards. The most frequently cited barrier to CDISC standards implementation was “significant human and financial effort and resource shortages” (62%, 56/91). A significant “supply and demand” gap was identified: while 66% (60/91) of institutions expected support (eg, materials, workshops) from other organizations, only 5% (5/91) were capable of providing such resources. Furthermore, 66% (60/91) of institutions had not yet decided on a future policy regarding CDISC compliance. Notably, institutions that had implemented CDISC standards had significantly higher numbers of data managers (P<.001) and biostatisticians (P<.001) compared to those that had not.

Conclusions: The implementation of CDISC standards in Japanese academia is limited, primarily due to resource shortages and a lack of specialized personnel. However, the discrepancy between the high demand for support and limited supply highlights a significant opportunity. Experienced academic research organizations, in cooperation with pharmaceutical companies and contract research organizations, can drive nationwide adoption by taking a strategic leadership role in providing educational resources and training, thereby addressing the critical shortage of specialized personnel and know-how.

(JMIR Med Inform 2026;14:e83774) doi:10.2196/83774

KEYWORDS

Clinical Data Interchange Standards Consortium; CDISC standards; academic research organization; ARO; Clinical Data Acquisition Standards Harmonization; CDASH; Study Data Tabulation Model; SDTM; Analysis Data Model; ADaM

Introduction

The Clinical Data Interchange Standards Consortium (CDISC) is a nonprofit organization dedicated to developing global electronic standards for clinical trial data. The “CDISC standards” refer to the set of standards developed by this consortium. Implementing these standards entails a series of technical activities throughout the clinical data lifecycle. Specifically, it involves designing case report forms based on Clinical Data Acquisition Standards Harmonization (CDASH), mapping collected raw data to Study Data Tabulation Model (SDTM) domains, and creating analysis-ready datasets according to Analysis Data Model (ADaM). These processes require writing programming scripts for data transformation, generating

XSL•FO
RenderX

metadata files such as define.xml, and rigorously validating the datasets and their conformance to the standards. In Japan, the Pharmaceuticals and Medical Devices Agency (PMDA) has been accepting electronic clinical trial data submissions compliant with CDISC standards for new drug marketing approval applications since October 2016 [1]. Furthermore, the Japan Agency for Medical Research and Development has expressed the view that compliance with CDISC standards from the planning and execution stages will become essential in academia-led clinical research and trials in the future [2].

Additionally, global data sharing platforms promote the sharing and reuse of clinical trial data [3-9], and several of these advance data standardization in collaboration with the CDISC [3-6,8].

The sharing of clinical trial data is expected to promote advancements in medicine and science [10,11]. Moreover, since July 1, 2018, the International Committee of Medical Journal Editors has mandated the inclusion of a data-sharing statement in manuscripts reporting the results of clinical trials submitted to International Committee of Medical Journal Editors member journals [12].

In 2017 in Japan, the University Hospital Medical Information Network Center and Japan Agency for Medical Research and Development cohosted the "CDISC Public Symposium" [2]. Furthermore, in 2018, University Hospital Medical Information Network, PMDA, and CDISC cohosted the "Workshop on CDISC Utilization in Academia" [13]. These activities aim to promote the dissemination of CDISC standards in Japanese academia.

The implementation of CDISC standards is expected to accelerate clinical research [14]. In particular, researchers and related institutions are beginning to recognize the importance of planning research and preparing case report forms based on the premise of electronic submissions compliant with CDISC standards throughout the processes of data collection, data management, and statistical analysis. These initiatives are expected to help reduce costs and workloads while maintaining data quality [15].

A previous study investigated the current status and challenges associated with the implementation of CDISC standards in Japanese academic medical centers [16]. However, in that study, the definition of "academia" was ambiguous.

Therefore, the present study aimed to comprehensively clarify the current status and challenges associated with implementing CDISC standards in Japanese academic medical centers.

Methods

Ethical Considerations

This study was a questionnaire survey targeting administrative personnel at academic medical centers regarding the operational status of CDISC standards. The study did not involve interventions on human participants or the collection of sensitive personal health information from patients. Therefore, an ethics review board assessment was not sought, in accordance with the ethical guidelines for medical and health research involving human subjects in Japan.

The purpose of the study was clearly explained in the invitation to the online questionnaire, and the completion and submission of the questionnaire were regarded as implied consent to participate. Participation was voluntary, and respondents could withdraw at any time by not submitting their responses.

To protect the privacy and confidentiality of the respondents, the collected data were stored securely and accessed only by the research team. The data were aggregated for analysis, and no personally identifiable information was reported. No compensation was provided to the participants for this study. This manuscript and the supplementary materials do not contain any images or data that could identify individual participants.

Participants and Survey Administration

In this study, "academia" is defined as institutions in Japan legally mandated to conduct research as part of their official duties. Based on this definition, the following 91 academic medical centers (hereinafter referred to as "institutions") were selected as survey targets: 83 university hospitals (43 national university hospitals, 8 public university hospitals, 31 private university hospitals, and 1 ministry-affiliated university hospital), 7 national centers for advanced and specialized medical care, and 1 national hospital organization. We contacted these institutions by email, requested their cooperation in the questionnaire survey, and sent them the questionnaire.

The survey period was from June 12 to August 16, 2024. The questionnaire survey was conducted online using Google Forms. As a general rule, 1 response per institution was accepted. No duplicate responses were received from any single institution; therefore, all submitted responses were included in the final analysis.

The survey was primarily distributed to the directors of clinical research centers or heads of data centers at each institution. In cases where direct contact information for these roles was unavailable, we identified the appropriate respondents through personal professional networks. For institutions where no direct contact could be established, the request was sent to the general administrative office of the Clinical Research Support Center, with instructions to forward the questionnaire to the staff member most knowledgeable about CDISC standards.

Questionnaire Items

The main questionnaire items were as follows:

Experience with contracted CDISC standards data creation

Implementation status of CDISC standards

Barriers to CDISC standards implementation

Future policies regarding CDISC standards-compliant data creation

Availability of CDISC standards-related documents and training

Ability to provide CDISC standards-related documents and training

Expectations for pharmaceutical companies, contract research organizations (CROs), and other academic research organizations (AROs)

XSL•FO
RenderX

Ability to provide to pharmaceutical companies, CROs, and other AROs

Number of data managers and biostatisticians

In this study, "implementation" was defined as CDISC data creation conducted internally, jointly with external partners, or externally with acceptance reviews performed by internal CDISC experts. "Experience with contracted CDISC standards data creation" was defined more broadly to encompass these implementation categories as well as external data creation performed without such internal expert-led acceptance procedures.

Further details regarding the questionnaire items are presented in Multimedia Appendix 1.

Statistical Analysis

Descriptive statistics were calculated using the total number of targeted institutions (n=91) as the denominator. This approach was adopted to establish a consistent baseline for future longitudinal surveys targeting the same population. Furthermore, using the total population as the denominator prevents the overestimation of implementation rates caused by selection bias—where institutions with greater interest in CDISC standards are more likely to respond—and ensures the results accurately reflect the status across all defined Japanese institutions. Categorical variables are presented as numbers and percentages. To assess the association of the number of data managers and biostatisticians with the CDISC implementation status, we performed the Mann-Whitney *U* test. In this analysis, the number of staff was treated as a continuous variable. Although staffing levels were categorized (0‐3, 4‐10, and ≥11) for visualization purposes, the statistical hypothesis testing was based on the continuous data. To examine potential confounders, we also analyzed the association between institution type (public/national vs private) and CDISC implementation using Fisher exact test. A *P* value <.05 was considered statistically significant. All statistical analyses were performed using R statistical software (version 4.4.1; R Foundation for Statistical Computing).

Results

Response Rate and Participant Characteristics

In this study, 91 institutions were targeted, among which responses were received from 84 (response rate: 92%). Of these, 7 (8%) institutions did not respond. The breakdown of the responding institutions was as follows: 42 national university hospitals, 27 private university hospitals, 7 public university hospitals, 7 national centers for advanced and specialized medical care, and 1 national hospital organization. No responses were received from ministry-affiliated university hospitals. The breakdown of respondents by job title is shown in Multimedia Appendix 2.

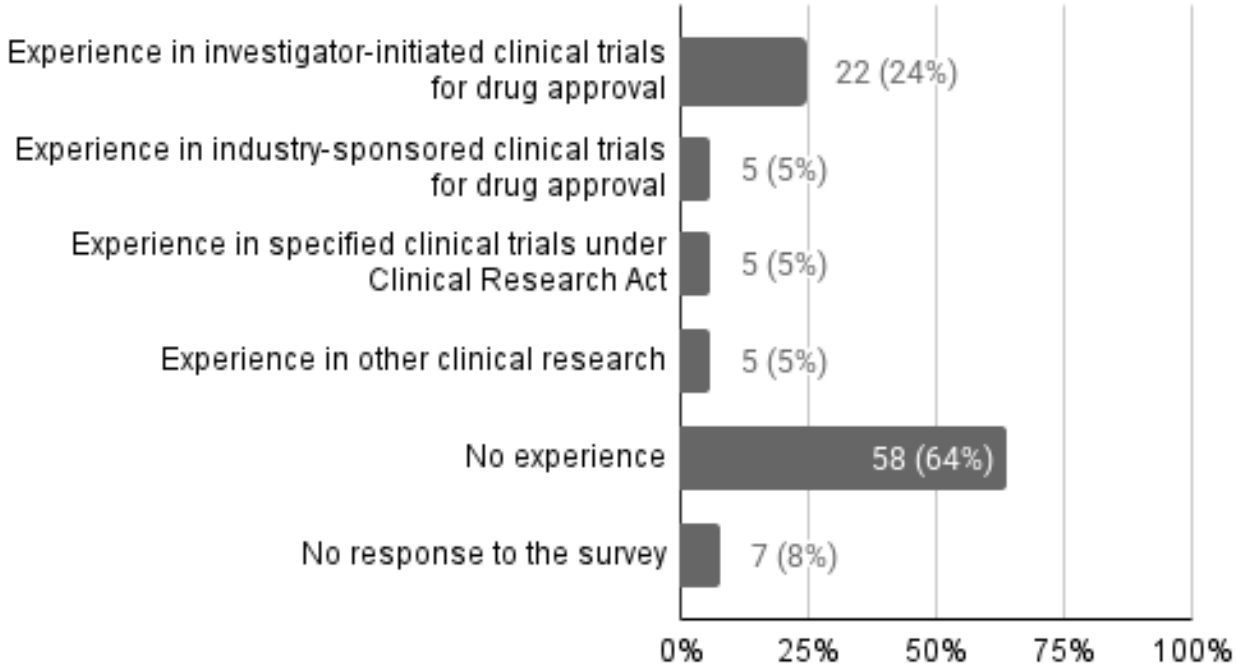
The percentages in these results were calculated with the total number of targeted institutions (N=91) as the denominator and rounded to the nearest integer, unless otherwise specified.

Experience With Contracted CDISC Standards Data Creation

Experience with contracted CDISC standards data creation refers to the track record of accepting projects for CDISC data creation. This includes cases where the actual data creation was outsourced to a third party. Regarding experience with contracted data creation using CDISC standards, 26 (29%) institutions reported having such experience, while 58 (64%) did not; 7 (8%) institutions did not respond to this question. Among the types of trials for which institutions had experience (multiple answers allowed), investigator-initiated clinical trials for drug approval were the most common, at 22 (24%), followed by industry-sponsored clinical trials for drug approval (n=5, 5%), specified clinical trials under Japan's Clinical Research Act [17] (n=5, 5%), and other clinical research (n=5, 5%) (Figure 1).

XSL•FO
RenderX

Figure 1. Experience with contracted Clinical Data Interchange Standards Consortium standards data creation (multiple responses allowed) (N=91).



These results indicate that slightly less than 30% (26/91) of the institutions had experience with contracted data creation using CDISC standards. Notably, this experience extends beyond trials for drug approval to include specified clinical trials and other clinical research.

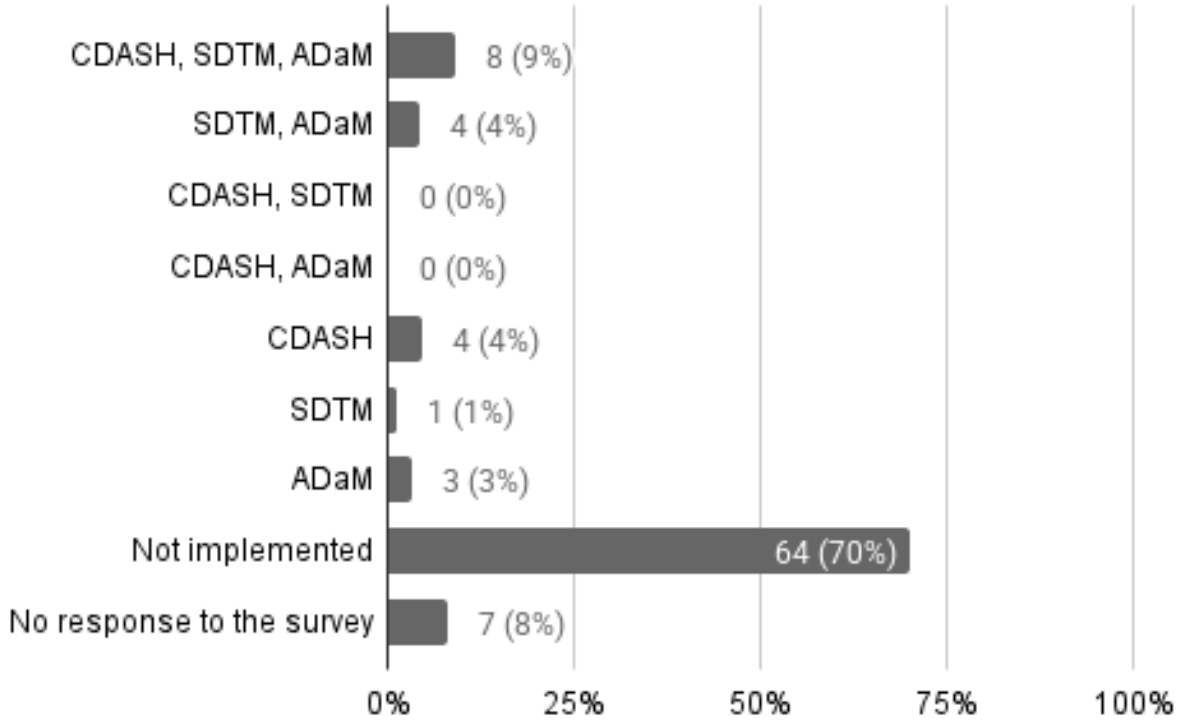
Implementation Status of CDISC Standards

Implementation of CDISC standards refers to creating CDISC data using internal resources or, in the case of outsourcing, conducting acceptance work by CDISC experts. Regarding the implementation status of CDISC standards, 20 institutions (22%) reported having implemented them, whereas 64 institutions (70%) reported not having implemented them; seven institutions (8%) did not respond.

Among the 20 institutions that had implemented CDISC standards, the most common combination of standards (multiple answers allowed) was the implementation of all three—CDASH, the SDTM, and the ADaM—reported by eight institutions (9% of the 91 institutions). This was followed by institutions implementing SDTM and ADaM (n=4, 4%), CDASH only (n=4, 4%), ADaM only (n=3, 3%), and SDTM only (n=1, 1%) (Figure 2).

XSL•FO
RenderX

Figure 2. Implementation status of the Clinical Data Interchange Standards Consortium standards (N=91). ADaM: Analysis Data Model; CDASH: Clinical Data Acquisition Standards Harmonization; SDTM: Study Data Tabulation Model.



Regarding individual standards, ADaM was implemented by 15 (16%) institutions, SDTM by 13 (14%), and CDASH by 12 (13%).

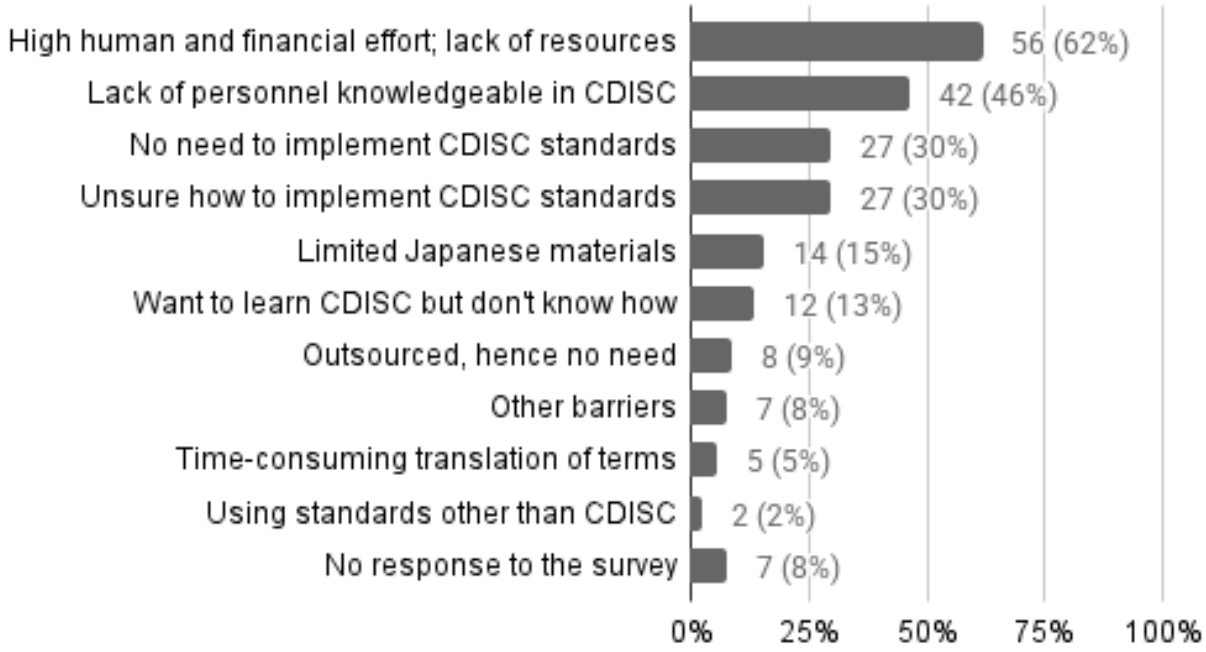
These results indicate that while only 22% (20/91) of all institutions had implemented CDISC standards, many of these implementing institutions use multiple standards (CDASH, SDTM, and ADaM) in combination.

Barriers to CDISC Standards Implementation

Barriers with CDISC implementation refer to the obstacles or challenges that institutions face when attempting to adopt or maintain CDISC standards. Institutions cited the following items as barriers to CDISC standards implementation (multiple answers allowed). The most frequently cited barrier was “high human and financial effort and lack of resources” (n=56, 62%), followed by “lack of personnel knowledgeable in CDISC” (n=42, 46%), “no need to implement CDISC standards” (n=27, 30%), and “unsure how to implement CDISC standards” (n=27, 30%) (Figure 3).

XSL•FO
RenderX

Figure 3. Barriers to CDISC standards implementation (multiple responses allowed). Note: Percentages were calculated based on the total number of survey targets (N=91). "No response to the survey" indicates institutions that did not answer this survey entirely. CDISC: Clinical Data Interchange Standards Consortium.



These results indicate that the main barriers to CDISC standards implementation are the significant human and financial effort required, resource shortages, and a lack of personnel knowledgeable about CDISC.

Future Policies Regarding CDISC Standards-Compliant Data Creation

Regarding future policies for CDISC standards-compliant data creation (single answer), the most common response, reported by 60 (66%) institutions, was "no decision made," followed by "will not undertake" (n=7, 8%), "will undertake, and external experts will perform the work with internal involvement" (n=6, 7%), "will undertake, and both internal and external personnel will perform the work" (n=5, 5%), "will undertake, but only external experts will perform the work without internal involvement" (n=5, 5%), and "will undertake, and the work will be performed internally" (n=1, 1%) (Table 1). These results indicated that 66% (60/91) of the institutions had not yet decided on their future policy; only 19% (17/91) had a clear policy for undertaking such work.

Table . Future policies regarding CDISC[a] standards-compliant data creation (N=91).[b]

Future policy regarding CDISC compliance	All institutions (N=91), n (%)	Currently implemented (n=20), n (%)	Currently not implemented (n=64), n (%)
Internal execution	1 (1)	1 (5)	0 (0)
Joint internal and external execution	5 (5)	3 (15)	2 (3)
External execution with internal involvement	6 (7)	5 (25)	1 (2)
External execution without internal involvement	5 (5)	1 (5)	4 (6)
Will not undertake	7 (8)	2 (10)	5 (8)
No decision made	60 (66)	8 (40)	52 (81)
No response to the survey	7 (8)	—[c]	—

[a]CDISC: Clinical Data Interchange Standards Consortium.

[b]Note: All values are n (%); percentages are calculated based on the total number of survey targets (N=91). "Implemented" and "not implemented" columns represent the current status at the time of the survey, where "implementation" is defined as internal CDISC data creation or expert-led acceptance work for outsourced tasks. "No response to the survey" refers to institutions that did not return the questionnaire.

[c]Not available.

The association between current implementation status and future policies is detailed in Table 1. Among the 20 institutions that had already implemented the standards, fewer than half (9 institutions, 45%) intended to continue "implementation"

XSL•FO
RenderX

(defined as work involving internal resources or internal expert-led acceptance). A similar lack of interest was observed among the 64 nonimplementing institutions, with only 3 (5%) planning for future implementation.

CDISC Standards Implementation Resources: Availability and Ability to Provide

Among all, 15 (16%) institutions reported having documents or training related to implementing CDISC standards. A breakdown of these resources (multiple answers allowed) indicated that "beginner-level training" was the most common, reported by 12 (13%) institutions, followed by "Japanese-language educational materials" (n=9, 10%), "documents on implementation know-how" (n=8, 9%), and "tools supporting CDISC" (n=8, 9%) (Table 2).

Table . Availability of and ability to provide CDISC[a] standards-related documents and training (multiple responses allowed) (N=91).[b]

Item	Availability, n (%)	Ability to provide, n (%)
Japanese-language educational materials on CDISC standards	9 (10)	4 (4)
Documents on implementation know-how for CDISC standards	8 (9)	4 (4)
Tools supporting CDISC standards implementation	8 (9)	4 (4)
Beginner-level training on CDISC standards	12 (13)	5 (5)
No response to the survey	7 (8)	7 (8)

[a]CDISC: Clinical Data Interchange Standards Consortium.

[b]Note: Percentages were calculated based on the total number of survey targets (N=91). "No response to the survey" indicates institutions that did not answer this survey entirely.

Next, regarding the question of whether these documents and training could be provided to other institutions, 5 (5%) institutions responded that they "could provide" them. A breakdown of providable resources is shown in Table 2.

These results indicate that while 16% (15/91) of institutions had CDISC standards implementation resources available, only 5% (5/91) were able to provide these resources to other institutions.

Support Expected From and Providable to Pharmaceutical Companies, CROs, and Other AROs

Regarding expectations for pharmaceutical companies, CROs, and other AROs (multiple answers allowed), 60 (66%) institutions responded that they had "some expectations." Among these 60 institutions, the most desired item was "materials and information on CDISC" (n=53, 58%), followed by "training sessions (seminars, workshops, etc) on CDISC" (n=45, 49%) and "on-the-job training related to CDISC" (n=33, 36%) (Table 3).

Table . Expectations for and ability to provide support to pharmaceutical companies, contract research organizations, and other academic research organizations (multiple responses allowed) (N=91).[a]

Item	Expectations, n (%)	Ability to provide, n (%)
Training sessions (seminars, workshops, etc) on CDISC[b]	45 (49)	2 (2)
On-the-job training related to CDISC	33 (36)	2 (2)
Materials and information on CDISC	53 (58)	4 (4)
No response to the survey	7 (8)	7 (8)

[a]Note: Percentages were calculated based on the total number of survey targets (N=91). "No response to the survey" indicates institutions that did not answer this survey entirely.

[b]CDISC: Clinical Data Interchange Standards Consortium.

Next, regarding support that could be provided to pharmaceutical companies, CROs, and other AROs (multiple answers allowed), 5 (5%) institutions responded that they "could provide" some support. A breakdown is shown in Table 3.

These findings revealed that 66% (60/91) of the institutions expected CDISC-related material provision, information sharing, and workshop hosting from pharmaceutical companies, CROs, and other AROs. On the other hand, only 5% (5/91) of the institutions were able to provide such support, indicating a limited capacity.

Number of Data Managers and Biostatisticians

Among the 84 responding institutions, the median number of data managers was 2 (mode: 0; range: 0 - 20) (Figure 4). Similarly, the median number of biostatisticians was 2 (mode: 1, range: 0 - 13) (Figure 5).

XSL•FO
RenderX

Figure 4. Number of data managers at responding facilities (n=84). Note: Analysis excluded the 7 institutions that did not respond to the survey.

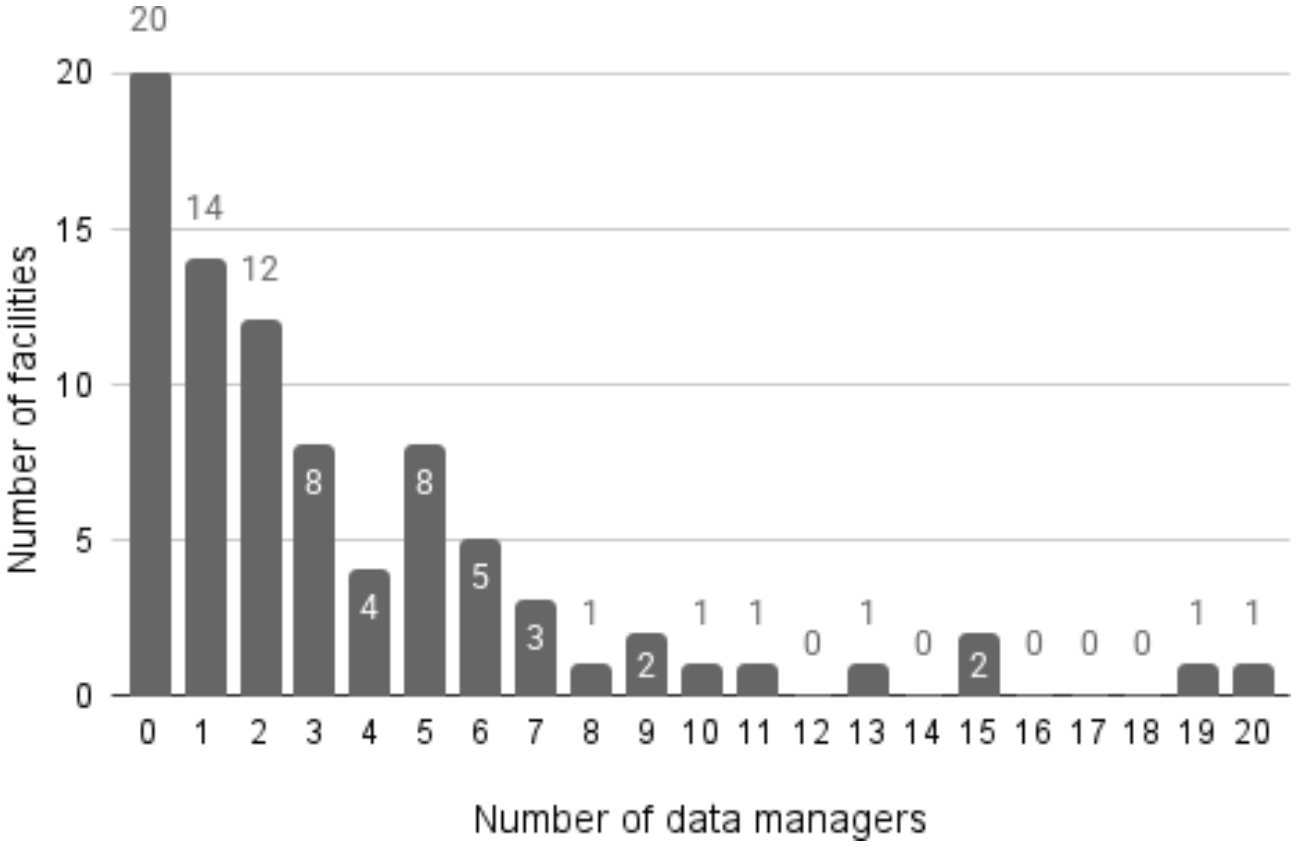
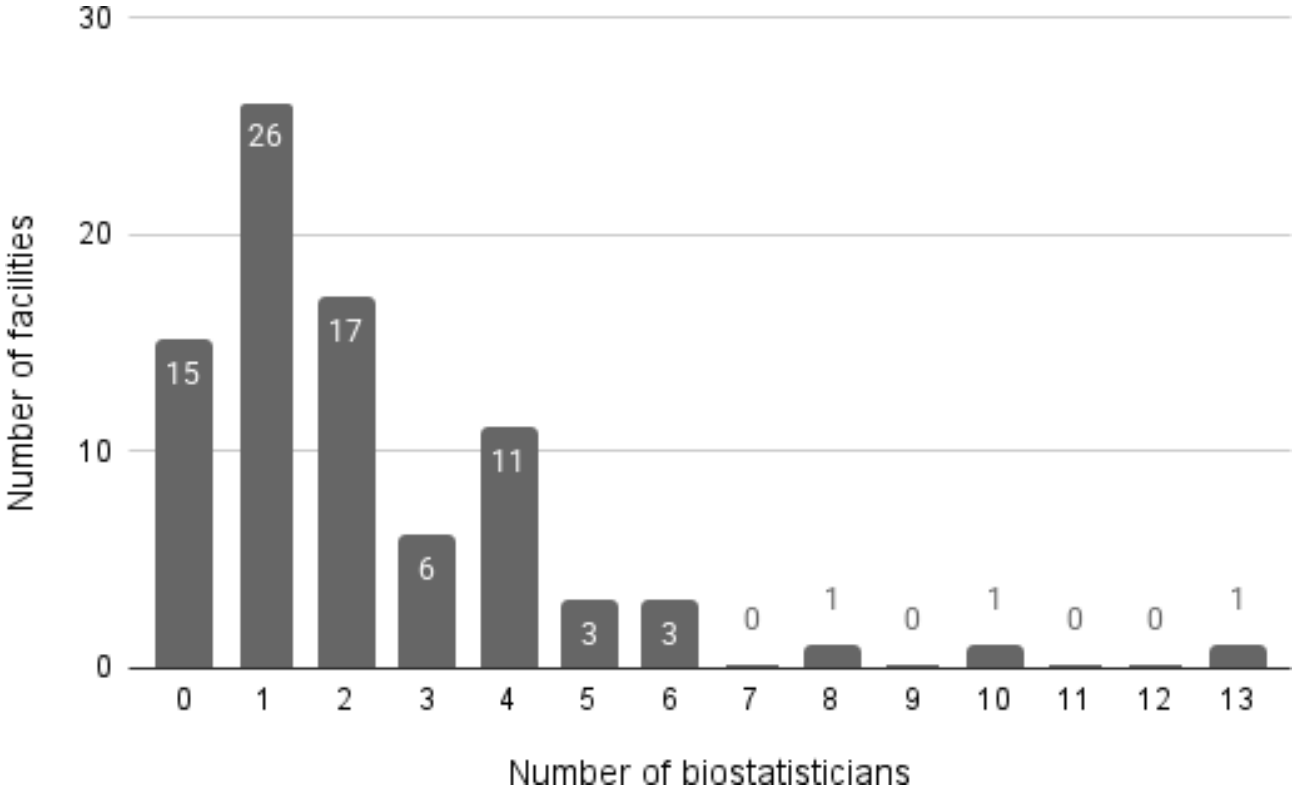


Figure 5. Number of biostatisticians at responding facilities (n=84). Note: Analysis excluded the 7 institutions that did not respond to the survey.



Association Between Staffing Levels and CDISC Standards Implementation Status

An analysis of the relationship between staffing levels and CDISC implementation showed that institutions with more data managers tended to be more advanced in their implementation (median 5, IQR 2 - 9.25 vs median 2, IQR 0 - 4; *P*<.001) (Figure 6). A similar trend was observed for biostatisticians (median 3.5, IQR 1.75 - 5 vs median 1, IQR 1-2; *P*<.001) (Figure 7).

XSL•FO
RenderX

Figure 6. Number of data managers and status of Clinical Data Interchange Standards Consortium standards implementation (n=84). Note: Analysis excluded the 7 institutions that did not respond to the survey. Data were categorized for visualization purposes, while the Mann-Whitney *U* test was performed on continuous variables (*P*<.001).

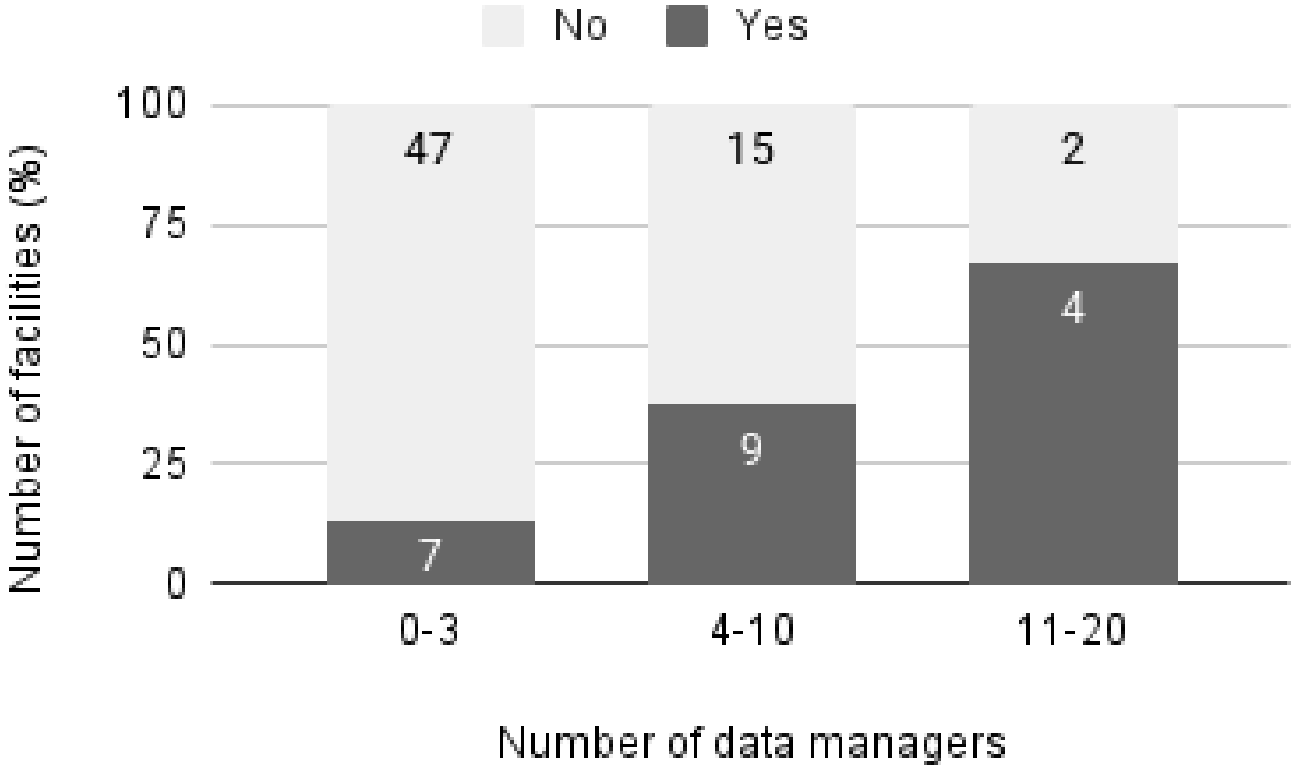


Figure 7. Number of biostatisticians and status of Clinical Data Interchange Standards Consortium standards implementation (n=84). Note: Analysis excluded the 7 institutions that did not respond to the survey. Data were categorized for visualization purposes, while the Mann-Whitney *U* test was performed on continuous variables (*P*<.001).

These findings suggest that institutions with more data managers and biostatisticians tend to be further along in implementing CDISC standards.

In addition to staffing levels, we evaluated the association between institution type and CDISC implementation. Implementation rates were 28.1% (16/57) in Public/National

XSL•FO
RenderX

institutions and 14.8% (4/27) in private institutions; however, no significant difference was observed (Fisher exact test, *P*=.27).

Discussion

Current Status and Significance of CDISC Standards Implementation

In a questionnaire survey conducted in 2020 to clarify the current status of and challenges associated with CDISC standards implementation in Japanese academic medical centers [16], the definition of "academia" was ambiguous, which posed challenges in interpreting the results and tracking changes over time. Therefore, in the present study, "academia" was more strictly defined to ensure the accuracy of longitudinal comparisons. By selecting institutions based on this clear definition, the present study could assess the current status and challenges associated with implementing CDISC standards more accurately, thereby providing foundational data to promote future adoption.

In a questionnaire survey conducted in 2020 to clarify the current status of and challenges associated with CDISC standards implementation in Japanese academic medical centers [16], the definition of "academia" was ambiguous, which posed challenges in interpreting the results and tracking changes over time. Therefore, in the present study, "academia" was more strictly defined to ensure the accuracy of longitudinal comparisons. By selecting institutions based on this clear definition, the present study could assess the current status and challenges associated with implementing CDISC standards more accurately, thereby providing foundational data to promote future adoption.

The findings revealed that only 29% (26/91) of institutions had experience with contracted CDISC standards data creation. Furthermore, institutions with experience in contracted data creation most frequently undertook investigator-initiated clinical trials for drug approval. This finding suggests that the application of CDISC standards in Japan, particularly that centered around investigator-initiated clinical trials for drug approval, is progressing. On the other hand, the present findings also confirmed that institutions had some experience in contracted CDISC standards data creation for specified clinical trials under the Clinical Research Act and other clinical research not intended for marketing approval applications. These findings suggest that CDISC standards are being used not only for regulatory submissions but also for standardizing and improving the efficiency of overall clinical research operations. In addition, it has been reported internationally that CDISC standards are also being used for Real-World Data [18]. We believe that clarifying the significance of standardization and efficiency and sharing practical case studies could lead to the promotion of CDISC standards implementation.

Regarding the implementation status of CDISC standards, the present study found that only 22% (20/91) of institutions had implemented them, indicating limited adoption. Notably, this study confirmed that while a high proportion of implementing institutions adopt multiple CDISC standards in combination, a certain number of institutions implement only individual standards. These results suggest that varying needs and resource situations at each institution may influence the form of CDISC standards implementation.

Barriers to CDISC Standards Implementation

Regarding implementation barriers, this study confirmed that "human and financial effort and resource shortages" remains the greatest challenge, followed by "a lack of personnel knowledgeable about CDISC." This finding suggests that securing human and financial resources and developing CDISC specialist personnel within institutions are crucial for promoting CDISC standards implementation. The findings also revealed that only 16% (15/91) of institutions possess documents or training related to CDISC standards, and furthermore, only 5% (5/91) of institutions can provide these resources to other institutions, which suggests that the expansion of CDISC-related educational resources within the country remains a challenge. We hypothesize that this discrepancy may stem from proprietary restrictions, copyright issues, or the lack of generalized templates that can be easily shared across institutions. In AROs in other countries as well, there is a demand not only for formal training by experts but also for open-source resources and "bite-sized" educational materials that are easily accessible online, such as the W3Schools model, as suggested by Jentoft et al [19]. Therefore, developing similar practical and accessible educational resources may serve as an important strategy for addressing the shortage of training materials in Japan.

Regarding barriers to implementation, 27 (30%) institutions cited "no need to implement CDISC standards." Within this group, some respondents cited "outsourced, hence no need." Furthermore, our detailed analysis revealed that the majority of these 27 (81%, 22/27) institutions selected "will not undertake" or "no decision made" regarding their future policies on CDISC standards-compliant data creation. The primary reason for these responses is likely that the PMDA requires CDISC compliance exclusively for new drug marketing approval applications, whereas many institutions primarily conduct research under the Clinical Research Act that falls outside this mandate. We included these institutions in our analysis to accurately reflect the current status of adoption, as excluding them would overlook the reality that a significant segment of academia operates outside the regulatory scope.

Regarding the respondents who cited outsourcing, it is important to distinguish this group from institutions that outsource but are considered to have "implemented" CDISC standards. As defined in our survey, "implementation" via outsourcing requires the institution to perform acceptance work using internal CDISC experts. In contrast, we interpret that respondents viewing outsourcing as a reason for nonimplementation may rely entirely on vendors without maintaining internal standards or oversight capabilities. However, as acknowledged in the Limitations section, the lack of explicit definitions in the survey may have led to variations in respondent interpretation regarding this distinction.

XSL•FO
RenderX

Future Policies and Opportunities for Strategic Support

Furthermore, the finding that a significant majority of institutions (66%, 60/91; Table 1) have made "no decision" regarding future CDISC policy suggests a deeper issue than mere resource constraints. Notably, this trend extends even to institutions that have already implemented CDISC standards. Half of these institutions (11/20) are uncertain about, or have decided against, continuing the implementation of CDISC-compliant data in the future, suggesting serious challenges in the sustainability of CDISC operations. This strategic inertia may indicate a systemic barrier, potentially stemming from the lack of regulatory incentives or clear mandates for the application of CDISC standards in clinical research not aimed at marketing approval. We considered that addressing this structural uncertainty is essential to enable institutions to commit to definitive standardization strategies.

While experienced AROs can play a pivotal role in driving adoption, the survey results indicate that 66% (60/91) of institutions specifically expect support from "pharmaceutical companies, CROs, and other AROs," with a primary desire for "materials and information." This highlights that academic institutions often cannot address the implementation gap on their own. Therefore, future strategies should not rely exclusively on peer-to-peer academic support but should also leverage industry resources. Active involvement and support from pharmaceutical companies and CROs were considered essential to provide the resources needed to bridge the gap between supply and demand.

Association Between Staffing Levels and Implementation

The present findings revealed significant variation among institutions in terms of the number of data managers and biostatisticians on staff. However, institutions with more of these specialists tended to be further along in regard to implementing CDISC standards. This result suggests that securing personnel with specialized knowledge is important for the effective implementation of CDISC standards, and as such, data managers and biostatisticians play a central role in the implementation process.

Limitations

A major limitation of this study is that adjustments for institutional resources such as hospital bed count or budget were not performed; these factors may be stronger determinants of CDISC adoption than staffing levels alone. The observed association between dedicated staffing and implementation may, therefore, be a reflection of overall institutional size or resource availability. Furthermore, the primary outcome variable, CDISC implementation status, was based solely on self-reported metrics from respondents and was not verified by external audit or inspection of the actual data standards used. This introduces the potential for response bias, such as the overestimation of institutional compliance. Finally, we acknowledge the nonresponse from specific institutional types; specifically, no responses were received from ministry-affiliated university hospitals. However, as this institutional category comprised only 1.1% (1/91) of the total institutions surveyed, the impact on the overall representativeness of Japanese academic medical centers is likely minimal. Furthermore, this study relies on self-reported data, which may be subject to respondent bias or misunderstanding of technical terms. For instance, a small number of sites reported implementing "ADaM only" without SDTM. Given that ADaM datasets are typically derived from SDTM, this discrepancy suggests a potential lack of accurate understanding of the standards among some respondents. Therefore, the results regarding technical proficiency should be interpreted with caution as they represent the respondents' subjective assessments rather than objectively verified capabilities. Furthermore, the questionnaire option labeled "outsourced, hence no need" was presented without explicit definitions distinguishing between outsourcing as a form of implementation and outsourcing as a reason for nonimplementation. Consequently, respondents may have selected this option based on their own interpretation. This lack of clarity may have led to misclassification, where institutions that had successfully implemented the system via outsourcing (with internal acceptance work) categorized themselves as having "no need."

Recommendations and Future Work

Therefore, developing these specialized human resources is likely essential for promoting their adoption. Expanding training opportunities and improving access to information regarding CDISC standards are key to promoting their wider adoption in Japanese academic medical centers.

Conclusions

The findings of the present study indicate that the implementation of CDISC standards in Japanese academic medical centers remains limited. The primary barrier identified was a lack of human and financial resources. On the other hand, many institutions are interested in implementing CDISC standards and seeking support such as workshops and the provision of materials. To bridge the gap between this high demand and the limited supply, future strategies should involve not only peer-to-peer academic support but also active resource sharing and cooperation from pharmaceutical companies and CROs. Notably, instances were observed where CDISC standards were being used even in clinical research not directly aimed at marketing approval applications, which is interpreted as a preliminary indication of the potential recognition within the research community of the operational efficiencies gained through standardization. Furthermore, the present findings suggest that the number of data managers and biostatisticians on staff is strongly associated with the implementation status of CDISC standards.

XSL•FO
RenderX

Acknowledgments

The authors would like to express their deep gratitude to the personnel of the 84 institutions that responded to the questionnaire, and also to the following individuals for their cooperation in this survey: Dr Hideki Hanaoka, Leader of TG3, National University Hospital Clinical Research Promotion Conference; Ms Mari Tsuchiya of its secretariat; and Ms Yoshiko Ito of the NHO Nagoya Medical Center. The authors used Google Gemini to assist in translating parts of the manuscript from Japanese to English. The authors thank FORTE Science Communications for English language editing.

Funding

This study was conducted as part of the Japan Agency for Medical Research and Development research and development project "Study of CDISC standards implementation in academia" (Principal Investigator: Toshiki Saito, NHO Nagoya Medical Center; Research Period: FY2022–FY2024) and supported by Japan Agency for Medical Research and Development under Grant Number JP22mk0101242 (https://www.amed.go.jp/koubo/11/03/1103C_00016.html). This study aims to comprehensively clarify the current status and challenges of implementing CDISC standards in Japanese academic medical centers.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KN, NI, YI, TIS
Data curation: KN
Formal analysis: KN
Funding acquisition: TIS
Investigation: KN, NI, TIS
Methodology: KN, NI, YI, TIS
Project administration: TIS
Resources: KN, NI, YI, TIS
Supervision: TIS
Validation: NI
Visualization: KN
Writing – original draft: KN
Writing – review & editing: KN, NI, YI, TIS
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire content.
[DOCX File, 11 KB - medinform_v14i1e83774_app1.docx]

Multimedia Appendix 2
Breakdown of the survey participants—roles and professions.
[DOCX File, 1898 KB - medinform_v14i1e83774_app2.docx]

References

1. Basic principles on electronic submission of study data for new drug applications. : Pharmaceuticals and Medical Devices Agency; 2014 URL: https://www.pmda.go.jp/files/000160019.pdf [accessed 2025-06-08]
2. Joining the Clinical Data Interchange Standards Consortium (CDISC) and activity status. Japan Agency for Medical Research and Development. URL: https://www.amed.go.jp/en/aboutus/collaboration/cdisc.html [accessed 2025-06-08]
3. Lessons learned from sharing clinical trial data using Vivli. : Clinical Data Interchange Standards Consortium; 2023 URL: https://www.cdisc.org/sites/default/files/2023-07/2023_CDISC_Japan_Tsukida_Vivli.pdf [accessed 2025-06-08]
4. Yin PT, Desmond J, Day J. Sharing historical trial data to accelerate clinical development. Clin Pharmacol Ther 2019 Dec;106(6):1177-1178. [doi: 10.1002/cpt.1608] [Medline: 31538661]
5. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Care Services; Board on Health Sciences Policy; Roundtable on Genomics and Precision Health; National Cancer Policy Forum; Forum on Neuroscience and Nervous System Disorders; Forum on Drug Discovery, Development, and Translation, Shore C, Hinners J, Khandekar E, Wizemann T. Data-sharing platforms. In: Reflections on Sharing Clinical Trial Data: Challenges

XSL•FO
RenderX

and a Way Forward: Proceedings of a Workshop: National Academies Press (US); 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK563333/ [accessed 2025-12-03]
6. ClinicalStudyDataRequest.com. URL: https://www.clinicalstudydatarequest.com/ [accessed 2025-12-03]
7. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015 May;20(5):464-e20. [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
8. Wold D, Walker S. CDISC BASIC Implementation Guide. Presented at: Presentation at CDISC 2022 US Interchange; Oct 26-27, 2022 URL: https://www.cdisc.org/sites/default/files/2022-11/2022-US-Interchange-Wold%2C%20Walker.pdf [accessed 2025-12-03]
9. Kato T. Clinical trial data sharing/utilization trends and company initiatives [Article in Japanese]. Clin Eval 2024;51(3):410-418 [FREE Full text]
10. For sharing individual participant/patient data (IPD) from clinical trials. : Japan Pharmaceutical Manufacturers Association; 2019 URL: https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc000000b2pz-att/ctds_nl_2019.pdf [accessed 2025-06-08]
11. Lo B. Sharing clinical trial data: maximizing benefits, minimizing risk. JAMA 2015 Feb 24;313(8):793-794. [doi: 10.1001/jama.2015.292] [Medline: 25594500]
12. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials. BMJ 2017 Jun 5;357(24):j2372. [doi: 10.1136/bmj.j2372] [Medline: 28584025]
13. CDISC standards [Article in Japanese]. University Hospital Medical Information Network (UMIN) Center. URL: https://www.umin.ac.jp/cdisc [accessed 2025-06-08]
14. CDISC: a global approach to accelerating medical research. : CDISC; 2011 URL: https://www.diaglobal.org/productfiles/25849/20110515/w5/01%20w5_rebecca%20d.kush.pdf [accessed 2025-06-08]
15. Kiuchi T. The application of CDISC standards to clinical research in Japanese academia and its significance. J Natl Inst Public Health 2019;68(3):202-211. [doi: 10.20683/jniph.68.3_202]
16. Nagai K, Ikeda Y, Ohno M, Saito TI, Hinotsu S. Current status and issues of CDISC standards implementation among academia in Japan [Article in Japanese with English abstract]. Jpn Pharmacol Ther 2022;50(Suppl 1):s51-s61 [FREE Full text]
17. Clinical trials act [Report in Japanese]. : Ministry of Health, Labour and Welfare (MHLW) of Japan; 2017 URL: https://www.mhlw.go.jp/file/06-Seisakujouhou-10800000-Iseikyoku/0000213334.pdf [accessed 2025-07-21]
18. Facile R, Muhlbradt EE, Gong M, et al. Use of Clinical Data Interchange Standards Consortium (CDISC) standards for real-world data: expert perspectives from a qualitative Delphi survey. JMIR Med Inform 2022 Jan 27;10(1):e30363. [doi: 10.2196/30363] [Medline: 35084343]
19. Jentoft K, Tustison E, Yu H. CDISC implementation in an academic research organization. J Soc Clin Data Manag 2022;2:3 [FREE Full text] [doi: 10.47912/jscdm.164]

Abbreviations

ADaM: Analysis Data Model
ARO: academic research organization
CDASH: Clinical Data Acquisition Standards Harmonization
CDISC: Clinical Data Interchange Standards Consortium
CRO: contract research organization
PMDA: Pharmaceuticals and Medical Devices Agency
SDTM: Study Data Tabulation Model

Edited by A Coristine; submitted 06.Oct.2025; peer-reviewed by A Cheng, C Decker; accepted 29.Jan.2026; published 06.Mar.2026.

Please cite as:
Nagai K, Ito N, Ikeda Y, Saito TI
Current Status, Challenges, and Opportunities Associated With Implementing Clinical Data Interchange Standards Consortium Standards in Japanese Academic Medical Centers: Cross-Sectional Survey
JMIR Med Inform 2026;14:e83774
URL: https://medinform.jmir.org/2026/1/e83774
doi:10.2196/83774

© Kaori Nagai, Noriko Ito, Yuya Ikeda, Toshiki I Saito. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated ICD-10–Anchored Classification of Primary Care Text Data: Development and Evaluation of a Custom Multilabel Classifier

Christina Haag[1,2], Dr. phil.; Thomas Grischott[3], PD, Dr med; Jakob M Burgstaller[3], Prof Dr med, Dr med dent, PhD, EMBA; Stefan Markun[3], PD, Dr med; Oliver Senn[3], Prof Dr med; Viktor von Wyl[1,2], Prof. Dr. sc. nat.

[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, Zurich, Switzerland

[2]Institute for Implementation Science in Health Care, University of Zurich, Universitätstrasse 84, Zurich, Switzerland

[3]Institute of Primary Care, University Hospital Zurich, University of Zurich, Zurich, Switzerland

Corresponding Author:
Viktor von Wyl, Prof. Dr. sc. nat.
Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, Zurich, Switzerland

Abstract

Background: Electronic medical records are a vast and valuable source of information, useful for tasks such as estimating disease prevalence. However, in routine primary care, much of this information is in free-text format rather than in a structured form and, therefore, not readily amenable to analysis. Manual coding of this textual data is both time-consuming and resource-intensive, making it impractical for large datasets. Although powerful open-source language models offer new opportunities for automated coding, their use on short heterogeneous primary care notes, particularly in German-language settings, remains insufficiently studied.

Objective: By providing hands-on guidance for applied health researchers, this study aims to demonstrate the effective and accurate automatic classification of free-text notes using a language model fine-tuned for automated *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) coding.

Methods: Building on the extensive Family Medicine Research Using Electronic Medical Records (FIRE) routine database from the Institute of Primary Care at the University Hospital Zurich and the University of Zurich, we trained a large language model–based multilabel classifier on a dataset of 38,728 free-text notes, which had been manually categorized into 47 classes using specific *ICD-10* codes and code ranges or nondiagnostic/ad hoc labels (eg, “unclear diagnosis,” “status post”). We stratified the labeled data into training (70%), validation (15%), and posttraining test (15%) sets, ensuring similar label distributions across these sets. Using the Transformers Python library, we trained the model over 10 epochs and evaluated it on the posttraining test dataset.

Results: Across 48 classes, the FIRE classifier achieved strong performance on the held-out posttraining set, with F_1-scores of 0.85 (micro, overall across all predictions), 0.86 (macro, mean of per-class scores treating classes equally), and 0.84 (weighted, per-class scores weighted by class frequency).

Conclusions: This study demonstrates steps for training open-source large language models and highlights the potential to streamline and scale the extraction of diagnostic information for practical applications. Our model can be robustly deployed, for example, for prescreening and labeling of free-text information, thus potentially reducing the burden of repetitive and error-prone manual handling.

(JMIR Med Inform 2026;14:e86533) doi:10.2196/86533

KEYWORDS

electronic medical records; free-text classification; ICD-10 coding; language models; natural language processing; automated coding; primary care; clinical text mining; health informatics

Introduction

Electronic medical records (EMRs) contain rich diagnostic information, but in routine primary care, much of this information is documented in short unstructured free-text format rather than as structured records. As a result, large-scale analysis and reuse of diagnostic data remains challenging, since manual coding of free-text notes is both time-consuming and resource-intensive, making it impractical on a large scale. Therefore, automated assignment of diagnostic codes, in particular from the *International Statistical Classification of*

XSL•FO
RenderX

Diseases, Tenth Revision (*ICD-10*), has long been of practical interest for research and health care.

The emergence of powerful open-source language models offers new opportunities to automate text-based diagnosis extraction for research or health care. Transformer-based models like bidirectional encoder representations from transformers (BERT) and generative pretrained transformers have gained traction for *ICD-10* coding of clinical texts, achieving strong performance on English language EMR datasets [1,2]. However, research on the classification of text data into *ICD-10*–anchored classes has been conducted mainly in English [1]. While recent studies demonstrate that domain-adapted transformer models can perform well on non-English clinical texts—including Portuguese [3], French [4], and Spanish [5,6]—the availability of data and pretrained models varies widely across languages. For German clinical texts, resources remain limited, and existing work has largely focused on well-structured, high-quality documents such as discharge summaries or surgery reports. In routine primary care documentation in particular, notes are often short, loosely structured, and heterogeneous. Diagnostic coding may also combine *ICD-10* codes with administrative or primary care–specific systems such as the *International Classification of Primary Care*. These characteristics limit the direct reuse of existing models and underscore the need for approaches that are robust to real-world data and adaptable to local coding practices.

Taken together, these factors—a scarcity of annotated German clinical text datasets from routine care, the predominance of English language research, and the short heterogeneous nature of primary care notes—leave little guidance on how to adapt open-source large language models (LLMs) for routine use in German primary care. To address this gap, our study describes the implementation of a flexible, German-language, multilabel LLM classifier for real-world free-text notes in primary care. Although focused on a German-language dataset, the methodological approach may serve as a blueprint in settings beyond English.

Methods

Data Source: The Family Medicine Research Using Electronic Medical Records Database

The Swiss Family Medicine Research Using Electronic Medical Records (FIRE) database was initiated in 2009 and has been designed as a basic knowledge base for primary care research through the use of EMRs [7]. So far, the FIRE project has enrolled around 450 general practitioners throughout Switzerland who are willing to contribute their patients' EMRs to the database. General practitioners in the FIRE project routinely document patient encounters—in clinics, at home, or by phone—and records are regularly exported to the FIRE database. The reasons for physician encounters are encoded using the *International Classification of Primary Care, 2nd edition*, albeit not systematically by all participating clinicians. Instead, many participating physicians often include unstructured free-text notes, particularly in the health problems and diagnoses fields, where diagnostic and health care use information is recorded.

Ethical Considerations

The local ethics committee of the Canton of Zurich waived approval for research with the FIRE database because patient data is fully anonymized and, therefore, outside the scope of the Swiss Human Research Act (BASEC-Nr. Req2017– 00797). The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines.

Diagnosis and Problem List Free-Text Notes

This study used manually annotated free-text notes from the diagnosis and problems lists in Swiss general practitioners' EMRs. In Swiss primary care, these lists function as longitudinal note-based records with brief entries summarizing patients' current and past health problems (eg, "hypertension," "low back pain") as well as related events or observations. Each note represents a single, concise observation that persists across consultations, as structured diagnostic coding is not routinely used in Swiss outpatient care.

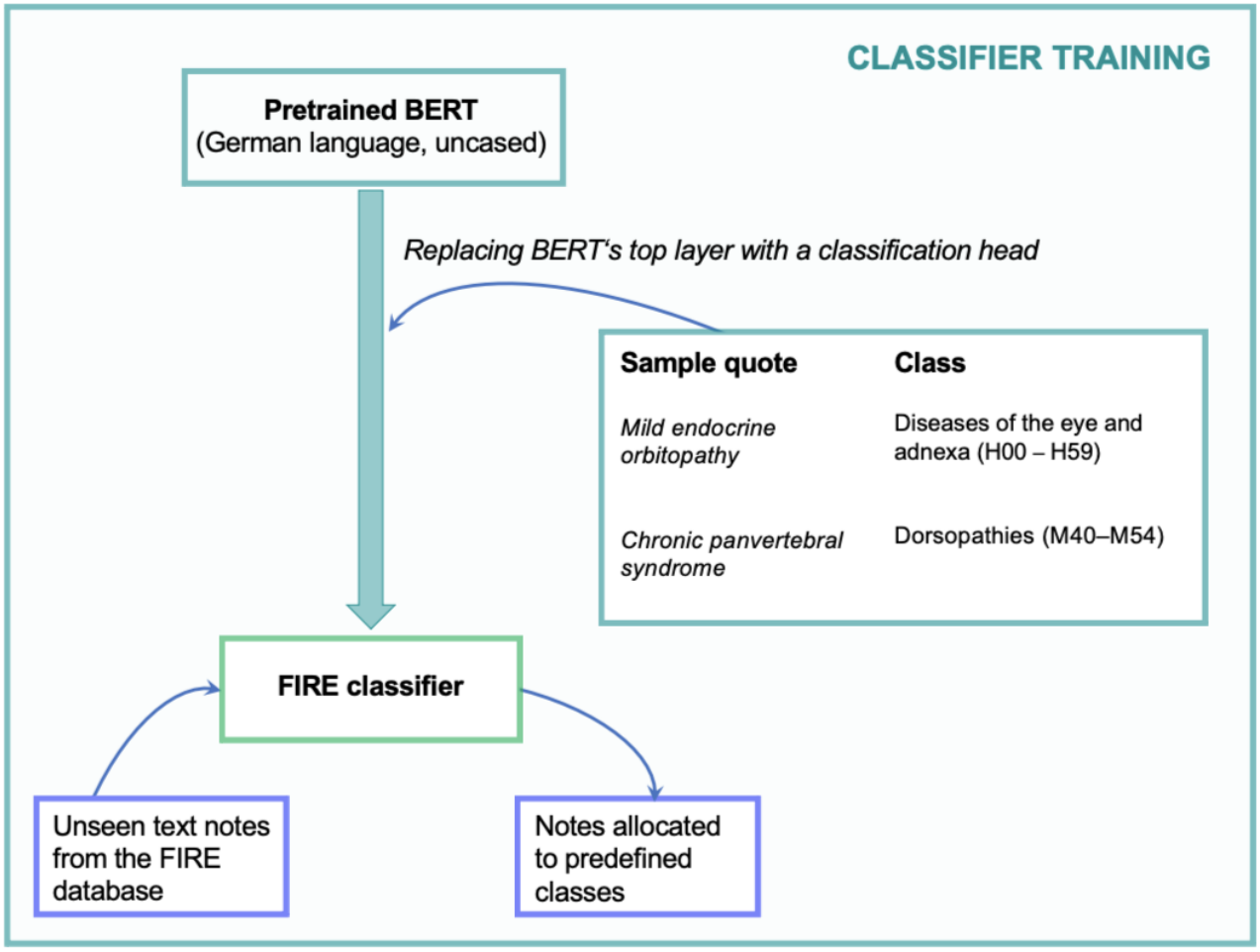
Training an LLM-Based Classifier for Automated Text Classification

Overview

All analyses and visualizations were performed in Python (version 3.11; Python Software Foundation), using the Jupyter Notebook environment (version 2024.3.2; Project Jupyter). Individual analysis steps and specific Python libraries are detailed in the following sections. The overall procedure is visualized in Figure 1. The Jupyter Notebook containing the Python code and complete model training output is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. The figure visualizes the modeling procedure to build the multilabel classifier based on the FIRE dataset. BERT: bidirectional encoder representations from transformers; FIRE: Family Medicine Research Using Electronic Medical Records.



Step 1: Ground Truth Data for Classifier Training

We worked with a total of 53,481 free-text notes, incorporating data from Wallnöfer et al [8] and additional notes annotated thereafter (n=26,501, 49.6%). Wallnöfer et al [8] developed an *ICD-10*–anchored framework comprising 105 diagnostic codes, including both individual *ICD-10* codes and code ranges, as well as a nonspecific "no diagnosis" category. For the present study, a subset of these codes was consolidated into broader groups by combining related diagnostic categories into higher-level ICD classifications, resulting in 44 categories. In addition, three less specific encounter codes not included in the original framework were introduced—"suspected diagnosis," "unclear," and "status post"—bringing the total to 47 categories. Reported interrater reliability for the original codes used in this study—either directly or in aggregated form—ranged from moderate to almost perfect (κ=0.53-0.98); however, these values are not directly transferable to the aggregated categories. For the Wallnöfer et al [8] dataset (n=26,980, 50.4%), two physician raters independently coded the notes (full details are reported in Wallnöfer et al [8]). The remaining 26,501 (49.6%) notes were annotated by a single trained physician.

Step 2: Data Preprocessing

To create a more balanced dataset, we reduced the number of entries for the most heavily overrepresented *ICD-10* code range Z00-Z99 "factors influencing health status and contact with health services," which occurred in 42.5% (n=22,714) of all free-text notes, either alone or in combination with other labels. Specifically, we removed all 14,316 free-text entries assigned exclusively to this code range, reducing the overall dataset to 38,728 free-text notes. As a result, this code range's representation was reduced to 8226 (21.2%), reflecting only those entries where Z00-Z99 appeared alongside other labels.

Step 3: Subdividing the Data Into Training, Validation, and Posttraining Test Data

We then divided the preprocessed dataset (n=38,728) into three parts: 70% (n=27,108) for training, 15% (n=5810) for validation during training, and 15% (n=5810) for final testing. To ensure each class label was proportionally represented in all three sets, we used a stratified splitting process based on the final list of 47 labels, using the *iterstrat* Python package and its MultilabelStratifiedShuffleSplit class for multilabel stratified splitting. In other words, if a particular class made up, for instance, 10% of the entire dataset, it also contained roughly 10% of the training set, 10% of the validation set, and 10% of the posttraining test set.

Step 4: Model Training

To build our classifier, we used a German BERT LLM (dbmdz/bert-base-german-uncased) from the Hugging Face

XSL•FO
RenderX

model hub [9] and replaced its top layer with a classification head to identify the 47 classes in unstructured medical free text. Tokenization was performed using the pretrained tokenizer associated with GermanBERT, a German-language–specific model; no multilingual or custom tokenization was applied (see executed Python script in Multimedia Appendix 1). The model's maximum input length was 512 tokens. No truncation was applied during training or inference, and the model always processed the complete note (median 15 tokens, mean 18 tokens, maximum 111 tokens). Since a single free-text note may belong to multiple categories, we used a multilabel classification approach to allow the assignment of several labels to each note. For model training, we used the Trainer class from the Transformers [10] Python library, which streamlines model training and metric calculation. We set the decision threshold for the binary predictions (true/false) at 0.5 per class.

Step 5: Classifier Evaluation During Training

We monitored evaluation metrics across all 10 epochs throughout training. During each epoch, the model iterated over all training batches, updating its weights via backpropagation based on the binary cross-entropy training loss.

Overall Model Evaluation

After each epoch, we evaluated model quality on the validation set. We first recorded training loss—the average discrepancy between predictions and ground truth on the training data; lower is better. We computed the same metric on the validation set (validation loss) to assess generalization: if validation loss falls and then rises while training loss keeps falling, the model is starting to memorize rather than generalize. We also tracked the microaveraged F_1-score (the harmonic mean of precision and recall), which gives more weight to more frequent classes. In addition, we measured the area under the receiver operating characteristic curve, summarizing how well the model separates positives from negatives across all thresholds (values near 1 indicate excellent separation; values near 0.5 indicate chance performance). Finally, because examples can have multiple labels, we reported subset (exact match) accuracy on the validation set—the fraction of examples for which every label is predicted correctly.

Per-Class Evaluation

Because each free-text note could belong to multiple classes, we examined model performance for each label individually using precision, recall, and F_1-score metrics. These metrics are particularly informative for multilabel classification because they capture how well the model identifies positive cases (recall), avoids false positives (precision), and balances both aspects (F_1-score). Accuracy was not used on a per-class basis, as it is dominated by true negatives in imbalanced datasets and may, therefore, provide a misleading indication of classifier performance.

Step 6: Classifier Evaluation on the Posttraining Test Data

The classifier was subsequently evaluated on the held-out 15% posttraining test set, which had not been used during model training or validation.

Overall Model Evaluation

To provide an overall measure of performance across all diagnostic labels, we calculated the micro-, macro-, and weighted averages of precision, recall, and F_1-score. The microaveraged metrics combine all true and false positives and negatives across labels, capturing the model's overall ability to classify instances correctly. The macroaveraged scores assign equal weight to each label, offering a balanced view that is not influenced by class frequency. In contrast, the weighted average metrics account for label prevalence, summarizing performance in proportion to how often each label occurs in the data.

Per-Class Evaluation

To complement the aggregate measures, the classifier's performance was also examined for each diagnostic category in the posttraining test set. For every label, precision, recall, and F_1-score were computed based on the corresponding ground truth and predicted binary vectors. These per-class metrics highlight how accurately the model distinguishes between individual diagnostic groups, allowing detailed insight into which classes were predicted most reliably and which proved more challenging.

Results

Overview

On average, each free-text note was assigned to 1.27 different classes, with the number of classes per free-text note ranging from 1 to 4. The class with the highest number of free-text notes was the umbrella class "factors influencing health status and contact with health services" (Z00-Z99; n=8226). Other classes with particularly high counts included "status post diagnosis" (n=3655) and "diseases of the musculoskeletal system and connective tissue" (M00-M99; n=2883). Conversely, classes such as "dementia" (F00-F03; n=97), "heart failure" (I50; n=116), and "acute upper respiratory infections" (J00-J06; n=163) were among those with the lowest number of free-text notes.

Monitoring Model Metrics

All model metrics during training and for the test data are provided in Multimedia Appendix 1.

Model Metrics During Training

Over the course of 10 training epochs, overall model performance reached a plateau, with the highest F_1-score reaching 0.85 achieved at epoch 10, indicating convergence of model learning. Overall model accuracy at this point was 0.75, and the area under the receiver operating characteristic curve equaled 0.92. The trained FIRE classifier demonstrated robust performance across the majority of classes; specifically, precision during training averaged 0.88 (SD 0.08), recall averaged 0.84 (SD 0.14), and the F_1-score averaged 0.86 (SD 0.11).

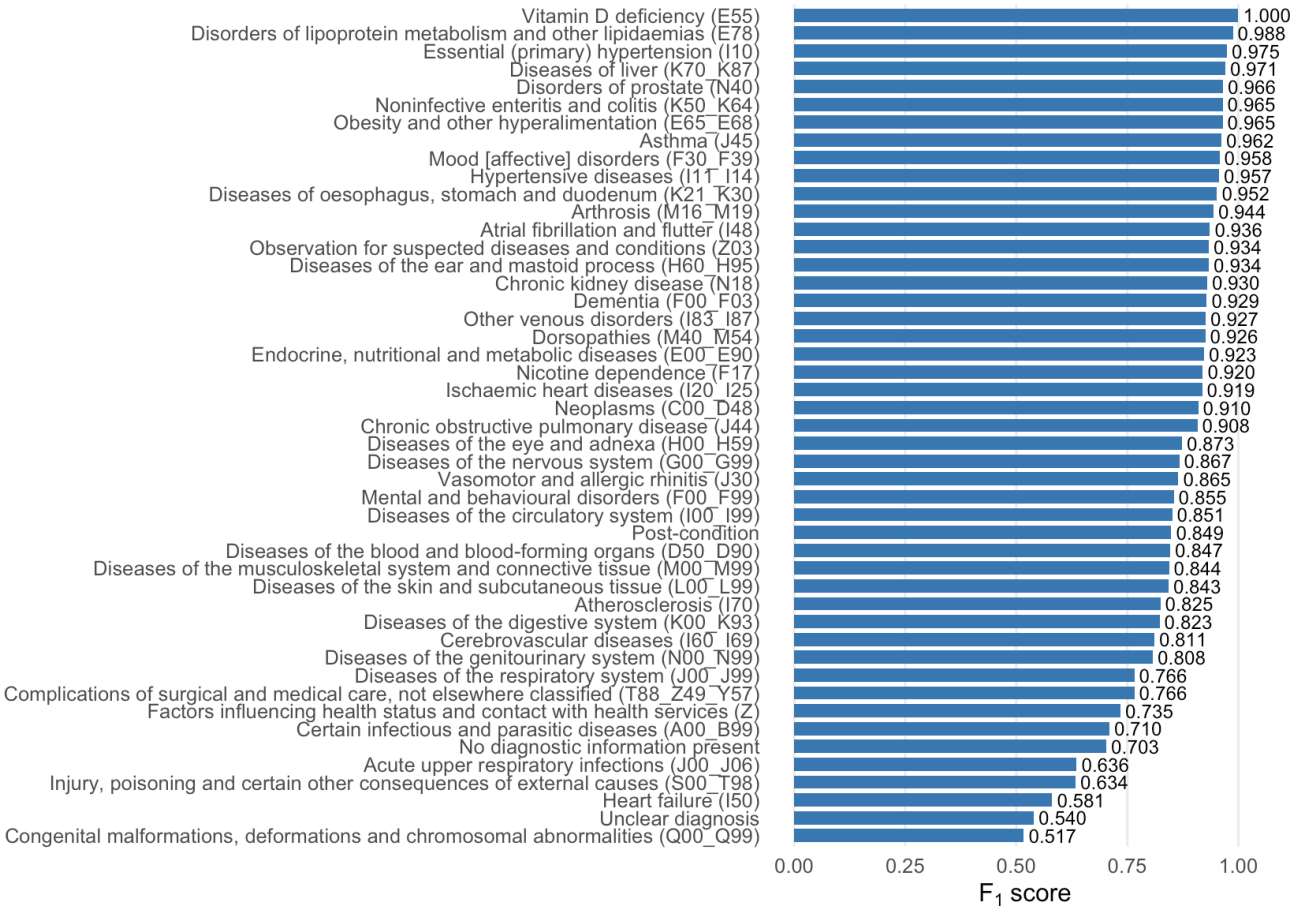
Performance on Posttraining Test Data

Across 47 classes, the FIRE classifier achieved strong performance on the held-out posttraining set, with F_1-scores of

XSL•FO
RenderX

0.85 (micro, overall across all predictions), 0.86 (macro, mean of per-class scores treating classes equally), and 0.84 (weighted, per-class scores weighted by class frequency). The classifier achieved the highest performance for "Vitamin D deficiency" (E55) with an F_1-score of 1.0, followed closely by "Disorders of lipoprotein metabolism and other lipidaemias" (E78), "Essential (primary) hypertension" (I10), and "Diseases of liver" (K70-K87), all exceeding 0.97. Strong results were also observed for a range of chronic and metabolic conditions such as obesity (E65-E68), "Asthma" (J45), and "Mood disorders" (F30-F39), with F_1-scores above 0.95. In contrast, performance declined for certain classes with fewer samples or greater diagnostic overlap, such as "Unclear diagnosis" (F_1-score 0.54) and "Congenital malformations, deformations and chromosomal abnormalities" (Q00-Q99; F_1-score 0.52). As displayed in Figure 2, F_1-scores for the posttraining test data were generally high, with variation across the different labels.

Figure 2. Performance of the Family Medicine Research Using Electronic Medical Records (FIRE) classifier on the posttraining test dataset. The bar chart displays the F_1-scores (x-axis) for each diagnostic class (y-axis) predicted by the FIRE classifier on the 15% test posttraining dataset, unseen during training. Larger bars indicate better classification performance.



Discussion

Extracting medical diagnoses or encounters from free-text EMRs or research data still poses critical practice challenges. By using a labeled dataset of ~53,000 records from a nationwide general practice research database, our study developed a classifier to predict appropriate codes in German free-text notes using a locally adapted diagnosis and medical encounter classification system. As such, our study offers insights into not only the challenges but also the potential rewards of creating locally adapted multiclassifier systems for medical free-text notes. Our multilabel classifier showed robust performance in classifying 47 different diagnostic and medical encounter categories during training and on unseen posttraining test data, with varying amounts of available labeled training data across categories. Our multilabel classifier was able to predict around half of the included diagnostic and medical encounter categories with high F_1-scores of 0.8 or greater. These findings showcase the potential usefulness of our classifier approach for systematic use in research.

Therefore, our study demonstrates the feasibility for health care providers to develop locally adapted text classification models designed to meet their unique data requirements and integration needs within operational workflows. We encountered challenges previously discussed in the literature [11], such as predicting a large number of classes simultaneously (a large label space) and dealing with an unbalanced label distribution, in particular with the predominant class "factors influencing health status and contact with health services." Despite preprocessing to mitigate its overrepresentation by removing free-text notes labeled with this class label, its high prevalence persisted due to co-occurrence with other classes, which we did not wish to thin out. This feature, while not ideal, reflects the real-world data scenario and underpins the relevance of our findings.

XSL•FO
RenderX

Importantly, despite these real-world challenges, our classifier maintained commendable performance on both common and rare classes, as well as on new, unseen data, providing valuable insights for health care providers interested in developing custom models amid similar data challenges.

Determining the optimal sample size for training multilabel classifiers remains a topic of debate, as there is no straightforward method such as the power analysis used in hypothesis testing. Our study shows that even for low-frequency classes (ie, with fewer than 100 labeled free-text notes available) such as hypertensive diseases, model performance can still be high. For health care providers looking to develop their own classifiers, our results suggest that the required sample size for each category could be fewer than 100 text samples when training a multilabel classifier of a similar size. However, it's important to note that the required sample size varies depending on how well the classes are defined by specific terms, such as "dementia," versus more ambiguous terms, such as "suspected diagnosis." Generally, our manual error investigation suggested that the quality of data labeling is of crucial importance.

There has been ongoing discussion about shifting toward prompt-based classification and moving beyond manually curated training data by leveraging industry-developed, general-purpose LLMs such as Google's Gemini or OpenAI's ChatGPT. However, a key drawback of currently available powerful general-purpose LLMs, such as OpenAI's GPT-4o, is that they are proprietary, closed-source, and tied to cloud-based paid services operated by large technology companies. This makes them unsuitable for sensitive clinical data, where local processing and data privacy are essential. Nonetheless, in the future, open-source general-purpose LLMs that can be run on local infrastructure are very likely to become more capable.

Our research has several limitations that merit consideration. We used a dataset specifically designed for this use case, which is characterized by its short text notes. Documents of different types, especially those with more extensive content such as admission or discharge summaries, might require alternative methods. For example, researchers dealing with longer documents may need to use models such as Longformer [12], which are designed to process longer text data. Most transformer-based models, including German BERT, typically have a default input constraint of 128 tokens, which can be extended to a maximum of 512 tokens—any text data longer than this is omitted. In the context of the German BERT tokenizer, a token is generally equivalent to a single word. Consequently, training LLMs on longer documents using Longformer models [12] could lead to different results. In addition, the text data was manually labeled by medical school graduates, and labeling may be prone to errors. Errors in data labeling can affect model performance and are a common challenge in the data labeling process. Finally, the predictive results of the model trained in this study are specifically tailored for use at the Institute of Primary Care.

An important consideration when interpreting our results is the *ICD-10* code range Z00-Z99 "Factors influencing health status and contact with health services," which primarily captures contextual or encounter-related information rather than explicit disease entities. In routine primary care documentation, such codes may function as a nonspecific "catch-all" category, particularly when diagnostic certainty is low. To mitigate potential bias, we excluded notes labeled exclusively with these encounter-related categories and retained them only when they co-occurred with disease-related labels, treating them as contextual modifiers rather than stand-alone outcomes. Nonetheless, its continued prevalence reflects real-world documentation practices. Notably, macro- and microaveraged F_1-scores were very similar, suggesting that performance was not driven solely by the dominant encounter-related class. Accordingly, overall performance should be interpreted with an emphasis on macroaveraged metrics and alongside per-class results (Figure 2 and Multimedia Appendix 1, which contains the executed analysis and full per-class results), providing a more balanced view across both common encounter codes and less frequent disease categories. Moreover, predictions of encounter-related codes should be interpreted as reflecting health care use context rather than disease burden, and the classifier should be viewed as a screening tool rather than an autonomous diagnostic system.

Beyond these methodological considerations, an important ethical issue is the potential for residual bias amplification given the imbalanced category distribution, mainly due to the overrepresented encounter-related codes. Although we excluded free-text notes labeled exclusively with encounter-related codes, the high prevalence of this class may still influence model behavior. For this reason, transparency regarding model design, training data, and class-specific performance is essential to support responsible use and clinician adoption. The classifier presented in this study is therefore best suited for screening and structuring heterogeneous clinical free-text notes, rather than as an autonomous diagnostic system, and it should not be used in settings requiring very high precision, such as downstream epidemiological analyses. Future research should focus on improving model explainability to support clinical acceptance and to facilitate the identification of potential biases during deployment.

This study uses supervised multilabel classification, which requires a sufficient number of positive examples per label to learn stable decision boundaries and produce reliable, interpretable performance estimates. We therefore focus on more prevalent *ICD-10* classes. Diagnostic categories with very low prevalence fall into a few-shot or zero-shot regime and cannot be robustly modeled within the same supervised fine-tuning framework. Although we explored prompt-based zero-shot label assignment using an open-source German LLM, performance was inconsistent and substantially inferior to supervised classification. Accordingly, this study concentrates on diagnostic categories with adequate training data for robust supervised model development. Rare conditions remain an important but distinct methodological challenge and may be better addressed through alternative paradigms such as few-shot retrieval-based methods or entailment-based classification.

For classification, we used *bert-base-german-uncased* as a well-established and widely used baseline model to demonstrate

XSL•FO
RenderX

supervised multilabel classification of German free-text notes in primary care. While alternative transformer architectures may offer performance gains, particularly through domain adaptation or different pretraining objectives, systematic comparisons were beyond the scope of this work. Relevant candidates for future evaluation include domain-adapted German biomedical models such as medBERT.de [13], GottBERT [14], and Me-LLaMA 70B [15], as well as alternative pretraining paradigms such as German ELECTRA [16], which has shown strong performance in text classification tasks. In addition, multilingual architectures (eg, the multilingual mBERT [17]) may be of interest for cross-lingual transfer scenarios. Future research would benefit from systematically assessing trade-offs between predictive performance, computational requirements, and deployment feasibility when applying these architectures to real-world clinical settings.

In summary, our study demonstrates the feasibility and benefits of training tailored multilabel text classifiers to categorize diagnostic physician notes in primary care. Combined, our work and documentation of the classifier development process lay a solid foundation for applied researchers and health care providers to leverage their own rich datasets for sophisticated statistical analyses and rigorous research. Moreover, after further validation, such multilabel classifiers could potentially also be used for generating more meaningful, precise, and actionable feedback to the participating general practitioners and thus help them improve the quality of care for their patients. We believe that our work is particularly relevant for health care organizations that accumulate vast amounts of data over time, including rich open text records that often exceed the amount of data typically collected by large research studies in a feasible time frame and budget. For example, many inpatient clinics have extensive longitudinal data, both quantitative and qualitative, including detailed patient histories across multiple admissions. However, our work also underscores the importance of high-quality data labeling and thoughtful integration of artificial intelligence into health care workflows to meaningfully enhance patient care.

Funding

This study is supported by a grant from the Federal Quality Commission c/o Federal Office of Public Health, Bern, Switzerland. The funder had no role in study design, data collection and analysis, preparation of the manuscript, and decision to publish.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Jupyter Notebook containing the Python code and complete model training output.
[PDF File, 766 KB - medinform_v14i1e86533_app1.pdf]

References

1. Kaur R, Ginige JA, Obst O. AI-based ICD coding and classification approaches using discharge summaries: a systematic literature review. Expert Syst Appl 2023 Mar;213:118997. [doi: 10.1016/j.eswa.2022.118997]
2. Liu L, Perez-Concha O, Nguyen A, Bennett V, Jorm L. Automated ICD coding using extreme multi-label long text transformer-based models. Artif Intell Med 2023 Oct;144:102662. [doi: 10.1016/j.artmed.2023.102662] [Medline: 37783551]
3. Coutinho I, Martins B. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. J Biomed Inform 2022 Dec;136:104232. [doi: 10.1016/j.jbi.2022.104232] [Medline: 36307020]
4. Arnaud É, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. Presented at: 15th International Joint Conference on Biomedical Engineering Systems and Technologies; Feb 9-11, 2022; Vienna, Austria p. 835-841. [doi: 10.5220/0011012800003123]
5. Padilla Cuevas J, Reyes-Ortiz JA, Cuevas-Rasgado AD, Mora-Gutiérrez RA, Bravo M. MédicoBERT: a medical language model for Spanish natural language processing tasks with a question-answering application using hyperparameter optimization. Appl Sci (Basel) 2024;14(16):7031. [doi: 10.3390/app14167031]
6. Iglesia I, Atutxa A, Gojenola K, Barrena A. EriBERTa: a bilingual pre-trained language model for clinical natural language processing. arXiv. Preprint posted online on Jun 12, 2023. [doi: 10.48550/ARXIV.2306.07373]
7. Chmiel C, Bhend H, Senn O, Zoller M, Rosemann T, FIRE study-group. The FIRE project: a milestone for research in primary care in Switzerland. Swiss Med Wkly 2011;140:w13142. [doi: 10.4414/smw.2011.13142] [Medline: 21279858]
8. Wallnöfer A, Burgstaller JM, Weiss K, Rosemann T, Senn O, Markun S. Developing and testing a framework for coding general practitioners' free-text diagnoses in electronic medical records - a reliability study for generating training data in natural language processing. BMC Prim Care 2024 Jul 16;25(1):257. [doi: 10.1186/s12875-024-02514-1] [Medline: 39014311]
9. MDZ Digital Library team. Bert-base-german-uncased. Hugging Face. 2023.
10. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv. Preprint posted online on Oct 9, 2019. [doi: 10.48550/arXiv.1910.03771]

XSL•FO
RenderX

11. Yan C, Fu X, Liu X, et al. A survey of automated International Classification of Diseases coding: development, challenges, and applications. Intell Med 2022 Aug;2(3):161-173. [doi: 10.1016/j.imed.2022.03.003]
12. Beltagy I, Peters ME, Cohan AL. The long-document transformer. arXiv. Preprint posted online on Apr 10, 2020. [doi: 10.48550/arXiv.2004.05150]
13. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
14. Scheible R, Thomczyk F, Tippmann P, Jaravine V, Boeker M. GottBERT: a pure German language model. arXiv. Preprint posted online on Dec 3, 2020. [doi: 10.48550/ARXIV.2012.02110]
15. Xie Q, Chen Q, Chen A, et al. Me LLaMA: foundation large language models for medical applications. arXiv. Preprint posted online on Feb 20, 2024. [doi: 10.48550/ARXIV.2402.12749]
16. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv. Preprint posted online on Mar 23, 2020. [doi: 10.48550/ARXIV.2003.10555]
17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]

Abbreviations

BERT: bidirectional encoder representations from transformers
EMR: electronic medical record
FIRE: Family Medicine Research Using Electronic Medical Records
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
LLM: large language model

Edited by A Benis; submitted 26.Oct.2025; peer-reviewed by K Triep, M Elbattah; revised version received 18.Feb.2026; accepted 06.Mar.2026; published 06.Apr.2026.

Please cite as:
Haag C, Grischott T, Burgstaller JM, Markun S, Senn O, von Wyl V
Automated ICD-10–Anchored Classification of Primary Care Text Data: Development and Evaluation of a Custom Multilabel Classifier
JMIR Med Inform 2026;14:e86533
URL: https://medinform.jmir.org/2026/1/e86533
doi:10.2196/86533

© Christina Haag, Thomas Grischott, Jakob M Burgstaller, Stefan Markun, Oliver Senn, Viktor von Wyl. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Sentence Classification–Based Medical Status Extraction Pipeline for Electronic Health Records: Institutional Case Study

Chuanming Dong[1], PhD; Boris Delange[1], MD; Alex Poiron[1], MSc; Mohamed El Azzouzi[1], PhD; Clément François[2], MSc; Guillaume Bouzillé[1], MD, PhD; Marc Cuggia[1], PhD, MD; Sandie Cabon[1], PhD

[1]Univ Rennes, CHU Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France, Rennes, France

[2]Kereval (France), Rennes, France

Corresponding Author:
Sandie Cabon, PhD
Univ Rennes, CHU Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France, Rennes, France

Abstract

Background: Clinical data warehouses store large volumes of unstructured text containing valuable information about patients' medical status. Traditional extraction systems based on named entity recognition (NER) identify medical terms but often fail to capture the contextual cues needed for accurate interpretation. Existing approaches to context-aware extraction differ in their reliance on expert annotation, computational power, and lexical resources, leading to uneven feasibility across institutions. Combined with heterogeneity in documentation practices and data-sharing restrictions, these limitations hinder the scalability and reuse of trained models. There is thus a need for practical frameworks that can be deployed and adapted locally within medical institutions.

Objective: This study aimed to introduce the Medical Status Extraction Pipeline (MSEP)**,** a methodological framework that extracts patients' medical status from clinical narratives through sentence classification and supports the local deployment of hybrid extractors, illustrated through an institutional case study.

Methods: MSEP extracts medical status by classifying sentences into predefined categories (presence, absence, or unknown) for each targeted condition. The pipeline combines modules for data selection, expert annotation, and model development, with parameters customizable to different settings. It was applied within our institutional environment on 6 conditions: smoking, hypertension, diabetes, heart failure, chronic obstructive pulmonary disease, and family history of cancer, using 12,119 manually annotated sentences from the eHOP Clinical Data Warehouse (Rennes University Hospital). Three types of extractors were compared: fine-tuned CamemBERT, large language model (LLM) prompt, and a rule-based baseline, evaluated through stratified 3-fold cross-validation, measuring precision, recall, specificity, macro *F*-score, balanced accuracy, as well as manual annotation time and model inference speed.

Results: Among the tested approaches, the CamemBERT-based extractor achieved the best overall performance, with macro *F*-scores above 0.94 for 5 of the 6 medical conditions. The study also highlights that when a medical status is very sparsely represented in the training data, rule-based extractors can outperform learned models (average macro *F*-score 0.94 vs 0.73 for family history of cancer). This shows the pragmatic value of choosing the extraction method according to data availability. Manual annotation time per sentence ranged from 1.2 to 2.9 seconds within the pipeline (2.23 to 4.25 seconds for informative sentences), compared with 7.8 to 16.5 seconds for named entity recognition–based systems. In our institutional experiments, the minimum time to complete all pipeline modules, from dataset construction to final extractor refinement, was 8 hours.

Conclusions: In our institutional case study, MSEP enabled rapid construction of datasets and extractors across multiple clinical conditions while reducing the effort required for local development. Its modular and configurable design allowed the adoption of hybrid extraction approaches and adaptation to different resource settings. These features highlight MSEP's value as a research tool and upstream component that facilitates local deployment of clinical information extraction workflows.

(JMIR Med Inform 2026;14:e77409) doi:10.2196/77409

KEYWORDS

sentence classification; information extraction; natural language processing; clinical data warehouse; deep learning; large language model; artificial intelligence

XSL•FO
RenderX

Introduction

In recent years, a substantial amount of clinical data has been collected from numerous patients and stored in clinical data warehouses (CDWs) [1]. These databases represent a valuable resource for clinical research. In many cases, a significant portion of their content exists as free unstructured texts, such as hospital discharge summaries and emergency reports, which can be difficult to process [2]. Nonetheless, these texts contain valuable information about patients' medical status that, once extracted, can be leveraged for clinical and epidemiological research [3-5]. Harnessing these unstructured data requires developing natural language processing (NLP) methods specifically adapted to the medical domain. Extensive efforts have been made in this area, as reflected in shared tasks such as the n2c2 (National NLP Clinical Challenges) competitions, which have addressed problems of extracting patients' family history regarding diseases, smoking, suicide, and drinking [6].

According to a recent review, most NLP approaches for medical information extraction rely on named entity recognition (NER) [2]. At first, rule-based approaches using dictionaries and terminologies to match terms and concepts were proposed [7,8]. However, their reliance on lexical resources and medical expertise for rule construction limits their long-term viability, since rules and vocabularies require regular manual updates to reflect evolving clinical practices. To cope with this, machine learning approaches arose. These methods take advantage of the large amount of medical knowledge in databases and can automate the construction and maintenance of medical information retrieval systems [9]. Its latest branch, deep learning approaches, has improved the ability to capture and use contextual information compared with rule-based and traditional machine learning methods, thanks to their neural network architectures; accordingly, this also makes their performance highly dependent on the availability of sufficient annotated data [10,11].

However, NER-based information extraction approaches present a significant limitation. While NER can effectively extract medical entities (eg, "cancer" and "hypertension"), it cannot capture the contextual information required for a full understanding of the patient's status. A comprehensive assessment involves interpreting the complete syntactic and semantic structure, including subject (patient or family member), predicate (having or not having), and object (the medical condition), to determine whether a condition is present, absent, or mentioned in a familial context [12-14]. Standard NER captures only the object component, necessitating additional processing by context qualifiers to detect negation, temporality, and subject attribution [6,15-17]. Clinical narratives further amplify these challenges, as they are typically unstructured, context-dependent, and institution-specific, with such cues often expressed at the sentence or discourse level [18,19]. For example, in "Tabagisme pendant 7 ans, reprise récente malgré un sevrage en 2020" (Smoking for 7 y, recent relapse despite cessation in 2020), successful interpretation requires resolving the temporal relation between cessation and relapse to infer that the patient is an active smoker, dependencies that standard NER pipelines, which treat entities and assertions separately, struggle to capture. These linguistic characteristics highlight the need for approaches more capable of sentence-level semantic interpretation.

At the same time, developing context-aware information extraction systems can be resource-intensive. The development and evaluation of NER models require extensive expert annotation to label entity boundaries and associated features within large-scale corpora [20], a process made even more laborious by the sparsity of clinically relevant content [21-23]. With the advancement of deep neural networks, embedding fusion techniques have been adopted to enhance context utilization and improve recognition accuracy in NER tasks [24]. However, their cross-domain applicability is limited by high computational cost and reliance on external resources for comprehensive feature engineering. To overcome the limitations of traditional NER pipelines and handcrafted feature engineering, recent research has shifted toward the use of transformer-based pretrained language models that inherently capture richer contextual representations [25]. Encoder-only language models such as BERT [26] and CamemBERT [27] have demonstrated high performance in domain-specific tasks through fine-tuning on relatively small annotated datasets [28]. Decoder-only large language models (LLMs), such as ChatGPT [29] and Mistral [30], can achieve similar goals via prompt engineering without requiring fine-tuning. However, both approaches still depend on domain-specific expertise for model adaptation [31,32], and decoder-only models introduce substantial computational costs that limit their feasibility for routine institutional deployment [33].

Beyond technical constraints, institutional heterogeneity, including differences in documentation practices, terminologies, and privacy regulations, presents major barriers to developing and sharing information extraction systems across institutions [34-36]. Clinical documentation varies widely across institutions, incorporating nonstandard abbreviations, local terminology, and inconsistent formatting [19,37], which hinders the generalizability and scalability of information extraction models, requiring substantial annotation and model adaptation for each use case [19,38]. The sensitive nature of patient data further restricts access to real-world electronic health records and limits the sharing of pretrained models, due to risks of unintended information leakage [39,40]. Consequently, research teams often need to rebuild extraction systems from scratch, even when addressing similar clinical objectives [16,41]. In light of these limitations, there is a critical need for practical, context-aware information extraction systems that can be developed and reused within institutional boundaries.

To address these limitations, we designed the Medical Status Extraction Pipeline (MSEP) as a methodological framework that supports the local development of hybrid approaches for information extraction. MSEP reframes medical status extraction as a sentence classification task, which directly leverages contextual cues such as negation, temporality, and subject attribution while reducing annotation complexity. In addition, by accommodating rule-based methods, fine-tuned language models, and LLM prompting, MSEP allows institutions to select extraction approaches suited to their data availability and computational resources.

XSL•FO
RenderX

To support its interest, the MSEP has been applied to build extractors of patients' status for 6 clinically relevant medical conditions [16,42,43] from a French CDW: smoking, hypertension, diabetes, heart failure, chronic obstructive pulmonary disease (COPD), and family history of cancer. This paper first outlines the steps, functions, and requirements of the MSEP pipeline, along with the materials and tools used. Next, we present the evaluation results of extractors built using MSEP and compare them with different medical information retrieval approaches (rule-based methods and prompt-based LLM extraction). Finally, we discuss the operational benefits of MSEP, particularly reduced annotation effort, and its potential use as an upstream research tool within institutional settings.

Methods

Overview

The MSEP pipeline classifies sentences in clinical documents as present, absent, or unknown regarding a patient's medical condition. Available as a Python package (repository link in Multimedia Appendix 1), the framework includes configurable parameters that allow its components to be adapted to local research needs and institutional constraints. In the following sections, we first describe the general design of the pipeline and then detail its implementation within our institutional setting.

Ethical Considerations

Ethics approval for data reuse was obtained from the Commission Nationale de l'Informatique et des Libertés (agreement number 2206739). This study was conducted on pseudonymized secondary clinical data provided by the Clinical Data Center of Rennes Hospital. In accordance with French regulations and the guidelines of the Commission Nationale de l'Informatique et des Libertés. Patients were informed of the potential use of their data for research purposes through the hospital's standard information notice. All data were pseudonymized by the Clinical Data Center before being transmitted to the research team. No identifying information was accessible at any stage of the analysis. No compensation was applicable or provided to participants.

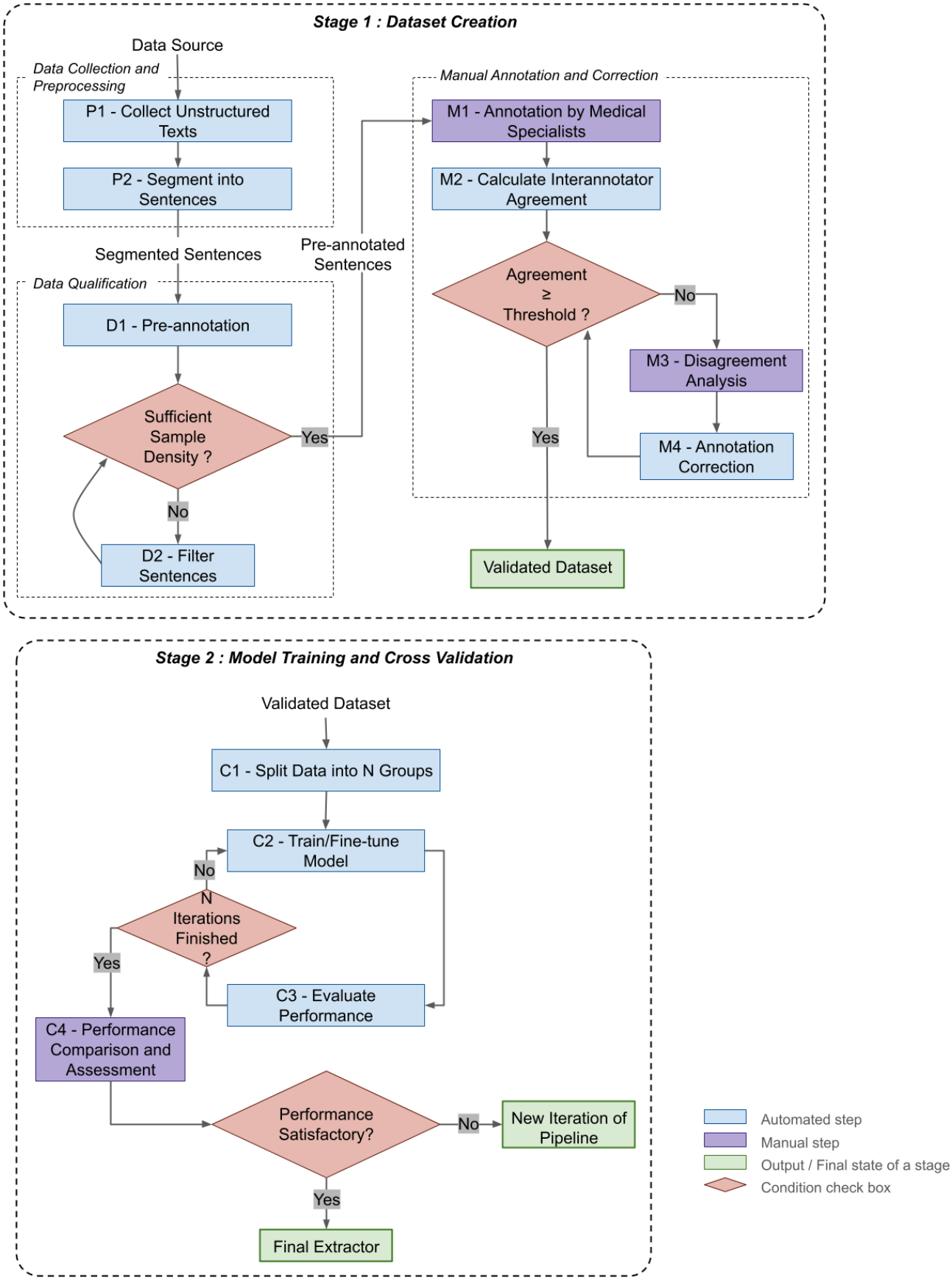
Generic Pipeline Design

Overview

The global framework of our method is depicted in Figure 1. The framework consists of two main stages: (1) dataset creation and (2) model training with cross-validation. In order to facilitate the comprehension of the framework, the abbreviations of the steps in the figure are also mentioned in the text.

XSL•FO
RenderX

Figure 1. Overview of the proposed iterative medical status extraction system. This pipeline is primarily designed to extract the status of one medical condition at a time; it may repeat multiple times (iterations) to improve the extractors. This diagram mainly shows the process of the first iteration.



XSL•FO
RenderX

Dataset Creation

Overview

The objective of this stage is to construct the datasets required for training and validating medical status extractors. It involves three processes: (1) data collection and preprocessing, (2) data qualification, and (3) manual annotation and correction.

Data Collection and Preprocessing

Data collection gathers unstructured texts from CDW (P1), which are preprocessed by segmenting into sentences for classification (P2). The output is a corpus of sentences with relevant metadata that maintains the relationship to their source documents.

Data Qualification

The input to the data qualification process is the preprocessed sentence corpus. This phase assesses whether the collected data have appropriate sample density and size for effective model training. Sentences undergo automatic preannotation (D1), preferably using quickly implemented rule-based approaches that match medical status keywords. When preannotating sentences from all documents is impractical due to time and resource constraints, a random subset may be selected. If sample density is low, based on preannotation results, a portion of sentences unlikely to contain medical status information is filtered out (D2). This supervised sampling approach is designed to enhance annotation efficiency and final data quality by optimizing the training sample structure. The output is a preannotated sentence dataset with optimized annotation density and a balanced representation of target medical statuses.

Manual Annotation and Correction

The preannotated sentences serve as input for manual annotation, which requires multiple medical specialists to classify them according to annotation guidelines (M1). Annotators should collaboratively annotate a subset of the sentences to calculate interannotator agreement (M2). If the agreement exceeds the predetermined threshold, the annotations form the training and validation datasets. Otherwise, disagreements are analyzed (M3) to improve guidelines, with corrections applied to the annotated data accordingly (M4). The correction approach depends on the type of disagreements, but preferably through manual review. The disagreements analysis-correction process can be repeated until the interannotator agreement exceeds the threshold. The output is a validated annotated dataset ready for model development.

Model Training and Cross-Validation

The annotated datasets become the input for training or fine-tuning neural network models for medical status extraction. Since the pipeline is designed to extract the status of one medical condition at a time, each corresponding model is trained in a multiclass classification setting, where the possible status labels (eg, present, absent, and unknown) are mutually exclusive. Model stability is assessed through cross-validation by dividing annotated data into several groups (C1), using one group for validation (C3) and the remainder for training (C2) in each session. The number of cross-validation sessions equals the number of groups of data. The data grouping strategy for cross-validation should align with study objectives; for example, testing model stability across different document types requires data partitioning by document type. The extractor's performance metrics across multiple validation sets are assessed and compared with extractors based on other approaches, if available, to decide whether it has stable and state-of-the-art performance (C4). The flexibility of the pipeline configuration allows for easy benchmarking of different models (such as BERT-based models, rule-based systems, or LLM prompts), facilitating comprehensive performance comparison with minimal additional setup.

If the assessment result is positive, the pipeline ends, and the output is a trained medical status extraction model with satisfactory performance.

Iterative Improvement Process

If, according to the assessment step (C4), the extractor does not achieve satisfactory performance (eg, a macro *F*-score below 90%), the pipeline allows for iterative refinement. The definition of "satisfactory" performance may vary depending on the objectives and standards of different research projects.

The iterative process involves both quantitative and qualitative improvements. Quantitatively, additional annotated sentences are incrementally introduced to expand the training and validation datasets. The number of added samples per iteration is not fixed and depends on multiple factors, including the time allocated for manual annotation, the number and availability of annotators, and computing resources.

Qualitatively, the pipeline may be refined by improving preannotation rules, updating annotation guidelines based on annotator feedback, or replacing rules with more accurate extractors from previous iterations for preannotation. The first iteration serves as an exploratory phase to identify bottlenecks, such as insufficient data for certain labels, weak vocabulary coverage in preannotation rules, or annotation inconsistencies, allowing for targeted improvements in subsequent iterations.

Study-Specific Implementation

Data Source and Working Corpus

We used the eHOP CDW from Rennes University Hospital, which contains 80 million unstructured and 430 million structured data elements covering over 1.8 million patients [44]. For this study, we extracted unstructured documents from our previous bladder cancer research dataset [3], obtaining 799,470 documents from 5398 patients. This corpus encompasses diverse medical document types including nursing care reports, medical reports, multidisciplinary follow-up & daily notes, imaging reports, prescriptions, and medication orders (Table 1).

XSL•FO
RenderX

Table . Numbers and ratios of the 9 most frequently found document types (as well as the rest of the 5020 documents in "Other") in our corpus.

Medical document type	Value, n (%)
Nursing care–procedures	329,359 (41)
Medical reports	187,370 (23)
Nursing care–monitoring and vitals	153,865 (19)
Multidisciplinary follow-up and daily notes	54,187 (7)
Imaging reports	31,430 (4)
Prescriptions and medication orders	13,163 (2)
Paramedical care	13,083 (2)
Administrative and admission	6839 (1)
Results and pathology reports	5225 (1)
Other	5020 (1)

Rationale for Extraction Method Design

MSEP is designed as a hybrid framework capable of accommodating different resource-availability scenarios. To assess this flexibility, we evaluated 3 extraction approaches that represent distinct requirements: fine-tuning an encoder-only transformer (high annotation demand and moderate computational cost), rule-based extraction (minimal computation and annotation requirements but substantial reliance on domain expertise), and LLM prompting (low annotation requirements but high computational and expertise demands) [45]. Our experiments focused primarily on encoder-only fine-tuning in order to assess how effectively MSEP can reduce the associated annotation burden in practice, while the rule-based and LLM-prompting approaches served as comparative baselines for low-annotation settings. The rule-based approach is used for preannotation (D1) before fine-tuning, reinforcing the hybrid nature of the MSEP framework.

Regarding the models and tools used in these approaches, CamemBERT was selected for encoder-only fine-tuning strategy because it aligns with the sentence-classification paradigm and has demonstrated strong performance in French NLP tasks [25,46,47]. Furthermore, we opted for the original CamemBERT rather than domain-adapted variants such as CamemBERT-bio [48], as the latter are pretrained on specialized biomedical corpora, which may bias them toward known biomedical tasks and entities [49,50]. Using a general-domain model allows us to evaluate the pipeline's components without relying on domain-specific pretraining and to examine how it behaves in our institutional corpus when applied to varied clinical conditions. For LLM prompting, we used Mixtral-8×7B-v0.1 [51], which offered the strongest performance that could be efficiently supported by our institutional infrastructure (an NVIDIA A100 40 GB GPU). The rule-based extractors were developed using regular expressions and terminology specific to each targeted medical condition, provided by medical experts.

Pipeline Configuration

The pipeline was independently evaluated on 6 medical conditions: smoking, hypertension, diabetes, heart failure, COPD, and family history of cancer. For each condition, we defined 3 possible statuses: "present," "absent," and "unknown," corresponding to confirmed presence, negation, and uncertainty regarding the patient's condition. For the smoking condition, an additional status, "former," was included to indicate that the patient previously smoked but has since quit (Table S1 in Multimedia Appendix 2 provides a detailed definition of medical statuses). Multimedia Appendix 2 contains the annotation guidelines, which give more details about the criteria for annotating the medical status.

We used Spacy's fr_core_news_md model [52] for sentence segmentation (P2) and manually crafted rules based on specialist-provided keywords for preannotation (D1), supplemented by EDS-NLP's negation and family context qualifiers [53] for smoking and family history of cancer. Two medical specialists performed annotations (M1) using the Prodigy interface [54], with Cohen kappa [55] measuring interannotator agreement (M2), which is defined as:

$$(1)\kappa=(Po-Pe)/(1-Pe)$$

where *Po* is the observed agreement and *Pe* is the expected agreement by chance. A threshold of 0.8 was used to determine acceptable agreement.

For CamemBERT-based extractors, we performed a 3-fold stratified cross-validation. The full annotated dataset was randomly partitioned into 3 equally sized folds (C1), preserving the proportion of annotated categories within each fold. Duplicated (identical) sentences were removed before the partition to prevent data leakage across folds. In each of the 3 cross-validation rounds, two folds were used for training and the remaining fold served as the validation set (C2). We trained the CamemBERT-large model [27] with loss weighting to address class imbalance. The weighted cross-entropy loss with class weighting is defined as:

$$(2)L=-\Sigma i\ wi\times\log\ p(yi|xi),wi=N/(C\times Nyi)$$

where $p(yi \mid xi)$ is the predicted probability for the true class *yi* of input *xi*, *wi* is the class-specific weight, *N* is the total number of samples, *C* is the number of classes, and *Nyi* is the number of samples in the class of example *i*. Model performance was evaluated (C3) using precision, recall, specificity, *F*-score on individual status, balanced accuracy, and macro *F*-score on all statuses of a medical condition. Balanced accuracy was defined as:

XSL•FO
RenderX

(3)Balancedaccuracy=(1/C)×Σc[((TPc/(TPc+FNc))+(TNc/(TNc+FPc)))/2]

and macro *F*-score was defined as:

(4)F1macro=(1/C)×ΣcF1(c)

where *C* is the number of classes, and *TPc*, *FNc*, *TNc*, and *FPc* are respectively the true positives, false negatives, true negatives, and false positives for class *c*. This set of metrics is supposed to cover all necessary aspects of evaluation for an information retrieval system according to [56]. For comparison (C4), we evaluated our rule-based extractors and Mixtral-8x7B-v0.1 LLM prompts on the same validation sets used for the cross-validation (C2+C3). To ensure a quite fair comparison, the Mixtral-8x7B prompts were designed using few-shot examples derived from the annotation guidelines, combining basic extraction instructions with expert-crafted example and counter-example sentences embedding specialist-provided keywords and clinical context narration.

Detailed configuration information for each step in MSEP appears in Multimedia Appendix 1. The MSEP python package (repository link in Multimedia Appendix 1) provides configurable modules to realize each step of the pipeline.

Results

Overview

We conducted 2 iterations of the pipeline in our institutional setting to create and refine the extractors, after which most targeted medical conditions reached satisfactory performance (average macro *F*-score>0.94) for our use case. This section presents the results of these iterations, describing implementation paths, dataset characteristics, cross-validation results, and comparisons between our extractors and alternative approaches.

Pipeline Execution and Time Requirements

Figure 2 details the implementation path and time consumption for each medical condition across both iterations.

XSL•FO
RenderX

Figure 2. Steps in the 2 iterations of the pipeline were realized to extract each medical status. The steps within the curly braces mean that they are repeatedly realized within one iteration, and the number after the * marks how many times these steps have been repeated. The color of the steps is consistent with Figure 1; blue means automated steps and purple means manual steps. The type of model used for preannotation (PA) and the interannotator agreement score calculated at step M2 are given underneath the list of steps. The last 4 columns show the total time consumption of preannotation (D1), manual annotation (M1) during both iterations, model fine-tuning during cross-validation (C2), and the total time consumed to go through the pipeline in 2 iterations (Ttl). The time consumption is calculated by hours (h) or by minutes (m), depending on the convenience of data representation. COPD: chronic obstructive pulmonary disease.

Medical condition	Steps (1st iter)	Steps (2nd iter)	Consumed time			
			D1	M1	C2	Ttl
Smoking	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.91	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: CamemBERT based Cohen kappa: 0.96	1st iter : 8 h 2nd iter : 11 h	10 h	66 m	30 h
Hyper-tension	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 1.0	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.96	4 h	5 h	123 m	11 h
Diabetes	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 1.0	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.90	3 h	4 h	42 m	8 h
Heart failure	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 1.0	P1-P2-D1-D2-M1-M2-M3-M4-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.57 (0.80 after the correction steps (M3, M4))	4 h	7 h	42 m	12 h
COPD	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.80	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 0.98	3 h	5 h	30 m	9 h
Family history of cancer	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 1.0	P1-P2-D1-D2-M1-M2-C1-{C2-C3}*3-C4 Preannotation: rule-based Cohen kappa: 1.0	24 h	4 h	33 m	29 h

Most medical conditions were preannotated using rule-based matching, except smoking status, which was preannotated in the second iteration using a CamemBERT-based extractor fine-tuned from the first, which outperformed the rules. Heart failure annotation during the second iteration initially yielded insufficient interannotator agreement (57%), necessitating disagreement analysis and correcting annotated data. The analysis revealed that the inconsistencies arose from concept confusion among annotators. A list of ambiguous terms was identified and clarified with a medical expert, incorporated into updated annotation guidelines, and used to automatically extract sentences that need revision from all annotations. These

XSL•FO
RenderX

sentences were then reviewed and corrected by the annotators based on the updated guidelines.

Time requirements varied by medical condition and process step. We focused on the 3 most time-intensive steps: preannotation (D1), manual annotation (M1), and extractor training during cross-validation (C2). Preannotation (D1) speed ranged from 0.02 to 0.17 seconds per sentence, with COPD and diabetes being fastest (0.02 s/sentence) and smoking and family history of cancer slowest (0.12 - 0.17 s/sentence). Manual annotation (M1) required 1.2 - 2.9 seconds per sentence when considering all status labels, with diabetes and family history of cancer being the most efficient (1.2 s/sentence) and smoking being the most time-intensive (2.9 s/sentence). After removing the speed bias introduced by quickly dismissed sentences labeled as unknown, annotation speed ranged from 2.23 to 4.25 s/sentence, with hypertension being the fastest and family history of cancer the slowest, indicating that annotation speed is partly driven by sentence complexity and semantic load. Cross-validation training (C2) averaged 5 - 22 min per session (of cross-validation), with COPD training being fastest (5 m/session) and hypertension slowest (22 m/session). Final extractor inference speed ranged from 0.09 to 0.12 s per sentence. Total implementation time across both iterations (Ttl) ranged from 8 hours (diabetes) to 30 hours (smoking). The time consumption and speed of all automatic steps (D1, C2, and inference of extractors) are calculated on a GPU of NVIDIA A100 40 GB graphics card.

Training and Validation Datasets

The first iteration involved quickly preannotating 9997 medical documents and selecting 991 sentences from them for manual annotation. In order to improve extractors' performance by training them with more medical status samples, the second iteration preannotated an additional 18,994 documents, yielding 11,128 sentences for manual annotation. Table 2 shows the distribution of medical status in sentences annotated during each iteration. The training and validation dataset for the 2nd iteration of the pipeline is the combination of all sentences annotated during both iterations (12,119 sentences in total).

Table . Number of samples of different medical status in sentences respectively annotated during the 1st (itr1) and the 2nd (itr2) iteration of the pipeline.

Medical condition	Present		Absent		Former		Unknown		Total	
	itr1	itr2	itr1	itr2	itr1	itr2	itr1	itr2	itr1	itr2
Smoking	75	646	42	211	87	417	787	9854	991	11,128
Diabetes	279	257	29	37	—[a]	—	683	10,834	991	11,128
Hypertension	217	1184	8	30	—	—	766	9914	991	11,128
Heart failure	46	183	41	217	—	—	904	10,728	991	11,128
COPD[b]	22	366	0[c]	1[c]	—	—	969	10,761	991	11,128
Family history of cancer	5	2	0[c]	1[c]	—	—	986	11,125	991	11,128

[a]Not applicable.

[b]COPD: chronic obstructive pulmonary disease.

[c]Status with too few samples for cross-validation.

Sample sparsity was evident for certain status classifications. During the first iteration, most conditions had fewer than 300 samples per status, with absence samples typically fewer than presence samples. No samples were found for the absence of COPD and family history of cancer. We then increased preannotated documents for the second iteration, which successfully increased sample counts for most status classifications, though samples remained limited due to the absence of COPD and both the presence and absence of family history of cancer.

Among the 12,119 sentences, approximately 10% (n=1200) were randomly selected from the medical documents without preannotation, yielding an unbiased subset of 1200 sentences (unfiltered sentences [US]). We compared the distribution of medical statuses in this subset with that of the remaining 10,919 sentences (preannotated filtered sentences [PS]) selected through preannotation to assess the extent of bias introduced by this data qualification process (D1, D2).

Table 3 shows the result of this comparison. For each medical condition, we calculated the proportion for each of its statuses within US and PS, which is defined as:

(5)proportion=c(status)/c(present,absent,former)

XSL•FO
RenderX

Table . Comparison of medical status samples' distribution (%) in 10,919 preannotated filtered sentences (PS) and 1200 unfiltered sentences (US).

Status	Smoking		Diab		Hyper		CHF[a]		COPD[b]		Fam	
	US	PS	US	PS	US	PS	US	PS	US	PS	US	PS
Present	44	49	100	89	100	97.3	33.3	47.1	100	99.7	0	88
Absent	12	17	0	11	0	2.7	66.7	52.9	0	0.3	0	12
Former	44	34	—[c]	—	—	—	—	—	—	—	—	—
Prevalence of relevant statuses	0.75	13.5	0.25	5.49	0.75	13.1	0.25	4.43	0.17	3.54	0	0.07

[a]CHF: congestive heart failure.

[b]COPD: chronic obstructive pulmonary disease.

[c]Not applicable.

where *c*() denotes the total number of samples of one or more statuses in the set. Since the purpose of the data qualification is specifically to reduce the proportion of sentences that do not express any status information (ie, the class unknown), this category was excluded from the computation to avoid inflating the proportions of relevant statuses (ie, present, absent, and former) in the PS and obscuring the assessment of distributional changes. Instead, we report the prevalence of relevant statuses in US and PS to quantify how effectively steps D1+D2 have increased relevant samples' density in PS by filtering out irrelevant sentences (class unknown). This measurement is defined as:

$$(6) \text{prevalence} = c(\text{present}, \text{absent}, \text{former}) / c(\text{present}, \text{absent}, \text{former}, \text{unknown})$$

Across most medical conditions, the distribution of status labels remained stable after the data filtering steps. The proportions of present status for smoking, diabetes, hypertension, and COPD, as well as the absent status for COPD and hypertension, showed only minor deviations between US and PS. A modest shift was observed for the former and absent statuses of smoking, whereas heart failure exhibited a more substantial change in the present status (from 33.3%, 1/3 in US to 47.1%, 228/484 in PS). Larger apparent shifts for rare categories, such as family history of cancer or the absent status of hypertension, reflect the extremely small number (or absence) of these samples in US, making baseline proportions unstable and limiting interpretability. Overall, despite shifts in certain conditions, the relative ordering of status frequencies remained consistent, suggesting preservation of the dominant patterns. Finally, the prevalence values show that the filtering steps substantially increased the density of relevant samples. For most conditions, the proportion of samples in PS was more than 10 times higher than in US (eg, from 0.75%, 9/1200 to 13.45%, 1469/10,919 for smoking, and from 0.25%, 3/1200 to 5.49%, 599/10,919 for diabetes), demonstrating the effectiveness of steps D1+D2 in removing irrelevant sentences while introducing only minor distributional bias.

CamemBERT-Based Medical Status Extractors Evaluation

As a reminder, each extractor was trained in a multiclass classification setting to predict the status of a single medical condition. Evaluation metrics include precision, recall, specificity, and *F*-score for each individual status, as well as balanced accuracy and macro *F*-score aggregated across all statuses. Table 4 presents the performance metrics for the best-performing extractor of each medical condition after the second iteration. Extractors for diabetes and heart failure achieved excellent overall performance (macro *F*-scores of 99% and 96%, respectively). The COPD extractor achieved a macro *F*-score of 99%, although evaluation was limited to the present and unknown classes due to the lack of absent samples. The diabetes, hypertension, and COPD extractors exceeded 98% *F*-score for presence detection. The smoking extractor achieved balanced performance across status classifications (93% - 94% *F*-scores), while the hypertension extractor showed imbalanced performance between presence (98% *F*-score) and absence (86% *F*-score). The extractor for family history of cancer underperformed with only 80% *F*-score for presence detection.

XSL•FO
RenderX

Table . Performance of the best CamemBERT-large based extractors obtained after the cross-validation of the 2nd iteration of the pipeline. For each of the studied 6 medical conditions—smoking, diabetes (diab), hypertension (hyper), congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), and family history of cancer (fam)—the extractor is evaluated on each of its status (unknown, absent, present, or former) with 4 metrics: *F*-score, precision, recall, and specificity. The macro *F*-score and balanced accuracy are calculated on (and only on) the evaluated status.

Extractor	Smoking	Diab	Hyper	CHF	COPD	Fam
Unknown *F*-score	0.98	0.99	0.98	0.97	0.99	0.99
Unknown Precision	0.99	1.0	0.98	0.96	1.0	1.0
Unknown Recall	0.97	0.99	0.99	0.99	0.99	0.99
Unknown Specificity	0.99	1.0	0.98	0.94	1.0	1.0
Absent *F*-score	0.94[a]	1.0[a]	0.86[b]	0.97[a]	—[c]	—
Absent Precision	0.94	1.0	0.86	0.98	—	—
Absent Recall	0.93	1.0	0.85	0.96	—	—
Absent Specificity	0.99	1.0	0.99	0.99	—	—
Present *F*-score	0.93[a]	0.99[a]	0.98[a]	0.92[a]	0.99[a]	0.80[b]
Present Precision	0.91	0.99	0.98	0.96	0.98	0.67
Present Recall	0.95	1.0	0.98	0.89	1.0	1.0
Present Specificity	0.97	0.99	0.99	0.99	0.99	0.99
Former *F*-score	0.94[a]	—	—	—	—	—
Former Precision	0.93	—	—	—	—	—
Former Recall	0.94	—	—	—	—	—
Former Specificity	0.99	—	—	—	—	—
Macro *F*-score	0.95[a]	0.99[a]	0.94[a]	0.96[a]	0.99[a]	0.89[b]
Balanced accuracy	0.95	0.99	0.94	0.95	0.99	0.99

[a]Status (other than Unknown) whose *F*-score has surpassed 90%.

[b]*F*-score has not surpassed 90%.

[c]Not applicable.

Table 5 presents the results of the 3-fold cross-validation from both iterations. For each medical condition, we evaluated the stability of the extractors by reporting the mean and SD of the *F*-score for each status category, as well as for the overall macro *F*-score, across the cross-validation folds. All extractors demonstrated improved performance after the second iteration. After the first iteration, only the diabetes extractor achieved a macro *F*-score exceeding 90%, while by the second iteration, most condition extractors surpassed this threshold. Hypertension, heart failure, and COPD extractors showed the most substantial improvement (>30%).

XSL•FO
RenderX

Table . Result (average *F*-score, SD) of 3-fold stratified cross-validation with CamemBERT-large based extractors trained during the 1st (iter 1) and the 2nd (iter 2) iterations of the pipeline.

Medical condition	Present *F*-score, mean (SD)	Absent *F*-score, mean (SD)	Former *F*-score, mean (SD)	Unknown *F*-score, mean (SD)	Macro *F*-score, mean (SD)
Iter 1					
Smoking	0.78 (0.02)	0.70 (0.03)	0.88 (0.005)	0.99 (0.02)	0.83 (0.02)
Diabetes	0.89 (0.04)	0.99 (0.01)[a]	—[b]	0.98 (0.004)	0.95 (0.03)[a]
Hypertension	0.94 (0.002)	0 (0)	—	0.97 (0.003)	0.63 (0.003)
Heart failure	0.24 (0.03)	0.14 (0.1)	—	0.71 (0.09)	0.36 (0.05)
COPD[c]	0 (0)	—	—	0.93 (0.04)	0.31 (0.01)
Family history of cancer[d]	0 (0)	—	—	0.99 (0.003)	0.33 (0.001)
Iter 2					
Smoking	0.93 (0.008)	0.93 (0.005)	0.92 (0.01)	0.98 (0.0004)	0.94 (0.003)
Diabetes	0.98 (0.008)[a]	0.98 (0.02)[a]	—	0.98 (0.004)	0.98 (0.01)[a]
Hypertension	0.98 (0.002)[a]	0.85 (0.01)[e]	—	0.98 (0.002)	0.96 (0.009)[a]
Heart failure	0.90 (0.02)	0.95 (0.02)[a]	—	0.97 (0.003)	0.94 (0.003)
COPD	0.99 (0.002)[a]	—	—	0.99 (0.003)	0.99 (0.002)[a]
Family history of cancer[d]	0.55 (0.37)	—	—	0.99 (0.003)	0.73 (0.17)

[a]Average *F*-score more than 95% at detecting medical status other than Unknown (either iteration).

[b]Not applicable.

[c]COPD: chronic obstructive pulmonary disease.

[d]For family history of cancer, with only 7 positive samples across 2 iterations, calculating a SD of 37% for the *F*-score is statistically unstable and potentially misleading regarding the extractor's reliability.

[e]Average *F*-score less than 90% after the second iteration.

Despite these improvements, the family history of cancer extractor achieved only 73% macro *F*-score with 17% instability, primarily due to poor performance in detecting presence (55% *F*-score, 37% instability). The hypertension extractor also showed suboptimal performance for absence detection (85% *F*-score, 1% instability). Due to insufficient samples, absence classification for COPD and family history of cancer could not be trained or evaluated.

To go one step deeper into the assessment of the extractor performances, a post hoc analysis according to document type was conducted. Table 6 reports the average *F*-scores with SD, obtained during three-fold cross-validation during the last iteration of the pipeline, for each medical status across the major clinical document types represented in the dataset.

XSL•FO
RenderX

Table . Average *F*-scores (SD) per medical status (present: pr, absent: ab, former: fm, and unknown: un) across clinical document types: Nursing care–Procedures (Nurs-Proc), Medical reports (Med report), Nursing care–Monitoring and vitals (Nurs-Monit), Multidisciplinary follow-up and daily notes (daily note), Imaging reports (image), Prescriptions and medication orders (Presc), Paramedical care (Paramed), Administrative and admission (Admin), Results and pathology reports (Patho), and Other.

Condition and status	Nurs-Proc	Med report	Nurs-Monit	Daily note	Image	Presc	Para-med	Admin	Patho	Other
Smoking, mean (SD)										
pr	1.0 (N/A[a])	0.95 (0.006)	—[b]	0.93 (0.02)	1.0 (N/A)	1.0 (N/A)	0.83 (0.236)	1.0 (N/A)	—	0.96 (0.064)
ab	—	0.92 (0.01)	—	1.0 (N/A)	1.0 (N/A)	—	—	1.0 (N/A)	—	—
fm	1.0 (N/A)	0.94 (0.01)	—	1.0 (N/A)	0.89 (0.1)	—	—	—	—	1.0 (N/A)
un	1.0 (N/A)	0.98 (0.004)	1.0 (N/A)	0.93 (0.016)	0.97 (0.029)	0.96 (0.064)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)
Diab, mean (SD)										
pr	1.0 (N/A)	0.97 (0.012)	—	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	0.97 (0.044)	—	—	—
ab	—	0.97 (0.02)	—	1.0 (N/A)	—	—	—	—	—	—
un	1.0 (N/A)	0.99 (0.0)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	0.93 (0.115)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)
Hyper, mean (SD)										
pr	1.0 (N/A)	0.99 (0.002)	0 (N/A)	0.97 (0.011)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	—	1.0 (N/A)
ab	—	0.84 (0.022)	—	1.0 (N/A)	—	—	—	—	—	—
un	1.0 (N/A)	0.98 (0.004)	0.99 (0.008)	0.94 (0.037)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)
CHF[c], mean (SD)										
pr	1.0 (N/A)	0.92 (0.029)	0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	—	—	1.0 (N/A)
ab	—	0.98 (0.02)	—	0.8 (0)	1.0 (N/A)	—	—	—	—	1.0 (N/A)
un	1.0 (N/A)	0.98 (0.002)	0.99 (0.009)	0.95 (0.014)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	—	1.0 (N/A)
COPD[d], mean (SD)										
pr	1.0 (N/A)	0.99 (0.002)	—	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	0.89 (0.192)	—	—
un	1.0 (N/A)	0.99 (0.001)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	0.94 (0.079)	1.0 (N/A)	1.0 (N/A)
Fam, mean (SD)										
pr	—	0.55 (0.37)	—	—	—	—	—	—	—	—
un	1.0 (N/A)	0.99 (0.003)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)	1.0 (N/A)

[a]N/A: not applicable.

[b]Not available.

[c]CHF: congestive heart failure.

[d]COPD: chronic obstructive pulmonary disease.

XSL•FO
RenderX

For 72 status–document-type combinations, no sample was retrieved in validation sets, and performance cannot be discussed (marked with an "-"). For the remaining 115 status document-type combinations, we can divide results into 2 categories.

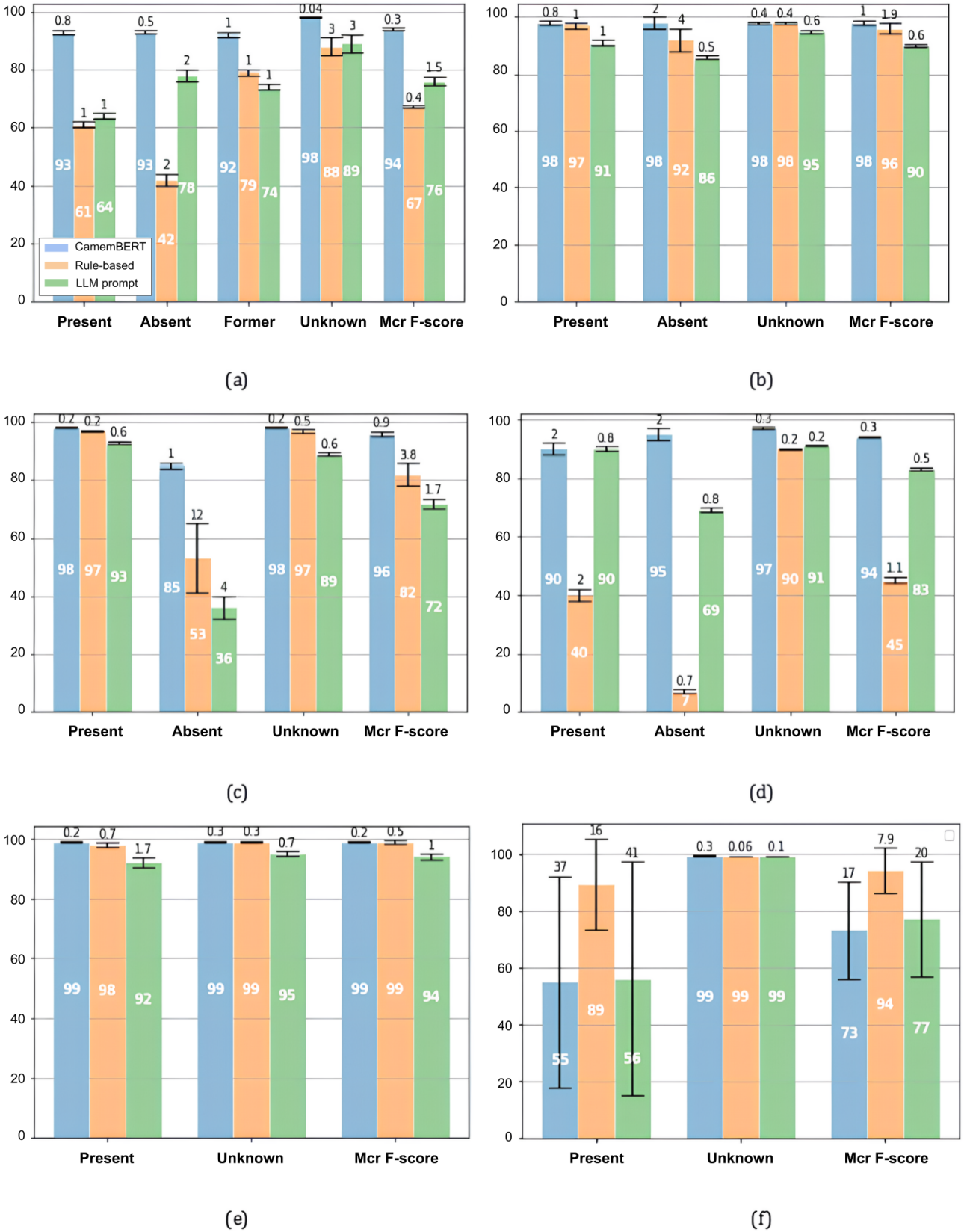
Those with comparable performance (less than 5% decrease) or improvement on specific documents (109/115, 94.8%) and those where a decrease is observed (6/115, 5.2%). In the first category, 79 instances show a score of 1.0 +/– N/A, indicating that only one-fold contains content for evaluation. This perfect performance also suggests that few samples were available for evaluation. In the second category, 2 instances were only evaluated in one fold, and extractors missed the status (ie, for hypertension presence and heart failure presence in the nurse report). For the 4 others (smoking presence and diabetes indifference in paramedical report, absence of heart failure in daily note, and presence of COPD in administration note), a decrease is observed, but performance remains acceptable (>80%). Examination of false positives and false negatives in these 4 cases suggests that most errors stem from the misleading narrative style of clinical notes. For example, in a paramedical report, "OH chronique sevré tabac syndrome dépressif" (Chronic alcohol use weaned tobacco use depressive disorder) was misclassified as former smoking because the boundary between alcohol withdrawal and smoking was unclear. Another sentence, "des atcd cardio,HTA,DT2non traité" (A history of cardiac disease, untreated hypertension, and type 2 diabetes) was misclassified as absence of diabetes because the negation "non" was attached to the diabetes "DT2" rather than to the treatment "traité." The extractor also produces some persistent errors due to the medical complexity of the content, which is challenging even for specialist reviewers; for instance, a sentence in a daily note was misclassified as absent of heart failure because it contains this expression "avec une fraction d'éjection ventriculaire conservée" (with a preserved ventricular ejection fraction), which does not exclude entirely the possibility of having a heart failure. In conclusion, although the representation of entities is very sparse, we observe overall good robustness of the extractors depending on the type of document encountered. To go further, the number of annotated sentences would need to be increased.

Comparison With Alternative Approaches

Figure 3 presents a comparison of the extractors developed using CamemBERT fine-tuning, rule-based methods, and LLM prompting (Mixtral-8×7B-v0.1). As a reminder, all extractors were evaluated on the same validation sets used for the 3-fold cross-validation of CamemBERT-based extractors. For conditions with ample annotated examples, such as smoking and heart failure, CamemBERT-based extractors achieved the highest performance (also confirmed by paired *t* test analysis; Multimedia Appendix 3). In contrast, for conditions with limited training data, rule-based extractors performed comparably to or better than CamemBERT-based models, reflecting their robustness in low-data scenarios. This was particularly evident for family history of cancer, where CamemBERT-based extractors achieved an average macro *F*-score of 0.73 with a cross-validation SD of 17%, while the rule-based approach reached an average macro *F*-score of 0.94 with a substantially lower SD of 7.9%.

XSL•FO
RenderX

Figure 3. Comparison of the performance of extractors based on 3 different approaches: fine-tuned CamemBERT, (**A**) rule-based extraction and large language model prompt query for smoking, (**B**) diabetes, (**C**) hypertension, (**D**) heart failure, (**E**) chronic obstructive pulmonary disease, and (**F**) family history of cancer. The white number on each bar marks the average macro *F*-score percentage, and the number above each bar represents the SD (also percentage) calculated from cross-validation. The performance is evaluated by testing all extractors on the same validation datasets used during the 3-fold cross-validation of the 2nd iteration of the pipeline.



Compared with CamemBERT-based and rule-based extractors, LLM-prompted extractors showed intermediate performance overall but were less efficient in our setting. Using Mixtral-8×7B-v0.1, sentence-level extraction required 12 - 24 seconds per sentence and 35 GB of GPU memory, which is

XSL•FO
RenderX

slower and more resource-intensive than CamemBERT (0.09 - 0.12 s per sentence; 2 GB GPU).

These findings indicate that CamemBERT-based and rule-based methods are the most suitable approaches for our institutional resource constraints, and that no single extraction approach is universally optimal across all clinical scenarios.

Discussion

Overview

This section examines the advantages and limitations of the MSEP pipeline within our institutional setting and explores how its design can support local development of clinical NLP extractors.

Technical Advantages

Efficiency Gains for Clinical Text Annotation

The MSEP pipeline addresses challenges posed by the low density of medical-status information in large institutional corpora, which can hinder the efficient local development of information extraction systems. Its preannotation step helps prioritize sentences with potentially relevant information, and reframing extraction as sentence classification allows annotators to make quicker labeling decisions. Together, these components streamline dataset construction and reduce the operational burden of annotation.

Although our annotators were unavailable for extensive medical entity annotation, preventing a direct comparison of annotation speed in our institution, prior work has shown that sentence classification generally requires less annotation time and effort than word-level NER [57-60]. This is expected, as sentence classification involves assigning a single label to an entire sentence, whereas NER requires annotators to identify multiple entities within a sentence and decide on their boundaries, increasing the complexity and time required for annotation [61]. This improvement in annotation speed is also supported by an empirical comparison with a previous study on NER efficiency [20], as our manual sentence classification ranged from 1.2 to 2.9 seconds per sentence (or 2.23 to 4.25 s per sentence when considering only informative sentences), significantly faster than the 7.82 to 16.48 s reported for NER.

The sentence classification-based extractors also demonstrated faster inference speed than the traditional composite extractors (NER + qualifier). In our study, extractors of rule-based NER combined with EDS-NLP qualifiers were tested for smoking and family history of cancer during the preannotation step of the first iteration of the pipeline. The composite approach (0.17 s/sentence) was notably slower than our CamemBERT-based extractor (0.09 - 0.12 s/sentence) and rules-only preannotation models (0.02 - 0.04 s/sentence). Importantly, this efficiency gain did not compromise performance. Compared with recent medical status extractors, our pipeline achieved comparable or superior results. Our extractors reached *F*-scores of 99%, 95%, 96%, and 99% for diabetes, smoking, heart failure, and COPD, respectively, matching or exceeding performance metrics reported in recent literature [16].

Hybrid, Configurable, and Locally Deployable Pipeline Architecture

The MSEP pipeline was designed as a flexible and configurable framework that enables the local deployment of hybrid extraction approaches across diverse institutional settings. The accompanying Python package (repository link in Multimedia Appendix 1) exposes all key components, such as preannotation models (D1), sentence filtering criteria (D2), cross-validation strategy (C2), and extractor choice, as configurable arguments. This modular design allows researchers to tailor each step of the workflow, including the choice of extraction approach, according to their resource constraints, characteristics of the targeted information, and institutional requirements. For instance, in our experiments, fine-tuned CamemBERT models produced satisfactory performance for most medical statuses, whereas rule-based extractors yielded more stable results for conditions with limited representation in the dataset. This means that the pipeline can be configured to favor rule-based extraction in low-example scenarios. Conversely, in settings with limited annotation resources or domain expertise, the pipeline can be configured to generate LLM prompt-based extractors, which can be readily improved by providing more detailed instructions and examples in the prompts [62].

Our institutional experiments further illustrated the practical feasibility of deploying MSEP within a local clinical setting. First, the full workflow, including preannotation, manual validation, model training, and validation, was applied independently to 6 medical conditions and completed within 99 hours in total (approximately 16.5 h per condition). This was achieved with 2 annotators and a single NVIDIA A100 40 GB GPU, a setup feasible for many medical research institutions. Second, the models were trained and evaluated on a heterogeneous set of clinical documents (including hospital discharge summaries, consultation notes, and reports from multiple specialties). The comparable performance observed across these diverse sources indicates that the workflow can be operationalized effectively within diverse documentation practices at a single institution.

While our current evaluation remains monocentric, MSEP's packaging and explicit parameterization are designed to facilitate reproducible deployment in other institutions. Applying identical configurations on comparable data would be expected to yield consistent results, and extending the approach to multicentric validation represents an important next step toward confirming generalizability and cross-site reproducibility.

Enhanced Monitoring and Optimization Mechanic

The pipeline provides means for evaluating and improving each of its mean processes (data qualification, manual annotation and correction, model training, and cross-validation). Preannotation (D1) allows estimation and supervision on corpus' sample density and medical status' distribution, which can be adjusted by sentence filtering (D2); interannotator agreement calculation (M2) and analysis of annotation disagreements (M3) not only revise the quality of annotated sentences, but also reveal problems in annotation guidelines and improve them accordingly (M4); extractors' assessment and comparison (C4) enables

XSL•FO
RenderX

comprehensive model evaluation and informs decisions on the need for and direction of further improvements.

The iterative nature of our pipeline provides a means for monitoring and optimizing dataset size. By accumulating annotated data gradually and evaluating extractors with cross-validation at each iteration, we can determine when to halt manual annotation, limiting resource expenditure to reasonable levels. We also considered active learning as an alternative approach, where the model selects samples difficult to predict for prioritized manual annotation and updates itself with the received annotation [63]. In theory, this approach can maximize annotation efficiency by prioritizing the most informative samples for model improvement. However, we ultimately rejected it because model updates were disproportionately influenced by status categories with the most samples, and uncertain sample distribution prevented effective loss function weighting. Nevertheless, active learning remains valuable when balanced datasets are available or when sample balancing can be implemented during data acquisition [64].

Importance for Clinical Research

Response to Clinical Text-Mining Challenges

Clinical text mining remains challenging due to the unstructured and context-dependent nature of medical narratives. Our evaluation shows that sentence-level classification, as implemented in MSEP, effectively addresses several of these challenges in practice. The approach demonstrated high accuracy across heterogeneous document types, confirming that the model captures contextual cues, such as negation or temporality, without relying on context qualifiers. Compared with traditional NER-based systems, which require token-level annotation and post hoc context detection, the sentence classification strategy substantially reduces annotation time and complexity. The high *F*-scores obtained for conditions such as smoking and diabetes illustrate that this formulation aligns well with the way medical status is expressed, which is typically at the sentence level rather than through isolated entities.

To our knowledge, there is currently no existing pipeline in the clinical NLP literature that applies sentence classification for medical status extraction. While sentence classification has been explored in other domains, such as sentiment analysis or social media mining, its application to clinical narratives remains underdeveloped and has not yet been systematically assessed for this purpose.

Easy Implementation as a Research Tool for Clinical Studies

Clinical text-mining efforts are often constrained by privacy regulations and institutional heterogeneity that prevent direct data or model sharing across sites [16,19,36-38]. The MSEP framework was designed to address these issues by enabling efficient local dataset construction, providing flexible configuration of extraction approaches according to institutional constraints, and supporting controlled resource expenditure through performance-monitored iteration.

Beyond initial model development, MSEP demonstrated utility as an upstream component in broader research workflows that can support more specialized downstream tasks. When applied first, MSEP extractors can identify only the documents or sentences that contain the targeted medical status, allowing researchers to exclude irrelevant material entirely and thus avoid expending effort on processing noninformative documents. In practice, this functionality has already been demonstrated within our institution, where a smoking-status classifier created by MSEP was applied to an independent corpus of 300 clinical reports, efficiently filtering for documents concerning patients who were smokers and achieving an *F*-score of 99.6% for detecting smoking-related content. Downstream tasks, such as extracting smoking duration or tobacco brand, can then be focused solely on these filtered documents, reducing workload and improving overall efficiency.

Current Limitations

Our study has several limitations worth noting. First, due to the availability of medical experts, we did not conduct a direct comparison between sentence classification and NER. Instead, we relied on empirical observations and general consensus to suggest that sentence classification offers greater annotation efficiency. While the improvement in annotation speed appears supported, the impact on annotation quality remains to be verified.

Second, clinical document type is a factor that can influence the interpretation of medical status occurrence in text, which was not fully addressed in our study. For instance, a mention of “hypertension” in a medical report, especially within the “Past Medical History” section, typically refers to a chronic condition. In contrast, if hypertension is noted in daily progress notes, it is more likely to reflect an acute episode under current observation or management. Therefore, when applying MSEP to clinical documents from a CDW, it may be necessary to incorporate a document selection step during data collection (P1), not only to include relevant document types, but also to focus on specific sections within documents, such as the “Past Medical History” section, when these are more likely to align with the study objectives.

Third, the construction and usage of the datasets are subject to several sources of bias. Preannotation and filtering were used to increase the density of informative samples in the dataset, but mislabeling by preannotation rules could have led to the exclusion of relevant sentences (eg, those labeled as “unknown”). To assess this risk, 10% of our dataset consisted of nonpreannotated sentences, which were used to evaluate potential bias introduced by the preannotation rules. While most conditions showed limited distortion, a notable shift was observed for the present status of heart failure, underscoring that preannotation can influence class distributions. A further concern is the risk of circular bias, which may be induced by using sentences initially selected by rules to evaluate those same rule-based extractors. To limit this risk, we constructed the annotated corpus by combining the preannotations from all descriptors rather than relying on a single rule-based extractor per condition, resulting in a relatively small proportion of validation sentences originating from the same rule set (eg, 3.4% - 3.6% for COPD). Evaluating extractors is a complex task in such large datasets. Ideally, most of the dataset would

XSL•FO
RenderX

need to be annotated to determine actual performance. However, this task requires too much annotation time for most noninformative sentences. Our approach to selection by multiple preannotations allows us to limit this time while verifying that little bias appears to be induced (eg, confirmed using sentences without preannotation). Finally, although identical sentences were deduplicated before cross-validation, dataset partitioning was not stratified by Patient_ID. As a result, sentences about the same patient could appear across validation folds, introducing a potential risk of data leakage. The proportion of validation sentences with patient overlap ranged from 3% to 8%. To assess the impact of this overlap, we revalidated the models after removing these sentences from the validation sets. Performance changes were minimal (less than 1% of drop for macro *F*-score of each cross-validation fold), suggesting that patient-level overlap did not artificially inflate extractor performance. Moving forward, these extractors should be used on other datasets while keeping in mind their potential limitations.

Fourth, the evaluation of certain medical statuses was limited by their very low prevalence in the corpus. For COPD and family history of cancer, the Absent status was represented by only a single instance, preventing meaningful assessment of the extractor's ability to distinguish absent from unknown. Consequently, both extractors effectively functioned as binary classifiers (present vs unknown). This limitation reflects typical documentation practices: the absence of a specific condition (eg, COPD) is rarely stated explicitly, as doing so would require exhaustively listing all conditions that are not present; the issue is somewhat different for family history of cancer, where the absence of this condition is generally more informative ("no family history of cancer") and therefore more likely to be mentioned, although such mentions remained infrequent in our dataset.

Finally, decoder-only LLMs were not fully used due to initial infrastructure limitations. Our computing environment, a secure server at the University Hospital of Rennes with 112 Intel(R) Xeon(R) Gold 6258R CPU cores and limited access to an NVIDIA A100 40 GB GPU, was insufficient for efficient, prompt-based inference (12 - 24 s/sentence when performing medical status extraction) or fine-tuning (requiring ≥64 GB VRAM).

Perspectives and Future Applications

Beyond its current implementation, the pipeline can be extended to broader medical information extraction tasks. Ongoing work includes detecting additional statuses relevant to bladder cancer [3] and extracting finer-grained information such as smoking duration and cigarette type. By first classifying sentences containing smoking mentions, the pipeline narrows the scope for subsequent extractions, improving efficiency.

Future studies will also address current limitations. First, a direct comparison of annotation speed and quality between sentence classification and NER will provide stronger evidence for our methodological choice. Second, it is well known that medical documents contain information, structures, and sometimes vocabulary that are unique to them. For our specific study, we have demonstrated a certain robustness of extractors to these phenomena. To reinforce this finding, more instances will need to be annotated, bearing in mind that we are working in a very sparse context. Third, synthetic datasets will be explored to mitigate preannotation bias, balance category distributions, and reduce reliance on sensitive records. They may also improve extractor performance on medical conditions with scarce samples (eg, family history of cancer). Prior studies have investigated GANs for generating synthetic structured EHRs [40]; while ensuring both utility and confidentiality remain challenging, synthetic data offer a promising path forward for clinical NLP.

Finally, the Ollama framework [65] will be used for a better exploration of LLM prompt potential. This framework enables accelerated local inference, improving medical status annotation speed with Mixtral-8x7B-v0.1 to 0.2 seconds per sentence. Although LLM prompting has not yet reached the same level of performance as CamemBERT fine-tuning for certain conditions, further exploration of prompt design and fine-tuning strategies may be pursued as infrastructure improves. However, it is important to keep in mind the computational and ecological cost of such a strategy when rule-based and CamemBERT-based approaches can provide satisfactory results.

Together, these perspectives aim to strengthen the robustness, reproducibility, and applicability of the pipeline across diverse clinical research settings.

Conclusions

We introduced MSEP, a modular and hybrid methodological framework for extracting medical status information from unstructured clinical text using sentence classification. In our institutional case study, the pipeline substantially reduced annotation effort: sentence-level labeling with preannotation required only 1.2 - 2.9 seconds per sentence. MSEP successfully supported the implementation of 3 extraction approaches: rule-based methods, fine-tuned CamemBERT model, and LLM prompting, each suited to different data and resource scenarios. Fine-tuned CamemBERT model achieved high performance for conditions with sufficient training examples (macro *F*-score>94%), whereas rule-based methods provided more stable results for sparsely represented conditions. Together, these findings highlight MSEP's value as a research tool that accelerates local dataset creation and enables flexible deployment of extraction systems. Importantly, all components of the pipeline are distributed as a ready-to-use Python package.

Acknowledgments

The authors thank VitaDX company for their collaborative contributions to this research. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. In accordance with the GAIDeT taxonomy (2025), GenAI tools were used under full human supervision for reformatting tasks. The GenAI tool used was ChatGPT-5. Responsibility for

XSL•FO
RenderX

the content and integrity of the final manuscript rests entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted under the collective responsibility of the authors.

Funding

This work received support from VitaDX (Rennes, France), which contributed to the project as an industrial partner. The funder had no influence on the study design, analysis, interpretation, or manuscript preparation.

Data Availability

The data used in this work contains pseudo-identifiable information and cannot be shared directly. However, a collaboration with the Clinical Data Center of Rennes Hospital can be considered for data exploitation while adhering to legal and regulatory frameworks. Please contact the corresponding author for more information. The Medical Status Extraction Pipeline is available as an open-source Python package (Multimedia Appendix 2 contains the repository link).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Pipeline configuration demonstration.
[DOCX File, 5577 KB - medinform_v14i1e77409_app1.docx]

Multimedia Appendix 2
Annotation guidelines.
[DOCX File, 4879 KB - medinform_v14i1e77409_app2.docx]

Multimedia Appendix 3
Complementary paired t-test evaluation of extractor performance across cross-validation folds.
[DOCX File, 3482 KB - medinform_v14i1e77409_app3.docx]

References

1. Wang Z, Syed M, Syed S, et al. Clinical data warehousing: a scoping review. J Soc Clin Data Manag 2024;4(2):8. [doi: 10.47912/jscdm.320] [Medline: 41602775]
2. Bazoge A, Morin E, Daille B, Gourraud PA. Applying natural language processing to textual data from clinical data warehouses: systematic review. JMIR Med Inform 2023 Dec 15;11:e42477. [doi: 10.2196/42477] [Medline: 38100200]
3. Cabon S, Brihi S, Fezzani R, Pierre-Jean M, Cuggia M, Bouzillé G. Combining a risk factor score designed from electronic health records with a digital cytology image scoring system to improve bladder cancer detection: proof-of-concept study. J Med Internet Res 2025 Jan 22;27:e56946. [doi: 10.2196/56946] [Medline: 39841985]
4. Pierre-Jean M, Marut B, Curtis E, et al. Phenotyping of heart failure with preserved ejection faction using electronic health records and echocardiography. Eur Heart J Open 2024 Jan;4(1):oead133. [doi: 10.1093/ehjopen/oead133] [Medline: 38196848]
5. Pardo I, Pierre-Jean M, Bouzillé G, et al. Safety of subcutaneous versus intravenous ceftriaxone administration in older patients: a retrospective study. J Am Geriatr Soc 2024 Apr;72(4):1060-1069. [doi: 10.1111/jgs.18786] [Medline: 38348519]
6. Shen F, Liu S, Fu S, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 national NLP clinical challenges (n2c2)/open health natural language processing (OHNLP) competition. JMIR Med Inform 2021 Jan 27;9(1):e24008. [doi: 10.2196/24008] [Medline: 33502329]
7. Weeks HL, Beck C, McNeer E, et al. medExtractR: a targeted, customizable approach to medication extraction from electronic health records. J Am Med Inform Assoc 2020 Mar 1;27(3):407-418. [doi: 10.1093/jamia/ocz207] [Medline: 31943012]
8. Lependu P, Liu Y, Iyer S, Udell MR, Shah NH. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Jt Summits Transl Sci Proc 2012;2012:63-70. [Medline: 22779054]
9. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform 2012 Feb;45(1):129-140. [doi: 10.1016/j.jbi.2011.10.007] [Medline: 22085698]
10. Lerner I, Jouffroy J, Burgun A. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv. Preprint posted online on Apr 24, 2020. [doi: 10.48550/arXiv.2004.11622]
11. Hahn U, Oleynik M. Medical information extraction in the age of deep learning. Yearb Med Inform 2020 Aug;29(1):208-220. [doi: 10.1055/s-0040-1702001] [Medline: 32823318]

XSL•FO
RenderX

12. Musen MA, Friedman C. Semantic text parsing for patient records. In: Kulikowski CA, editor. Medical Informatics: Knowledge Management and Data Mining in Biomedicine: Springer; 2006:423-448. [doi: 10.1007/0-387-25739-X_15]
13. Oxford University Press. Medical status - oxford reference. Oxford Reference. 2011. URL: https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100146893 [accessed 2025-02-28]
14. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 1;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
15. Bhatia P, Celikkaya B, Khalilia M. Joint entity extraction and assertion detection for clinical text. : Association for Computational Linguistics Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy p. 954-959 URL: https://aclanthology.org/P19-1 [doi: 10.18653/v1/P19-1091]
16. Petit-Jean T, Gérardin C, Berthelot E, et al. Collaborative and privacy-enhancing workflows on a clinical data warehouse: an example developing natural language processing pipelines to detect medical conditions. J Am Med Inform Assoc 2024 May 20;31(6):1280-1290. [doi: 10.1093/jamia/ocae069] [Medline: 38573195]
17. Neuraz A, Lerner I, Digan W, et al. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. J Med Internet Res 2020;22(8):e20773. [doi: 10.2196/20773]
18. Li I, Pan J, Goldwasser J, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev 2022 Nov;46:100511. [doi: 10.1016/j.cosrev.2022.100511]
19. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. In: Sarmiento RF, editor. PLOS Digit Health 2023 Oct;2(10):e0000347. [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
20. Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation - what factors are associated with the cost of time? AMIA Annu Symp Proc 2018;2018:1552-1560. [Medline: 30815201]
21. Moscato V, Postiglione M, Sperlì G, Vignali A. ALDANER: active learning based data augmentation for named entity recognition. Knowl Based Syst 2024 Dec;305:112682. [doi: 10.1016/j.knosys.2024.112682]
22. Tsuruoka Y, Tsujii JI, Ananiadou S. Accelerating the annotation of sparse named entities by dynamic sentence selection. BMC Bioinformatics 2008 Nov 19;9 Suppl 11(Suppl 11):1-10. [doi: 10.1186/1471-2105-9-S11-S8] [Medline: 19025694]
23. Settles B, Craven M, Friedland L. Active learning with real annotation costs. Presented at: Proceedings of the NIPS Workshop on Cost-Sensitive Learning; Dec 13, 2008.
24. Long K, Zhao H, Shao Z, et al. Deep neural network with embedding fusion for chinese named entity recognition. ACM Trans Asian Low-Resour Lang Inf Process 2023 Mar 31;22(3):1-16. [doi: 10.1145/3570328]
25. Kalyan KS, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv. Preprint posted online on 2021. [doi: 10.48550/arXiv.2108.05542]
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North; Jun 2-7, 2019. [doi: 10.18653/v1/N19-1423]
27. Martin L, Muller B, Ortiz Suárez PJ, et al. CamemBERT: a tasty french language model. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 6-8, 2020. [doi: 10.18653/v1/2020.acl-main.645]
28. Grießhaber D, Maucher J, Vu NT. Fine-tuning BERT for low-resource natural language understanding via active learning. Presented at: Proceedings of the 28th International Conference on Computational Linguistics; Dec 8-13, 2020 URL: https://www.aclweb.org/anthology/2020.coling-main [doi: 10.18653/v1/2020.coling-main.100]
29. OpenAI. Introducing chatgpt. OpenAI. 2022. URL: https://openai.com/blog/chatgpt [accessed 2025-02-28]
30. Mistral AI. Mistral 7B: a new paradigm for open-weight language models. Mistral AI. 2023. URL: https://mistral.ai/news/announcing-mistral-7b/ [accessed 2025-02-28]
31. LLMs accelerate annotation for medical information extraction. Presented at: Proceedings of the 3rd Machine Learning for Health Symposium; Dec 10, 2023.
32. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
33. Posada A, Rueckert D, Meissen F, Müller P. Evaluation of language models in the medical context under resource-constrained settings. arXiv. Preprint posted online on Jun 24, 2024. [doi: 10.48550/arXiv.2406.16611]
34. Ademola A, George C, Mapp G. Addressing the interoperability of electronic health records: the technical and semantic interoperability, preserving privacy and security framework. ASI 2024;7(6):116. [doi: 10.3390/asi7060116]
35. Hu Y, Zuo X, Zhou Y, et al. Information extraction from clinical notes: are we ready to switch to large language models. arXiv. Preprint posted online on Nov 15, 2024. [doi: 10.48550/arXiv.2411.10020]
36. Ghosheh GO, Li J, Zhu T. A survey of generative adversarial networks for synthesizing structured electronic health records. ACM Comput Surv 2024 Jun 30;56(6):1-34. [doi: 10.1145/3636424]
37. Hier DB, Carrithers MD, Do TS, Obafemi-Ajayi T. Efficient standardization of clinical notes using large language models. arXiv. Preprint posted online on Dec 31, 2024. [doi: 10.48550/ARXIV.2501.00644]
38. Lahiri A, Shukla S, Stear B, et al. Benchmarking transformer embedding models for biomedical terminology standardization. Mach Learn Appl 2025 Sep;21:100683. [doi: 10.1016/j.mlwa.2025.100683] [Medline: 40718094]

XSL•FO
RenderX

39. Zhao H, Sui D, Wang Y, Ma L, Wang L. Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction. Sensors (Basel) 2025 Sep;25(8):2374. [doi: 10.3390/s25082374]
40. Fowl L, Geiping J, Reich S, et al. Decepticons: corrupted transformers breach privacy in federated learning for language models. arXiv. Preprint posted online on Jan 29, 2022. [doi: 10.48550/ARXIV.2201.12675]
41. Villena F, Bravo-Marquez F, Dunstan J. NLP modeling recommendations for restricted data availability in clinical settings. BMC Med Inform Decis Mak 2025 Mar 7;25(1):116. [doi: 10.1186/s12911-025-02948-2] [Medline: 40055634]
42. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. J Biomed Inform 2020 Sep;109:103526. [doi: 10.1016/j.jbi.2020.103526] [Medline: 32768446]
43. Beaney T, Jha S, Alaa A, et al. Comparing natural language processing representations of coded disease sequences for prediction in electronic health records. J Am Med Inform Assoc 2024 Jun 20;31(7):1451-1462. [doi: 10.1093/jamia/ocae091] [Medline: 38719204]
44. Madec J, Bouzillé G, Riou C, et al. eHOP clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019 Aug 21;264:1536-1537. [doi: 10.3233/SHTI190522] [Medline: 31438219]
45. Karanikolas N, Manga E, Samaridi N, Stergiopoulos V, Tousidou E, Vassilakopoulos M. Strengths and weaknesses of LLM-based and rule-based NLP technologies and their potential synergies. Electronics (Basel) 2025 Jul 31;14(15):3064. [doi: 10.3390/electronics14153064]
46. Jupin-Delevaux É, Djahnine A, Talbot F, et al. BERT-based natural language processing analysis of French CT reports: application to the measurement of the positivity rate for pulmonary embolism. Res Diagn Interv Imaging 2023 Jun;6:100027. [doi: 10.1016/j.redii.2023.100027] [Medline: 39077547]
47. MacLean C, Cavallucci D. Assessing fine-tuned NER models with limited data in French: automating detection of new technologies, technological domains, and startup names in renewable energy. MAKE 2024 Aug 27;6(3):1953-1968. [doi: 10.3390/make6030096]
48. Touchent R, Romary L, Clergerie E. CamemBERT-bio: leveraging continual pre-training for cost-effective models on french biomedical data. arXiv. Preprint posted online on Jun 27, 2023. [doi: 10.48550/ARXIV.2306.15550]
49. Baldini I, Wei D, Natesan Ramamurthy K, Singh M, Yurochkin M. Your fairness may vary: pretrained language model fairness in toxic text classification. Presented at: Findings of the Association for Computational Linguistics; May 22-27, 2022 URL: https://aclanthology.org/2022.findings-acl [doi: 10.18653/v1/2022.findings-acl.176]
50. Hendrycks D, Liu X, Wallace E, Dziedzic A, Krishnan R, Song D. Pretrained transformers improve out-of-distribution robustness. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020 URL: https://www.aclweb.org/anthology/2020.acl-main [doi: 10.18653/v1/2020.acl-main.244]
51. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of experts. arXiv. Preprint posted online on Jan 8, 2024. [doi: 10.48550/arXiv.2401.04088]
52. Honnibal M, Montani I. SpaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Research. 2017. URL: https://sentometrics-research.com/publication/72/ [accessed 2025-02-28]
53. Wajsburt P, Petit-Jean T, Dura B, Cohen A, Jean C, Bey R. EDS-NLP: efficient information extraction from french clinical notes. GitHub. URL: http://aphp.github.io/edsnlp [accessed 2025-04-22]
54. Explosion. Prodigy. 2022. URL: https://prodi.gy/ [accessed 2025-04-22]
55. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973 Oct;33(3):613-619. [doi: 10.1177/001316447303300309]
56. Poiron A, Cabon S, Cuggia M. How trueness of clinical decision support systems based on machine learning is assessed? Stud Health Technol Inform 2024 Aug 22;316:813-817. [doi: 10.3233/SHTI240535] [Medline: 39176916]
57. Komiya K, Suzuki M, Iwakura T, Sasaki M, Shinnou H. Comparison of methods to annotate named entity corpora. ACM Trans Asian Low-Resour Lang Inf Process 2018 Dec 31;17(4):1-16. [doi: 10.1145/3218820]
58. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform 2021 Jan 18;22(1):146-163. [doi: 10.1093/bib/bbz130] [Medline: 31838514]
59. Galileo AI. What is NER and why it's hard to get right. Galileo AI. URL: https://www.galileo.ai/blog/what-is-ner-and-why-it-s-hard-to-get-right [accessed 2025-04-22]
60. Explosion AI. Named entity recognition. Prodigy. URL: https://prodi.gy/docs/named-entity-recognition [accessed 2025-04-22]
61. Herman Bernardim Andrade G, Yada S, Aramaki E. Is boundary annotation necessary? evaluating boundary-free approaches to improve clinical named entity annotation efficiency: case study. JMIR Med Inform 2024;12:e59680. [doi: 10.2196/59680]
62. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Sep 1;31(9):1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
63. Baldridge J, Osborne M. Active learning and the total cost of annotation. Presented at: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; 2004.
64. Aggarwal U, Popescu A, Hudelot C. Active learning for imbalanced datasets. Presented at: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); Mar 1-5, 2020. [doi: 10.1109/WACV45572.2020.9093475]
65. Ollama. Ollama. URL: https://ollama.com/ [accessed 2025-04-16]

XSL•FO
RenderX

Abbreviations

CDW: clinical data warehouse
COPD: chronic obstructive pulmonary disease
LLM: large language model
MSEP: Medical Status Extraction Pipeline
n2c2: National NLP Clinical Challenges
NER: named entity recognition
NLP: natural language processing
PS: preannotated filtered sentences
US: unfiltered sentences

Edited by A Coristine; submitted 13.May.2025; peer-reviewed by Á García-Barragán, L Ma, S Matos; revised version received 26.Dec.2025; accepted 27.Dec.2025; published 26.Mar.2026.

Please cite as:
Dong C, Delange B, Poiron A, El Azzouzi M, François C, Bouzillé G, Cuggia M, Cabon S
A Sentence Classification–Based Medical Status Extraction Pipeline for Electronic Health Records: Institutional Case Study
JMIR Med Inform 2026;14:e77409
URL: https://medinform.jmir.org/2026/1/e77409
doi:10.2196/77409

© Chuanming Dong, Boris Delange, Alex Poiron, Mohamed El Azzouzi, Clément François, Guillaume Bouzillé, Marc Cuggia, Sandie Cabon. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Iterative Large Language Model–Guided Sampling and Expert-Annotated Benchmark Corpus for Harmful Suicide Content Detection: Development and Validation Study

Kyumin Park[1*], MS; Myung Jae Baik[2*], MD; YeongJun Hwang[3*], MS; Yen Shin[4], BS; HoJae Lee[4], BS; Ruda Lee[5], MA; Sang Min Lee[2], PhD; Je Young Hannah Sun[2], MD; Ah Rah Lee[2], PhD; Si Yeun Yoon[2], MA; Dong-ho Lee[1], PhD; Jihyung Moon[1], MS; JinYeong Bak[3], PhD; Kyunghyun Cho[6], PhD; Jong-Woo Paik[2], PhD; Sungjoon Park[1], PhD

[1]SoftlyAI, Seoul, Republic of Korea

[2]Department of Psychiatry, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[3]Department of Artificial Intelligence, Sungkyunkwan University, Office 27306, Engineering Building 2, 2066 Seobu-ro Jangan-gu, Suwon-si, Gyeonggi-do, Republic of Korea

[4]KAIST, Daejeon, Republic of Korea

[5]Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

[6]Department of Computer Science, New York University, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
JinYeong Bak, PhD
Department of Artificial Intelligence, Sungkyunkwan University, Office 27306, Engineering Building 2, 2066 Seobu-ro Jangan-gu, Suwon-si, Gyeonggi-do, Republic of Korea

Abstract

Background: Harmful suicide content on the internet poses significant risks, as it can induce suicidal thoughts and behaviors, particularly among vulnerable populations. Despite global efforts, existing moderation approaches remain insufficient, especially in high-risk regions such as South Korea, which has the highest suicide rate among Organisation for Economic Co-operation and Development countries. Previous research has primarily focused on assessing the suicide risk of the authors who wrote the content rather than the harmfulness of content itself which potentially leads the readers to self-harm or suicide, highlighting a critical gap in current approaches. Our study addresses this gap by shifting the focus from assessing the suicide risk of content authors to evaluating the harmfulness of the content itself and its potential to induce suicide risk among readers.

Objective: This study aimed to develop an artificial intelligence (AI)–driven system for classifying online suicide-related content into 5 levels: illegal, harmful, potentially harmful, harmless, and non–suicide-related. In addition, the researchers construct a multimodal benchmark dataset with expert annotations to improve content moderation and assist AI models in detecting and regulating harmful content more effectively.

Methods: We collected 43,244 user-generated posts from various online sources, including social media, question and answer (Q&A) platforms, and online communities. To reduce the workload on human annotators, GPT-4 was used for preannotation, filtering, and categorizing content before manual review by medical professionals. A task description document ensured consistency in classification. Ultimately, a benchmark dataset of 452 manually labeled entries was developed, including both Korean and English versions, to support AI-based moderation. The study also evaluated zero-shot and few-shot learning to determine the best AI approach for detecting harmful content.

Results: The multimodal benchmark dataset showed that GPT-4 achieved the highest F_1-scores (66.46 for illegal and 77.09 for harmful content detection). Image descriptions improved classification accuracy, while directly using raw images slightly decreased performance. Few-shot learning significantly enhanced detection, demonstrating that small but high-quality datasets could improve AI-driven moderation. However, translation challenges were observed, particularly in suicide-related slang and abbreviations, which were sometimes inaccurately conveyed in the English benchmark.

Conclusions: This study provides a high-quality benchmark for AI-based suicide content detection, proving that large language models can effectively assist in content moderation while reducing the burden on human moderators. Future work will focus on enhancing real-time detection and improving the handling of subtle or disguised harmful content.

(JMIR Med Inform 2026;14:e73725) doi:10.2196/73725

XSL•FO
RenderX

KEYWORDS

artificial intelligence; dataset; suicide; suicide-related content; large language models

Introduction

Harmful suicide content on the internet poses a significant risk because it can induce suicidal thoughts in readers, potentially leading to self-harm or suicide [1,2]. The harmful suicide content includes materials that encourage or glorify suicide [3], making it appear as an attractive option and sharing suicide methods or instilling suicide knowledge in individuals with suicidal thoughts, thereby increasing the likelihood of actual suicide attempts [4]. In some cases, exposure to such harmful suicide content has led middle school students to commit suicide [5]. An analysis of adolescent suicide cases reveals that this age group, particularly female adolescents, is more vulnerable to the influence of triggering content [6,7]. Therefore, it is crucial to moderate such harmful suicide content before it spreads extensively.

Therefore, efforts to moderate harmful suicide content are intensifying. In the United States, initiatives focus on raising public awareness and safe content distribution, aligning with the World Health Organization guidelines [8]. Meanwhile, in 2022, the United Kingdom passed a law that makes such content illegal, emphasizing its serious commitment to addressing this issue [9]. In the Republic of Korea, which has the highest suicide rates among Organisation for Economic Co-operation and Development countries [10], the National Assembly of the Republic of Korea amended the Suicide Prevention Act, and the government has declared the dissemination of such content as illegal since 2019 [2]. Despite the increasing spread of harmful suicide content, its moderation is currently handled by only a single official and fewer than a thousand volunteers [11,12]. Considering the extensive use of social media in Korea [13], monitoring the large amounts of content is extremely challenging. In addition, moderating suicide content often leads to a high level of mental stress, hindering their ability to consistently and effectively monitor such content. Therefore, the need for an automatic harmful suicide content moderation system is urgent. The system can efficiently manage a growing volume of the content and ease the burden on human moderators.

However, previous research on online suicide content primarily focused on predicting the suicide risk of the authors who wrote the content. Zirikly et al [14] classified the suicide risk of authors based on content posted online (Reddit) into 4 levels. Similarly, Milne et al [15] and Yates et al [16] conducted research to predict the suicide and self-harm risks of online content authors. Subsequently, Yang et al [17] and Sawhney et al [18] used weakly supervised learning to enhance detection performance or collaborated with clinicians. Furthermore, Rawat et al [19] and Sawhney et al [20] performed tasks to detect suicide ideation and suicide events. All these studies focused on detecting the suicide risk of the author presented in the posts; they did not consider the potential propagation of suicide risk to readers by increasing their susceptibility to self-harm or suicidal behaviors. Other studies focused on understanding the negative effects of suicide content [6,21], or identifying the individuals who are most affected by the content [22-25]. Studies among Chinese adolescents have shown significant correlations between digital media usage and suicide or self-harm [23], and a meaningful relationship between suicide cases among Korean youths and searches related to suicide or self-harm [25]. In addition, three-quarters of young adults who have attempted suicide have reported using the internet for suicide or self-harm–related reasons [26], highlighting the risk posed by information that can induce suicide or self-harm.

Therefore, we introduce a *harmful suicide content detection* task that determines the level of harmfulness of the harmful suicide content to viewers. We then develop a multimodal *harmful suicide content detection benchmark* and a *task description document*. This document contains detailed instructions for annotators on how to assess the harmfulness of suicide-related content, which could also be useful for building instructions for large language models (LLMs). The benchmark and the document are developed by medical professionals, because such content might involve harmful visual language information that requires the judgment of the professionals (eg, self-harm photographs, or name of illegal drugs that can be used for suicide). Because labeling harmful suicide content causes mental stress, we focus on creating a small yet high-quality dataset. Furthermore, we demonstrate various methods using LLMs that can be effectively performed with few-shot examples.

Our contributions are as follows: (1) We propose a harmful suicide content detection task that classifies multimodal harmful suicide content as illegal, harmful, potentially harmful, harmless, or non–suicide-related. (2) We build a harmful suicide content detection benchmark that is multimodal and a Korean benchmark of 452 curated user-generated contents with corresponding medical expert annotations and a detailed task description document including the task details and instructions for annotators. (3) We create an English harmful suicide content detection benchmark translated from the Korean benchmark using a model (GPT-4) and analyze the quality and issues of translating suicide contents. (5) We demonstrate strategies to use LLMs to detect harmful suicide content by using the task description document and a small yet high-quality harmful suicide content detection benchmark. We test various closed and open-sourced LLMs using the machine-translated English benchmark. We observe that GPT-4 achieves F_1-scores of 66.46 and 77.09 in detecting illegal and harmful suicide content, respectively.

Methods

Study Design

In this section, we describe the design process for the harmful suicide content detection task and the steps to construct a harmful suicide content detection benchmark for this task. First, in the “Task Design” subsection, we outline the process of designing the task with emphasis on real-world implementation for detecting harmful suicide content in real time from online sources. This design phase includes detailed considerations for building the dataset.

XSL•FO
RenderX

Next, the "Data Processing Workflow" subsection describes the end-to-end process, from collecting harmful suicide content to developing the final harmful suicide content detection benchmark. Harmful suicide content is gathered from various online sources, and the preannotation process [27,28], which uses LLMs, is used to reduce annotators' exposure to vast quantities of harmful suicide content. Feedback generated from the annotation process by medical experts is used to update the task description document, which includes the categories and descriptions of harmful suicide content. This task description document is then used as an instruction for the preannotation process, creating an iterative process that builds and refines the harmful suicide content detection benchmark. Through this process, we ensure the inclusion of diverse, accurately labeled real-world harmful suicide content in the harmful suicide content detection benchmark while constructing a task description document that provides precise and comprehensive classification and explanations of the harmful suicide content.

As a result, from a collection of 43,244 harmful suicide content data points, we created a harmful suicide content detection benchmark consisting of 452 data points annotated by medical experts with suicide content categories, subcategories, and rationales for their classifications.

Ethical Considerations

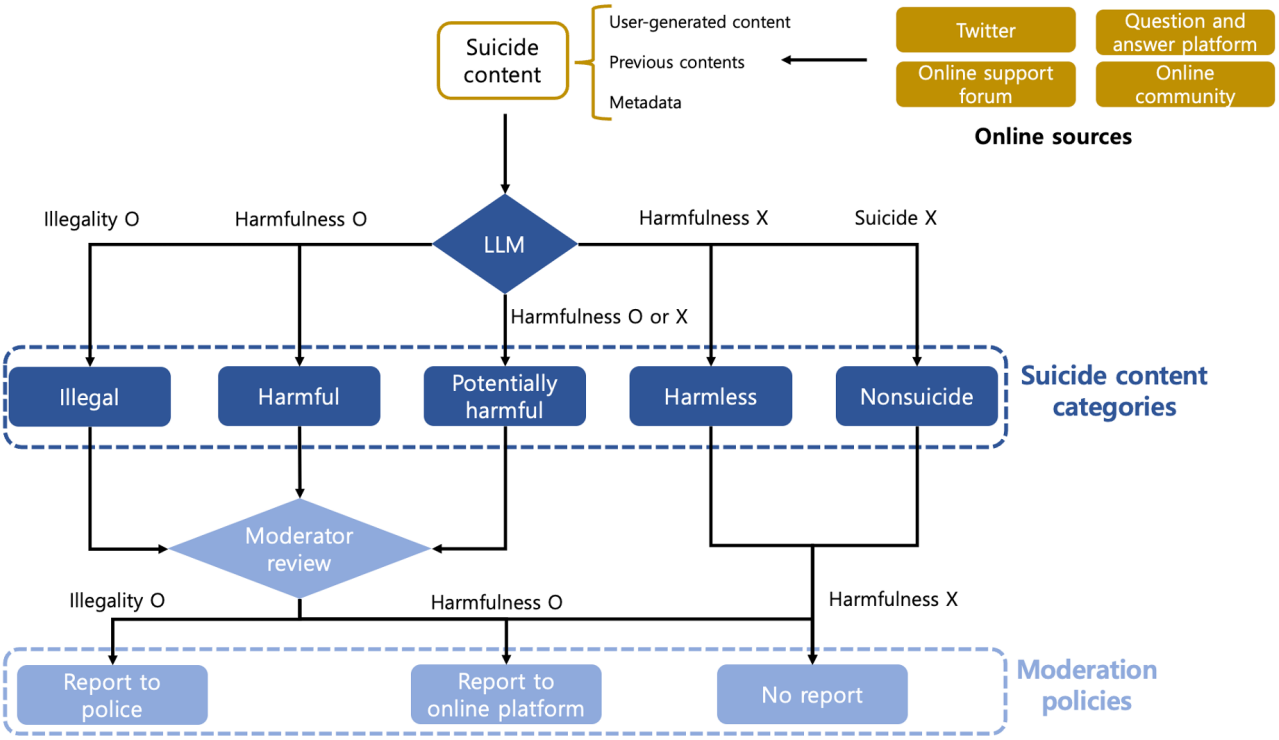
We obtained institutional review board (IRB) approval (approval number: KHUH 2023-04-072) from the IRB of Kyung-Hee University Hospital, South Korea, based on the principles of the Declaration of Helsinki, ensuring adherence to ethical standards. The IRB approval includes the collection of suicide-related data, the development of guidelines for data classification, and data labeling. It also specifies the acquisition of data through the collection of publicly available posts from online spaces targeting the general public, and it outlines the development of an artificial intelligence model for detecting suicide-related content using these data. All personally identifiable information (PII) (eg, user IDs, names, addresses, and phone numbers) was anonymized and replaced with standardized tags. The benchmark contains extremely disturbing text and images, including self-harming photographs, blood, tools used for suicide and self-harm, and drug information. Even among medical professionals and researchers, prolonged exposure to such images can lead to severe mental stress. Therefore, we have deliberately chosen not to aim for the creation of a large-scale dataset but rather to limit the workload to prevent further intensifying mental stress. All these processes were conducted with IRB approval obtained prior to data collection. Given the nature of harmful suicide content and the legal restrictions against its unrestricted distribution, it is challenging to share the benchmark dataset openly. We understand the legal implications of distributing data containing information that potentially induces suicide. Despite these concerns, we believe that collecting such data to build a benchmark and conducting research to prevent its spread on the internet outweigh these legal issues. Access to the benchmark will be strictly limited, allowing only those researchers with IRB approval and a commitment not to distribute the content further, ensuring responsible use for research purposes only and adherence to legal standards. We believe that our work contributes significantly to the ongoing international effort against harmful suicide content and hope to aid in preventing its spread on the internet.

Task Design

Figure 1 illustrates the concept of using harmful suicide content detection in a real-world *moderation system*. The moderation system uses a model to automatically detect harmful suicide content and checks for illegal or harmful content and implements the appropriate *moderation policy* through a moderator's review. As this study introduces the harmful suicide content detection for the first time, our focus is on developing a model to automatically detect harmful suicide content rather than implementing an end-to-end moderation system. Therefore, this study focused on developing a harmful suicide content detection with this moderation system in mind, leaving the implementation of an end-to-end moderation system for future work. Specifically, we considered the inputs and outputs of harmful suicide content detection, considering the various real-world information on harmful suicide content that a moderation system might encounter as well as the moderation actions by a moderator.

XSL•FO
RenderX

Figure 1. Moderation system for harmful suicide content detection, categorizing online user-generated content into 5 classes by legality, harmfulness, and suicide relation. A moderator reviews content with potential illegality or harm, leading to legal reporting or content removal requests. No action is taken if no risks are found. "O" indicates that the corresponding value exists, and "X" indicates that the value does not exist. "O/X" denotes the potentiality of whether the corresponding value exists or not. LLM: large language model.



The model classifies the content into 5 distinct *suicide categories* based on illegality, harmfulness, and suicide-related aspects. For categories identified as having minimal harmfulness, a moderator validates the harmfulness through a review. Finally, the moderator moderates harmful suicide content using a moderation policy suitable for the identified harmfulness and illegality of the content, thereby minimizing human intervention and effectively moderating harmful suicide content.

To ensure that the detection system is applicable to real-world online data, the task targets various data from diverse sources to encompass a broad spectrum of user-generated content encountered on the internet. This approach ensures that the system effectively addresses the complexities and nuances of online posts ("Task Input" section).

The classification results produced by the harmful suicide content detection were crafted considering the functionality of a moderation system in mind. This means that the categories into which the content is sorted are specifically designed to facilitate the practical use of these results in moderating the content, ensuring that the system can serve as an effective tool for maintaining online safety and supporting mental well-being ("Task Output" section).

A moderator review was designed to validate the model's classification results and implement appropriate moderation. Specifically, for content categorized under the harmful categories, the process validates the harmfulness to ensure the reliability of the moderation system. In addition, because this process involves confirming the results of the model rather than newly identifying and classifying suicide content, it is more efficient in terms of reducing mental stress on moderators. Finally, suitable moderation policies were implemented based on the validation results. This system was developed to regulate suicide content and execute different moderation policies based on the illegality and harmfulness, thus preventing the spread of harmful suicide content online ("Moderator Review" section).

In summary, the entire process involves the model classifying online content into suicide categories, the moderator reviewing the results, and then implementing the corresponding moderation policy to ensure that the moderation system functions effectively. Throughout this process, multimodality information such as text and image data are used to reflect various aspects of the content in the model's input. Metadata from diverse sources and previous content serve as context. The model's output comprises 5 suicide categories, each differentiated by the presence or absence of illegality, harmfulness, and suicide-related aspects. Finally, the moderator review efficiently uses the model's classification results for validation, and effective moderation policies are implemented based on the content's illegality and harmfulness, thereby preventing the spread of harmful suicide content online.

Task Input

Considerations

We consider the following for designing the input of the task. (1) *Multimodality*: Because 50% of harmful suicide content contain images or videos [29], we consider text and images as inputs. (2) *Source diversity*: Harmful suicide content appears across various platforms, from social media to online communities [29]. We collected data from diverse sources for comprehensive coverage. (3) *Context information*: We also

XSL•FO
RenderX

incorporate previous content and metadata into our inputs. Previous content reveals the context of the target content, whereas metadata, such as user descriptions and view counts, provides additional insights, aiding in accurate harmfulness assessment.

Inputs

Given the considerations, the inputs for the task are as follows: (1) *User-generated content*: Contents created by users. The content includes text, and possibly images and URLs. Images and URLs are converted to text manually or using machine learning models (eg, image captioning and summarization). (2) *Previous content*: Previous content is often required because it provides context, clarifies references, and provides background information essential for the full comprehension of user-generated content. (3) *Metadata*: Other contextual information about the user-generated content such as view counts, like counts, creation time, user self-description, and so on.

Task Output

Considerations

We consider the following for designing the outputs of the task. (1) *Expert judgment*: Harmful suicide content involves specific terminology related to suicide, such as professional drug names, slang, and abbreviations. Thus, clinical expertise is required to accurately determine the legality and harmfulness of such suicide content and to decide on an appropriate response to the content. (2) *Moderation policy*: If an automatic harmful suicide content detection model is developed, it should be part of a moderation system and collaborate with human moderators or domain experts [18]. This implies that once the model detects harmful suicide content, it is necessary to consider appropriate actions. Therefore, the response of each output was considered when defining the output.

Outputs

We develop five harmful suicide content categories. Content should be mapped to one of the following categories: (1) *illegal* content that encourages or assists suicidal behavior; (2) *legal but harmful* content that, while not illegal, significantly induces suicide; (3) *potentially harmful* content that could be triggering for certain individuals, whereas it may be benign for others; (4) *harmless* content that is either neutral or positive for suicide; and (5) *nonsuicide* content that is not related to suicide.

Moderator Review

Moderator review includes the process of reexamining the suicide content classified by the model using moderator (eg, a clinical expert) and implementing the appropriate moderation policy. The moderation system identifies the harmfulness and illegality of suicide content and implements a corresponding moderation policy to block the spread of such content online. Thus, through moderator review, the moderator (1) verifies the classification result of the model and (2) implements the corresponding moderation policy. The moderator review is conducted for results classified as illegal, harmful, and potentially harmful because most online information is unrelated to suicide and reviewing all the information would increase moderator fatigue. Hence, reviews are conducted only for suicide information that may cause harm. The moderator verifies the illegality and harmfulness of content within these categories and conducts the corresponding moderation policy. Therefore, the moderator requires knowledge to comprehend and understand the content and distinctions of suicide content.

Moderation Policies

The moderator reviews the model's classification results and implements a corresponding moderation policy. The moderation policies are as follows: (1) *Report to police*: This is the strongest form of moderation policy intended to subject content creators to legal regulations by reporting to legal institutions. (2) *Report to online source*: Reporting the content to the online source where it is posted intends to prevent the spread of harmful information by requesting the deletion of the content. (3) *No report*: No additional actions, such as reporting the posts, are taken, allowing it to circulate online.

Report to police responds to content within the illegal suicide category containing illegal information. According to Korean law, certain types of content related to suicide are defined as illegal. Such content often includes illegal activities, such as the sale of illegal drugs; hence, reporting to legal institutions (eg, the police) imposes legal sanctions on the poster of such content.

Report to online source prevents the online spread of content containing or potentially containing harmful information related to suicide, such as illegal, harmful, and potentially harmful content. Because information spreads quickly online, it is reported to the online source where it was posted, and its removal is requested to prevent dissemination. For potentially harmful information, the harmfulness of which can vary depending on the reader, the moderator assesses the degree of harmfulness and reports whether it is severe.

No report is for harmless or nonsuicide content that poses no problem when posted online. Most online content is unrelated to suicide; therefore, it does not require reporting.

Table 1 shows the features of each category in terms of legality, harmfulness, and association with suicide. *Illegal suicide content* contains the most dangerous information, explicitly encouraging or facilitating suicidal behaviors. This category is critical for immediate intervention, embodying content that can actively propel individuals toward self-harm or suicide. *Harmful suicide content*, while not directly inciting suicide, significantly affects the audience by portraying suicide or self-harm in a manner that can trigger such actions among vulnerable individuals. The specificity of the depiction, whether through graphic imagery or detailed descriptions, amplifies its potential harm, making it a crucial target for moderating content. *Potentially harmful suicide content* traverses a gray area, with content that might not universally trigger harmful behaviors but could potentially do so in susceptible populations. This category underscores the complex challenge of content moderation, in which the impact of the content is not universally harmful and may vary significantly among individuals. *Harmless suicide content* focuses on providing support, hope, or neutral information regarding suicide without posing a risk of harm. This category

XSL•FO
RenderX

plays an essential role in suicide prevention by offering resources, support, and information to reduce suicide rates. Finally, *nonsuicide content* serves as a catchall for material that does not pertain to suicide or self-harm, highlighting the importance of distinguishing between genuinely harmful content and content unrelated to suicide.

Table . Illegality, harmfulness, and suicide relativity of categories and the moderation protocols[a].

Category	Illegality	Harmfulness	Suicide-related	Moderator review	Moderation policy
Illegal	O	O	O	O	Report to police and online source
Harmful	X	O	O	O	Report to online source
Potentially harmful	X	Δ	O	O	Report to online source or no report
Harmless	X	X	O	X	No report
Nonsuicide	X	N/A[b]	X	X	No report

[a]The (Δ) symbol represents a state that is in a gray area, indicating that the characteristic is neither fully present nor completely absent. "O" denotes the presence or applicability of a given attribute, whereas "X" denotes its absence. For instance, data categorized as "Harmful" are characterized by the absence of illegality (X), the presence of harmfulness (O), suicide-related (O), and the moderator review (O).

[b]N/A: not applicable.

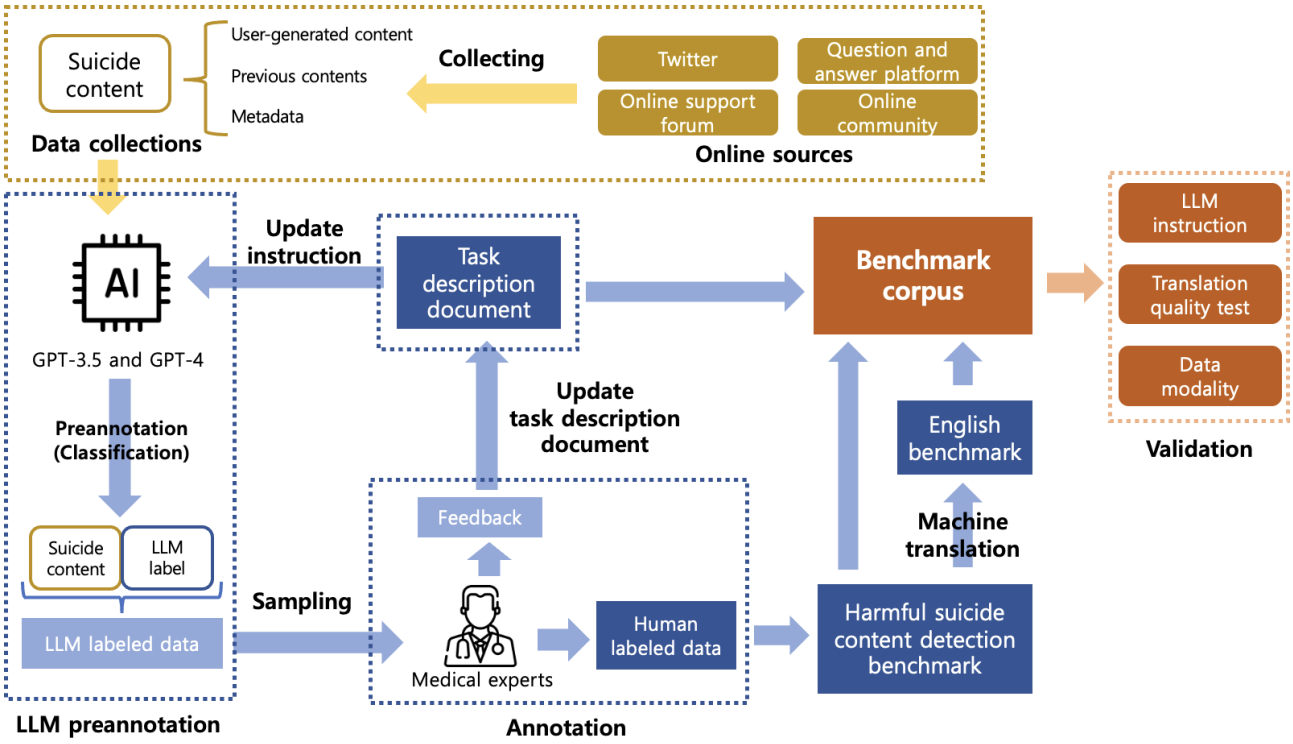
Data Processing Workflow

This section describes the data processing workflow used to create a high-quality harmful suicide content detection benchmark for detecting harmful suicide content. As illustrated in Figure 2, the workflow starts from data collection, detailed in “Data Collections” subsection, and outlines the process to develop a robust and diverse benchmark for harmful suicide content detection, described in “Data Annotation” subsection. The primary goal is to create a small but high-quality benchmark that can comprehensively capture real-world scenarios and task description document that can be effectively used for LLM instruction. Key challenges include the stressful nature of the annotation process, the extremely low prevalence of illegal suicide content, and the need to ensure the benchmark’s completeness and use, given that this is a newly proposed task that must reflect various aspects of real-world data.

The methods used to address these challenges include (1) LLM sampling to reduce annotator exposure to a large volume of harmful suicide content while increasing the representation of rare, illegal suicide content; (2) an iterative annotation approach to continuously update the task description document and incorporate real-world scenarios, aiming to reflect diverse data until no further updates are required, while also considering modality (image and context) to adapt annotators effectively; and (3) filtering and validation, which involve updating the harmful suicide content detection benchmark to align with the latest task description document, categorizing and subcategorizing data accurately, and removing outliers such as incomprehensible content.

XSL•FO
RenderX

Figure 2. Overview of the data processing workflow, which consists of data collection, benchmark construction, and benchmark validation. We collect suicide content and annotate it with medical experts, leveraging LLM-based preannotation to balance categories, and we iteratively refine the task description document through feedback. We then construct the benchmark corpus with human-labeled data, an English-translated version, and the updated task description, and we validate it for applicability to LLMs. AI: artificial intelligence; LLM: large language model.



Data Collections

Developing a large-scale harmful suicide content dataset is highly challenging. Harmful suicide content is infrequently encountered in real-world scenarios [30], and the distressing nature of such content can cause mental strain for annotators. In addition, obtaining annotations from medical experts is expensive. Therefore, we focus on developing a high-quality curated benchmark dataset. Prior to the dataset collection, we obtained approval from the IRB of Kyung-Hee University Hospital, South Korea.

To cover the diverse source domains of the content, we collect user-generated content related to suicide from social media, Q&A platforms, online support forums, and online communities. Table 2 lists the number of raw data, benchmark data, descriptions of content, previous content, and metadata of each source.

Table . Number of collected harmful suicide content and harmful suicide content detection benchmark dataset for each domain[a].

Source	Raw data (no. of images)	Benchmark (no. of images)	Content	Previous content	Metadata
Twitter	12,125 (3,671)	359 (78)	Tweets written by users	Previous tweets in the thread where the content is written	User description, view count, like count, etc
Online community	794 (794)	48 (48)	Title and bodies of post written by users	N/A[b]	User nickname, view count
Q&A[c] platform	13,104 (0)	33 (0)	Question or answers written by users	Questions (if the content is an answer)	N/A
Online support forum	17,325 (0)	23 (0)	Counseling request posts or responses written by users or counselors	Counseling request posts (if the content is a response)	N/A
Total	43,244 (4,429)	452 (126)	N/A	N/A	N/A

[a]User-generated suicide content was collected along with source-specific metadata and previous content as context.

[b]N/A: not applicable.

[c]Q&A: question and answer.

XSL•FO
RenderX

Twitter

Twitter constitutes the majority of social media posts flagged for containing suicide-inducing information, with a substantial share of 74.69% [29]. To collect data related to suicide from Twitter, we used the Twitter application programming interface (API) v2 to gather posts that include suicide-related keywords in their text or hashtags. These suicide-related keywords were collected from previous research [31] and the guidelines of the "Korean Suicide Inducing Information Monitoring Group" [32]. We gathered 12,021 tweets, including 3635 with images, from May to August 2023 using the Twitter API. The suicide-related keywords used in the Twitter API are summarized in Table 3.

XSL•FO
RenderX

Table . Suicide-related keywords[a].

Classification of search terms	Search terms	Synonyms	Korean version of search terms	Korean version of synonyms
Suicide-related terms	Suicide	• Suicide	• 자살	• 자살
	Suicide method	• Suicide method • How to suicide • How to commit suicide	• 자살방법	• 자살방법 • 자살하는법 • 자살하는방법
	Dying method	• Dying method • How to die	• 죽는방법	• 죽는방법 • 죽는법
	Suicidal idea	• Suicidal idea • Suicidal thinking	• 자살충동	• 자살충동 • 자살생각
	Fall down	• Fall down • Suicide by jumping from a height	• 투신	• 투신 • 투신자살
	Hanging	• Hanging • Hanging suicide • Neck hanging	• 목맴	• 목맴 • 목맴자살 • 목매달기
	Will	• Will • ·How to write will	• 유서	• 유서 • 유서쓰는법
Self-harm–related terms	Self-harm	• Self-harm	• 자해	• 자해
	Self-harm method	• Self-harm method • How to self-harm	• 자해방법	• 자해방법 • 자해하는법
	Wrist cutting	• Wrist cutting • How to cut my wrist • Wrist cutting method	• 손목자해	• 손목자해 • 손목자해하는법 • 손목자해방법
	Self-harm wound	• Self-harm wound • Self-harm mark • Treatment for self-harm wound	• 자해흉터	• 자해흉터 • 자해자국 • 자해흉터치료
	Drug overdose	• Drug overdose • Drug lethal dose	• 약물과다복용	• 약물과다복용 • 약물치사량
	Acetaminophen	• Acetaminophen overdose • Acetaminophen lethal dose	• 타이레놀	• 타이레놀과다복용 • 타이레놀치사량

XSL•FO
RenderX

Classification of search terms	Search terms	Synonyms	Korean version of search terms	Korean version of synonyms
Suicide risk factor terms	Academic score	• Academic concern • Academic score	• 성적	• 성적 • 성적고민
	Bullying	• Bullying • Covert bullying • Outcast	• 왕따	• 왕따 • 은따 • 따돌림
	School violence	• School violence	• 학교폭력	• 학교폭력 • 학폭
	Family troubles	• Family troubles	• 가족문제	• 가족문제
	Domestic violence	• Domestic violence	• 가정폭력	• 가정폭력
	Dropout	• Dropout • How to drop out • Dropout method	• 자퇴	• 자퇴 • 자퇴하는법 • 자퇴방법
	Career	• Career • Career concern	• 진로	• 진로 • 진로고민
Suicide prevention terms	Suicide prevention	• Suicide prevention	• 자살예방	• ·자살예방
	Call for life	• Call for life • Call for life of Korea	• 생명의전화	• 생명의전화 • 한국생명의전화
	Suicide prevention center	• Suicide prevention center • 1393	• 자살예방센터	• 자살예방센터 • 1393
	Psychiatry	• ·Psychiatry • ·Neuropsychiatry • Psychiatry department • Mental hospital	• 정신과	• ·정신과 • ·신경정신과 • ·정신건강의학과 • 정신병원
	Mental health center	• Mental health center	• 정신건강복지센터	• 정신건강복지센터
Depression-related terms	Depression	• ·Depression • ·Depressed • ·Depressive disorder • ·Depressive symptom	• 우울증	• 우울증 • 우울 • ·우울장애 • ·우울증상

[a] Search terms and synonyms are defined in the study by Park et al [33].

Q&A Platform

On Q&A platforms, users often post questions about suicide-related issues, such as suicide methods, or respond to these queries. We collected questions and answers containing suicide-related keywords from Naver Knowledge In (a Korean Q&A platform) [33]. We collected data from March 2022 to March 2023, using the same keywords as those used for Twitter, resulting in 13,104 content items.

Online Support Forum

In online support forums, people write about their suicide-related concerns, and counselors provide responses to support them [31]. We collected posts from Lifeline Korea [34] and the Companions of Life Suicide Prevention Counselling [35]. We collected 17,325 pieces of content posted from March 2021 to June 2023.

Online Community

DCinside [36], a widely used online community in Korea comparable to Reddit, includes boards that function similarly to subreddits. We collected posts from 2 depression-focused boards (depression-minor and depression-mini boards) on the DCinside, known to contain suicide-related posts and where actual suicide incidents have been reported [37]. We collected posts including those containing images, resulting in a total of 794 data entries. Overall, we collected 43,244 entries of harmful suicide content, in total, from 4 online sources, with each harmful suicide content entry comprising content, previous content, and metadata.

Data Preprocessing

To remove private information from the benchmark and generate information that requires human intervention, such as link or image descriptions, we took the following steps. First, we

XSL•FO
RenderX

removed all PII. This involves replacing URLs, names, locations, phone numbers, emails, and IDs within the text with corresponding tags. Thereafter, we provided supplementary descriptions for the contents of the external links. Given that these links may contain significant information for accurately understanding the content, we manually reviewed the links and summarized their content. Third, we added text descriptions to the images whenever they were included in the content. We used GPT-4 to generate initial descriptions, which were subsequently reviewed and refined for accuracy by the researchers. Consequently, all PII values were removed from the text of the data, and we created link descriptions that summarized the content of any URLs present in the content text, along with text descriptions for the images.

Data Annotation

Task Description Document

The task description document was designed to explain the harmful suicide content detection and to provide guidance to the annotators. It contains vital information, including the purpose of identifying harmful suicide content and a detailed guide for annotating the content. In addition, it outlines the categories and subcategories of harmful suicide content, supplemented with real-world examples.

Our basis for understanding the definitions, categories, and examples of harmful suicide content was the “Korean Suicide Prevention Law” [2] and documents published by the “Korea Life Respect Hope Foundation’s suicide/harmful information monitoring team” [32]. We found that certain category names and descriptions were unclear or overlapped, thus requiring more distinct clarifications. To address this, we involved medical professionals in the data annotation process, which led to significant revisions and refinements of the categories and their descriptions, as well as the expansion of examples for each category. Following the studies by Fiesler et al [38] and Moon et al [39], we used an iterative coding process such that the medical experts individually annotate the real-world content, come together to refine the task description document, and then repeat the coding process individually. This updating process was iterative and performed 3 times to ensure comprehensive refinement. Further details of the iterative process are presented in the “Annotation Process” section. We demonstrate each category and its description in Table 4.

Table . Name and description of each suicide category.

Name	Description
Illegal suicide content	Content that can actively encourage others to commit suicide or assist suicide behavior.
Harmful suicide content	Harmful content that is not as harmful as illegal suicide content but clearly has the effect of causing suicide or self-harm in the general public.
Potentially harmful suicide content	Content that may trigger suicide or self-harm in some people but may not cause it in others or may rather have a positive effect in others.
Harmless suicide content	Content that is not harmful, such as content that helps prevent suicide to the general public or provides neutral information related to suicide.
Nonsuicide content	Content unrelated to suicide.

Annotation Process

The annotation process was divided into 3 phases. In each phase, medical experts (a clinical expert with an MD degree and a psychiatry professor with a PhD degree) annotated real-world harmful suicide content, using the task description document as a reference. At the end of each phase, the authors and annotators reviewed and enhanced the task description document through discussions, before proceeding to the next phase.

In the first phase, medical professionals annotated harmful suicide content by referring to the initial task description document. Before starting the annotation, we preannotated and sampled the contents to be annotated from the data collections, as illustrated in Figure 3. Although the contents are gathered using suicide-related keywords, only a small fraction is actually harmful suicide content that can cause harm to others. Therefore, we used the task description document as instruction for the LLMs, allowing them to preliminarily categorize the content into predefined categories. This approach enhances the efficiency of the annotation process for medical experts and reduces mental strain and costs. Consequently, we used the OpenAI GPT API to preannotation and sampled 196 harmful suicide content for human annotation from the collected 2272 Twitter data, 17,325 online forum data, and 13,104 Q&A data. Medical professionals then proceeded to annotate the sampled 196 harmful suicide content by following the annotation protocol and the initial task description document. The annotation protocol is described in the later part of this section. During the annotation process, they did not refer to the LLM labels generated by the LLM. Following the annotation, both the categories and subcategories were updated, leading to a revision of the task description document. Specifically, we refined 7 subcategories, added 2 new ones, and removed one.

In the second phase, we diversified the harmful suicide content in the benchmark and refined the task description document. Before annotation, we further preannotated and sampled 175 harmful suicide content for annotation from a pool of 8408 Twitter data points collected between May and June 2023. Similar to the first phase, we preannotated them using OpenAI GPT API with instructions written based on the task description document. Subsequently, medical professionals began the annotation of sampled harmful suicide content, strictly adhering to the annotation protocol and using the revised version of the

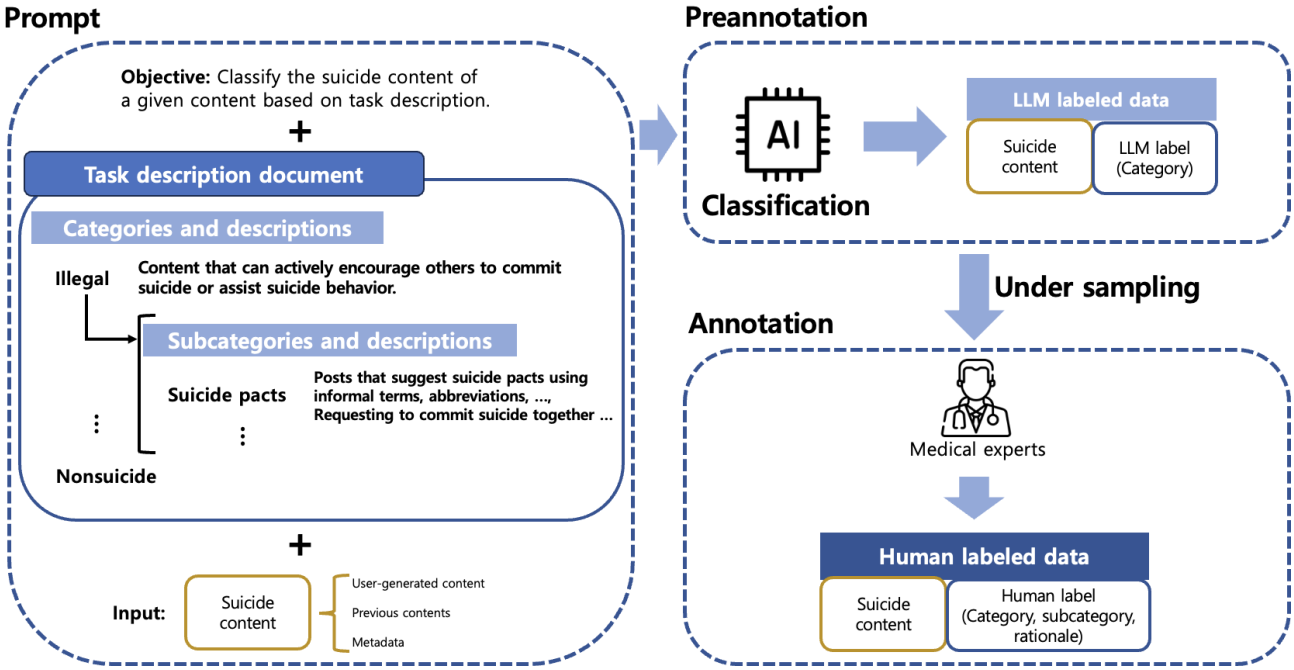
XSL•FO
RenderX

task description document as their guide. Once the annotation process was completed, we merged the 4 subcategories into 2.

In the final phase, we added multimodal (text and image) harmful suicide content to the benchmark dataset and included online communities as an additional source domain. For the image content, we initially generated textual descriptions of harmful images using visual language LLMs ("Data Preprocessing" subsection). These initial descriptions were then revised to correct any inaccuracies or fill in missing details. The refined descriptions were subsequently used to preannotate the content into categories and subcategories, as defined in the task description document from the second phase. Following this process, we preannotated and sampled 95 multimodal harmful suicide content items for annotation. Medical professionals then annotated based on the annotation protocol, and the task description document was finalized by revising the previous version.

Finally, we manually verified the entire benchmark dataset. This involved identifying and eliminating any remaining PIIs from all harmful suicide content and validating the final labels. During the finalization process, 14 content items were excluded from the benchmark. These contents deal with subcultures (such as games and comics) and, therefore, are incomprehensible to all annotators and cannot be categorized into any suicide category, leading to their exclusion. In addition, the task description document was completed, providing comprehensive information on the 5 categories and 25 subcategories, including their harmful category names, descriptions, and illustrative examples.

Figure 3. Illustration of the preannotation process. We construct prompts from the task description document, which includes categories, detailed descriptions, and subcategories (Table 4), and we use LLMs such as GPT-3.5 and GPT-4 to generate LLM-labeled data. We then sample the data to balance category distributions and have medical experts annotate them to build the human-labeled benchmark. AI: artificial intelligence; LLM: large language model.



Annotation Protocol

In every phase, we adopted a consensus-based method for biomedical research and clinical practice [40,41]. For each harmful suicide content, 2 separate medical professionals (a clinical expert with an MD degree and a psychiatry professor with a PhD degree) independently labeled the category, subcategory, and rationale for their decisions regarding both the category and the subcategory. Each individual annotator assigned the label based on a comprehensive review of the user-generated content (text and image), previous content, and metadata associated with the harmful suicide content. The Inter-Annotator Agreement for category labels reached a high agreement of 0.77 (Cohen κ) after the second phase of the annotation process. In cases where there is a discrepancy in the category label assigned by individual annotators, a consensus is established through the input of 3 annotators, which includes an additional clinical expert (a psychiatry professor with a PhD degree). During this consensus, rationales written by the 2 individual annotators are combined into a single rationale. In addition, annotators comment on any data whose association with suicide content is uncertain, as well as on instances that imply a potential need to revise the task description. These comments were used at the end of each annotation phase to refine and update the task description document.

Construction of English Benchmark

We further created an English benchmark by translating all input attributes (content text, link or image descriptions, and metadata such as user descriptions). To enhance the usability of the Korean harmful suicide content detection benchmark, it was translated into English using machine translation with GPT-4 - 0613 for each attribute. As a result, we produced an English harmful suicide content detection benchmark containing the same 452 data entries as the original Korean harmful suicide content detection benchmark. The quality of the English

XSL•FO
RenderX

benchmark is evaluated in detail in the "English Benchmark Validation" subsection.

Results

Overview

After constructing the harmful suicide content detection benchmark through the data processing workflow, we evaluated the dataset from 3 perspectives. First, we assess whether the categories and descriptions of harmful suicide content outlined in the task description document are appropriate for providing sufficient information as instructions to LLMs to classify suicide contents ("Leveraging Task Description" subsection).

Next, we evaluate whether the data modalities included in the benchmark (eg, images) are suitable for classification by LLMs ("Leveraging Multimodality" subsection). In addition, we verify whether the LLMs can perform classification using only few-shot examples without further training through few-shot experiments ("Leveraging Few-shot Examples" subsection).

Finally, for the English harmful suicide content detection benchmark, we conduct experiments using various LLMs to compare performance with the original Korean benchmark to verify the use and carry out a direct analysis of data quality to verify the translation's quality ("Comparison Between LLMs" subsection and "English Benchmark Validation" subsection).

Setup

We used the GPT-3.5-turbo-16k API with a temperature of 0.0 and default hyperparameters, conducting 3-8 runs to calculate the average and standard error. For the few-shot experiments, we adopted an *N*-way *K*-shot approach by selecting *K* samples from each of the 5 classes (n=5) in the training dataset ("Data Annotation" subsection).

Metrics

We used the following metrics:

1. Macro F_1 measures the overall performance across the 5 categories.
2. *Mean Absolute Error (MAE) of Harmfulness* measures the model's deviation in predicting harmfulness and is categorized into four levels: 3 (most harmful: Illegal Suicide Content), 2 (harmful: Harmful Suicide Content), 1 (potentially harmful: Potentially Harmful Suicide Content), and 0 (not harmful: Harmless Suicide Content and Nonsuicide Content). This metric assesses the extent of the error in terms of harmfulness.
3. *Illegal* identifies illegal suicide content, and it is crucial for prompt regulation.
4. Harmful separates illegal or harmful suicide content from noncritical content, which is essential for moderating the content that poses harm.

In alignment with our objective of identifying and moderating as much harmful content as possible, our model is designed to initially detect harmful content, after which the results are carefully reviewed by a human moderator or expert. Hence, recall is prioritized over precision. This focus is particularly important for the Illegal and Harmful content categories, where missing instances of harmful suicide content pose a far greater risk than misclassifying content that is actually safe. Within our proposed framework, human moderators or clinical experts make the final judgment, ensuring that such misclassified cases can be corrected while minimizing the chance of harmful content going undetected. This design reflects how the system can function in real-world moderation scenarios, where recall is prioritized to maximize safety without disregarding precision.

Leveraging Task Description

We investigated the formulation of a task description document with diverse and extensive information into instructions because instruction construction significantly influences LLM performance [42-44]. The task description document for the harmful suicide content detection task contains crucial details, including the names and descriptions of 5 suicide categories as well as the names and explanations of 25 subcategories constituting up to 60% of the instruction at maximum. Category information includes detailed names and descriptions of the categories and subcategories. Thus, our experiments were designed to determine the details that most significantly impact performance by varying the granularity of the information.

Setup

We evaluate harmful suicide content detection performance by varying the detailed category information levels as follows: (1) category name; (2) category name and description; (3) category name with category description and subcategory name; and (4) category name with category description and subcategory name with subcategory description.

- Category name and description
- Category name with category description, and subcategory name
- Category name with category description, subcategory name with subcategory description

Results

Table 5 shows that the performance improves with more category information, with the most comprehensive level yielding the highest F_1-scores. Specifically, the macro F_1-score increases by 89% (from 18.86 to 35.75), and the illegal F_1 and harmful F_1-scores increase by 200% (from 11.87 to 35.80) and 57% (from 37.63 to 59.10), respectively, as compared with when using only the category name.

The increasing trend in macro F_1 and illegal or harmful F_1-scores suggests that more detailed information enhances the model's detection capabilities. However, adding only category descriptions decreased illegal or harmful recall (from 27.27 and 65.77 to 20.00 and 57.66, respectively).

XSL•FO
RenderX

Table . Results from the category and subcategory information detail experiment[a].

Category information	Subcategory information	Macro F_1	MAE[b]	Illegal F_1	Illegal recall	Harmful F_1	Harmful recall
Name	N/A[c]	18.86 (0.08)	1.4809 (0.0044)	11.87 (0.04)	27.27 (0.00)	37.63 (0.11)	65.77 (0.00)
Name and description	N/A	25.13 (0.16)	1.2173 (0.0042)	13.23 (0.17)	20.00 (0.00)	39.18 (0.04)	57.66 (0.00)
Name and description	Name	32.18 (0.06)	0.9867 (0.0016)	26.36 (0.00)	30.91 (0.00)	45.26 (0.08)	66.67 (0.00)
Name and description	Name and description	35.75 (0.29)[a]	0.8549 (0.0079)[a]	35.80 (0.87)[a]	58.79 (1.21)[a]	59.10 (0.41)[a]	86.79 (0.60)[a]

[a]A consistent increase in macro F_1, illegal F_1, and harmful F_1-scores is observed as the amount of information increases. Values are mean (SE). These values indicate the best performance in each column.

[b]MAE:mean absolute error.

[c]N/A: not applicable.

Formulating LLM Inputs

We assessed performance changes by incorporating images and training examples as inputs. We focused on the impact of images as multimodal data (“Leveraging Multimodality” subsection) and the effect of using training data with the annotation guide as postinstruction when combined with instruction (“Leveraging Few-shot Examples” subsection).

Leveraging Multimodality

Setup

The objective of this experiment was to determine the effect of image information on the classification performance of the model. We used 2 methods of conveying image information and compared their performances: the first method converts images into text descriptions, referred to as image description, whereas the second uses the images directly as inputs, referred to as vision. Three settings were tested for image descriptions: the first did not provide any image information, the second generated image descriptions using a model (gpt-4 - 1106), and the third involved human modifications to the descriptions created by the model. This allowed for a comparison of the performance of the models based on the generation of text-based image descriptions. In addition, we examined the impact of images (vision) when paired with the same image descriptions to observe their influence on performance. This involved adding the original image to each image description experiment for comparison. Overall, this setup evaluates the model’s performance in terms of the modality of suicide content through image descriptions and assesses the model’s multimodal capabilities through vision. Notably, during the annotation process, the annotators labeled the suicide category of the content based on both the text and the original images. We used GPT-4-turbo-2024-04-09, which can use both text and image inputs, for this experiment. We conducted an experiment on 113 test data entries that included images, among which only 3 belonged to the illegal suicide category; thus, illegal metrics were excluded from the results.

Results

Table 6 shows the impact of multimodal information on harmful suicide content detection tasks. In experiments regarding image descriptions without visual information, providing image details leads to superior performance compared with omitting them. Specifically, when using GPT-4–generated image descriptions, macro F_1 increased by 9.16% (from 50.46 to 55.08) and MAE decreased by 15.93% (from 0.3894 to 0.3333), indicating enhanced classification performance across all suicide categories. In addition, harmful F_1 and recall both increased by 8.00% (from 68.50 to 73.98 and from 75.76 to 81.82), suggesting that image information significantly aids in identifying harmfulness within suicide content. Comparing GPT-4 and human-modified image descriptions, using human-modified descriptions results in a reduction of macro F_1 by 3.55% (from 55.08 to 53.19) and increases MAE by 19.4% (from 0.3333 to 0.3982), although harmful F_1 increases by 2.61% (from 73.98 to 75.91) and harmful recall by 16.66% (from 81.82 to 95.45), indicating that human modifications enhance clarity and detection of harmfulness in content while decreasing overall category performance.

XSL•FO
RenderX

Table . Results of the input modality experiment on a subset of the benchmark that includes images[a].

Image description (text)	Vision (image)	Macro F_1	MAE[b]	Illegal F_1	Illegal recall	Harmful	Harmful recall
No description	X	50.46 (0.12)	0.3864 (0.0135)	N/A[c]	N/A	68.50 (0.21)	75.76 (0.26)
GPT-4 description	X	55.08 (0.09)[a]	0.3333 (0.0102)[a]	N/A	N/A	73.98 (0.09)	81.82 (0.00)
Human description	X	53.19 (0.13)	0.3982 (0.0154)	N/A	N/A	75.91 (0.08)[a]	95.45 (0.00)[a]
No description	O	47.78 (0.05)	0.4189 (0.0102)	N/A	N/A	67.42 (0.07)	90.91 (0.00)
GPT-4 description	O	50.08 (0.16)	0.3894 (0.0234)	N/A	N/A	69.22 (0.34)	83.33 (0.26)
Human description	O	46.55 (0.15)	0.4631 (0.0270)	N/A	N/A	69.37 (0.01)	90.91 (0.00)

[a]We compare 3 types of image descriptions (none, GPT-4-generated, and human-modified) with and without vision input. Without vision, GPT-4 descriptions achieve the best macro F_1 and MAE, while human descriptions perform best on harmful metrics. With vision, we observe an overall decrease in performance. Values are mean (SE). These values indicate the best performance in each column. "O" indicates that the column is used (vision) and "X" indicates that the column is not used.

[b]MAE: mean absolute error.

[c]N/A: not applicable.

In experiments using image information as visual input, we found a general decrease in overall performance across all settings, with reductions in macro F_1, MAE, and harmful F_1. Even in scenarios without image descriptions, macro F_1 decreased by 5.31% (from 50.46 to 47.78), and MAE increased by 8.41% (from 0.3864 to 0.4189). Particularly, the F_1-scores of potentially harmful content decreased significantly (from 54.46 to 42.77). This is owing to the model's sensitive reaction to certain images of potentially harmful suicide content, overestimating their harmfulness and classifying them as harmful.

However, in settings where we used images only as vision (no image description), the harmful recall score was 90.91, which was higher than when no image information was used (75.76); the score was 83.33 when using images with GPT-4 image description, which was higher than when using GPT-4 image description alone (81.82). This suggests that despite a decrease in the overall model performance owing to multimodality, using image information improves the identification of harmfulness in suicide content. In addition, vision can convey more information about harmfulness than text descriptions when human modification does not explicitly note the harmfulness of an image. Overall, the experiments with image descriptions confirmed that the information contained in an image enhances model performance in the harmful suicide content detection task, whereas adding vision information in a multimodal format decreases performance. However, the increase in harmful recall when using vision supports the potential of using vision as an effective tool for enhancing model capabilities in identifying harmful content, paving the way for future improvements in multimodal model performance.

Leveraging Few-Shot Examples

Few-shot refers to a setting where the model is provided with a small number of labeled examples as demonstrations before classifying new instances [45]. The goal of our few-shot experiments is to assess whether LLMs can effectively learn from only a handful of examples in high-stakes domains, where large-scale annotation is often insufficient to train models.

Setup

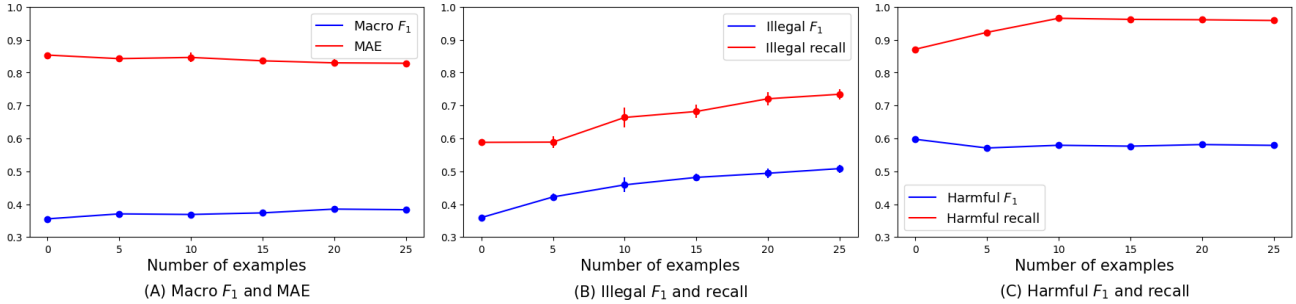
We examined 1- to 5-shot configurations, corresponding to 1-5 examples per category, totaling 5-25 examples. To construct the pool of demonstrations, medical experts who participated in the benchmark annotation selected 10 representative samples per category (50 in total). These expert-selected samples were used as the source for training demonstrations, while the remaining annotated data served as the test set. From the pool, 5 examples per category were randomly drawn for each experiment. To ensure robustness, we repeated the same experiment with 3 different random seeds, each sampling distinct few-shot samples, and we report the averaged results across these runs.

Results

Figure 4 shows how the performance metrics changed with the number of demonstration examples, with the standard error represented by vertical bars for each few-shot case. As the number of examples increased, macro F_1, MAE, and illegal metrics improved (Figure 4A), specifically illegal F_1 and recall in the 5-shot (Figure 4B). Although the F_1-score for harmful effects remained relatively stable, recall increased but plateaued after a certain threshold (2-shot) (Figure 4C).

XSL•FO
RenderX

Figure 4. Results from the few-shot example experiment. Increasing examples increases illegal F_1 and recall, with 5-shot setting achieving peak performance in the illegal metric. (A) Macro F_1 and mean absolute error; (B) illegal F_1 and recall; and (C) Harmful F_1 and recall. MAE: mean absolute error.



Comparison Between LLMs

We compared the performance of various LLMs in identifying harmful suicide content. Because open-sourced LLMs have instruction-following capabilities that depend on the language they have seen in the instruction tuning phase, we conducted experiments with different models for Korean and English benchmarks to address language barriers.

Setup

We categorized the selected LLMs into closed and open-sourced models. For the Korean benchmark, we used closed models because of the lack of open-source or multilingual LLMs that can properly follow the task's instructions in Korean. We also included a random baseline that arbitrarily categorized content into one of the 5 categories.

Closed Models

We used OpenAI's GPT-3.5 (gpt-3.5-turbo-16k-0613) and GPT-4 (gpt-4 - 1106-preview), which are accessed through the OpenAI API and capable of handling a context length of 1,28,000 characters. In addition, we experimented with Clova X, an LLM trained on Korean, using the Naver API [46].

Open-Sourced Models

We used the Zephyr-7B-beta model [47], an enhanced version of mistral-7B, which supports a context length of up to 32,000 characters. We also use LongChat-7B-16k [48] and Vicuna-7B-v1.5 - 16k [49], which are both fine-tuned Large Language Model Meta artificial intelligence (AI) models with a maximum context length of 16,000 characters.

To explore the model's adaptability of the model in few-shot learning contexts, we conducted experiments in both the zero-shot and 5-shot scenarios ("Leveraging Few-Shot Examples" subsection). However, for models unable to accept the context length of 12k tokens required for the 5-shot experiments, such as Clova X (4096), we limited our analysis to the zero-shot trials.

Results

Figure 5 shows the performance of the GPT models and Clova X on the Korean benchmark. GPT-4 outperformed all other models in every metric except for harmful recall (Figure 5A-E). GPT-3.5 follows GPT-4 in terms of performance across all metrics, except for harmful recall. Clova X showed lower performance than the GPT models but achieved the highest score in harmful recall (Figure 5F), indicating its high sensitivity to harmful content.

Figure 6 shows the performance of the GPTs and open-sourced LLMs on the translated English benchmark. GPT-4 exhibited the highest performance in differentiating categories in both the zero-shot and 5-shot settings across various metrics (Figure 6A-C and E). It leads the performance charts with macro F_1-scores of 46.37 in zero-shot and 52.59 in 5-shot. Notably, GPT-4 showed a significant MAE difference (0.5655 in zero-shot and 0.5755 in 5-shot), indicating that even when the category predictions were incorrect, they tended to be within similar levels of harmfulness.

GPT-3.5 ranked second to GPT-4 in category distinction performance (Figure 6A, B, and E) in zero-shot settings and showed comparable performance to open-sourced models in 5-shot settings. Its recall was relatively higher than GPT-4 in 5-shot settings (Figure 6C-F), indicating a more sensitive response to harmful information than the GPT-4.

Zephyr outperforms random in zero-shot settings with a macro F_1 of 20.99 and MAE of 1.2711; in addition, it achieves a comparable performance to GPT-3.5 in 5-shot settings with macro F_1 of 37.52 and MAE of 0.8217 (Figure 6A and B). LongChat exhibits the largest standard error in the illegal and harmful metrics (Figure 6C-F), indicating that few-shot examples significantly impact performance compared with other models. It recorded the highest standard errors in illegal F_1 and harmful F_1 at 4.09 and 1.67, respectively (Figure 6C and E). LongChat also showed the lowest recall for illegal and harmful content, particularly in harmful content, suggesting that it is less sensitive to harmful information (Figure 6D and F). Vicuna recorded the lowest performance in category classification (Figure 6A) among all models but achieved high recall for illegal and harmful content (Figure 6D and F). Notably, it scored the highest illegal recall of 76.36 and a harmful recall of 83.78, comparable with GPT-3.5's 85.89.

XSL•FO
RenderX

Figure 5. Results from the Korean benchmark experiment. Hatched bars indicate the Korean large language model (Clova X). Although Clova X has lower overall performance than generative pretrained transformers, it excels in harmful recall. (A) Macro F_1; (B) mean absolute error; (C) illegal F_1; (D) illegal recall; (E) harmful F_1; and (F) harmful recall. MAE: mean absolute error.

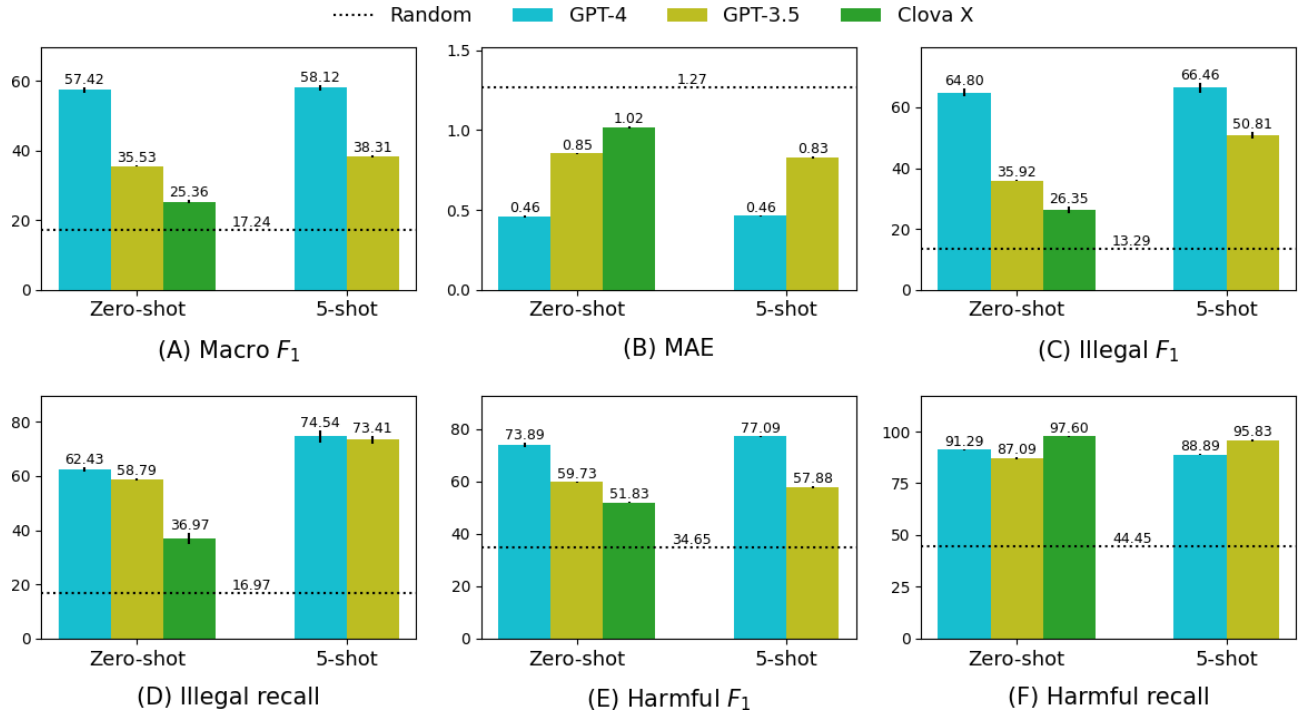
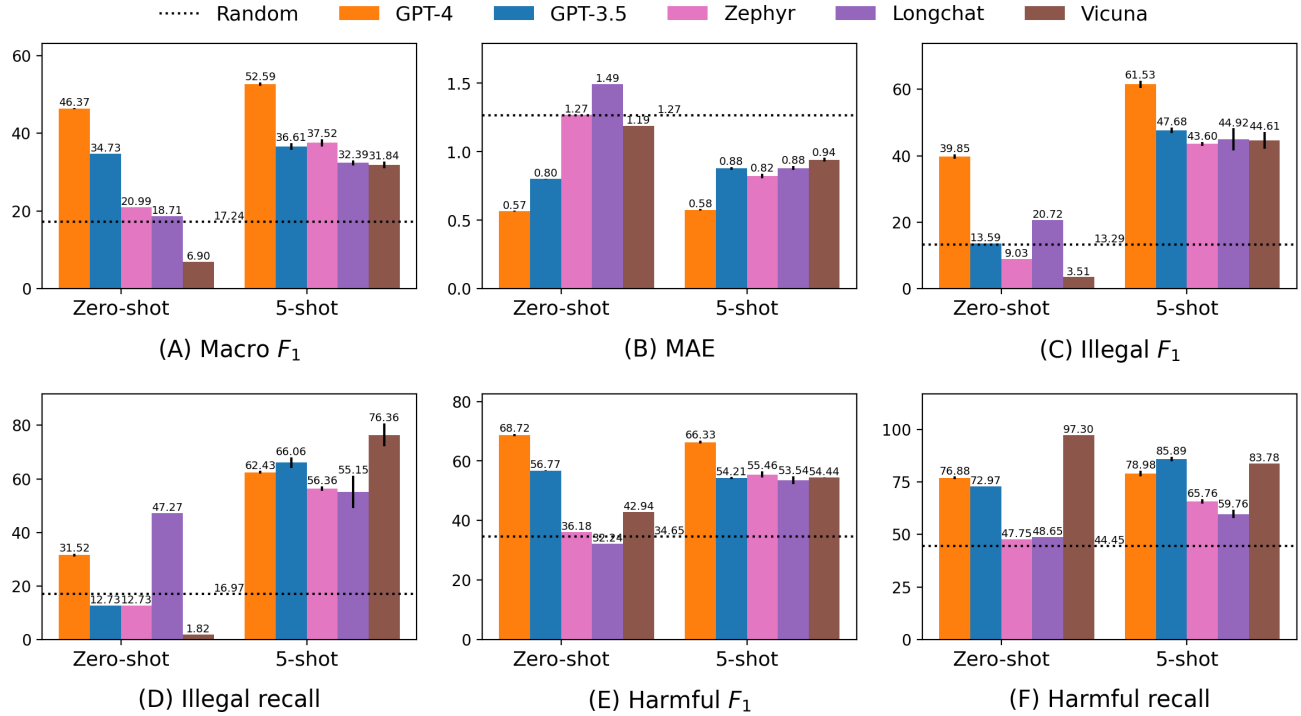


Figure 6. Results from the translated English benchmark experiment. Closed models (GPT-4 and GPT-3.5) show superior performance in the zero-shot setting compared with open-sourced models, whereas open-sourced models reach comparable performance to GPT-3.5 in 5-shot. (A) Macro F_1; (B) mean absolute error; (C) illegal F_1; (D) illegal recall; (E) harmful F_1; and (F) harmful recall. MAE: mean absolute error.



Analysis

Open-Sourced Versus Closed LLMs

GPT-4 recorded the highest performance across all accuracy metrics (macro F_1, MAE, illegal F_1, and harmful F_1) for all few-shot settings. In 5-shot settings, open-sourced models achieve a similar performance to GPT-3.5. However, in zero-shot settings, they struggled to understand lengthy instructions, resulting in random predictions (eg, LongChat) or biased predictions toward specific categories (eg, Vicuna), with Zephyr slightly outperforming random. In 5-shot scenarios, Zephyr matches GPT-3.5 in macro F_1, MAE, and harmful F_1, whereas LongChat and Vicuna show comparable performance in their respective metrics. Except for Vicuna, open-sourced

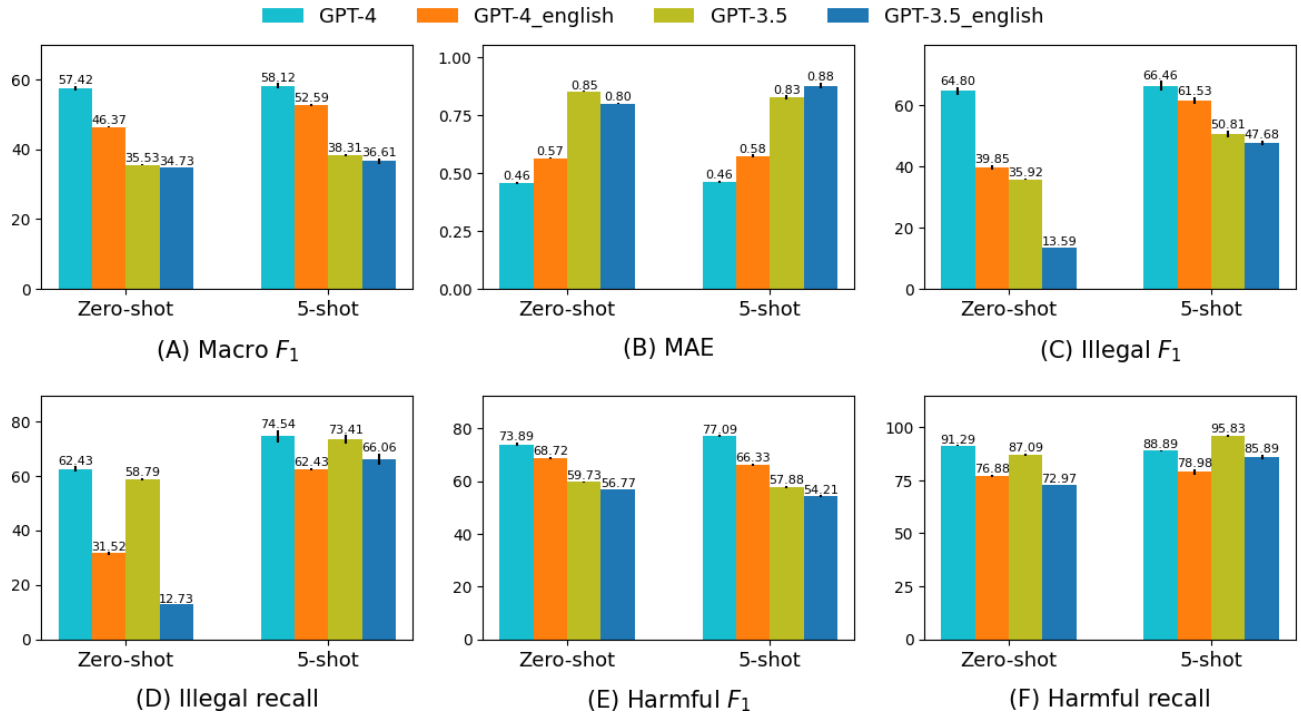
XSL•FO
RenderX

models generally showed lower recall than closed models in terms of illegal and harmful content.

Original Korean Versus Translated English

Figure 7 shows an analysis of GPT-3.5 and GPT-4's performance on the Korean and translated English benchmarks. Both models performed better on the Korean benchmark across all F_1 metrics (Figure 7A, C, and E). However, GPT-4 shows a decrease in macro F_1 from the English to the Korean benchmark by 19.24% in zero-shot (from 57.42 to 46.37) and 9.51% in 5-shot (from 58.12 to 52.59), with the largest decrease in illegal F_1 by 36.17% in zero-shot (from 64.80 to 39.85). GPT-3.5 also showed a considerable reduction in zero-shot illegal F_1 by 62.17% (from 35.92 to 13.59). Illegal recall decreases considerably, with GPT-4 decreasing by 49.51% (from 62.43 to 31.52) and GPT-3.5 by 78.34% in illegal recall (from 58.79 to 12.73), indicating a larger decrease than in F_1-scores (Figure 7D). The decrease in harmful F_1 is less significant, with GPT-4 decreasing by 7.00% (from 73.89 to 68.72) in zero-shot, and GPT-3.5 decreasing by 4.96% (from 59.73 to 56.77). This indicates that while translating the benchmark using GPT-4 does not significantly affect the overall quality, it may lead to issues in specific categories, notably illegal.

Figure 7. Performance comparison of closed models (GPT-3.5 and GPT-4) on the Korean and translated English benchmarks. Closed models exhibit better classification performance on the Korean benchmark than on the English benchmark, with the most significant difference noted in classifying the illegal category. (A) Macro F_1; (B) mean absolute error; (C) illegal F_1; (D) illegal recall; (E) harmful F_1; and (F) harmful recall. MAE: mean absolute error.



English Benchmark Validation

Considering the use of the English benchmark for experimenting with open-sourced models, it is necessary to evaluate the translation results. In particular, for content containing words related to suicide and harmful information, it is crucial to assess both the overall translation quality and how well the content has been translated. Because OpenAI's use policy potentially refuses to respond to harmful content, there may be instances in which proper translation has not been achieved. Therefore, we analyzed the following 2 aspects: (1) overall translation quality (quantitative analysis), and (2) translation of harmful content (qualitative analysis).

We evaluated the text content (CONTENT_TEXT) of every instance in the benchmark because every instance contains content text and it constitutes the largest proportion of text. We performed a quantitative analysis to assess translation quality using models (GPT-4 - 0613) and a qualitative analysis of the translation of harmful content by the authors.

Translation quality (quantitative analysis): Translation quality assesses how similar the translated content (in English) is to the original content (in Korean). We evaluated translation quality using Scalar Quality Metric (SQM) and Direct Assessment (DA) methods [50] through the GPT-4 - 0613 API, which aligns closely with human evaluations. SQM evaluates the translation quality of the source text (Korean) and target text (English) on a scale of 0-100, with descriptions provided for "no meaning preserved," "some meaning preserved," "most meaning preserved and few grammar mistakes," and "perfect meaning and grammar." DA, such as SQM, rates translation quality on a scale of 0-100 but provides descriptions only for "no meaning preserved" and "perfect meaning and grammar." On average, the translated contents scored 79.55 on SQM and 78.10 on DA, indicating that most instances of the benchmark translation results fall under "most meaning preserved and few grammar mistakes," successfully retaining the original meaning.

Translation of harmful content (qualitative analysis): Illegal and harmful suicide content includes harmful words and

XSL•FO
RenderX

expressions related to suicide and self-harm, encompassing abbreviations, drug names related to suicide, and expressions of methods for suicide and self-harm. Moreover, owing to OpenAI's use policy, there are cases in which harmful content is not translated or translation is refused. Thus, to determine how well such content was translated, we analyzed translation error cases for expressions related to suicide: (1) expressions related to suicide (abbreviations and words), and (2) OpenAI moderation.

Expressions related to suicide: After analyzing 55 instances of illegal suicide content and 56 instances of harmful suicide content, we identified the following types of translation errors: (1) abbreviation translation error, (2) translation of substances used for suicide, and (3) translation of slang related to suicide and self-harm.

Table 7 shows examples of translation errors for each category and error type. *Abbreviation translation error* occurs when abbreviations related to suicide and self-harm are incorrectly interpreted. To evade online platform moderation, abbreviations related to suicide are often used. In these cases, the translation process incorrectly translates these abbreviations into entirely different words. The Korean abbreviation means "commit double suicide and death leap"; however, the English translation misinterprets it entirely. In this study, 12 benchmark instances were identified. *Translation of Substances used for suicide* refers to cases in which drugs related to suicide and self-harm were incorrectly translated. Substances used for suicide are often referred to by abbreviations to avoid online platform moderation, and the actual drug names are often translated into general names for drugs during translation. In this example, the drug "zolpidem" was translated as "SleepingPill," which translates to the purpose of the drug (zolpidem is a type of sleeping pill) rather than the actual name of the drug. However, such translations result in the inability of the model to correctly identify the sale of specific drugs (illegal suicide category) during the category classification process. In this study, 15 instances were identified. *Translation of slang related to suicide and self-harm* refers to errors in the translation of clear expressions of suicide. For example, the Korean expression for "bloodletting self-harm" was incorrectly translated as "blood donation" in English, which changed the meaning of the text. Three instances were identified for this case.

Table . Qualitative analysis of benchmark translation results. The source text is the content text from the Korean benchmark data, and the proper translation is the result translated by a human while preserving the meaning. The translation result is obtained using a model and has been applied to the English benchmark. The underlined words indicate parts where translation errors occurred in the model's output.

Type	Content text
Translation error in the abbreviation (illegal suicide content)	
Source text	ㄷㅂㅈㅅㅌㅅ 하실 분 자차나 면허 있으신 분으로요
Proper translation (human)	Looking for someone who will commit double suicide and death leap, preferably someone with a car or a driver's license.
Translation result (model)	Looking for someone who can do delivery and sales, preferably someone with a car or a driver's license.
Translation error in substances that are used for suicide (illegal suicide content)	
Source text	정품#졸프ㅣ뎀판ㅁㅐ 하는곳 가짜없고 사기없는 솔직한 비밀상담 관심空 #여성최음ㅈㅔ ㄱㅜ입하는 곳. 라인: [ID] 카톡: [ID] 텔레: [ID]
Proper translation (human)	#GenuinezolpidemSellingPlace A place where there are no fakes, no scams, honest confidential consultation. Interested in #FemaleAphrodisiac Purchase place. Line: [ID] Kakaotalk: [ID] Telegram: [ID]
Translation result (model)	#GenuineSpleepingPillSellingPlace A place where there are no fakes, no scams, honest confidential consultation. Interested in #FemaleAphrodisiac Purchase place. Line: [ID] Kakaotalk: [ID] Telegram: [ID]
Translation of slang related to suicide or self-harm (harmful suicide content)	
Source text	사혈자해를 해보고싶다는 생각이 들어
Proper translation (human)	I'm thinking about wanting to try bleeding self-harm.
Translation result (model)	I'm thinking about wanting to try blood donation.

OpenAI Moderation

During the translation process of the benchmark data, we found a few instances where different translation errors occur from those related to expressions of suicide; some content are not translated at all. Unlike parts of the text that are translated incorrectly, the OpenAI API (gpt-4‐1106) refuses to translate and generates content (translated output) that has no relation to the original text (input text). Two instances occurred for this case; one involving content asking for help in committing suicide or self-harm and another involving direct sexual content.

XSL•FO
RenderX

Discussion

English Benchmark Integrity

The English benchmark was generated through machine translation using GPT-4. Quantitative evaluation showed relatively high scores (SQM: 79.55, DA: 78.10), indicating that most meanings were preserved. However, qualitative analysis revealed errors in suicide-related slang, abbreviations, and substance names (eg, bleeding self-harm translated as blood donation, zolpidem simplified to sleeping pill). Such inaccuracies may affect model performance and limit the benchmark's validity for multilingual generalization. Therefore, results on the English benchmark should be interpreted with caution. Future work will involve human validation of high-risk categories and improved translation methods for suicide-specific terminology.

Benchmark Size

Although the harmful suicide content benchmark is an essential step toward understanding and moderating online suicide-related content, it encompasses 452 data entries. This relatively small benchmark size is largely attributable to the fact that posts related to suicide comprise a small fraction of the total online content. In addition, filtering and deleting such content by online sources inherently limit the volume of data available for collection. Nevertheless, a carefully controlled annotation process that incorporates the knowledge of clinical experts supports the credibility of the benchmark and ensures a reliable set of labels. In addition, the task description document details 25 different subcategories of suicide content, and the benchmark comprises a wide array of suicide content, including actual data for each subcategory. Therefore, our detailed task description document and the data within our benchmark lay the groundwork for future efforts to create a large-scale suicide content dataset using the annotations described for the suicide content.

Dataset Usage

The harmful suicide content benchmark includes a task description document containing categories and subcategories of suicide content, along with a dataset of 452 annotated entries. Users can leverage the contents of the task description document as instructions for LLMs. In addition, data labeled for training within the benchmark can be used as few-shot examples, while data labeled for testing can serve as evaluation data.

Future Work

In clinical practice, detecting that a patient who has attempted suicide or self-harm has been exposed to suicide-related content can be highly valuable for preventing further self-harm or suicide attempts. This is because restricting harmful online content represents a relatively modifiable factor among the many risk factors associated with suicide and self-harm [51,52]. Building on this perspective, in future work, we plan to further validate the practical effectiveness of our suicide content detection model. Specifically, we aim to conduct a comparative study between individuals who have attempted suicide or self-harm and a control group of individuals without such experiences. By analyzing their online activity, we will investigate whether those with a history of suicidal or self-harming behaviors are more frequently exposed to harmful suicide-related content. Such a study would not only provide empirical evidence for the real-world use of our detection framework but also offer insights into the role of online harmful content in influencing vulnerable populations.

Funding

This research was partially supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) and funded by the Korean government (MSIT) (NRF-2021M3A9E4080780 and 2021M3A9E4080784) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190421, AI Graduate School Support Program(Sungkyunkwan University), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (no. RS-2024 - 00509258 and no. RS-2024-00469482, Global AI Frontier Lab).

Data Availability

Our benchmark dataset contains sensitive and, in some cases, illegal information related to suicide. Due to the sensitivity of suicide-related content, the full benchmark dataset cannot be publicly released and is granted only through a controlled approval process [53]. In addition, since the dataset also contains illegal information, it cannot be made publicly available; however, it may be shared under appropriate requests and upon review by the authors [54]. We provide the website [55] to obtain the benchmark dataset and the codes used in the experiments. Applicants must follow the outlined process, after which the authors will review the requests and share the dataset as appropriate. This version includes only references to the original data (eg, tweet IDs, and URLs) and annotations by medical experts, ensuring that no legal or ethical violations occur upon public release.

Authors' Contributions

KP, MJB, RL, JM, JWP, and SP contributed to conception. KP, MJB, RL, JM, KC, JWP, and SP contributed to idea and theory. KP, MJB, YS, SML, YHS, ARL, SYY, JWP, and SP contributed to task design. KP, YH, YS, HL, and SP contributed to building data collections. MJB, RL, and JWP contributed to data annotation. KP, YH, YS, and SP contributed to data processing. KP, YH, YS, SML, YHS, ARL, SYY, and SP contributed to building the benchmark. KP, YH, DL, JB, KC, and SP contributed to technical validation. All authors contributed to the final manuscript and reviewed and approved the final manuscript. JB (jy.bak@skku.edu),

XSL•FO
RenderX

KC (kyunghyun.cho@nyu.edu), JWP (paikjw@khu.ac.kr), and SP (sungjoon.park@softly.ai, psj6512@gmail.com) are equal co-corresponding authors.

Conflicts of Interest

KP, DL, JM, and SP are affiliated with SoftlyAI. MJB and ARL are affiliated with Kyung Hee University Medical Center. YH and JB are affiliated with SungKyunKwan University. YS and HL are affiliated with KAIST. RL is affiliated with the University of Pennsylvania. MJB, SML, YHS, ARL, SYY, and JWP are affiliated with Kyung Hee University. KC is affiliated with New York University.

References

1. Understanding self-harm and suicide content online. Samaritans. 2020. URL: https://www.samaritans.org/about-samaritans/research-policy/internet-suicide/guidelines-tech-industry/understanding-self-harm-and-suicide-content/ [accessed 2025-12-20]
2. 자살예방 및 생명존중문화 조성을 위한 법률 [The act on the prevention of suicide and the creation of a culture of respect for life]. MOHW. 2019. URL: https://tinyurl.com/42xe7rwy [accessed 2025-12-20]
3. Zdanow C, Wright B. The representation of self injury and suicide on emo social networking groups. Afr Sociol Rev Afr de Sociol 2012;16:81-101 [FREE Full text]
4. Biddle L, Gunnell D, Owen-Smith A, et al. Information sources used by the suicidal to inform choice of method. J Affect Disord 2012 Feb;136(3):702-709. [doi: 10.1016/j.jad.2011.10.004]
5. Milmo D. 'The bleakest of worlds': how Molly Russell fell into a vortex of despair on social media. The Guardian. 2022. URL: https://www.theguardian.com/technology/2022/sep/30/how-molly-russell-fell-into-a-vortex-of-despair-on-social-media [accessed 2025-12-20]
6. Balt E, Mérelle S, Robinson J, et al. Social media use of adolescents who died by suicide: lessons from a psychological autopsy study. Child Adolesc Psychiatry Ment Health 2023 Apr 7;17(1):48. [doi: 10.1186/s13034-023-00597-9] [Medline: 37029395]
7. Twenge JM, Haidt J, Lozano J, Cummins KM. Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. Acta Psychol (Amst) 2022 Apr;224:103512. [doi: 10.1016/j.actpsy.2022.103512]
8. National suicide prevention strategies. World Health Organization. 2018 URL: https://iris.who.int/server/api/core/bitstreams/9b9ea82a-b798-4c63-b90e-74bf8bdbf6b1/content [accessed 2025-12-20]
9. Donelan M. Secretary of State for Science, Innovation and Technology, the Lord Parkinson of Whitley Bay & Parliamentary Undersecretary of State for Arts and Heritage. Online Safety Act 2023. 2023. URL: https://www.legislation.gov.uk/ukpga/2023/50 [accessed 2025-12-20]
10. World health statistics 2023: monitoring health for the SDGS, sustainable development goals. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240074323 [accessed 2025-12-20]
11. Jung K. 처벌조항 유명무실 [Ministry of Health and Welfare requests zero investigations on suicide-inducing information on SNS, rendering punitive clauses ineffective]. 2022. URL: http://www.bosa.co.kr/news/articleView.html?idxno=2182323 [accessed 2025-12-20]
12. Min S. 자살유발정보 4년간 7배 늘었는데 · · · 모니터링 전담인력은 10년째 1명뿐 [Incidents of suicide-inducing information increase sevenfold in four years, yet monitoring staff remains solely one for a decade]. 2023. URL: https://www.khan.co.kr/article/202309241417001 [accessed 2025-12-20]
13. 2022 internet usage survey summary report (2022년도 인터넷이용실태조사 요약보고서). 2022 URL: https://www.msit.go.kr/bbs/view.do?sCode=user&bbsSeqNo=79&nttSeqNo=3173583 [accessed 2025-12-20]
14. Zirikly A, Resnik P, Uzuner Ö, Hollingshead K. CLPsych 2019 shared task: predicting the degree of suicide risk in reddit posts. In: Niederhoffer K, Hollingshead K, Resnik P, Resnik R, Loveys K, editors. 2019 Presented at: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; Jun 6, 2019; Minneapolis, Minnesota p. 24-33. [doi: 10.18653/v1/W19-3003]
15. Milne DN, Pink G, Hachey B, Calvo RA. CLPsych 2016 shared task: triaging content in online peer-support forums. In: Hollingshead K, Ungar L, editors. 2016 Presented at: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology; Jun 16, 2016; San Diego, CA, USA p. 118-127. [doi: 10.18653/v1/W16-0312]
16. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: Palmer M, Hwa R, Riedel S, editors. 2017 Presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Sep 9-11, 2017; Copenhagen, Denmark p. 2968-2978. [doi: 10.18653/v1/D17-1322]
17. Yang C, Zhang Y, Muresan S. Weakly-supervised methods for suicide risk assessment: role of related domains. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 22: Short Papers); Aug 1-6, 2021. [doi: 10.18653/v1/2021.acl-short.133]
18. Sawhney R, Neerkaje A, Gaur M. A risk-averse mechanism for suicidality assessment on social media. In: Muresan S, Nakov P, Villavicencio A, editors. 2022 Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); May 22-27, 2022; Dublin, Ireland p. 628-635. [doi: 10.18653/v1/2022.acl-short.70]

XSL•FO
RenderX

19. Rawat BPS, Kovaly S, Yu H, Pigeon W. ScAN: suicide attempt and ideation events dataset. Presented at: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics; Jul 10-15, 2022; Seattle, United States p. 1029. [doi: 10.18653/v1/2022.naacl-main.75]
20. Sawhney R, Joshi H, Shah RR, Flek L. Suicide ideation detection via social and temporal user representations using hyperbolic learning. 2021 Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 6-11, 2021; Online p. 2176-2190 URL: https://aclanthology.org/2021.naacl-main [doi: 10.18653/v1/2021.naacl-main.176]
21. Marchant A, Hawton K, Stewart A, et al. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: the good, the bad and the unknown. PLoS One 2017;12(8):e0181722. [doi: 10.1371/journal.pone.0181722] [Medline: 28813437]
22. Sedgwick R, Epstein S, Dutta R, Ougrin D. Social media, internet use and suicide attempts in adolescents. Curr Opin Psychiatry 2019;32(6):534-541. [doi: 10.1097/YCO.0000000000000547]
23. Wang L, Liu X, Liu ZZ, Jia CX. Digital media use and subsequent self-harm during a 1-year follow-up of Chinese adolescents. J Affect Disord 2020 Dec;277:279-286. [doi: 10.1016/j.jad.2020.05.066]
24. Patchin JW, Hinduja S, Meldrum RC. Digital self-harm and suicidality among adolescents. Child Adolesc Ment Health 2023 Feb;28(1):52-59. [doi: 10.1111/camh.12574] [Medline: 35811440]
25. Choi WS, Han J, Hong HJ. Association Between internet searches related to suicide/self-harm and adolescent suicide death in South Korea in 2016-2020: secondary data analysis. J Med Internet Res 2023 Apr 20;25:e46254. [doi: 10.2196/46254] [Medline: 37079349]
26. Mars B, Heron J, Biddle L, et al. Exposure to, and searching for, information about suicide and self-harm on the Internet: prevalence and predictors in a population based cohort of young adults. J Affect Disord 2015 Oct;185:239-245. [doi: 10.1016/j.jad.2015.06.001]
27. Wang S, Liu Y, Xu Y, Zhu C, Zeng M. Want to reduce labeling cost? GPT-3 can help. In: Moens MF, Huang X, Specia L, Yih ST, editors. 2021 Presented at: Findings of the Association for Computational Linguistics; Nov 7-11, 2021; Punta Cana, Dominican Republic p. 4195-4205. [doi: 10.18653/v1/2021.findings-emnlp.354]
28. Wang X, Kim H, Rahman S, Mitra K, Miao Z. Human-LLM collaborative annotation through effective verification of LLM labels. 2024 May 11 Presented at: Proceedings of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu HI USA p. 1-21. [doi: 10.1145/3613904.3641960]
29. 온라인 자살유발정보, 국민이 직접 찾아내고 삭제한다! [Citizens directly identify and remove online suicide- inducing information!]. KPHN. 2020. URL: https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=359157 [accessed 2025-12-20]
30. Markov T, Zhang C, Agarwal S, et al. A holistic approach to undesired content detection in the real world. AAAI 2023;37(12):15009-15018. [doi: 10.1609/aaai.v37i12.26752]
31. Lee D, Park S, Kang J, Choi D, Han J. Cross-lingual suicidal-oriented word embedding toward suicide prevention. 2020 Nov Presented at: Findings of the Association for Computational Linguistics; Nov 16-20, 2020 URL: https://www.aclweb.org/anthology/2020.findings-emnlp [doi: 10.18653/v1/2020.findings-emnlp.200]
32. 미디어 자살정보 모니터링 시스템 [Media suicide information monitoring system]. KFSP. 2023. URL: https://sims.kfsp.or.kr/ [accessed 2025-12-20]
33. Park S, Park K, Ahn J, Oh A. Suicidal risk detection for military personnel. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020 URL: https://www.aclweb.org/anthology/2020.emnlp-main [doi: 10.18653/v1/2020.emnlp-main.198]
34. 사이버상담 [cyber counselling]. Lifeline Korea. 2023. URL: https://www.lifeline.or.kr/business/opening_consult.php [accessed 2025-12-20]
35. 공개상담실 [public counselling room]. KSPCC. 2023. URL: http://auru.counselling.or.kr/contents/sub0201.php?PHPSESSID=4f99b850a26d94de69fe936d68006e51 [accessed 2025-12-20]
36. 디시인사이드 [DCinside]. DCInside. 2023. URL: https://www.dcinside.com/ [accessed 2025-12-20]
37. Jo J. 잇단 사고에 말 많은 '우울증 갤러리'‥"폐쇄는 안 한다"는 정부 [following a series of incidents, the government states it will not shut down the controversial 'depression gallery' despite the ongoing discussions]. MBC. 2023. URL: https://imnews.imbc.com/replay/2023/nwdesk/article/6490952_36199.html [accessed 2026-01-29]
38. Fiesler C, Jiang J, McCann J, Frye K, Brubaker JR. Reddit Rules! Characterizing an Ecosystem of Governance. ICWSM 2018;12(1). [doi: 10.1609/icwsm.v12i1.15033]
39. Moon J, Lee DH, Cho H, et al. Analyzing norm violations in live-stream chat. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore p. 852-868. [doi: 10.18653/v1/2023.emnlp-main.55]
40. Gattrell WT, Hungin AP, Price A, et al. ACCORD guideline for reporting consensus-based methods in biomedical research and clinical practice: a study protocol. Res Integr Peer Rev 2022 Jun 7;7(1):3. [doi: 10.1186/s41073-022-00122-0] [Medline: 35672782]
41. Vakil N. Consensus guidelines: method or madness? Am J Gastroenterol 2011 Feb;106(2):225-227. [doi: 10.1038/ajg.2010.504] [Medline: 21301451]

XSL•FO
RenderX

42. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist 2024 Feb 23;12:157-173. [doi: 10.1162/tacl_a_00638]
43. Wu Y, Iso H, Pezeshkpour P, Bhutani N, Hruschka E. Less is more for long document summary evaluation by LLMs. 2023 Presented at: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2; Mar 17-22, 2024; St Julian's, Malta p. 330-343. [doi: 10.18653/v1/2024.eacl-short.29]
44. Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. In: Meila M, Zhang T, editors. Presented at: Proceedings of the 38th International Conference on Machine Learning; Jul 18-24, 2021; Virtual p. 12697-12706 URL: https://proceedings.mlr.press/v139/zhao21c.html [accessed 2025-12-20]
45. Brown T, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901 [FREE Full text]
46. Kim B, Kim H, Lee SW, et al. What changes can large-scale language models bring? intensive study on hyperclova: billions-scale Korean generative pretrained transformers. In: Moens MF, Huang X, Specia L, Yih ST, editors. Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Nov 7-11, 2021; Online and Punta Cana, Dominican Republic p. 3405-3424. [doi: 10.18653/v1/2021.emnlp-main.274]
47. Tunstall L, et al. Zephyr: direct distillation of LM alignment. arXiv. Preprint posted online on Oct 25, 2023. [doi: 10.48550/arXiv.2310.16944]
48. Li D, et al. How long can open-source llms truly promise on context length. LMSYS Org. 2023. URL: https://lmsys.org/blog/2023-06-29-longchat/ [accessed 2025-12-29]
49. Chiang WL, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS Org. 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/ [accessed 2025-12-20]
50. Kocmi T, Federmann C. Large language models are state-of-the-art evaluators of translation quality. arXiv. Preprint posted online on Feb 28, 2023. [doi: 10.48550/arXiv.2302.14520]
51. Hawton K, Saunders KEA, O'Connor RC. Self-harm and suicide in adolescents. Lancet 2012 Jun 23;379(9834):2373-2382. [doi: 10.1016/S0140-6736(12)60322-5] [Medline: 22726518]
52. Skegg K. Self-harm. Lancet 2005;366(9495):1471-1483. [doi: 10.1016/S0140-6736(05)67600-3] [Medline: 16243093]
53. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé III H, Resnik P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. 2018 Presented at: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; Jun 5, 2018; New Orleans, LA p. 25-36. [doi: 10.18653/v1/W18-0603]
54. Choshen L, Eldad D, Hershcovich D, Sulem E, Abend O. The language of legal and illegal activity on the darknet. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy p. 4271-4279. [doi: 10.18653/v1/P19-1419]
55. Bak J. Iterative large language model-guided sampling and expert-annotated benchmark corpus for harmful suicide content detection: development and validation study. GitHub. URL: https://github.com/Human-Language-Intelligence/Harmful-Suicide-Content-Detection [accessed 2025-12-20]

Abbreviations

AI: artificial intelligence
API: application programming interface
DA: Direct Assessment
IRB: institutional review board
MAE: mean absolute error
PII: Personally Identifiable Information
Q&A: question and answer
SQM: Scalar Quality Metric

Edited by A Benis; submitted 12.Mar.2025; peer-reviewed by A Gupta, L Wreyford, S Chowdhury, S Banerjee; accepted 16.Nov.2025; published 05.Feb.2026.

Please cite as:
Park K, Baik MJ, Hwang Y, Shin Y, Lee H, Lee R, Lee SM, Sun JYH, Lee AR, Yoon SY, Lee DH, Moon J, Bak J, Cho K, Paik JW, Park S
Iterative Large Language Model–Guided Sampling and Expert-Annotated Benchmark Corpus for Harmful Suicide Content Detection: Development and Validation Study
JMIR Med Inform 2026;14:e73725
URL: https://medinform.jmir.org/2026/1/e73725
doi:10.2196/73725

XSL•FO
RenderX

© Kyumin Park, Myung Jae Baik, YeongJun Hwang, Yen Shin, HoJae Lee, Ruda Lee, Sang Min Lee, Je Young Hannah Sun, Ah Rah Lee, Si Yeun Yoon, Dong-ho Lee, Jihyung Moon, JinYeong Bak, Kyunghyun Cho, Jong-Woo Paik, Sungjoon Park. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 5.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Evaluation of SNOMED CT Automated Mapping Tool: Advancing Terminology Standardization and Semantic Interoperability

Youngsun Park[1], MSN; Hannah Kang[1], PhD; Jiwon Kim[1], BSN; Soo-Yong Shin[1], PhD; Dosang Cho[2], PhD; Sang Youl Rhee[3,4], PhD; Hong Seok Park[5], PhD; Kyung-Jae Lee[6], PhD; Sungchul Bae[1], PhD

[1]Kakao Healthcare Corp, Seongnam-si, Republic of Korea

[2]Department of Neurosurgery, Ewha Womans University Medical Center, Seoul, Republic of Korea

[3]Center for Digital Health, Kyung Hee University, Seoul, Republic of Korea

[4]Department of Endocrinology and Metabolism, Kyung Hee University College of Medicine, Seoul, Republic of Korea

[5]Department of Urology, Korea University College of Medicine, Seoul, Republic of Korea

[6]Department of Orthopedic Surgery, Keimyung University Dongsan Hospital, Daegu, Republic of Korea

Corresponding Author:
Sungchul Bae, PhD
Kakao Healthcare Corp
166, Pangyoyeok-ro, Bundang-gu, Gyeonggi-do
Seongnam-si, 13529
Republic of Korea
Phone: 82 1088569908
Email: utdwest13@gmail.com

Abstract

Background: Effective secondary use of healthcare data is hindered by fragmentation and a lack of semantic interoperability due to heterogeneous local terminologies. Standardizing clinical terms using SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is essential but remains a manual, labor-intensive, and inconsistent process, especially across multiple institutions. Automated, scalable solutions are needed to support reliable mapping and new concept authoring for large-scale research.

Objective: We aimed to develop a large language model (LLM)-assisted tool that streamlines SNOMED CT terminology mapping and concept authoring, which enables seamless, standardized data integration across multi-institutional clinical datasets.

Methods: The mapping pipeline included preprocessing local terms, syntactic and LLM-based vector similarity mapping, and iterative enrichment based on validated results. Translation and semantic representation used GPT-4o (OpenAI). New concepts were authored through a structured postcoordination process, and both the efficiency and quality of authoring (including duplicate rate and Machine Readable Concept Model validation violations) were quantitatively evaluated. Performance was evaluated using diagnostic and surgical procedural terms from 4 major hospital networks (9 university hospitals) in South Korea, with additional usability feedback gathered from clinical terminologists.

Results: Using reference terms, top-5 accuracy for diagnostic mapping reached 98.7%, 89.7%, 98.5%, and 92.8% across the 4 institutions and 99.2%, 82.6%, 98.7%, and 84.7% for surgical procedural mapping. Implementation of the tool reduced manual mapping rates by 30% and overall manual workload by up to 90%. The proposed tool reduced average mapping and new concept creation time by approximately 75%, while decreasing the final mapping table processing time by 90%. New concept authoring errors also decreased, with duplicate concepts reduced by 83% and modeling rule violations by 72%.

Conclusions: This study developed and validated an automated, LLM-assisted SNOMED CT mapping tool that significantly improved efficiency, mapping accuracy, and new concept quality. Limitations include technical integration challenges and dependency on translation quality. Future directions involve leveraging SNOMED CT's ontology structure and knowledge graphs, enhancing sustainability through ongoing maintenance and quality assurance, and further advancing new concept authoring with automated Machine Readable Concept Model rule enforcement and inactivation processes to achieve robust and scalable terminology standardization.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e82670) doi:10.2196/82670

KEYWORDS

large language model; health information interoperability; systematized nomenclature of medicine; terminology; standardization

Introduction

Background

The secondary use of healthcare data holds tremendous promise for advancing medical research and improving patient outcomes [1,2]. However, several critical barriers—such as data fragmentation, the absence of standardized medical terminologies, and semantic interoperability—continue to impede its effective reuse, particularly across multiple institutions [3]. Data generated in different health care settings are often siloed and recorded using heterogeneous systems and local coding practices, resulting in substantial misinterpretation and incomplete integration [4]. Furthermore, these challenges are compounded by the unstructured nature of much clinical data and stringent privacy and security requirements [5,6].

Achieving semantic interoperability—the ability of systems and organizations to exchange data with unambiguous, shared meaning—is essential to overcoming these challenges and unlocking the true value of multi-institutional clinical research [7,8]. At the core of semantic interoperability is terminology standardization, which enables consistent representation and interpretation of clinical concepts across diverse sources [9]. Without such standardization, variations in medical terminology across institutions can lead to inconsistencies, miscommunications, and errors that ultimately undermine the quality and reliability of biomedical research [10]. This challenge has driven international efforts, such as the World Health Organization Digital Health Interoperability Frameworks and Centers for Disease Control and Prevention National Electronic Disease Surveillance System, to prioritize standardized terminologies and interoperable health data infrastructure on a global scale [11,12].

In line with these global initiatives and the growing demands for healthcare data interoperability, Kakao Healthcare has developed the Healthcare Data Research Suite (HRS)—a platform supporting multi-institutional research via standardized data models, advanced artificial intelligence (AI) tools, and robust data privacy features. Notably, HRS places terminology standardization at its core, reflecting both the operational realities and strategic priorities identified by international guidelines [11,12]. By integrating domain-specific standard terminologies, HRS aims to bridge the gaps between fragmented datasets and promote data sharing for research.

To realize this goal, HRS adopts SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) as its primary terminology. SNOMED CT is recognized worldwide as one of the most comprehensive clinical terminology systems, encompassing a wide range of domains including disorders and procedures [13,14]. Its hierarchical structure, postcoordination capabilities, and rich attribute relationships support the precise representation of clinical concepts and sophisticated querying across large datasets [15]. Despite these advantages, mapping local clinical terms to SNOMED CT remains a significant bottleneck: the process is labor-intensive, demands expertise, and often results in inconsistent mappings both within and across institutions [16].

Recent studies have explored automated SNOMED CT mapping using various AI approaches. Davidson et al [17], 2025, benchmarked SNOMED CT entity linking models by organizing a challenge using annotated clinical notes. Gehrmann et al [18], 2025, evaluated how variable formulation impacts MedCAT's SNOMED CT mapping performance. Kulyabin et al [19], 2024, developed and tested SNOBERT, a BERT-based entity linking model for clinical notes. While these solutions have made meaningful progress in the field, the requirements of multi-institutional research in our context further necessitated several key capabilities: (1) consistent mapping of diverse local terms, both within and across institutions; (2) management of bilingual language inputs, such as the simultaneous use of Korean and English; (3) incorporation of domain-specific curation rules; and (4) the ability to author new SNOMED CT concepts, to achieve accurate mapping that preserves the detailed meaning of source terms.

To meet these requirements and enable reliable standardization across multi-institutional studies, we have developed a large language model (LLM)–assisted terminology standardization tool that integrates automated SNOMED CT mapping and postcoordination in a unified workflow. This tool is specifically designed to overcome the challenges of mapping local terms from heterogeneous sources, handle multilingual data, and facilitate the flexible authoring of new concepts as required for large-scale clinical research.

Objectives

The objective of this work was to create an LLM-assisted tool that enhances terminology standardization for multi-institutional research by seamlessly integrating automated SNOMED CT mapping and postcoordination within a single solution.

Methods

Requirement Collection

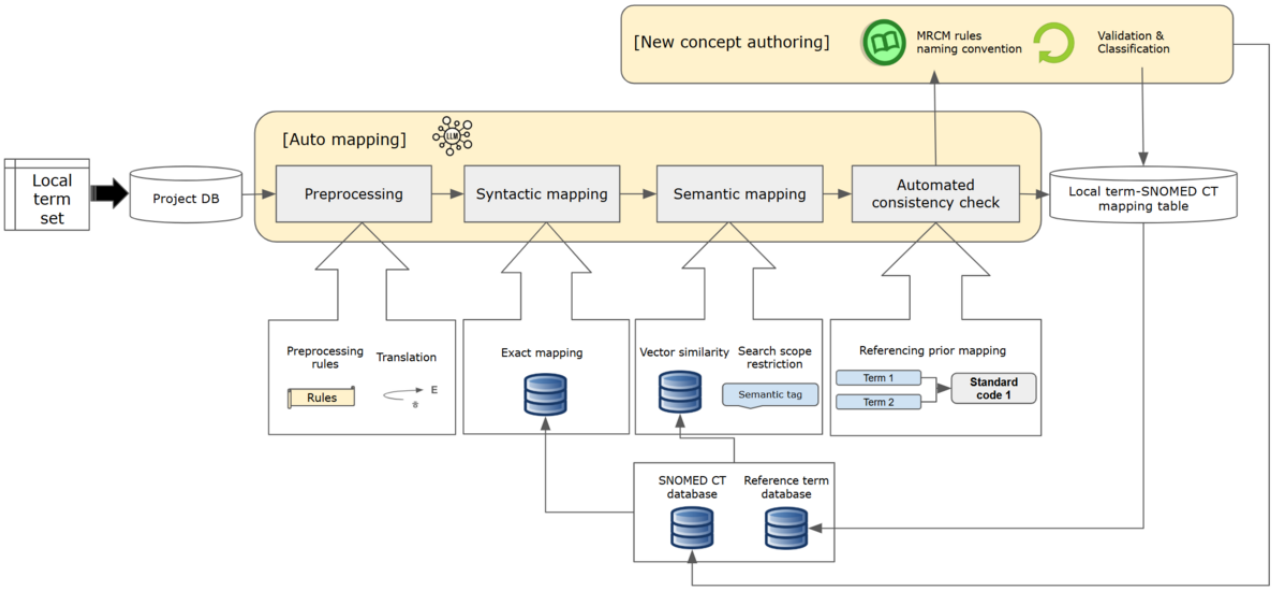
The participants involved in user requirements collection included 10 experienced clinical terminologists and 3 data engineers. Participants were asked to provide feedback on essential features through an open-ended questionnaire.

System Architecture and Implementation

The overall process of automatic mapping is illustrated in Figure 1. The workflow consists of 2 main processes: auto-mapping and new concept authoring. The final mappings are progressively incorporated into the local term-SNOMED CT mapping database and integrated into the custom reference term database, which includes the previous mapping results of local codes.

XSL•FO
RenderX

Figure 1. Overall workflow of the LLM-assisted terminology standardization tool. LLM: large language model; MRCM: Machine Readable Concept Model; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

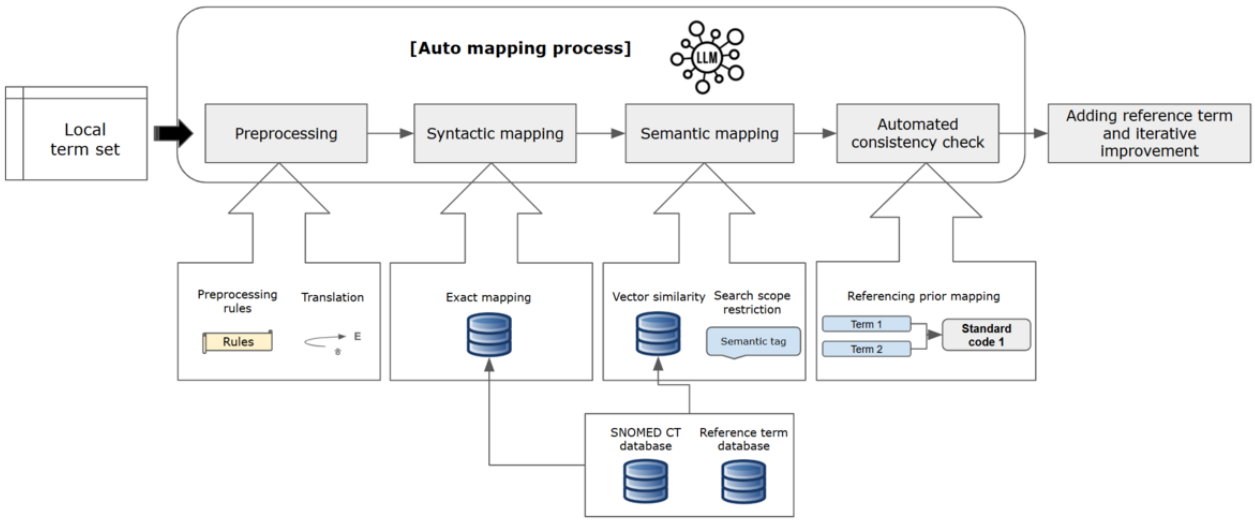


Auto-Mapping Method

Overview

The automated mapping process, as illustrated in Figure 2, begins with receiving a local term set and sequentially integrates four distinct components: (1) data preprocessing, (2) syntactic mapping, (3) semantic mapping, and (4) automated consistency check.

Figure 2. Overview of the LLM-assisted automated SNOMED CT mapping process for multi-institutional standardization. LLM: large language model; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Data Preprocessing

Data preprocessing consists of three main steps: (1) removing unnecessary symbols, (2) expanding abbreviations to full terms, and (3) translating non-English terms such as Korean into English with LLM tools (eg, GPT-4o (OpenAI) or Gemini 2.0 Flash (Google)). In the first step, unnecessary symbols are removed based on professional review. Second, abbreviations are converted to their full forms through manual curation, as context-dependent abbreviations such as “MS” can have multiple meanings (“multiple sclerosis” in neurology or “mitral stenosis” in cardiology); automatic expansion is therefore avoided. Third, any local terms in Korean are translated to English using LLMs. Although GPT-4o was used in this study, the system is designed to flexibly substitute other models (eg, Gemini 2.0 Flash) to accommodate institutional IT or cloud service restrictions.

Furthermore, to maximize long-term accuracy, we implemented a postmapping rule-based refinement mechanism. This system analyzes the automated mapping output and removes nonessential terms or extraneous symbols that do not affect similarity scores. This quality assurance step ensures that these refined terms are immediately incorporated into the custom

XSL•FO
RenderX

reference term database for iterative use, thereby boosting subsequent mapping performance.

To enhance the reproducibility and institutional specificity, the system allows each participating institution to define and apply custom rule-based preprocessing before automated mapping. These user-defined rules are automatically applied within the mapping pipeline, allowing each site to enforce its own normalization standards as needed.

Syntactic Mapping

Following the data preprocessing steps, the syntactic mapping component performs exact mapping by identifying standard codes that are textually identical to the preprocessed local term. To achieve this, the component prioritizes a search that is restricted within the SNOMED CT description table. In addition, the search leverages custom reference terms, which function similarly to the SNOMED CT description and were developed based on previously confirmed mapping results. If a textual identity between the normalized local term and one of the entries in the reference corpus is found, it is recommended to the user as an exact match.

Semantic Mapping

The semantic mapping component is used when the syntactic mapping component fails to find an exact match, using vector similarity search techniques to automatically map local terms to standard codes. This process leverages LLM-based embedding models (such as the text-embedding-3-small model for OpenAI deployments) to encode all reference terms, which include SNOMED CT descriptions and previously mapped local terms. These reference embedding results are stored in the Faiss vector database [20].

To ensure the efficiency and domain relevance of semantic search, the component applies a crucial constraint developed from empirical insights. First, the search is restricted to the SNOMED CT description table and custom reference terms validated through previous mappings. This prioritizes exact and highly relevant matches within established terminology sources. Second, the results are further refined by limiting the search scope to specific semantic tags and datasets relevant to each domain. For instance, the imaging and functional test domain is limited to “Procedure” and “Regime/Therapy” semantic tags; the diagnosis domain exclusively uses “Disorder,” “Finding,” “Situation,” “Event,” and “Person” tags. Additionally, certain cases restricted searches to descendants of specific concepts, allowing for targeted searches within specific datasets. For example, when mapping datasets for culture medium, searches were limited not only to “substance” but specifically to descendants of the concept 421955000 |Culture medium (substance)|, enhancing both speed and accuracy.

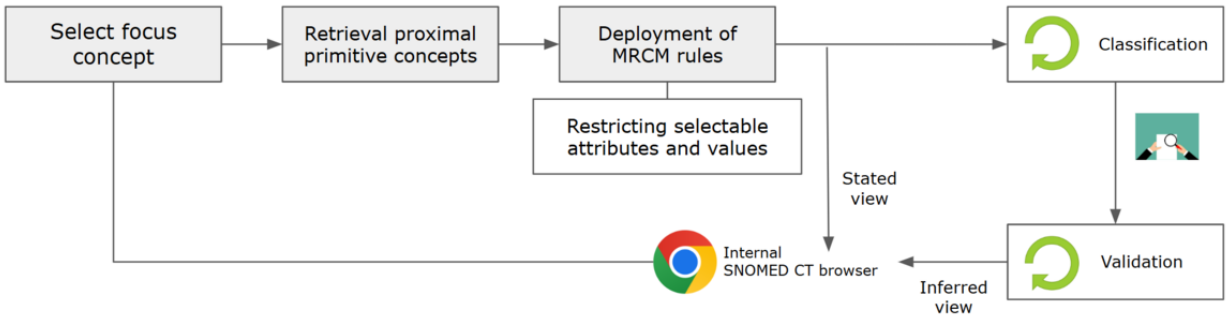
Automated Consistency Check

The automated consistency check component serves as the final stage in the automated mapping process. Its primary function is to ensure consistency and adherence to established mapping principles across all mapping candidates generated by the preceding components. This component achieves its goal primarily through referencing a prior mapping rule. This rule is specifically designed to maintain consistency across terms. When duplicate source terms are present, a validated mapping for 1 instance is automatically applied to all others. This rule is applied across different institutions, increasing efficiency and ensuring uniformity, except when the datasets belong to different domains. A successful automated consistency check leads directly to the final system step: adding reference terms and iterative improvement enriching the knowledge base for future mapping runs.

New Concept Authoring Method

Our new concept authoring procedure closely follows the principles and steps outlined in the SNOMED CT Editorial Guide [21]. Subsequently, Figure 3 illustrates the process of the authoring of new concepts through postcoordination. The process consists of four steps: (1) the selection of a focus concept, supported by proximal primitive concept retrieval; (2) concept modeling performed in alignment with Machine Readable Concept Model (MRCM) rules [22], followed by terming; (3) classification and validation; and (4) integration with the mapping table.

Figure 3. New concept authoring process using postcoordination within the LLM-assisted terminology standardization tool. LLM: large language model; MRCM: Machine Readable Concept Model; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



In the current version of our tool, we have operationalized 2 core rules from the editorial guide: the proximal primitive modeling principle and enforcement of MRCM rules. Other elements of the editorial guidance, such as cardinality, concept inactivation workflow, and modeling templates, are not yet implemented and are under consideration for future development.

The first implemented editorial rule is proximal primitive modeling. Adhering to this principle is recommended when

XSL•FO
RenderX

authoring new concepts in SNOMED CT [23]. The principle of proximal primitive modeling is vital for ensuring that classifiers accurately infer the most relevant proximal supertypes, which enhances precise subsumption [24,25]. When a user selects a focus concept required for modeling, not only the concept ID and fully specified name (FSN) of the focus concept, but also the corresponding proximal primitive concept are retrieved via the authoring-form API (Application Programming Interface) provided by Snowstorm [26].

The second implemented rule is the automated enforcement of MRCM rules. The MRCM defines the rules within the SNOMED CT concept model in a computer-readable format, allowing for validation of concept definitions and expressions to ensure compliance [27]. We have integrated these rules into the user interface (UI) by using the Snowstorm API. The tool restricts selectable attributes based on the domain of the focus concept. Using the GET /mrcm/{branch}/domain-attributes API, the focus concept ID is passed as parent IDs to retrieve only attributes relevant to the specific domain.

For naming guidance, our tool supports authors by automatically loading the FSN of the selected focus concept—retrieved via the authoring-form API—as a terming template. This allows users to customize the FSN, ensuring that the created concept precisely reflects the intended meaning, while maintaining overall consistency with SNOMED CT naming conventions.

Classification and validation are crucial steps in authoring new SNOMED CT terms for maintaining the precision and comprehensiveness of SNOMED CT and supporting accurate cohort queries over electronic health records [28]. We used the SNOMED (Systematized Nomenclature of Medicine) Release Validation Framework and the SNOMED OWL (Web Ontology Language) Toolkit in parallel to ensure the quality and consistency of our SNOMED CT editions, generate OWL expressions, and perform classification [29,30].

Once new concepts are authored, they are systematically stored in the mapping table, which is a structured database that includes the source terminology and corresponding target SNOMED CT identifiers [31]. We managed a Kakao Healthcare edition in our database that included all SNOMED International concepts, supplemented with new concepts assigned to a unique namespace identifier of the affiliated institution.

As the SNOMED CT Editorial Guide and modeling rules are subject to periodic updates, we monitor release notes and editorial guide updates on a monthly basis. While we conduct regular follow-up reviews each month, updates are applied to our edition on an annual basis through scheduled manual audits for the new concept authoring to comply with recent editorial changes.

System Implementation

To support the key functionalities of automatic mapping and new concept authoring, the system was implemented as 2 main components: a web application and an automatic mapping server. The web application, developed using Next.js (version 14; Vercel) and TypeScript (version 5; Microsoft), serves as the central platform that integrates various operations such as project management, data submission, mapping review, and concept authoring. React Query (version 5.28.14; TanStack) is used for managing asynchronous data interactions, and Tailwind CSS (version 3.3.0; Tailwind Labs) is used for styling. The application communicates with the backend services and the database to manage workflows and interface with the automatic mapping server. The automatic mapping server, built in Python (version 3.10; Python Software Foundation), performs the core mapping logic, including preprocessing of local terms, translation using LLM, and vector-based similarity search using FAISS (faiss-cpu version 1.9.0.post1; Meta AI Research). The server generates mapping results and stores them in a PostgreSQL (version 14.12; PostgreSQL Global Development Group) database. FastAPI (version 0.110.3; FastAPI framework) is used to serve the backend API.

Performance Evaluation

Method for Evaluating Accuracy

Dataset

The dataset for evaluating performance was collected from 4 hospital networks (a total of 9 university hospitals) in South Korea: Ewha Womans University Medical Center (EUMC) with 2 branches, Korea University Medical Center (KUMC) with 3 branches, DongSan Medical Center (DSMC) with 2 branches, and KyungHee University Medical Center (KHMC) with 2 branches. The selection of diagnostic and surgical procedure terms is strategically made due to their clear and distinct alignment with specific SNOMED CT semantic tags, such as “Disorder” and “Procedure,” which facilitates precise mapping. Furthermore, these codes encompass nearly all medical specialties, ensuring the heterogeneity of local medical practice in South Korea, thereby providing a robust dataset for evaluating the mapping tool’s performance. EUMC contributed 24,554 diagnostic and 8053 surgical procedure codes, KUMC provided 44,646 and 13,734, respectively, DSMC supplied 24,146 and 9432, while KHMC offered 27,842 diagnostic codes and 20,759 surgical procedure codes.

Preprocessing Rules

All performance evaluations were conducted using a common, standardized set of preprocessing rules applied uniformly across all 4 participating hospital networks (EUMC, KUMC, DSMC, and KHMC). The system’s “user-defined rules” functionality, while supported for production deployment, was intentionally not used during the formal testing in this study. This ensured that the comparison of mapping accuracy reflected the system’s performance under standardized, reproducible conditions.

LLM Configuration

All translation and concept generation tasks in this study were performed using GPT-4o (OpenAI, 2024) and its corresponding text-embedding-3-small embedding model for vector-based similarity mapping. The selection of GPT-4o was fixed across all institutions to ensure methodological consistency and reproducibility. All reported accuracy, efficiency, and usability metrics are therefore based exclusively on results generated from the GPT-4o-based pipeline.

XSL•FO
RenderX

Accuracy Assessment Strategy

The performance was measured by comparing its outputs against manually mapped results, which have been established as the gold standard. The manual mapping results were verified for accuracy and reliability through at least 2 rounds of internal review by 5 clinical terminology experts and 1 physician, and were further refined by external review with 2 physicians.

Accuracy was evaluated using both "top-1" and "top-5" approaches. The top-1 accuracy refers to the highest similarity score of a candidate being correct. The top-5 accuracy includes having the correct mapping among the 5 candidates with the highest similarity scores. Furthermore, to better characterize the contribution of the shared knowledge base versus the AI components, we stratified the "accuracy with reference terms" metric. We specifically calculated the "exact match" rate to distinguish successful mappings achieved via syntactic reference identity lookups (memorization from partner hospitals) from those achieved via vector-based semantic similarity search.

Specifically, we used a 2-pronged approach to measure mapping accuracy. First, we assessed the baseline performance at each participating hospital—EUMC, KUMC, DSMC, and KHMC—by using only the descriptions of the SNOMED CT Kakao Healthcare edition. The SNOMED CT Kakao Healthcare edition serves as the core internal database, managing all SNOMED International concepts and supplemented with new concepts assigned a unique namespace identifier of the affiliated institution. Crucially, the baseline analysis intentionally excludes any reference terms (previously confirmed mappings) that were accumulated or validated from the other participating hospital networks.

Second, we enhanced our analysis by adding reference terms from the other participating hospitals into the mapping process. For instance, when assessing EUMC, we introduced reference terms from KUMC, DSMC, and KHMC to gauge any changes in mapping accuracy. This comparative methodology was implemented in stages to allow us to compare the tool's performance before and after the addition of reference terms, as mapping data from other hospitals accumulated. The benefit measured in the subsequent metric, "accuracy with reference terms," therefore represents the isolated effect of integrating this shared, cross-institutional mapping knowledge.

Handling of "Not Mapped" Terms

To ensure comprehensive performance accounting, we clarified the treatment of local terms designated as "not mapped" within the accuracy calculation. These "not mapped" terms were established as part of the gold standard definition by the clinical terminology review committee. Such terms often represent local entries that human experts determined lacked a valid or appropriate SNOMED CT concept (eg, referencing only body structures or materials without a corresponding procedure).

Crucially, the system was not explicitly designed to predict a "not mapped" status. Therefore, when a term is validated by experts as "not mapped" (the gold standard outcome), the system's failure to provide any correct SNOMED CT prediction within the top-1 or top-5 candidates is inherently reflected in the overall accuracy calculation as a system prediction failure or error.

Method for Evaluating Efficiency

As a baseline, all processes were manually executed to assess the time required for each step. Data preprocessing time is the average time required for data cleaning and translation. New concept authoring time refers to the average time it takes for a person to author a new concept in compliance with SNOMED CT's principles. The final processing steps of the mapping table refer to the average time taken per dataset for integrating mapping tables and verifying data integrity. This observation is based on a 6-month operational period of a clinical terminology team consisting of 6 members. During this time, detailed logs and time-tracking data were collected across all stages of the terminology mapping process.

The results from the tool-based workflows were compared with the baseline data to evaluate the improvements: (1) efficiency gains are the reduction in time for each phase of the process, expressed as a percentage improvement over baseline; and (2) resource optimization evaluates the changes in human resource consumption.

Quality Assessment of New Concept Authoring

Quality was evaluated pre- and posttool adoption using 2 metrics derived from the full SNOMED CT authoring dataset as of August 27, 2024. First, the proportion of duplicate concepts was calculated by dividing the number of equivalent errors identified after classification by the total number of authored concepts at each time point. Second, the number of MRCM validation violations was measured by counting rule violations across all authored concepts before and after tool implementation.

System Usability Evaluation by Users

The system usability evaluation was conducted with 5 mappers who had previous experience with mapping methodologies before the tool development. These participants were introduced to the main features of the newly developed tool and were asked to perform approximately 100 sample mappings. After using the tool, they were encouraged to provide feedback in an open-ended format regarding their experience with the tool.

Ethical Considerations

This study was conducted as part of the research project "Development of an LLM-based Named Entity Recognition (NER) tool and a standardization support tool for using multicenter clinical free text." As no human participants or identifiable personal information were involved, institutional review board (IRB) approval was not required for this study itself. However, IRB approval or exemption was obtained from each institution for project administration purposes as follows: Ewha Womans University Medical Center (SEUMC IRB 2025-02-016, approved), DongSan Medical Center (DSMC IRB 2025-01-006, approved), Korea University Medical Center Anam (KUMC IRB 2025AN0239, approved), Guro (KUMC IRB 2025GR0222, approved), Ansan (KUMC IRB 2025AS0122, approved), and KyungHee University Medical Center (KHUH IRB 2025-01-047, exempted). This study was

XSL•FO
RenderX

conducted in compliance with institutional policies, the Korean Bioethics and Safety Act [32], and the Declaration of Helsinki [33]. Informed consent was not required due to the nature of the study.

Results

System Development and Deployment

System Development Requirements

Based on the users' survey, we found the following 11 key requirements as shown in Table 1. The UI of the auto-mapping enables users to map terms with a single click in a familiar table-based layout. Figure 4 displays the primary UI for the automated SNOMED CT mapping function. The interface is organized in a familiar table-based layout to facilitate the review and review of local diagnostic and surgical procedure terms. To maximize efficiency and minimize manual errors, the UI provides single-click mapping and automatic retrieval of FSNs immediately upon entering a standard concept ID. Furthermore, the interface supports essential collaborative functions for large-scale projects, including comprehensive filtering options and dedicated fields for review tracking and quality assurance by terminology experts.

Table 1. Key requirements for the LLM[a]-assisted automated SNOMED CT[b] mapping and new concept authoring tool, defined by clinical terminologists and data engineers for multi-institutional research environments.

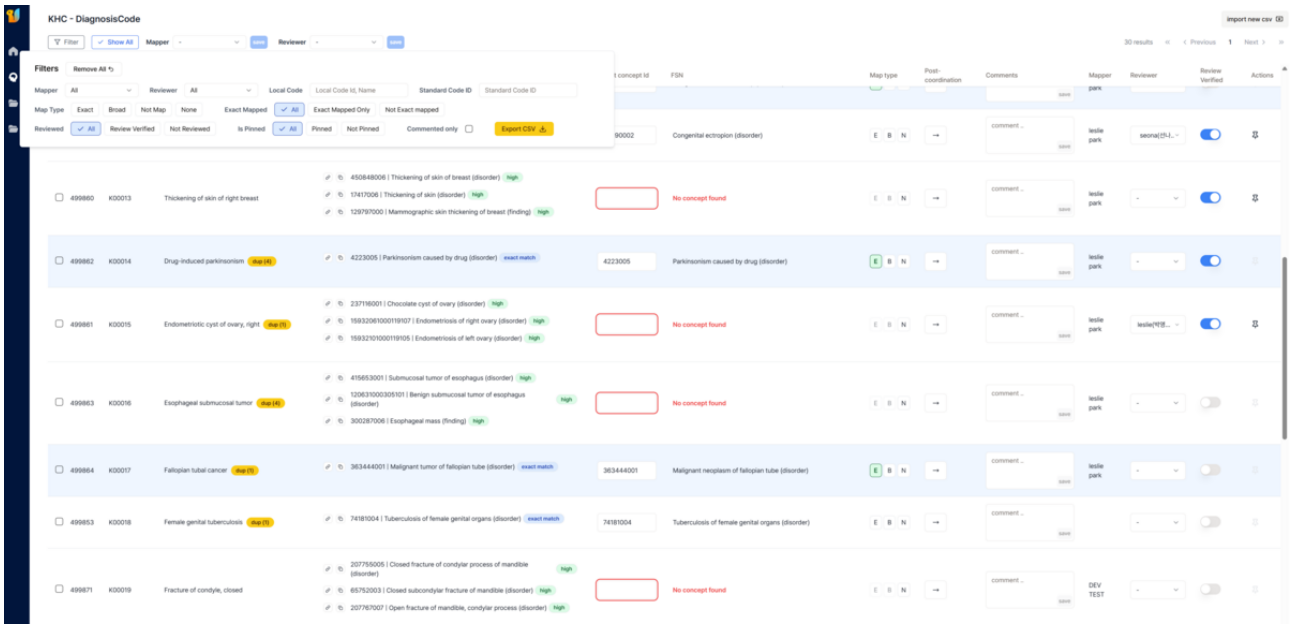
Category	Key requirements
Project and dataset version management	• The ability to create projects and manage their versions for terminology datasets by institution and domain. • Support for uploading source files and downloading mapping table result files in the computable format (CSV and JSON)
Automated mapping and terminology support	• Automated mapping to eliminate the need for manual browser searches for each local term. • Translation functionalities to support multilingual datasets. • Management tools for mapping tables, including cross-checking mapping results.
New concept authoring and integration	• Functionality to author new concepts during the mapping process. • A user interface applying MRCM[c] rules during authoring. • Application of the classification of newly authored concepts. • Immediate incorporation of newly authored concepts into the browser. • Automatic integration of newly authored concepts into the mapping tables.
Review and quality assurance	• Features for human review of newly authored concepts for naming inconsistencies and semantic modeling errors.

[a]LLM: large language model.

[b]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[c]MRCM: Machine Readable Concept Model.

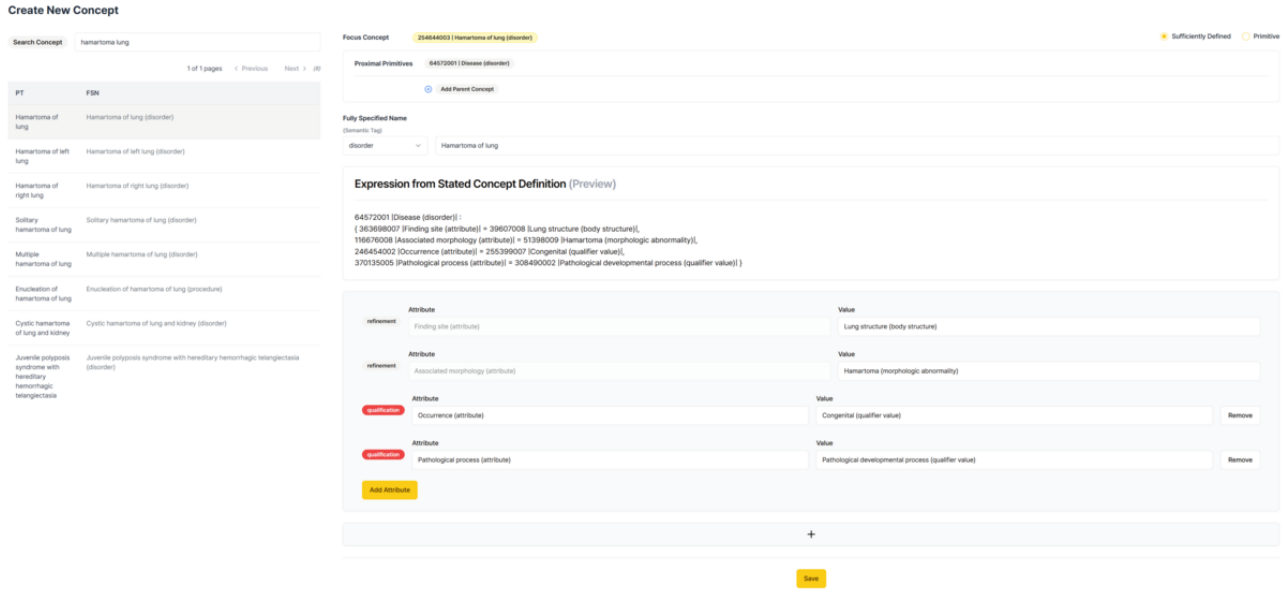
Figure 4. User interface of the automated mapping page within the LLM-assisted terminology standardization tool. KHC: Kakao Healthcare; LLM: large language model.



XSL•FO
RenderX

The new concept authoring tool, as shown in Figure 5, makes it easier to author new concepts with an intuitive visual editor that supports real-time updates. In illustrating Figure 5, the concept "254644003 |Hamartoma of lung (disorder)|" is used as the focus concept. Its modeling was cloned and then customized for the illustration, as duplicating the modeling of the focus concept is often an efficient and practical method when modeling new concepts. Then, the proximal primitive concept of selected focus concept, "64572001 |Disease (disorder)|," is automatically retrieved. Users can perform authoring within the UI through refinement or qualification methods.

Figure 5. User interface of the new concept authoring page within the LLM-assisted terminology standardization tool. FSN: fully specified name; LLM: large language model; PT: preferred term.



System Deployment

The system is actively deployed for terminology standardization at the Research Alliance participating hospitals, which is organized by Kakao Healthcare. To date, it has processed terminology from a total of 6 hospital networks, encompassing 12 university hospitals and 189,455 terms. The developed tool has been actively used to standardize terms across various domains, including diagnoses, surgical procedures, chief complaints, and laboratory tests.

System Evaluation

Result of Performance Evaluation

Table 2 displays the successful mappings achieved through syntactic matching (exact match lookups) using the custom reference database (baseline) vs those achieved after the iterative enrichment via shared reference terms (accuracy with reference terms) across 4 major hospital networks. The reported percentage represents the proportion of local terms that were successfully mapped by exact identity lookup. The analysis revealed that the inclusion of shared reference terms universally increased the proportion of terms resolved via exact match, reducing the reliance on vector-based search. However, the magnitude of this increase varied significantly across hospital networks.

XSL•FO
RenderX

Table 2. Syntactic mapping performance: exact match accuracy for diagnosis and surgical procedure domains.

Hospital networks, domain	Metric	Baseline accuracy (%)	Accuracy with reference terms (%)	Δ accuracy (%)
EUMC[a]				
Diagnosis	Exact match	36.3	94.5	58.2
Surgical procedure	Exact match	18.5	97.6	79.1
KUMC[b]				
Diagnosis	Exact match	23.4	36.6	13.2
Surgical procedure	Exact match	17.7	29.8	12.1
DSMC[c]				
Diagnosis	Exact match	34.9	94.9	60
Surgical procedure	Exact match	19.8	96.9	77.1
KHMC[d]				
Diagnosis	Exact match	21.9	55.2	33.3
Surgical procedure	Exact match	12.8	28.3	15.5

[a]EUMC: Ewha Womans University Medical Center.

[b]KUMC: Korea University Medical Center.

[c]DSMC: DongSan Medical Center.

[d]KHMC: KyungHee University Medical Center.

Notably, EUMC and DSMC exhibited a dramatic surge in exact match rates. For the diagnosis domain, exact match accuracy increased from 36.3% to 94.5% for EUMC and from 34.9% to 94.9% for DSMC. A similar trend was observed in the surgical procedure domain, where exact match rates reached 97.6% and 96.9%, respectively. In contrast, KUMC and KHMC showed more moderate increases. For the diagnosis domain, KUMC's exact match rate improved from 23.4% to 36.6%, and KHMC's improved from 21.9% to 55.2%. In the surgical domain, the exact match rates remained below 30% for both institutions (29.8% for KUMC and 28.3% for KHMC), though this still represented a gain of over 10% for each.

Table 3 summarizes the comparative performance evaluation of mapping accuracy for diagnostic and surgical procedure domains across 4 hospital networks. For each hospital and domain, mapping performance was measured as top-1 and top-5 accuracy, both at baseline (before incorporating reference terms) and after the inclusion of reference terms. To ensure statistical rigor, 95% CIs were calculated for all accuracy estimates. The absolute improvement in accuracy (Δ accuracy) was also calculated to quantify the benefit of leveraging shared reference mappings.

XSL•FO
RenderX

Table 3. Comparative performance evaluation of automated SNOMED CTa[a] mapping accuracy for diagnostic and surgical procedure domains across 4 South Korean multi-institutional networks. All improvements were statistically significant (McNemar test, $P<.01$ for all cases). 95% CIs were calculated using the Wald method.

Hospital network, domain, not mapped rate (%), metric	Baseline accuracy (%)	Accuracy with reference terms (95% CI)	Δ accuracy[b] (%)
EUMC[c]			
Diagnosis			
0			
Top-1	73.4	97.4 (97.2–97.6)	24.0
Top-5	88.9	98.7 (98.6–98.8)	9.8
Surgical procedure			
0			
Top-1	64.9	98.5 (98.2–98.8)	33.6
Top-5	84.7	99.2 (99.0–99.4)	14.5
KUMC [d]			
Diagnosis			
0.4			
Top-1	69.3	73.2 (72.8–73.6)	3.9
Top-5	87.0	89.7 (89.4–90.0)	2.7
Surgical procedure			
1.7			
Top-1	61.6	66.1 (65.3–66.9)	4.5
Top-5	79.0	82.6 (82.3–82.9)	3.6
DSMC [e]			
Diagnosis			
0			
Top-1	72.7	97.2 (97.0–97.4)	24.5
Top-5	88.2	98.5 (98.3–98.7)	10.3
Surgical procedure			
0			
Top-1	66.6	97.9 (97.6–98.2)	31.3
Top-5	85.2	98.7 (98.5–98.9)	13.5
KHMC [f]			
Diagnosis			
0			
Top-1	67.9	81.4 (80.9–81.9)	13.5
Top-5	85.5	92.8 (92.5–93.1)	7.3
Surgical procedure			
0			
Top-1	59.7	65.1 (64.5–65.7)	5.4
Top-5	81.1	84.7 (84.2–85.2)	3.6

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[b]Δ accuracy: accuracy with reference terms−baseline accuracy.

[c]EUMC: Ewha Womans University Medical Center.

[d]KUMC: Korea University Medical Center.

XSL•FO
RenderX

[e]DSMC: DongSan Medical Center.

[f]KHMC: KyungHee University Medical Center.

At baseline, top-5 accuracy for diagnosis was similar across hospitals, ranging from 85.5% to 88.9%. In contrast, surgical procedures showed greater variation, with baseline top-5 accuracy from 79% to 85.2%. After reference terms were incorporated, top-5 accuracy for diagnosis increased to between 89.7% and 98.7%, and for surgical procedures it rose to between 82.6% and 99.2%. EUMC and DSMC showed the highest postreference top-5 accuracy in both domains and achieved the largest absolute improvement in surgical procedures (up to 14.5% at EUMC and 13.5% at DSMC).

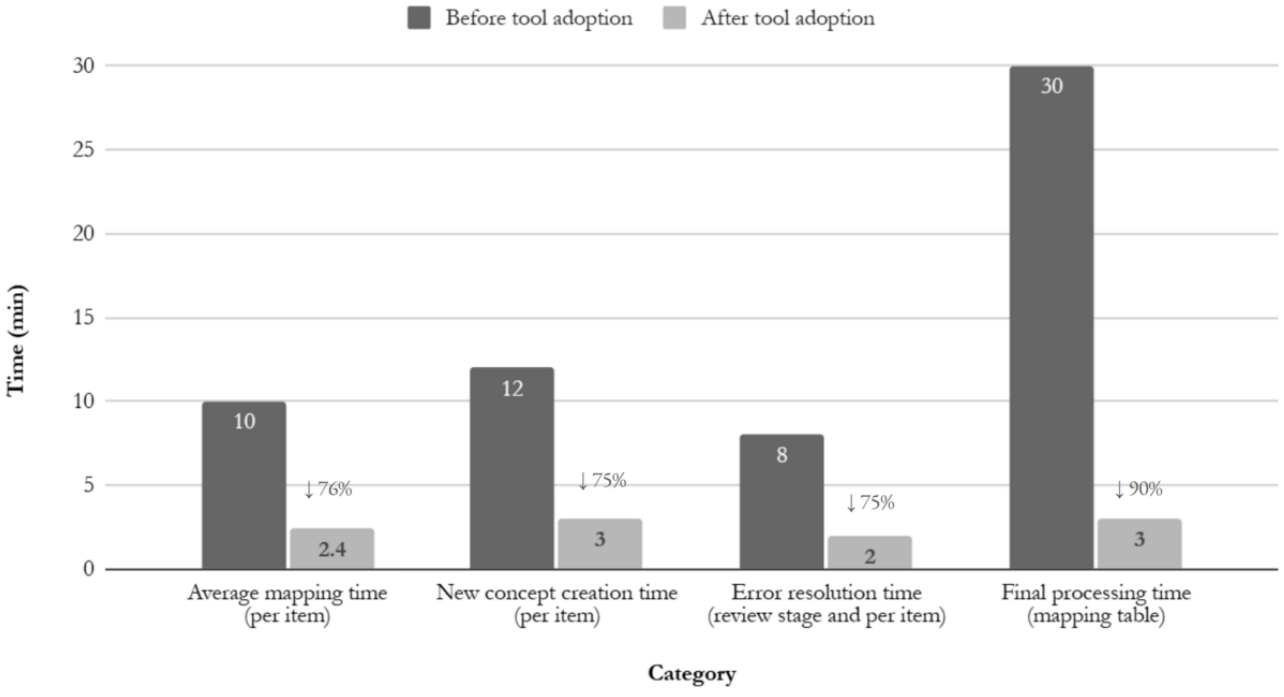
For top-1 accuracy, the improvement after reference terms addition was even more pronounced, especially in surgical procedures, with EUMC reaching a 33.6% increase. Top-5 accuracy remained consistently higher than top-1 in all hospitals and domains, but the relative improvement with reference terms was greater for top-1 than for top-5. All gains in accuracy were statistically significant (McNemar test, $P<.01$). In a supplementary evaluation, the exclusion of the syntactic mapping (exact match) step resulted in negligible variations in top-5 accuracy, suggesting that the vector-based semantic search effectively identifies these identical terms.

To provide context for performance ceilings, we analyzed the prevalence of unmappable terms. The gold standard dataset included terms categorized as "not mapped," representing local codes for which experts determined no valid SNOMED CT concept existed. These instances were treated as errors in the accuracy calculation. The prevalence of "not mapped" terms varied by institution and domain: EUMC (diagnosis: n=0, 0%; surgical procedure: n=0, 0%), KUMC (diagnosis: n=174, 0.4%; surgical procedure: n=219, 1.7%), DSMC (diagnosis: n=1, 0%; surgical procedure: n=0, 0%), and KHMC (diagnosis: n=0, 0%; surgical procedure: n=0, 0%).

Result of Efficiency Evaluation

Figure 6 clearly demonstrates the efficiency evaluation based on a 6-month operational period, showing significant reductions in manual workload across 4 key stages of the SNOMED CT mapping process. The manual mapping rate decreased from 100% to 70%, indicating a 30% reduction and highlighting the tool's effectiveness in enabling automated mapping. More broadly, the overall manual workload was significantly reduced, with marked enhancements in efficiency throughout the entire mapping process. Specifically, the average mapping time per item decreased from 10 to 2.4 minutes, representing a 76% reduction. The time required to create a new concept similarly dropped from 12 to 3 minutes (75% reduction), while the time needed to resolve errors during the review stage was reduced from 8 to 2 minutes (75% reduction). Most notably, the final processing time of the mapping table per dataset decreased from 30 to 3 minutes, achieving a 90% reduction.

Figure 6. Efficiency evaluation of the SNOMED CT mapping workflow, comparing manual baseline time to performance after adopting the LLM-assisted automated tool across 4 key stages: mapping, new concept authoring, error resolution, and final processing of the mapping table. LLM: large language model; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Results of New Concept Authoring Quality Evaluation

After implementation of the tool, both classification and MRCM validation errors decreased substantially, even as the total number of authored concepts increased. The proportion of duplicate concepts dropped from 0.8% (122/15,637) with manual authoring to 0.1% (46/34,401) with the tool, representing an 83% reduction. Likewise, MRCM validation violations

XSL•FO
RenderX

decreased by 72%, from 264 in the manual process to 75 with the tool-assisted workflow.

Result of System Usability Evaluation

Users highlighted several benefits of the system. First, the tool reduced effort during the mapping process by automatically suggesting candidate target concepts, which minimized the number of manual typing and clicking steps required. Second, users noted a significant decrease in errors, as the system eliminated mistakes that previously occurred when characters were deleted or data types were changed during manual copying and pasting of IDs or FSNs. The tool now conveniently enters the FSN when a valid ID is provided, further improving accuracy. Finally, the new system offered enhanced usability compared to the previous spreadsheet-based screens, which users found visually straining and inflexible. With the new interface, users reported a more comfortable and convenient experience.

Discussion

Principal Findings

This study developed an LLM-assisted SNOMED CT automatic mapping tool and evaluated it across multiple real-world hospital settings. Notably, the tool supports postcoordination, which minimizes semantic loss from source terms and leverages the strengths of SNOMED CT. In evaluation, the tool demonstrated high top-5 accuracy rates—92.6% for diagnostic concepts and 93.6% for surgical procedures. Incorporating reference mappings further increased accuracy by over 6% in both areas. Workflow efficiency was markedly enhanced, with mapping and concept authoring times reduced by approximately 75% and the proportion of cases requiring manual mapping dropping from 100% to 70%. Importantly, the tool reduced human modeling errors in new concept authoring, as evidenced by an 83% decrease in duplicate concept rate and a 72% reduction in MRCM validation violations. Error resolution and final processing of the mapping table were also significantly streamlined.

When evaluating site-specific performance, we found substantial variation across different hospital networks, especially after integrating the shared reference knowledge base. Performance in networks such as EUMC and DSMC was exceptionally high, with exact match rates exceeding 94% in diagnosis and 96% in surgical domains—likely reflecting the use of shared electronic medical record systems, which allowed most mappings to be resolved through simple identity lookups.

To analyze the actual contribution of the exact match step, we performed an internal assessment comparing mapping accuracy with and without the preliminary exact match stage. This assessment revealed that the difference in top-5 accuracy between the 2 configurations was negligible. Thus, the primary value of the exact match component lies not in enhancing end-to-end mapping quality but in workflow optimization: by quickly resolving about 30% of cases, it helps bypass computationally and financially intensive processes, contributing to pipeline efficiency and sustainability.

The accumulation and reuse of reference mappings across institutions is also central to our approach and reflects both the realities and strengths of terminology standardization initiatives in Korea; as more hospitals join and contribute to the shared repository, we anticipate continued improvements in mapping consistency and efficiency across various settings.

By contrast, KUMC and KHMC exhibited more modest improvements, as a substantial portion of their mappings could not be resolved through exact identity lookup alone. For these institutions, the shared knowledge base contributed to performance gains but was insufficient on its own; thus, many mappings still required the use of vector-based semantic similarity search. This highlights the ongoing need for advanced AI components when handling diverse or nonstandardized local terminologies that may not be well represented within a shared repository.

Further analysis of auto-mapping errors revealed 3 main sources of reduced accuracy: (1) the presence of local terms that reference only body structures or materials without an associated procedure—instances best categorized as “not mapped”; (2) terms requiring new concept creation via postcoordination—particularly in the surgical procedure domain, where postcoordination was frequently necessary; and (3) unresolved or ambiguous abbreviations. Understanding these error patterns provides important directions for future refinement of the system.

Comparison With Prior Work

A consistent set of challenges emerges across this and prior work, including the persistent difficulty of mapping rare or institution-specific terms, handling ambiguous abbreviations, and resolving semantically unclear expressions [17-19]. Beyond algorithmic advancements, real-world clinical deployment presents unique integration hurdles, such as harmonizing diverse IT infrastructures, maintaining synchronization between terminology servers and electronic medical record systems, and managing evolving local terminologies [18]. Importantly, our study reaffirms that, regardless of automation level, human review remains indispensable—not only for establishing and maintaining gold standard mappings that underpin system performance, but also for defining and updating mapping rules to reflect evolving clinical realities [17,18].

Limitations

First, we encountered various technical difficulties during implementation. The internal PostgreSQL database and the ElasticSearch (Elasticsearch B.V.) engine supporting SNOMED CT search (via SNOMED Snowstorm API and Browser) required frequent manual synchronization due to architectural differences. This was especially challenging since ElasticSearch lacks built-in transactional support. Accordingly, script-based data reconciliation was required to maintain consistency, but it remains to be fully automated.

Second, a significant limitation of our methodology involves the Korean-to-English translation step required for processing bilingual input data. As the accuracy of the subsequent vector similarity mapping is entirely dependent on the quality of the translated term, we acknowledge that the lack of gold-standard

XSL•FO
RenderX

comparisons or error analysis for this step is a methodological limitation.

Third, translation was handled as a fully automated preprocessing step to minimize human intervention and maximize efficiency within the mapping pipeline. To control quality when using models such as GPT-4o or Gemini 2.0 Flash, we used carefully designed, domain-specific prompts (Multimedia Appendix 1). These prompts enforced specific constraints: they defined clinical context (eg, diagnosis type), removed noise, and ensured that terms were translated strictly into the official medical vocabulary. However, the accuracy of subsequent mapping is entirely dependent on translation quality. Despite these controlled prompts, we currently lack external gold-standard metrics or formal error analysis for this critical phase.

We note that the volume of local terms requiring Korean translation was low, at approximately 1.5% of the total mapped terms. While this proportion is small, we acknowledge that translation inaccuracies in even a limited subset of terms could introduce minor bias into the overall mapping accuracy. Accordingly, we recognize the necessity of integrating translation-specific metrics into the overall evaluation framework.

Finally, while the gold standard for mapping underwent extensive review by experienced terminology experts both internally and externally, we did not collect or report an objective measure of interrater agreement, such as Cohen κ. Although this expert involvement likely improved reliability, the absence of a quantitative assessment remains a limitation and should be considered when interpreting our results.

Future Directions

Although retrieval-augmented generation was explored in our preliminary model, we observed limitations where the current general-purpose LLM failed to accurately leverage the inherent ontological structure of SNOMED CT, often yielding results identical to simple vector similarity search [34]. Therefore, further improvement is expected by using LLMs that are specifically designed to leverage both contextual information and the intrinsic structure of SNOMED CT in future research. Further, the use of SNOMED CT's ontology structure and knowledge graphs (eg, GraphRAG) is expected to enhance semantic retrieval and disambiguation capabilities, increasing accuracy and robustness across broader domains [35]. To improve authoring of new concepts, automation features applying detailed MRCM cardinality rules and an advanced inactivation process will be developed, reducing modeling errors and ensuring SNOMED CT compliance.

From a sustainability perspective, widespread adoption will require continued resource investment for tool maintenance and updates, as well as collaboration across clinicians, informaticians, terminology experts, and policymakers. Keeping standard codes current, ensuring ongoing quality assurance, and adapting to multilingual and ICD-11 (International Classification of Diseases, 11th Revision) or International Classification of Health Interventions-integrated environments—particularly for mapping complex surgical procedures—are essential for scalable and future-proof deployment. Despite concerns about the potential for increased operational costs when deploying commercial LLMs, our experimental results indicate that these expenses were negligible relative to the benefits achieved. For example, embedding 1.6 million tokens costs just US $0.03, and translating 1779 cases incurred only minimal costs based on standard token pricing. These findings highlight the economic viability and sustainability of incorporating LLM components into automated pipelines.

Our long-term vision is to deliver this solution as a cloud-based software-as-a-service platform, supporting agile terminology standardization across diverse medical institutions both domestically and worldwide. For effective adoption in new environments, local verification and adaptation may be necessary to address differences in data standards and clinical practices. Notably, the system's methodological approach aligns with the World Health Organization's and Centers for Disease Control and Prevention's international guidelines, supporting its potential generalizability across health systems worldwide [11,12].

Conclusions

We developed and validated the first large-scale, multi-institutional automatic terminology mapping tool supporting both pre- and postcoordinated SNOMED CT concepts. The tool substantially reduces manual workload, improves mapping accuracy, and enhances interoperability for clinical data across the evaluated sites. While our current evaluation was limited to internal implementation, further work will focus on advancing commercialization and expanding deployment. We anticipate that, with continued refinement and local verification, this tool could support terminology standardization and interoperability for broader national and international healthcare data integration in the future.

Acknowledgments

The authors would like to acknowledge the valuable contributions of the Ewha Womans University Medical Center, Kyung Hee University Medical Center, Korea University Medical Center, and Keimyung University DongSan Medical Center, particularly the IT departments at each hospital, for their cooperation. Our deepest gratitude goes to the Clinical Terminology Team and the Standards Technology Team at Kakao Healthcare for their outstanding support and technical expertise throughout the development of this study. We also used the generative artificial intelligence (AI) tool ChatGPT (GPT-4o, OpenAI; used in 2025) solely to assist with English language editing and grammar. The AI was not used to generate scientific content, analyze data, or create figures, tables, or references. All AI-assisted text was reviewed, verified, and edited by the authors, who take full responsibility for the content.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the hospital policies.

Funding

This work was supported by the Industrial Strategic Technology Development Program-Digital Healthcare (RS-2024-00467897, for the development of an LLM-based Named Entity Recognition [NER] tool and a standardization support tool for using multicenter clinical free text) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). The funder was involved in this study's design and data collection, but had no role in the analysis, interpretation, or writing of this paper.

Authors' Contributions

Conceptualization: YP (equal), SB (equal)
Formal analysis: SB (lead)
Investigation: YP (lead), SB (lead), HK (supporting)
Methodology: YP (equal), SB (equal), HK (equal)
Project administration: S-YS (lead)
Resources: DC (equal), SYR (equal), HSP (equal), K-JL (equal)
Supervision: S-YS (lead)
Visualization: HK (lead)
Writing – original draft: YP (equal), SB (equal), HK (equal), JK (equal)
Writing – review & editing YP (lead), SB (lead), HK (supporting), JK (supporting), S-YS (supporting)
SB and JK developed the software.

Conflicts of Interest

Several authors (YP, SB, HK, JK, and S-YS) are employees of Kakao Healthcare Corp. The technology described in this paper is the subject of 2 pending Korean patent applications assigned to Kakao Healthcare Corp: (1) "Apparatus and Method for Mapping Local Term of Medical Institution Terms to Medical Standard Terms" (Korean patent application number 10-2024-0085332; filed June 28, 2024), and (2) "Apparatus and Method for Generating Medical Standard Terms" (Korean patent application number 10-2024-0085330; filed June 28, 2024). Some of the authors are named inventors on these applications and may be entitled to royalties or other financial benefits if the patents are granted.

Multimedia Appendix 1
GPT-4o translation prompt.
[PDF File (Adobe PDF File), 109 KB - medinform_v14i1e82670_app1.pdf]

References

1. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: a systematic review. eClinicalMedicine 2024;71:102551 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102551] [Medline: 38533128]
2. Premanandan S, Ahmad A, Jaranka A, Üreten E, Lee E, Gross J, Cajander, et al. Opportunities and challenges for designing in connected health: insights from an expert workshop. J Med Internet Res 2025;27:e72446 [FREE Full text] [doi: 10.2196/72446] [Medline: 41172290]
3. Dinov ID. Volume and value of big healthcare data. J Med Stat Inform 2016;4:3 [FREE Full text] [doi: 10.7243/2053-7662-4-3] [Medline: 26998309]
4. Ambalavanan R, Snead RS, Marczika J, Towett G, Malioukis A, Mbogori-Kairichi M. Challenges and strategies in building a foundational digital health data integration ecosystem: a systematic review and thematic synthesis. Front Health Serv 2025;5:1600689 [FREE Full text] [doi: 10.3389/frhs.2025.1600689] [Medline: 40621432]
5. Torab-Miandoab A, Samad-Soltani T, Jodati A, Akbarzadeh F, Rezaei-Hachesu P. A unified component-based data-driven framework to support interoperability in the healthcare systems. Heliyon 2024;10(15):e35036 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e35036] [Medline: 39161828]
6. Prakash R, Dupre M, Østbye T, Xu H. Extracting critical information from unstructured clinicians' notes data to identify dementia severity using a rule-based approach: feasibility study. JMIR Aging 2024;7:e57926 [FREE Full text] [doi: 10.2196/57926] [Medline: 39316421]
7. Park HA, Hardiker N. Clinical terminologies: a solution for semantic interoperability. J Korean Soc Med Inform 2009;15(1):1. [doi: 10.4258/jksmi.2009.15.1.1]
8. Park HA, Kim HY, Min YH. Use of clinical terminology for semantic interoperability of electronic health records. J Korean Med Assoc 2012;55(8):720-728. [doi: 10.5124/jkma.2012.55.8.720]
9. Interoperability in healthcare. HIMSS. URL: https://nationalcapitalarea.himss.org/resources/interoperability-healthcare [accessed 2026-02-20]

XSL•FO
RenderX

10. Fennelly O, Grogan L, Reed A, Hardiker NR. Use of standardized terminologies in clinical practice: a scoping review. Int J Med Inform 2021;149:104431 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104431] [Medline: 33713915]
11. Global Strategy on Digital Health 2020-2025. World Health Organization. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2026-02-13]
12. National Notifiable Diseases Surveillance System (NNDSS): Integrated surveillance information systems/NEDSS. CDC. URL: https://www.cdc.gov/nndss/what-is-case-surveillance/nedss.html [accessed 2026-02-13]
13. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthcare Inf Res 2021;27(1):3-10 [FREE Full text] [doi: 10.4258/hir.2021.27.1.3] [Medline: 33611871]
14. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290. [Medline: 17095826]
15. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc 2021;28(9):2017-2026 [FREE Full text] [doi: 10.1093/jamia/ocab084] [Medline: 34151978]
16. Isaradech N, Khumrin P. Auto-mapping clinical documents to ICD-10 using SNOMED-CT. AMIA Jt Summits Transl Sci Proc 2021;2021:296-304 [FREE Full text] [Medline: 34457144]
17. Davidson R, Hardman W, Amit G, Bilu Y, Della Mea V, Galaida A, et al. SNOMED CT entity linking challenge. J Am Med Inform Assoc Oxford University Press 2025;32(9):1397-1406. [doi: 10.1093/jamia/ocaf104] [Medline: 40657868]
18. Gehrmann J, Dogan A, Hagelschuer L, Quakulinski L, Koy A, Beyan O. Catnip for MedCAT: optimizing the input for automated SNOMED CT mapping of clinical variables. Stud Health Technol Inform 2025;331:142-152. [doi: 10.3233/SHTI251390] [Medline: 40899536]
19. Kulyabin M, Sokolov G, Galaida A, Maier A, Arias-Vergara T. SNOBERT: a benchmark for clinical notes entity linking in the SNOMED CT clinical terminology. Springer 2024:154-163. [doi: 10.1007/978-3-031-78119-3_11]
20. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The Faiss library. arXiv. Preprint posted online on Oct 23, 2025 2025 Oct 23. [doi: 10.48550/arXiv.2401.08281]
21. SNOMED CT Editorial Guide. SNOMED International. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide [accessed 2026-02-13]
22. SNOMED CT machine readable concept model. SNOMED International. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-machine-readable-concept-model [accessed 2026-02-13]
23. Proximal primitive parent. SNOMED International. URL: https://docs.snomed.org/snomed-international-documents/snomed-ct-glossary/p/proximal-primitive-parent [accessed 2026-02-13]
24. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403 [FREE Full text] [Medline: 9865037]
25. Proximal primitive modeling. SNOMED CT Editorial Guide. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide/readme/authoring/general-modeling/proximal-primitive-modeling [accessed 2026-02-20]
26. IHTSDO/snowstorm. GitHub. URL: https://github.com/IHTSDO/snowstorm [accessed 2026-02-13]
27. Machine readable concept model (MRCM). SNOMED International. URL: https://docs.snomed.org/snomed-ct-user-guides/authoring-user-guide/faqs/machine-readable-concept-model-mrcm [accessed 2026-02-13]
28. Hao X, Li X, Huang Y, Shi J, Abeysinghe R, Tao C, et al. Quantitatively assessing the impact of the quality of SNOMED CT subtype hierarchy on cohort queries. J Am Med Inform Assoc 2025;32(1):89-96. [doi: 10.1093/jamia/ocae272] [Medline: 39520711]
29. IHTSDO/release-validation-framework. GitHub. URL: https://github.com/IHTSDO/release-validation-framework [accessed 2026-02-13]
30. IHTSDO/snomed-owl-toolkit. GitHub. URL: https://github.com/IHTSDO/snomed-owl-toolkit [accessed 2026-02-13]
31. HINS healthcare information standard. URL: https://www.hins.or.kr [accessed 2026-02-13]
32. Korean Law Information Center. URL: https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2&query=bioethics#liBgcolor0 [accessed 2026-02-27]
33. WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2026-02-27]
34. Bratanič T. The limitations of text embeddings in RAG applications. neo4j. 2024. URL: https://neo4j.com/blog/developer/rag-text-embeddings-limitations/ [accessed 2026-02-20]
35. Choudhary S, Luthra T, Mittal A, Singh R. A survey of knowledge graph embedding and their applications. arXiv. Preprint posted online on Jul 16, 2021 2021 Jul 16. [doi: 10.48550/ARXIV.2107.07842]

Abbreviations

AI: artificial intelligence
API: application programming interface
DSMC: DongSan Medical Center
EUMC: Ewha Womans University Medical Center
FSN: fully specified name

XSL•FO
RenderX

HRS: Healthcare Data Research Suite
ICD-11: International Classification of Diseases, 11th Revision
IRB: institutional review board
KHMC: KyungHee University Medical Center
KUMC: Korea University Medical Center
LLM: large language model
MRCM: Machine Readable Concept Model
OWL: Web Ontology Language
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
UI: user interface

Edited by A Coristine; submitted 20.Aug.2025; peer-reviewed by S Sung, K Adegoke, K Triep, D Chrimes; comments to author 27.Oct.2025; revised version received 21.Jan.2026; accepted 23.Jan.2026; published 09.Mar.2026.

Please cite as:
Park Y, Kang H, Kim J, Shin SY, Cho D, Rhee SY, Park HS, Lee KJ, Bae S
Development and Evaluation of SNOMED CT Automated Mapping Tool: Advancing Terminology Standardization and Semantic Interoperability
JMIR Med Inform 2026;14:e82670
URL: https://medinform.jmir.org/2026/1/e82670
doi:10.2196/82670
PMID:

©Youngsun Park, Hannah Kang, Jiwon Kim, Soo-Yong Shin, Dosang Cho, Sang Youl Rhee, Hong Seok Park, Kyung-Jae Lee, Sungchul Bae. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Analysis Model for Infant Incubator Adverse Events Using Retrieval-Augmented Generation Combined With Dual-Adapter Fine-Tuning: Development and Evaluation Study

Wenke Xia[1], MEng; Wanting Zhu[1], MEng; Tianchun Li[1], MEng; Li Wang[2], MSc; Weiqi Li[1], PhD; Peiming Zhang[1], PhD

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China

[2]Center for Drug Reevaluation of Henan, Zhengzhou, China

Corresponding Author:
Peiming Zhang, PhD
School of Health Science and Engineering
University of Shanghai for Science and Technology
Number 516, Jungong Road
Yangpu District
Shanghai, 200093
China
Phone: 86 13524693176
Email: zpmking@163.com

Abstract

Background: Infant incubator adverse events refer to various harmful incidents that occur during the normal use of marketed infant incubators and result in, or may result in, bodily harm. In recent years, however, the number of reported adverse events has continued to rise. This trend has made the monitoring of infant incubator adverse events time-consuming and labor-intensive when relying solely on manual processing by medical device adverse-event monitoring personnel. Meanwhile, general-purpose large language models (LLMs) still face domain knowledge gaps and hallucination issues in specialized fields. Through fine-tuning, LLMs can be adapted to specific application scenarios, while retrieval-augmented generation (RAG) enhances their ability to handle knowledge-intensive tasks. Therefore, LLMs that integrate these 2 technologies hold significant potential for addressing monitoring challenges.

Objective: This study aims to enhance the intelligent monitoring of adverse events related to medical devices by integrating RAG with dual-adapter fine-tuning to construct an adverse-event analysis model for infant incubators. The model enables an integrated workflow centered on extraction and analysis, encompassing structured extraction of adverse events, narrative analysis, and regulatory question answering.

Methods: This study leveraged adverse-event data from Chinese infant incubators to construct a high-quality dataset through prompt engineering. Technologically, it combined 2 parameter-efficient fine-tuning methods—low-rank adaptation and infused adapter by inhibiting and amplifying inner activations—to achieve efficient adaptation on the Qwen2-7B base model. Simultaneously, it introduced the FINBGE embedding model with supervised contrastive semantic optimization to build a knowledge retrieval system that mitigates hallucination.

Results: This study comprises 1565 pediatric disease question-and-answer entries from PediaBench, 2530 specific corpora on adverse events in infant incubators, and 1488 regulatory corpora. Extensive experiments demonstrate the superiority of this analytical model across various metrics. Under the same experimental conditions, the element recall rate reaches 0.815, the accuracy of infant incubator adverse event analysis is 0.898, and the accuracy in regulatory clause question-answering tasks attains 0.938.

Conclusions: The analytical model proposed in this study demonstrates significant advantages in analyzing adverse events related to infant incubators, while also achieving substantial improvements in text generation metrics. When combined with RAG, it not only effectively mitigates hallucination but also enhances knowledge timeliness. This study employs the Qwen model as its foundational framework, leveraging large-model fine-tuning and RAG to achieve intelligent analysis of high-risk medical device adverse-event monitoring data. It demonstrates the feasibility of implementing intelligent medical device regulation within the Qwen ecosystem.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e83745) doi:10.2196/83745

KEYWORDS

adverse events in medical devices; infant incubators; large language models; fine-tuning techniques; retrieval-augmented generation

Introduction

Background

Collecting and analyzing adverse event monitoring data for high-risk medical devices is a crucial means of ensuring their safety and effectiveness [1]. Currently, medical device regulatory authorities have widely adopted electronic monitoring systems, which not only enhance data collection efficiency and convenience but also lead to a sustained increase in the volume of adverse event reports submitted [2]. Badnjević et al [3] found that despite clear regulatory mandates for postmarket surveillance obligations across countries, issues persist in adverse event reporting, including inconsistent standards, noncomparable data, and fragmented oversight. In China, infant incubators are classified as class III high-risk medical devices. Adverse event monitoring for these products involves interdisciplinary knowledge integration across multiple fields. Challenges include large volumes of adverse event data, predominantly in unstructured and nonstandardized formats [4], as well as significant variation in data quality reported by health care institutions [5]. These factors substantially increase the workload and complexity of analyzing and processing reports. Recent innovations in natural language processing technology offer new possibilities for balancing efficiency and professionalism. Ruseva et al [6] proposed a deep learning model based on the TensorFlow framework, demonstrating strong performance in assessing adverse drug reaction risks and enhancing postmarketing drug surveillance capabilities. Hauben [7] developed artificial intelligence (AI)–based methods to enhance drug-drug interaction detection and prediction. Borjali et al [8] integrated deep learning to accurately and promptly detect adverse medical events from free-text medical narratives.

While traditional deep learning models can perform text classification, they lack long-range contextual associations [9] and causal reasoning capabilities [10], exhibit weak generalization, and demonstrate poor transferability between events [11]. Lu et al [12] found that Transformer models consistently outperformed traditional convolutional neural network models in nearly all medical text classification scenarios, with convolutional neural networks demonstrating acceptable performance only when categories were balanced. Naik et al [13] highlighted that adapting event extractors to new medical domains suffers from covariate shift issues, rendering direct transfer ineffective and necessitating specialized domain adaptation techniques.

Although large language models (LLMs) can provide solutions to core challenges in medical device adverse event monitoring, pretraining foundational models from scratch for vertical domains incurs prohibitively high costs. For instance, first, training a model with tens of billions of parameters requires multiple graphics processing units and months of computational time [14], demanding substantial computational resources and time investment. Second, pretraining often requires datasets in the gigabyte or even terabyte range [15], while vertical domain data are typically strictly confidential [16]. Publicly available data may constitute less than 0.1% of pretraining requirements, and the processes of data collection, cleaning, annotation, and storage involve substantial human and material resources. Finally, pretraining imposes steep technical barriers, demanding both deep algorithmic engineering expertise and the development of large-scale distributed training frameworks [17]—a burden beyond the reach of most small-to-medium teams.

Therefore, this study constructs an infant incubator adverse event analysis model by fine-tuning LLMs. Methods such as low-rank adaptation (LoRA) and infused adapter by inhibiting and amplifying inner activations, that is, $(IA)^3$, require only minor weight updates, preserving general language prior knowledge while incorporating domain-specific expertise [18]. However, for tasks involving the analysis of adverse events related to infant incubators—which are complex, densely populated with specialized terminology, and characterized by imbalanced samples—a single method has inherent limitations. LoRA excels at capturing long-range contextual dependencies and temporal event correlations through low-rank updates, but its global low-rank matrix may lack sufficient feature resolution when handling rare fault descriptions or subtle terminological variations. Meanwhile, $(IA)^3$ enables fine-grained control over the strength of specific features through dynamic scaling of activation values, but it struggles to model long-range semantic relationships across sentences or paragraphs. Neither approach alone can comprehensively address these challenges.

To this end, this study proposes a “dual-adapter joint fine-tuning” architecture. Its novelty lies not in simple superposition but in a hierarchical division of labor and collaboration: the LoRA component focuses on modeling logical chains and long-range dependencies within adverse events, while the $(IA)^3$ component performs local feature enhancement and normalization on terminology and key entities to suppress noise. Coupled through a dynamic weight transfer mechanism, this approach enables the model to achieve deep adaptation to vertical domain knowledge while preserving general medical semantic understanding. Although vertical domains often lack extensive data, Wang et al [19] demonstrated the feasibility of studying small-sample adverse event data in medical devices when investigating knowledge transfer and calibration for small-sample medical signals (epileptic electroencephalography).

High-quality and task-relevant data form the cornerstone for ensuring that fine-tuned models achieve the expected performance, accuracy, and specific behavioral patterns. Data quality directly determines the nuances learned by the model, the biases it may inherit or mitigate, and its generalization capabilities [20]. The data source used in this study comprises nonpublic adverse event data from infant incubators within

XSL•FO
RenderX

China's Medical Device Adverse Event Monitoring System (the details of adverse events are shown in Multimedia Appendix 1). This dataset authentically captures adverse events generated by infant incubator use in medical institutions in recent years, encompassing complex and diverse semantic expressions within the domain. Real-world noise reflects the challenges of practical application scenarios; mild noise can actually compel the model to learn deeper semantic features rather than relying on spurious patterns in the data, thereby enhancing model robustness [21].

It is important to note that even after model fine-tuning, hallucination persists. From a computational theory perspective, all AI models—including current mainstream LLMs—are fundamentally probabilistic mathematical function approximators [22]. These models map inputs to output spaces through complex parameter matrix operations, generating outputs that are essentially responses sampled from high-dimensional probability distributions. Retrieval-augmented generation (RAG) significantly mitigates this issue. It is an AI technique that combines information retrieval technology with language generation models [23]. This approach enhances the model's capability to handle knowledge-intensive tasks by retrieving relevant information from external knowledge bases and feeding it as prompts to LLMs.

Related Work

In recent years, the rapid advancement of AI technology has increasingly highlighted the application value of large-scale pretrained models across multiple professional domains. Advanced models such as GPT and Bidirectional Encoder Representations from Transformers (BERT), with their exceptional language comprehension and generation capabilities, have demonstrated significant application potential in the health care sector. Research data indicate that models based on the Transformer architecture possess unique advantages in medical text processing: Kim et al [24] confirmed that domain-adapted BERT variants perform well in predicting appropriate medical departments based on patient symptom descriptions and in answering specialized medical questions. Zhang et al [25] utilized LLMs for medical record generation tasks, saving physicians approximately 0.5-1 hour of work time daily.

Notably, the application scope of such models has expanded beyond text processing into medical devices and regulatory domains. Tang et al [26] highlighted that the global AI medical device industry is experiencing robust growth, while the urgent need for regulatory oversight is becoming increasingly apparent—a trend expected to intensify. For instance, Kim et al [27] developed a deep learning predictive model to rapidly and accurately detect mechanical ventilation requirements in neonatal intensive care units using electronic health records; Matheny et al [28] systematically explored how applying LLMs to the US Food and Drug Administration's (FDA) medical product safety surveillance system enhances the identification and assessment of adverse drug events.

Regarding fine-tuning medical LLMs, multiple studies have explored application methods. For instance, Jia et al [29] employed LoRA for instruction fine-tuning of pretrained LLMs, ultimately enhancing the automated generation of children's health science resource labels. Abbasian et al [30] utilized electronic health records from patients with diabetes for P-tuning and LoRA fine-tuning of large models, opening new avenues for diabetes management and improving therapeutic efficacy. Integrating large models with RAG has also become a research hotspot. Hu et al [31] introduced a pioneering personalized dementia care approach by integrating GPT-4, fine-tuned and RAG-enabled, with a vector database. Kulshreshtha et al [32] enhanced a medical chatbot based on the Llama source model using RAG methods to generate reliable, contextually relevant responses, thereby improving clinical judgment and reducing health care workers' workload.

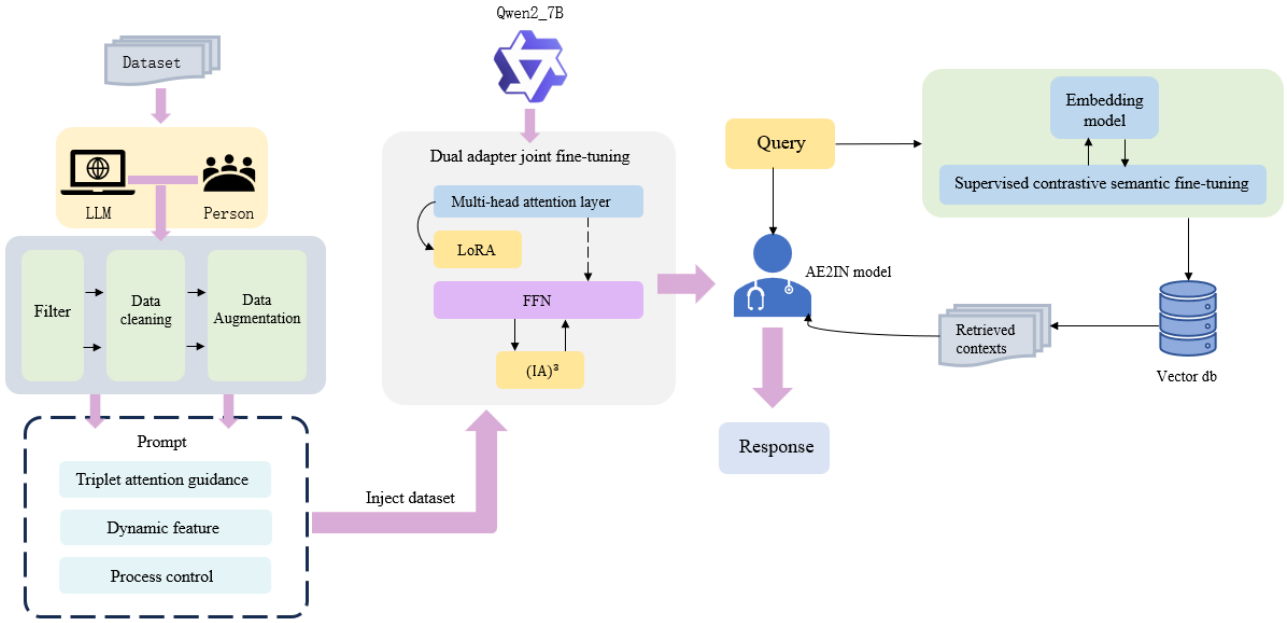
Methods

Overview

The adverse event analysis model AE2IN proposed in this study for infant incubators comprises 2 core modules: a data preprocessing layer and a knowledge-enhanced analysis layer. The specific framework is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Framework diagram of the AE2IN adverse event analysis model for infant incubators. The preprocessing layer cleans the data, normalizes terminology, and enhances adverse event reports. The knowledge enhancement layer enables collaborative analysis through external knowledge injection and large-model semantic reasoning, using a Qwen model fine-tuned with dual adapters alongside a fine-tuned embedding model. FFN: feedforward network; (IA)3: Infused Adapter by Inhibiting and Amplifying Inner Activations; LLM: large language model; LoRA: low-rank adaptation.



At the data preprocessing layer, the system first standardizes input adverse event reports for infant incubators through data filtering, cleaning, and augmentation. The data standardization, cleaning, and augmentation processes for this study's training dataset leveraged GPT-4 as a teacher model for auxiliary processing. Through knowledge distillation in a teacher-student paradigm, it provided high-quality training samples for subsequent domain-specific fine-tuning of the Qwen2_7B model. Standardization during conventional inference fully utilizes specialized medical dictionaries and regular expression rules to ensure accurate recognition of diverse terminology. The preprocessed structured data are subsequently fed into the knowledge-enhanced analysis layer.

The knowledge-enhanced analysis layer serves as the system's core, employing a dual-path parallel processing mechanism. On the one hand, supervised comparison-based semantic optimization fine-tunes the embedded model FINBGE to vectorize queries and retrieve the most relevant clause content from the medical device regulatory knowledge base. On the other hand, the Qwen2_7B model, jointly fine-tuned via dual adapters, performs deep semantic understanding of the input text. Notably, this model innovatively combines 2 parameter-efficient fine-tuning techniques: LoRA and $(IA)^3$.

Fine-Tuning

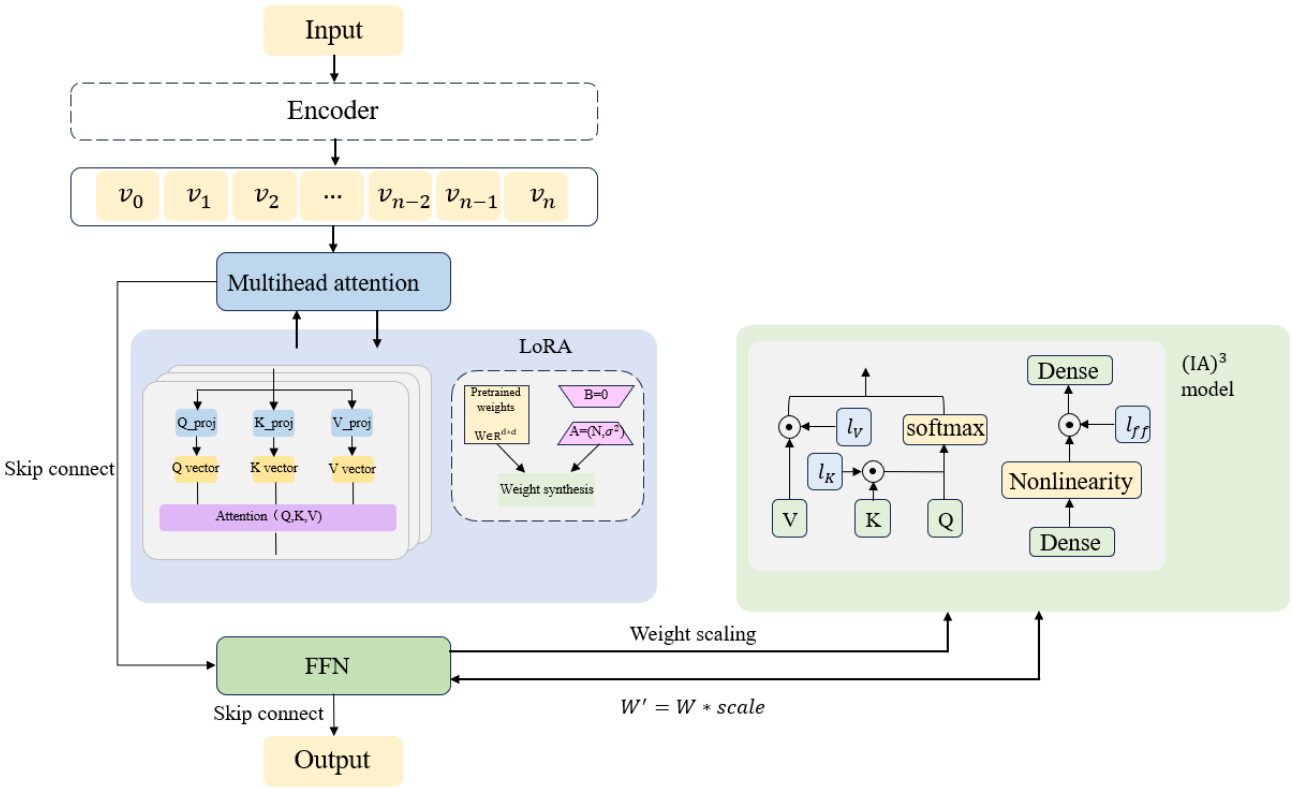
Dual-Adapter Fine-Tuning

This study designed a dual-adapter joint fine-tuning method that combines LoRA and $(IA)^3$, 2 parameter-efficient fine-tuning techniques. LoRA focuses on optimizing the Q (query)/V (value) projection matrices within the attention mechanism through low-rank matrix updates (r=32), effectively capturing contextual dependencies in descriptions of adverse events in infant incubators. Meanwhile, $(IA)^3$ refines the feature representations of the feedforward network by learning feature scaling coefficients without introducing additional parameters, thereby enhancing the representation of abnormal features and the generation of handling suggestions. Mathematically, this can be expressed as a composite transformation:

where W represents the original pretrained weights, ΔW is the low-rank update matrix introduced by LoRA, composed of the product of 2 small matrices A and B, and l is the feature scaling vector learned by $(IA)^3$. This architecture retains LoRA's advantage in pattern discovery—learning domain-specific attention patterns through low-rank space—while integrating the precise feature selection control of $(IA)^3$. The overall framework is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Dual-adapter fine-tuning framework diagram. The architecture is built on the Qwen2_7B foundation model and integrates 2 parameter-efficient methods: low-rank adaptation (LoRA) and Infused Adapter by Inhibiting and Amplifying Inner Activations (IA)3. The diagram illustrates the core design, with LoRA applied to the query/value projection matrices in the Transformer attention mechanism and (IA)3 applied to the feedforward network (FFN). It also depicts the fine-tuning strategy, where pretrained weights are frozen and only adapter parameters are updated, highlighting key processes such as attention computation, weight composition, and feature scaling.



The domain-adaptive training strategy employs a dynamic weighting mechanism to balance domain-specific data with general-purpose corpora. Its objective function is defined as follows:

where λ_d is the domain data weight coefficient; λ_g is the general data weight coefficient, with a ratio of 3:1; D_d and D_g represent the domain dataset and general dataset, respectively; denotes the loss function for a single sample; y_t is the tth token in the target sequence; $y_{<t}$ indicates the historical output sequence; and T is the target sequence length.

By employing a 3:1 weighted sampling mechanism, domain data receive high-frequency exposure to accelerate the model's mastery of infant incubator adverse event norms, while continuous infusion of general data maintains foundational medical semantic understanding.

This method can be divided into 2 stages. The first stage focuses on optimizing the attention mechanism. By utilizing low-rank matrix updates, it establishes cross-layer semantic associations for factors in the Q/V projection matrix that contribute to adverse events in infant incubators. This step notably enhances the model's ability to understand temporal patterns in infant incubator adverse events. It can be conceptualized as the model's capability to accurately capture causal chains such as "temperature alarm → skin temperature probe failure → component inspection." The second stage specifically adjusts the feature representations in the feedforward network. By learning scaling coefficient vectors, it finely controls the generation intensity of infant incubator adverse event terminology and the standardization of subsequent handling recommendations. For example, operations such as "Immediately replace the incubator, monitor the infant's temperature and pulse oxygen saturation, and report to the equipment department for repair" fall under the follow-up actions category, while information such as "Equipment model YP-90B, root cause analysis of the fault" is accurately categorized under the additional information category. This ensures that the model's output maintains both professional precision and compliance with medical documentation standards.

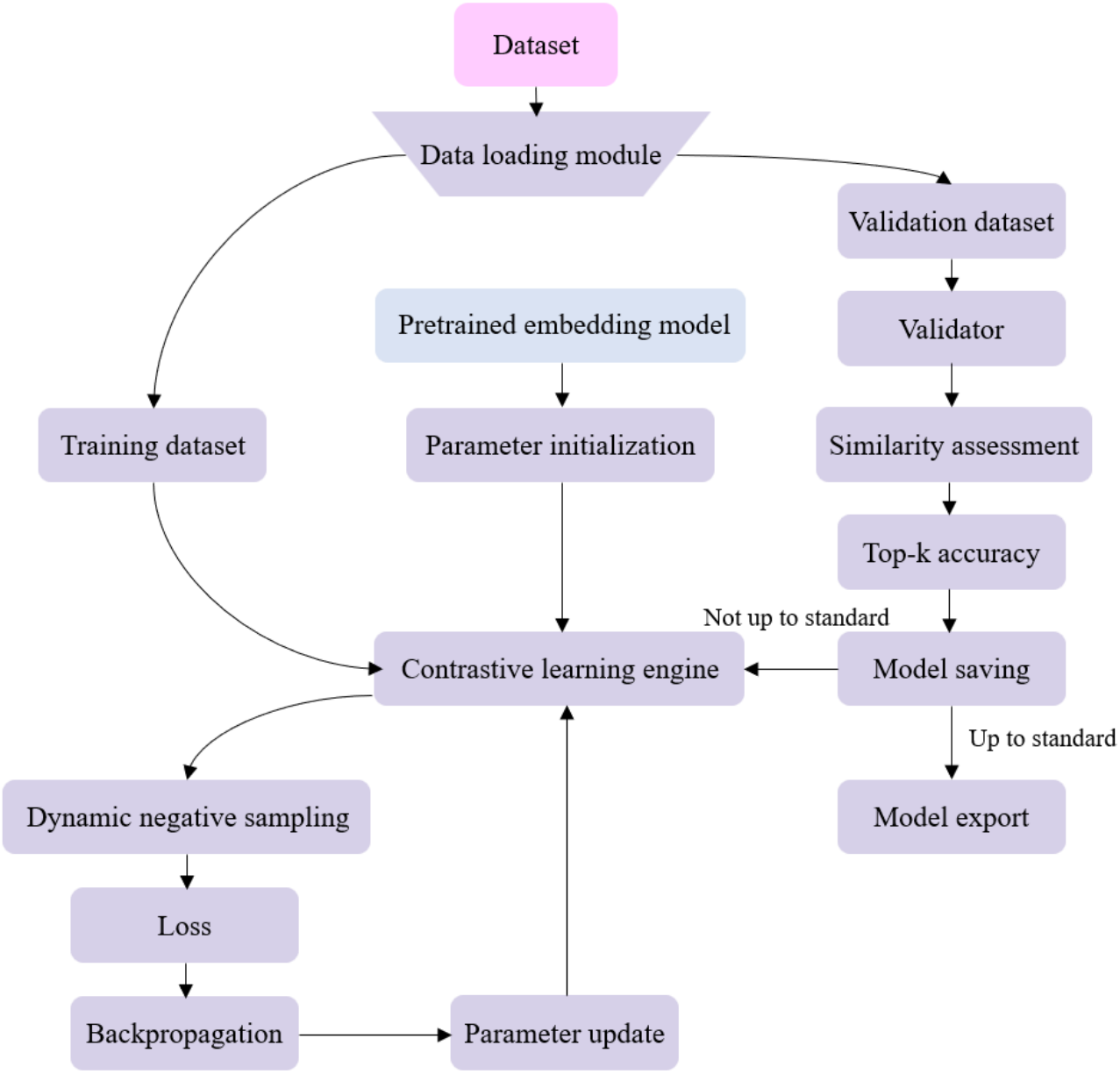
Knowledge inheritance between stages is achieved through dynamic weight transfer—using attention patterns obtained from LoRA training as initialization priors for the $(IA)^3$ stage. This strategy converges faster than parallel training while preserving model generalization capabilities and notably enhancing domain-specific task performance. It effectively overcomes the limitations of either approach: fine-grained feature variations that are difficult for LoRA to capture are compensated by $(IA)^3$, while long-range dependency modeling—a weakness of $(IA)^3$—is supplemented by LoRA.

XSL•FO
RenderX

Supervised Contrastive Semantic Optimization Fine-Tuning

In specialized domains, general-purpose embedding models often struggle to accurately capture subtle distinctions between professional terminology, leading to similar yet unrelated concepts being overly close in vector space [33]. Therefore, this paper proposes a supervised contrastive semantic optimization method to fine-tune embedding models. The fine-tuning workflow framework is illustrated in Figure 3.

Figure 3. Framework diagram of the supervised contrastive semantic optimization fine-tuning process. The model performs comparative learning and parameter updates using both the training and validation datasets.



By constructing an end-to-end fine-tuning pipeline, we achieve efficient domain adaptation for pretrained embedding models. Utilizing the Sentence Transformers Finetune Engine framework, we combine contrastive learning mechanisms with domain knowledge injection. This enables the model to accurately capture domain-specific semantic association patterns while preserving its original semantic understanding capabilities. The core training objective is mathematically expressed as follows:

where q_i is the embedding vector for query i; p_i^+ and p_i^- are the positive and negative samples associated with q_i, respectively; τ is the temperature coefficient controlling the sharpness of the distribution; N denotes the number of queries in the batch; and K represents the number of negative samples corresponding to each query. Cosine similarity establishes the relevance metric between queries and documents. The temperature parameter τ controls the sensitivity of similarity scores. The numerator term $e[\text{sim}(q_i, p_i^+)/\tau]$ generates centripetal force, drawing relevant documents closer to the query point in vector space. The denominator term generates a repulsive field, pushing

XSL•FO
RenderX

irrelevant documents away from the query point's semantic neighborhood.

Beyond reconstructing the semantic space, feature space calibration requires an embedding projection transformation, formulated as:

$$h = \text{LayerNorm}[W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2]$$

W_1 and W_2 form a parameterized transformation matrix that projects the original embedding onto a subspace more suitable for the current task. The GELU function introduces a smooth threshold, preserving important features while filtering out noise. LayerNorm maintains numerical stability, ensuring comparability between embeddings from different domains.

Infant Incubator Adverse Event Evaluation Indicators

When analyzing adverse events involving infant incubators, core elements are of paramount importance. Yan and Huifang [34] proposed analyzing both the manifestations and root causes within adverse event data for infant incubators. Kaur et al [35] suggested describing the implemented control measures. Additionally, according to the writing requirements of the guidelines issued by the National Medical Products Administration, adverse event reports typically require information such as location, time, source, injury manifestation, device malfunction, cause analysis, and specific control measures. Accordingly, we summarize the core and secondary elements of infant incubator adverse events. The core elements are the "occurrence date" of the adverse event, the "abnormal condition" of the infant incubator, and the "follow-up measures" for risk control. Secondary elements include the "cause of failure" and the "risk triggered" by the infant incubator. Beyond these core and secondary elements, any additional information captured from the original text is categorized as supplementary notes.

Therefore, during testing experiments, in addition to employing general evaluation metrics, we have developed specialized evaluation indicators for adverse events in infant incubators: the element recall rate and the information density index of adverse events with correction factor (IDIAE). The formulas are as follows:

where K denotes the core feature set; E_k^{ref} represents feature k in the reference text; E_k^{pred} denotes feature k in the predicted text; and I is the indicator function. During computation, text segments E_k^{ref} and E_k^{pred} corresponding to the 3 core elements are extracted from the reference text and prediction text, respectively. Subsequently, a hierarchical matching assessment is performed: an exact match is determined by checking whether the normalized reference text is fully contained within the prediction text. If the exact match fails, the similarity ratio based on edit distance between the 2 texts is calculated. If this ratio exceeds the threshold of 0.8, the segment is deemed a match.

$$\beta = 1 + \alpha \cdot (N_{\text{red}}/10) + \gamma \cdot M_{\text{error}}$$

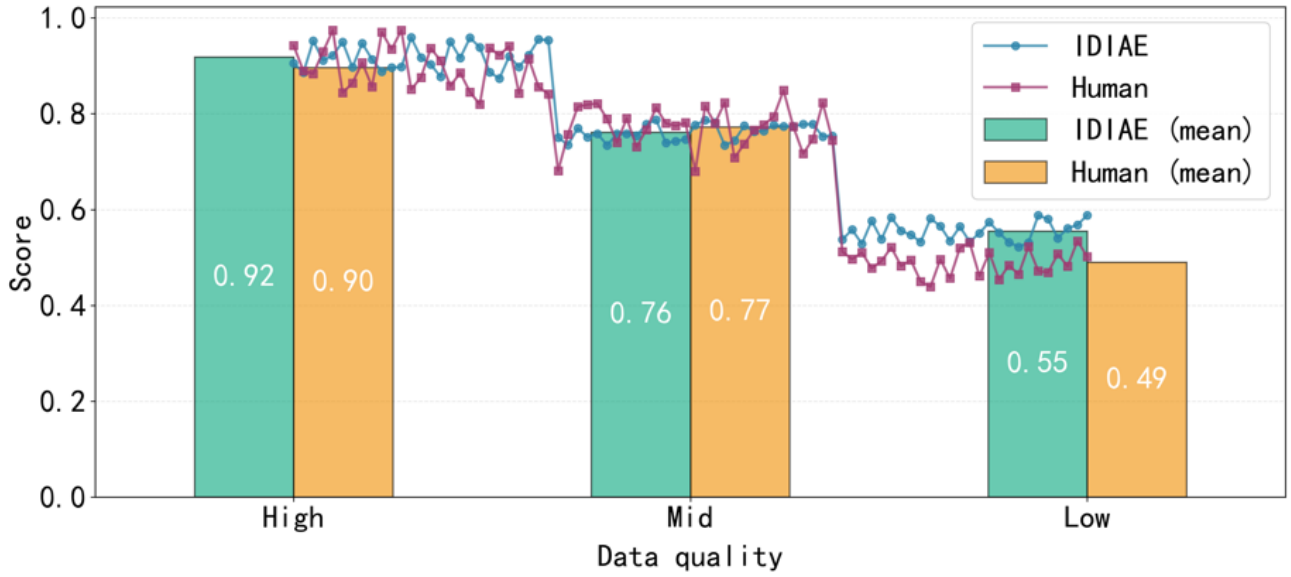
where C_j represents the coverage status of the jth standard report field; m denotes the total number of required fields; w_i is the weight coefficient for the ith category of information (core element weight is 1.0, secondary requirement weight is 0.8); β is the text redundancy coefficient, and the redundancy coefficient incorporates a redundancy word penalty weight α to assess whether the model's response is excessively verbose; N_{red} indicates the number of irrelevant characters; γ is the terminology error penalty weight used to evaluate whether the model captures professional terminology correctly; and M_{error} represents the number of errors.

To determine the globally optimal combination of α and γ and to validate the reliability of the IDIAE metric, this study randomly selected 80 adverse event reports for infant incubators. These were stratified into high-, medium-, and low-quality grades, with 26, 28, and 26 reports, respectively. Experts were invited to rescore the reports under blinded conditions. As shown in Figure 4, varying only the α and γ values revealed that the intraclass correlation coefficient peaked at 0.78 (95% CI 0.72-0.83) when α=.1 and γ=0.2. This indicates a strong positive correlation between IDIAE scores and expert ratings under this weighting configuration (correlation coefficient r=0.82, P<.001), representing the globally optimal combination. To further validate IDIAE reliability, after fixing the weights (α=.1, γ=0.2), analyses across high-, medium-, and low-quality grades showed that the mean IDIAE scores for all 3 groups in Figures 5 and 6 consistently aligned with expert-rated mean trends for both completeness and accuracy. The classification consistency κ coefficient reached 0.71. These results demonstrate that the IDIAE metric effectively reflects expert criteria for assessing report completeness, exhibiting good reliability and clinical applicability.

XSL•FO
RenderX

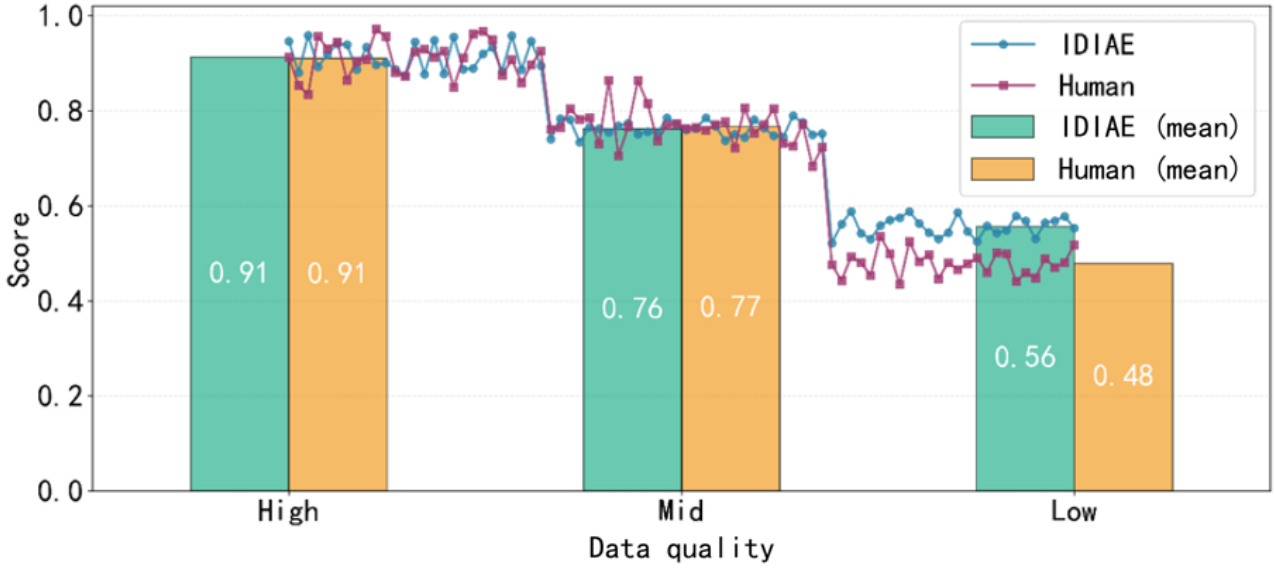
Figure 4. Sensitivity heatmap of Information Density Index of Adverse Events (IDIAE) metric weights (intraclass correlation coefficient [ICC] values). The experiment used 80 randomly selected adverse event reports from infant incubators, with expert blinded scoring as the gold standard. The x- and y-axes represent the redundancy penalty weight (α) and the terminology error penalty weight (γ), respectively. Color blocks indicate the ICC values for each weight combination, enabling identification of the globally optimal parameter setting.

Figure 5. Comparison of Information Density Index of Adverse Events (IDIAE) and manual scoring integrity. The figure compares completeness scores and their mean values between the IDIAE metric and human expert ratings across 3 data groups with varying quality levels.



XSL•FO
RenderX

Figure 6. Accuracy comparison between Information Density Index of Adverse Events (IDIAE) and human scoring. The figure compares accuracy scores and their mean values between the IDIAE metric and human expert ratings across 3 data groups with varying quality levels.



Prompt Engineering

To efficiently align the model with the desired training outcomes, this study designs 3 prompts for different scenarios, as shown in Figures 7-9. Figure 7 is used to assist in generating a suitable fine-tuning dataset. Figure 8 presents the system prompt embedded in the inference stage, aiming to standardize the model output. The retrieval prompt integration employs a structured constraint template, as shown in Figure 9. System instructions enforce that the model must base its responses solely on retrieved content, annotating each sentence with a source identifier [Ref-N] after interpretation. Each retrieved chunk is organized according to a fixed format. The prompt template constrains model behavior through 3 core rules:

- It explicitly prohibits vague references such as “Article X,” mandating explicit linkage to specific regulatory document names and clause numbers to prevent cross-document clause confusion.
- It requires declarations of “insufficient basis” for uncertain content, thereby eliminating model speculation or the fabrication of nonretrieved clauses.
- Regulatory requirement sections must alternate between verbatim citations and interpretations, with each interpretive sentence immediately followed by a [Ref-N] annotation.

Figure 7. Prompt for generating fine-tuning datasets for embedding models.

Prompt	Original text
<|im_start|>system As a regulatory expert at the National Medical Products Administration, please generate {num_questions} natural questions based on the following text: Important Rules: 1.The format "Article XX of [Regulation Name]" can only be used when the original text explicitly mentions the name of the regulation. 2.When the original text does not mention the regulation name, generating questions that include "Article XX" is strictly prohibited. 3.Questions must not be generated using header or footer content. 4.Ensure that the answers to the questions can be found in the original text. Examples (Correct): -"What are the basic requirements for online sales of medical devices?" -"According to Article 20 of the 'Regulations on the Supervision and Administration of Medical Devices,' what are the requirements for production licenses?" Examples (Incorrect): -"What materials are needed according to Article 5?" (Lacks regulation name) -"What are the requirements referenced in Article 18?" (Lacks regulation name) <|im_end|>	Regulations on the Administrative Law Enforcement Responsibility System of Market Supervision and Administration (Promulgated by Order No. 41 of the State Administration for Market Regulation on May 26, 2021, and effective from July 15, 2021) ..... Article 8: Staff of the market supervision and administration department shall exercise their powers within the scope of their legal authority and in accordance with legal procedures, ensuring strict, standardized, fair, and civilized law enforcement. They must not neglect their duties, exceed their powers, or abuse their authority. Article 9: If staff of the market supervision and administration department, due to intentional actions or serious negligence, unlawfully perform their administrative law enforcement duties, resulting in harmful consequences or negative impacts, this constitutes an administrative law enforcement fault, and they shall bear administrative law enforcement responsibilities in accordance with the law.
	Correct answer
Generated question	Article 9: If staff of the market supervision and administration department, due to intentional actions or serious negligence, unlawfully perform their administrative law enforcement duties, resulting in harmful consequences or negative impacts, this constitutes an administrative law enforcement fault, and they shall bear administrative law enforcement responsibilities in accordance with the law.
How should serious negligence by staff of the market supervision and administration department be handled?	

XSL•FO
RenderX

Figure 8. System prompt for the inference phase. The template defines the model's role, output specifications, and core constraints for adverse event analysis and regulatory questions and answers in infant incubators. It governs output format, professional compliance, and content boundaries to ensure responses meet specialized regulatory requirements for medical device adverse events.

Prompt

System Role and Task Instructions
You are a professional intelligent assistant for analyzing adverse events in medical devices. Your core task is to perform structured analysis on **raw text reports of infant incubator adverse events** submitted by users, strictly adhering to the following output format.

Core Analysis Element Definitions
You must extract or summarize the following elements from the report:
1. **Occurrence Date**: The specific date or time of the event explicitly stated in the report.
Key Rules
1. **Strictly Evidence-Based**: All output must derive solely from the input text. **Do not fabricate, infer, or add information not present in the original text**.
2. **Original Text Priority**: For precise details like dates or model numbers, reproduce the original text whenever possible. For descriptive text, provide professional, concise summaries.
3. **Handling of Gaps**: If an element is completely absent from the original text, populate the corresponding JSON field with "Not Mentioned" or an empty string.
2. **Anomalies**: Malfunctions, abnormal phenomena, or alarm messages exhibited by the infant incubator. Summarize using standard medical terminology (e.g., "incubator temperature probe failure," "abnormal fan noise").
3. **Follow-up Actions**: Specific risk control measures taken by operators (e.g., "immediately replaced the incubator," "reported for repair").
4. **Root Cause**: (If inferable) The probable underlying cause of the anomaly.
5. **Potential Risk**: (If inferable) The potential safety risk this anomaly may pose to the infant.
6. **Additional Notes**: Other important details, such as equipment model, personnel involved, or associated medical devices.

XSL•FO
RenderX

Figure 9. Retrieval-augmented generation retrieval prompt template. Designed for a regulatory expert role aligned with the National Medical Products Administration, it defines 4 mandatory rules and standardized output requirements: conclusions must be based solely on retrieved regulatory provisions; each interpretation must include source citations; fabrication of provisions is prohibited; and any uncertainty must be labeled as "insufficient basis." The template takes user queries and retrieved regulatory texts as input and produces standardized regulatory question-and-answer outputs.

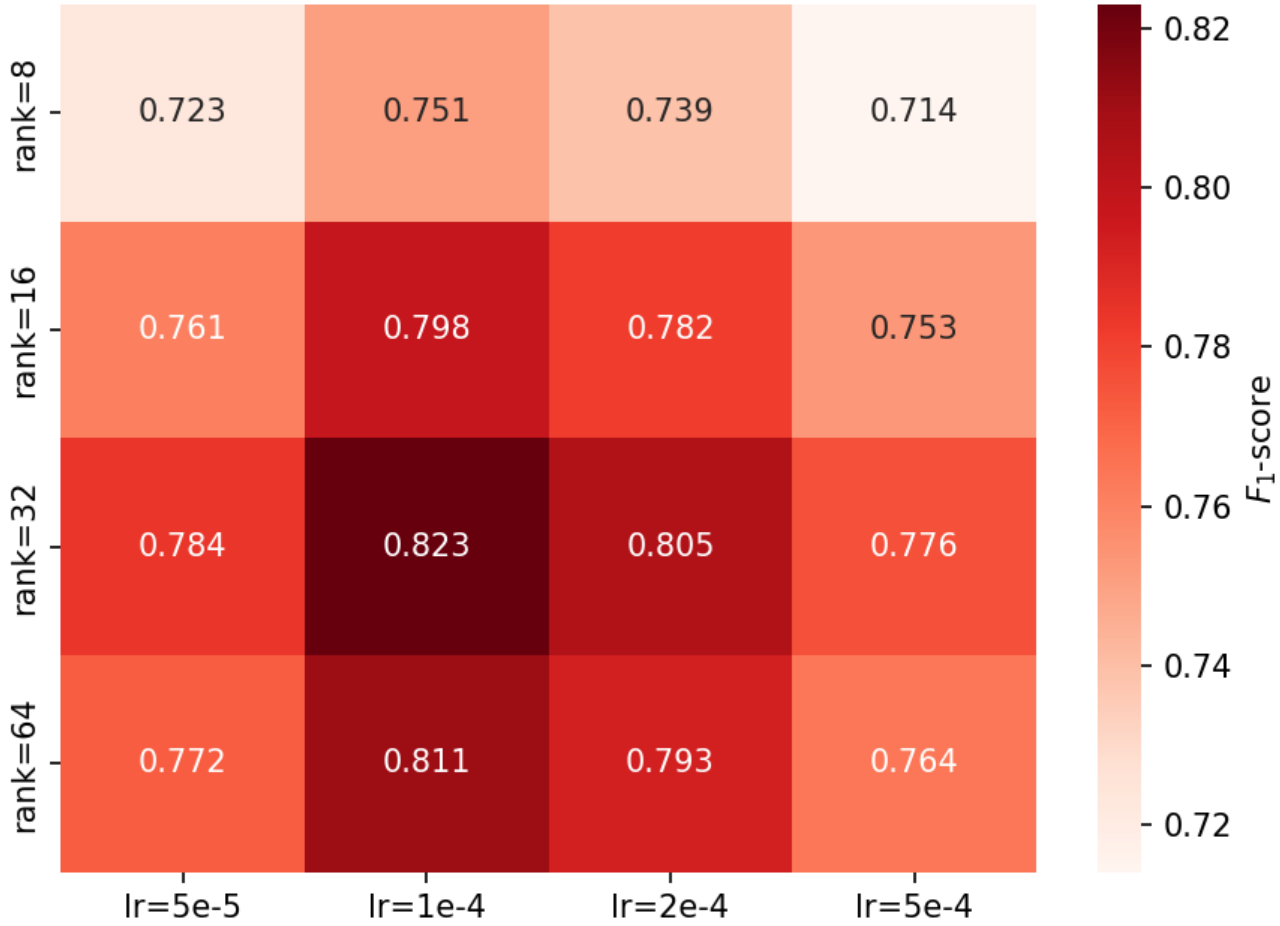
```
Prompt

<|im_start|>system
As a regulatory expert at the National Medical Products Administration, you must answer
user questions **based solely on the retrieved regulatory provisions below**. No
unsupported interpretations are permitted.
Mandatory Rules:
1. Citation Required: Each regulatory reference must be marked with [Ref-N], formatted as:
[Ref-N] "Regulation Title" Article X
2. Prohibition on Inventing Provisions: Use only explicitly stated content from retrieved
regulatory clauses.
3. Prohibition on Ambiguous Language: Avoid non-traceable expressions like "according to
relevant regulations."
4. Uncertainty Disclaimer: When retrieved content cannot fully address the query, clearly
state "insufficient basis."
Output Requirements:
1. Applicable Provisions: [List all referenced Ref numbers]
2. Regulatory Requirements: [Original text + interpretation, each sentence followed by Ref]
3. Operational Recommendations: [Bulleted list, each point supported by a
provision]<|im_end|>
<|im_start|>user
User query: {query}
Retrieved regulatory provisions
{chunks_with_refs}
End of retrieved content <|im_end|>
<|im_start|>assistant
"""
```

Optimal Parameter Configuration

In typical small-sample medical data analysis tasks, r=32 has been validated as an effective LoRA range [36]. Before experimentation, we initially favored selecting r=32. However, to rigorously validate the impact of rank-learning rate combinations on the performance of the infant incubator adverse event question-answering task, we conducted a 4 × 4 grid scan while fixing all other hyperparameters. The resulting F_1-scores are visualized as a heatmap, as shown in Figure 10. The heatmap shows that the F_1-score peaks at 0.823 at rank=32 and lr=1 × 10^{-4}, then decreases toward the surrounding edges, exhibiting an overall bell-shaped "medium-high-medium" distribution. This phenomenon occurs because when the rank is too low, the low-rank matrix lacks expressive power and cannot fully capture the long-range dependencies and medical terminology details in event descriptions; when the rank is too high, the number of parameters increases, which can lead to overfitting on domain data and weaken generalization. Regarding the learning rate, 5 × 10^{-5} converges too slowly, whereas 5 × 10^{-4} has an excessively large step size, making it difficult for the model to stabilize at the optimal weights. By contrast, 1 × 10^{-4} strikes a balance between convergence speed and stability. Therefore, the experiment ultimately adopted the combination of rank=32 and lr=1 × 10^{-4}.

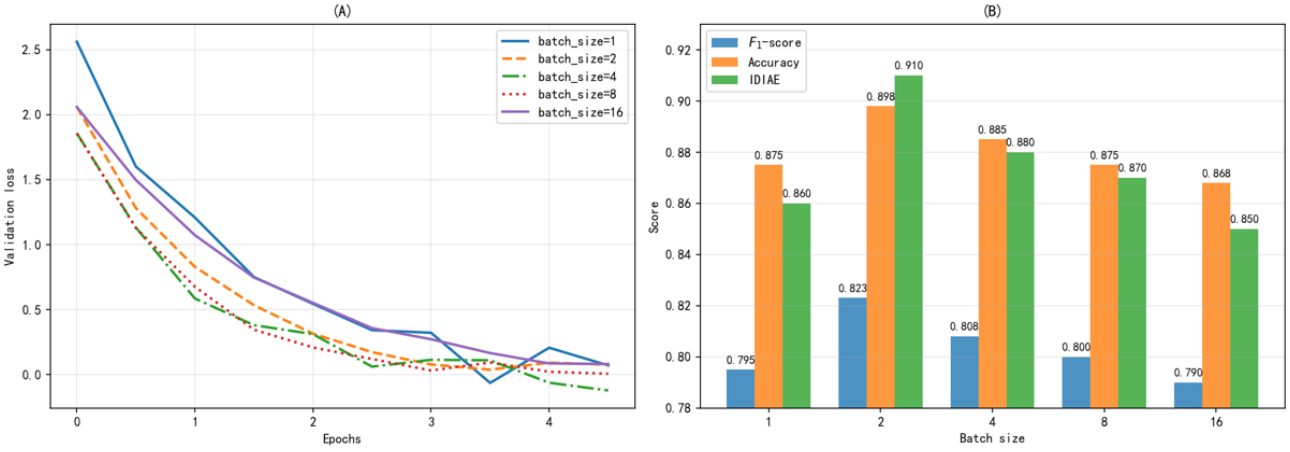
XSL•FO
RenderX

Figure 10. Hyperparameter sensitivity heatmap. The figure shows variation in F1-scores on the test set for the infant incubator adverse event question-answering task as a function of low-rank adaptation rank and learning rate, evaluated via a 4 × 4 grid search.



To investigate the impact of batch size on training stability and performance, we conducted experiments while fixing the other hyperparameters. As shown in Figure 11A, when batch size is 1, the loss curve oscillates notably and converges slowly due to high gradient update noise; a batch size of 2 (the configuration selected for this study) achieves an optimal balance between training stability and convergence speed, with a smoothly declining loss curve and the lowest final convergence value; increasing the batch size to 4 or 8 yields a more stable training process but slightly slower convergence; a batch size of 16 fails to iterate sufficiently due to graphics processing unit memory constraints. Figure 11B further quantifies the final performance metrics across different batch sizes: a batch size of 2 achieves peak values for F_1-score, accuracy, and IDIAE, showing average improvements of 4.0% and 3.4%, compared with batch sizes of 1 and 8, respectively. Therefore, the final experimental configuration adopts a batch size of 2.

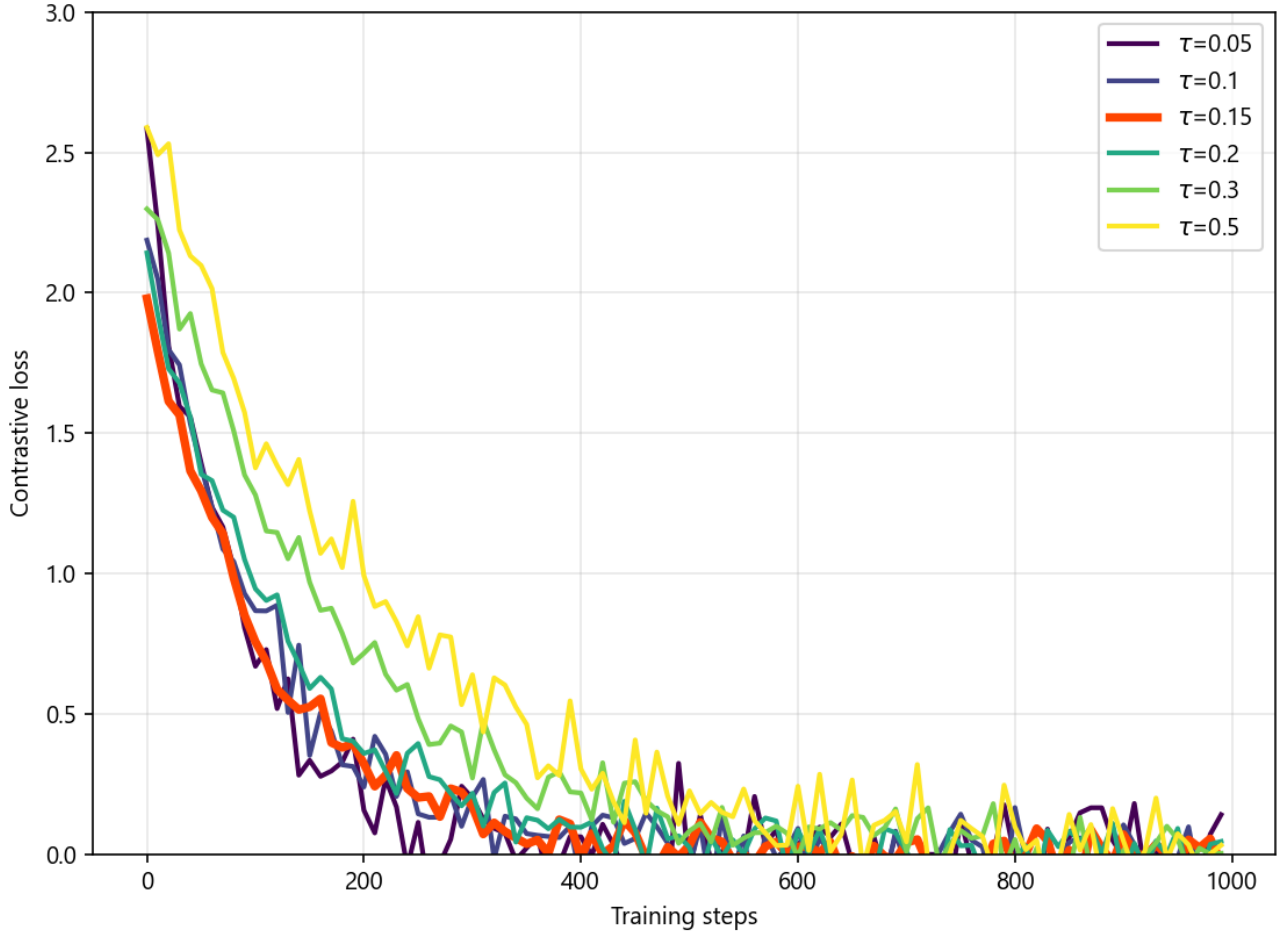
Figure 11. Batch size impact analysis. The figure examines how different training batch sizes affect the AE2IN model's training stability and final performance. (A) Training loss curves across epochs for various batch sizes, with epochs on the x-axis and loss on the y-axis. (B) The final F1-score, accuracy, and Information Density Index of Adverse Events (IDIAE) metric for each batch size, with batch size on the x-axis and metric values on the y-axis.



XSL•FO
RenderX

In supervised contrastive semantic optimization, the temperature coefficient τ controls the sharpness of the similarity distribution. To achieve optimal results, we fixed the other parameters and conducted experiments on the temperature coefficient. As shown in Figure 12, when τ=0.05, the similarity distribution becomes excessively sharp, causing the loss curve to oscillate and converge slowly. At τ=0.15 (the value selected for this study), the loss curve declines smoothly and rapidly to a stable region. When τ increases beyond 0.3, the similarity distribution becomes overly smooth, notably slowing loss convergence. Therefore, τ=0.15 is determined to be the optimal configuration for medical device regulatory retrieval tasks.

Figure 12. Effect of the temperature coefficient (τ) on supervised contrastive learning. The figure shows how different τ values influence convergence of the training loss during supervised contrastive semantic optimization fine-tuning of the embedding model. The x-axis represents training steps, and the y-axis shows the contrastive learning loss.



Ethical Considerations

Our data originate from China's medical device regulatory authorities. The data consist of medical device adverse event reports (similar to the US FDA's Manufacturer and User Facility Device Experience [MAUDE] database). The content primarily includes failure information recorded during device use. These data contain no patient information; therefore, formal ethics approval is not required. This project was jointly completed with regulatory agencies under government-funded research initiatives. All data were authorized, and the research complies with relevant regulations.

Results

Experimental Data

This large-model experimental dataset comprises general-domain data and domain-specific data. The general-domain data originate from the PediaBench dataset publicly released by the State Key Laboratory of Public Big Data at the School of Computer Science and Technology, Guizhou University, China. It contains 1565 question-and-answer entries covering common pediatric diseases such as neonatal disorders, pediatric infectious diseases, pediatric respiratory diseases, and pediatric primary care and developmental behavioral disorders. Domain-specific data originate from private adverse event data on infant incubators within China's Key Monitoring Information System for Medical Device Adverse Events. After manual and GPT-4 screening, data cleaning, and augmentation, combined with prompt engineering, the data were processed into a specialized question-and-answer dataset on infant incubator adverse events, totaling 2530 entries (Table 1).

XSL•FO
RenderX

Table 1. Fine-tuning the Qwen model dataset sources and scale.

Dataset	Data source	Data volume
PediaBench	Public dataset from the State Key Laboratory of Public Big Data, School of Computer Science and Technology, Guizhou University, China	1565
Infant Incubator Adverse Event Questions and Answers	Adverse event data on private infant incubators in China's Key Monitoring Information System for Medical Device Adverse Events	2530

We split the data at the adverse-event report level: all question-and-answer pairs derived from the same report were assigned entirely to either the training, validation, or test set, ensuring that no report appears in more than one set. This experimental dataset allocates 60% (2457/4095) to the training set, 30% (1229/4095) to the test set, and 10% (409/4095) to the validation set.

The knowledge base data originate from regulatory documents issued by the China National Medical Products Administration. We selected administrative regulations and departmental rules to form the regulatory provisions knowledge base, comprising a total of 29 documents. The fine-tuning dataset for the embedding model is derived from these regulatory documents. We utilized documents from this knowledge base and, combined with prompt engineering, leveraged GPT-4 to generate fine-tuning datasets. This dataset comprises 1488 entries, with 60% (893/1488) allocated to the training set, 30% (446/1488) to the test set, and 10% (149/1488) to the validation set (Table 2).

Table 2. Knowledge base and fine-tuning the embedding model dataset.

Dataset	Data source	Data volume
Regulatory Knowledge Base	Regulatory documents issued by the China National Medical Products Administration	29 documents
Embedding Model Fine-Tuning Dataset	Generated using regulatory documents issued by the China National Medical Products Administration	1488

To mitigate the "student-teacher" loop bias introduced by large-model rewriting, we collaborated with the Henan Drug Administration to implement a manual quality control process for all data generated by large models. First, we established a 3-tier annotation standard (perfect equivalence/minor drift/significant drift) and trained annotators ($\kappa \geq 0.78$). Next, data were grouped into batches of 100 items for double-blind annotation by pairs of annotators, with a third party serving as the arbitrator. Only when the "significant drift" rate was controlled at ≤5% were the data released. Finally, senior experts centrally reviewed and refined the "significant drift" data.

Experimental Environment

This work selected Qwen2_7B as the source model and employed a dual-adapter joint fine-tuning approach for training, with specific configurations detailed in Tables 3-5.

Table 3. Environmental configuration.

Category	Configuration information
Central processing unit	Xeon(R) Platinum 8352V
Graphics processing unit	RTX 4090 × 2
Graphics memory	24 GB
Distributed training strategy	Data parallelism
Memory	90 GB
Compute Unified Device Architecture	12.2

Table 4. Hyperparameter configuration.

Hyperparameter	Value
Train epochs	3
Learning rate	1.00×10^{-4}
Batch size	2
Lora_rank	32
Logging_steps	3
Gradient accumulation	4

XSL•FO
RenderX

Table 5. Retrieval-augmented generation parameter configuration.

Parameters	Value
Chunk size	512
Overlap	80
Top-*k*	5
Cosine	0.72
Nlist	128

Base Model Testing

This study selected BGE-Large-ZH-V1.5 as the embedding source model. Here, BGE-Large-ZH-V1.5 was compared with M3E-Large, Text2Vec-Large, and Luotuo on the Massive Text Embedding Benchmark (MTEB) and the Chinese Massive Text Embedding Benchmark (C-MTEB). The embedding dimension was unified to 1024, and the evaluation metrics selected were retrieval, semantic textual similarity, pairwise classification, and classification.

Table 6 compares BGE-Large-ZH-V1.5 with mainstream models, demonstrating that it achieves leading performance on both MTEB and C-MTEB. Its high recall score of 70.46 indicates that it is particularly well-suited as a foundation for RAG systems, demonstrating a superior ability to accurately match specialized terminology within knowledge bases compared with other models.

For large models, this paper compares base-model performance by evaluating Qwen2_7B, Llama3_8B, and GLM4_9B across 8 metrics. The results are shown in Table 7.

Table 6. Embedding base model test results. This table compares the performance of the BGE-Large-ZH-V1.5 embedding model selected for this work against 3 mainstream Chinese embedding models across multiple dimensions on the MTEBa and C-MTEBb standard evaluation datasets.

Model	Embedded dimension	Retrospective retrieval	Semantic textual similarity	Pairwise classification	Classification
BGE-Large-ZH-V1.5	1024	70.46	56.25	81.6	69.13
M3e-Large	1024	54.75	50.42	64.3	68.2
Text2Vec-Large	1024	41.94	44.97	70.86	60.66
luotuo	1024	44.4	42.78	66.62	61

[a]MTEB: Massive Text Embedding Benchmark.

[b]C-MTEB: Chinese Massive Text Embedding Benchmark.

Table 7. Base model test results. This table presents the comparison results of general-purpose models, specialized medical models (HuatuoGPT-7B and DISC-MedLLM), and AE2IN on benchmark metric tests. This evaluation verifies AE2IN's ability to adapt to specialized domains while preserving foundational general capabilities and assesses its mitigation of catastrophic forgetting.

Indicator/model	AlignBench	MT-Bench	MMLU[a]	GSM8K	MATH	HumanEval	C-Eval	PediaBench
Qwen2_7B	7.21	8.41	70.5	82.3	49.6	79.9	77.2	79.7
Llama3_8B	6.20	8.05	68.4	79.6	30.0	62.2	45.9	59.4
GLM4_9B	7.01	8.35	72.4	79.6	50.6	71.8	75.6	71.5
HuatuoGPT-7B	6.09	8.01	65.1	77.5	28.3	59.3	80.7	88.7
DISC-MedLLM	7.30	8.22	67.3	76.1	45.4	62.1	82.4	90.6
AE2IN	7.32	8.44	67.8	79.2	46.3	64.8	75.1	80.1
Only fine-tuning for specific domains	7.08	8.21	62.5	75.3	39.8	59.4	71.8	73.5

[a]MMLU: Massive Multitask Language Understanding.

Table 7 shows that the Qwen2 model has stronger overall performance than the other 2 models. Qwen2 performs well on the Massive Multitask Language Understanding (MMLU) medical subset and C-Eval medical questions, indicating that the model is stable in diagnostic reasoning and clinical terminology comprehension, but exhibits slight errors in complex calculation questions. Meanwhile, the fine-tuned AE2IN model shows score decreases ranging from 2.1 to 2.7 on the MMLU medical subset and C-Eval medical tasks. Because of the incorporation of PediaBench data during fine-tuning, its scores slightly surpass those of the base model by 0.4 points. By contrast, models fine-tuned solely on domain-specific datasets experienced score declines ranging from 5.4 to 8.0 on the MMLU medical subset and C-Eval medical tasks, along with a 6.2-point drop on PediaBench. These

XSL•FO
RenderX

comparative results validate that AE2IN effectively mitigates catastrophic forgetting.

Comparative Experiment

We selected 5 models for comparative testing with the fine-tuned model. The results are shown in Table 8.

Table 8. Comparative evaluation of the fine-tuned model AE2IN against general models and specialized medical models (HuatuoGPT, DISC-MedLLM) in the infant incubator adverse event domain. This table compares the overall performance of the base large model, the base model combined with the same RAGa framework model, mainstream medical-specific large models, and the AE2IN model proposed in this study on the infant incubator adverse event analysis task. All experiments were replicated 3 times independently. Results are presented as mean (SD), supplemented with 2-tailed (unpaired and independent) samples t test *P* values and statistical significance annotations for each model versus AE2IN.

Model	BLEU[b]-4	*P* value	ROUGE-L[c]	*P* value	Accuracy	*P* value	Element recall rate	*P* value	Average IDIAE[d]	*P* value
Qwen2_7B	0.137 (0.031)	<.001	0.208 (0.025)	<.001	0.785 (0.028)	<.001	0.691 (0.023)	<.001	0.61 (0.018)	<.001
Qwen2_7B+RAG	0.162 (0.025)	.003	0.315 (0.022)	.001	0.800 (0.024)	.002	0.715 (0.019)	.002	0.75 (0.015)	.002
Llama3_8B	0.108 (0.026)	<.001	0.237 (0.033)	<.001	0.722 (0.030)	<.001	0.651 (0.028)	<.001	0.49 (0.025)	<.001
Llama3_8B+RAG	0.131 (0.022)	.003	0.302 (0.028)	.002	0.756 (0.026)	.003	0.688 (0.024)	.002	0.62 (0.019)	.003
GLM4_9B	0.105 (0.025)	<.001	0.282 (0.033)	<.001	0.763 (0.031)	<.001	0.636 (0.025)	<.001	0.55 (0.020)	<.001
GLM4_9B+RAG	0.158 (0.021)	.002	0.356 (0.025)	.001	0.801 (0.023)	.002	0.702 (0.021)	.001	0.71 (0.017)	.002
HuatuoGPT-7B	0.155 (0.021)	.03	0.381 (0.022)	.03	0.803 (0.025)	.03	0.722 (0.031)	.03	0.80 (0.022)	.03
DISC-MedLLM	0.174 (0.020)	.02	0.411 (0.024)	.02	0.816 (0.022)	.02	0.740 (0.020)	.02	0.82 (0.019)	.02
AE2IN	0.215 (0.018)	N/A[e]	0.484 (0.021)	N/A	0.898 (0.022)	N/A	0.815 (0.016)	N/A	0.91 (0.013)	N/A

[a]RAG: retrieval-augmented generation.

[b]BLEU: Bilingual Evaluation Understudy.

[c]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence.

[d]IDIAE: Information Density Index of Adverse Events.

[e]N/A: not applicable.

Through systematic quantitative evaluation, the AE2IN model demonstrates notable advantages across all key performance indicators in the domain of infant incubator adverse events. In terms of text generation quality, AE2IN attained a BLEU-4 score of 0.215, reflecting a 56.9% improvement relative to the baseline model and a 32.7% improvement over the baseline equipped with the same RAG framework. Its Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (ROUGE-L) score reached 0.484, corresponding to a 53.7% increase. For core element analysis of infant incubator adverse events, the element recall rate reached 0.815, representing a 17.9% improvement over the baseline model and a 14.0% improvement over the baseline integrated with the same RAG framework. In the experiments, the dual-adapter architecture exhibited strong causal reasoning capacity when processing complex alarm system failure events.

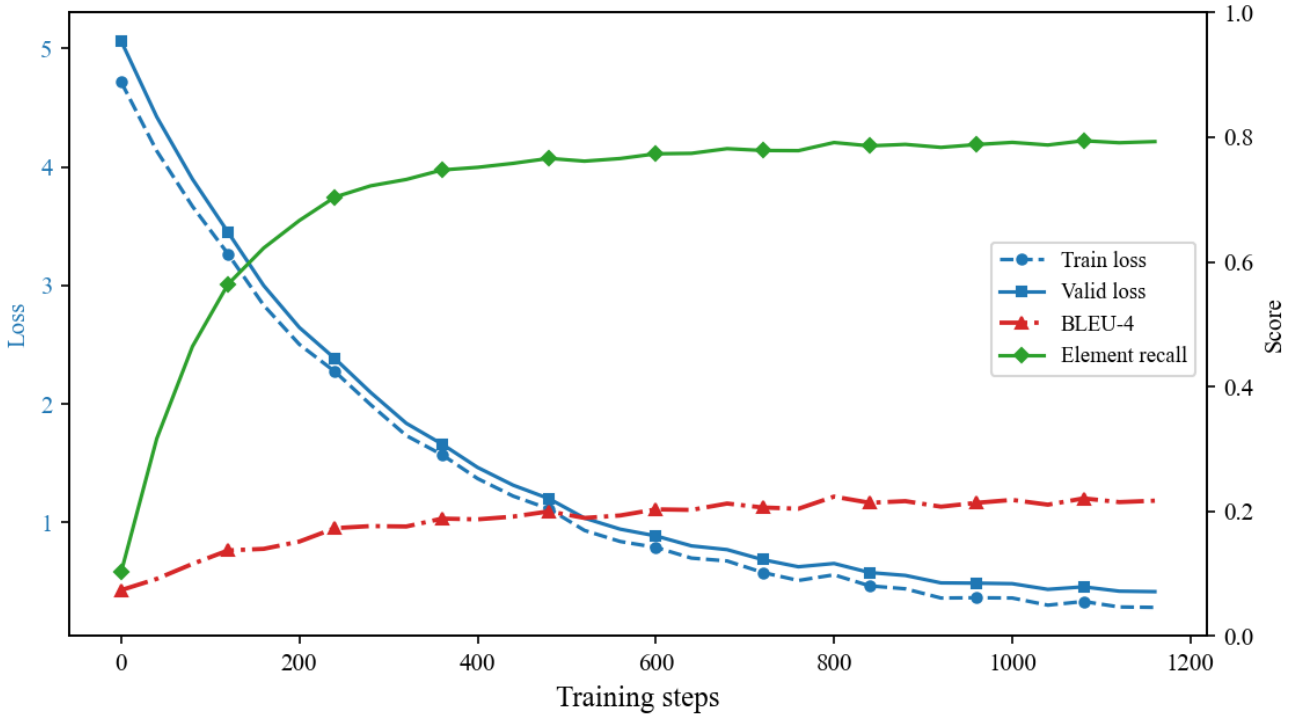
To conduct a more comprehensive analysis of the AE2IN model, we also selected cases in which accuracy, element recall, and IDIAE were all below 0.5. We found that misjudgment of core elements directly affected completeness conclusions and notably impacted both accuracy and IDIAE values. In failed cases, we observed that for rare fault descriptions—such as the onomatopoeic “pushing clack-clack sound”—low frequency and susceptibility to semantic integrity disruption caused the inferred Query-Key similarity to be drowned out by the attention distribution of primary terms after softmax. This prevented the activation of corresponding neurons during value vector weighting. Furthermore, the scaling coefficient update for $(IA)^3$ relies on batch statistics, leading to inconsistent weighting across cases. When encountering such descriptions, the update process may become stuck at a saddle point, failing to amplify this feature.

We continuously record training and validation loss, along with key evaluation metrics, to determine in real time whether the model is in a state of normal convergence, overfitting, or underfitting. As shown in Figure 13, validation loss and training loss decrease smoothly and synchronously over the first 600 steps, maintaining a small difference of less than 0.15, indicating that the learning rate of 1×10^{-4} effectively balances model capacity and regularization. BLEU-4 increases monotonically from 0.07 to 0.22, only 0.005 away from the final test set score of 0.215, confirming that the validation and test sets are identically distributed and that the model has fully converged.

XSL•FO
RenderX

Feature recall also increases from 0.60 to 0.815 over the first 600 steps, rising in parallel with BLEU-4 without intersecting, indicating that the model improves fluency without sacrificing its ability to extract key information. All 3 metrics plateau simultaneously, demonstrating stable convergence without the need for early stopping or secondary parameter tuning. This also indicates that the final test score is not a random peak but the result of stable convergence.

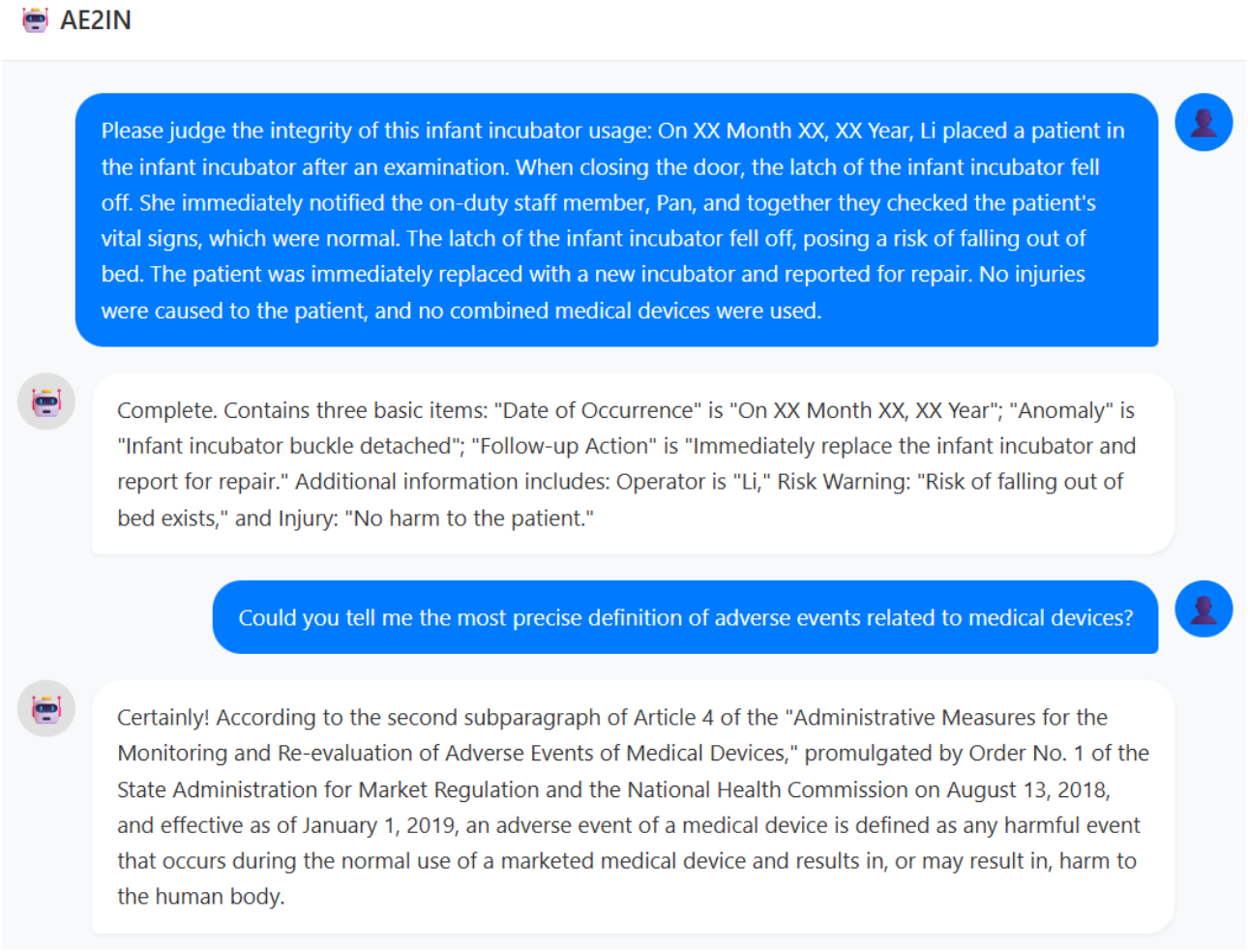
Figure 13. Convergence of dual-adapter fine-tuning (AE2IN). The figure shows the joint evolution of training loss, validation loss, Bilingual Evaluation Understudy-4 score, and ERR metric over training steps. Core hyperparameters were fixed (low-rank adaptation rank=32, learning rate=1 × 10–4, batch size=2). The x-axis represents training steps; the left y-axis shows loss values, and the right y-axis shows text generation and key element extraction scores.



As shown in Figure 14, when determining the completeness of adverse events related to infant incubators, AE2IN directly extracts core elements—incident date, nature of the abnormality, and subsequent corrective actions—while supplementing them with details such as operator information, risk warnings, and injury status. This approach concisely highlights essential information.

XSL•FO
RenderX

Figure 14. AE2IN response examples. The figure presents sample outputs from AE2IN, including assessments of adverse event report completeness for infant incubators and responses to medical device adverse event definitions.



However, compared with other models, AE2IN's responses tend to be more concise. To investigate whether this brevity affects safety, a supplementary risk recall assessment experiment was conducted to quantify the quality of model outputs from a safety perspective. First, a core safety item list was established through expert consultation and alignment with the "Measures for the Administration of Medical Device Adverse Event Monitoring and Re-evaluation" and the national standard "GB 9706.219-2021." This list encompasses adverse event phenomenon identification, safety risk identification, immediate patient safety measures, equipment disposal measures, adverse event reporting measures, and risk cause localization. Next, 100 high-quality data entries, each containing core safety items, were extracted and annotated, and the process was repeated 5 times. The results of the risk recall experiment are shown in Table 9.

Table 9. Risk recall rate comparison. This table displays the risk recall rates and the average number of core safety items identified by AE2IN across various models.

Model	Risk recall, mean (SD)	Average number of core safety items identified
AE2IN	0.92 (0.012)	5.52
Qwen2_7B	0.77 (0.024)	4.62
Llama3_8B	0.75 (0.022)	4.50
GLM4_9B	0.80 (0.019)	4.80
HuatuoGPT-7B	0.82 (0.016)	4.92
DISC-MedLLM	0.84 (0.017)	5.04

AE2IN demonstrates the highest risk recall rate, achieving an improvement of approximately 15% over other models. Although some data may be underreported due to rare fault descriptions, its overall risk recall rate confirms that AE2IN effectively maximizes risk extraction. By outputting structured information that highlights core details, it helps ensure safety.

We used the AE2IN model with the same RAG framework and different embedding models for regulatory question answering.

XSL•FO
RenderX

The experiments used a regulatory clause validation set to compare the performance of the fine-tuned embedding models. The comparative results are shown in Table 10.

Table 10. Comparative evaluation of regulatory question-answering performance between the AE2IN model with different embedding models and general-purpose models.

Configuration/indicators	AE2IN					General model		
	FINBGE[a] (ours)	BGE-Large-ZH-V1.5	M3e-Large	Text2Vec-Large	Luotuo	Qwen2_7B	Llama3_8B	GLM4_9B
Accuracy, mean (SD)	0.938 (0.011)	0.863 (0.015)	0.804 (0.015)	0.780 (0.016)	0.761 (0.017)	0.715 (0.038)	0.652 (0.040)	0.727 0.042)
Recall, mean (SD)	0.894 (0.013)	0.807 (0.016)	0.763 (0.017)	0.759 (0.017)	0.685 (0.019)	0.652 (0.042)	0.594 (0.044)	0.671 (0.050)
F_1-score, mean (SD)	0.915 (0.013)	0.834 (0.016)	0.783 (0.016)	0.769 (0.017)	0.721 (0.018)	0.682 (0.041)	0.622 (0.038)	0.698 (0.04)

[a]FINBGE is a BGE model fine-tuned using supervised contrastive learning.

The experimental results in Table 11 show that the AE2IN model, when equipped with different embedding models, exhibits a clear performance gradient in the regulatory question-answering task. The FINBGE model leads by a notable margin, achieving the highest levels across the 3 core metrics—accuracy, recall, and F_1-score—compared with the source model BGE-Large-ZH-V1.5. This result validates the effectiveness of the supervised contrastive semantic optimization fine-tuning strategy, which successfully narrows the vector distance between the question and the correct answer. Furthermore, comparison with common models reveals that Llama3_8B achieves the lowest scores, likely because the test dataset consists entirely of Chinese regulatory question-answer pairs. The best-performing general model, GLM4_9B, achieves scores of 0.727, 0.671, and 0.698 in accuracy, recall, and F_1-score, respectively, but still lags behind AE2IN. This demonstrates the necessity of domain adaptation and the RAG architecture, which can mitigate model hallucinations, address knowledge timeliness, and improve accuracy.

Table 11. AE2IN model performance ablation study.

Indicators/configuration	BLEU[a]-4, mean (SD)	ROUGE-L[b], mean (SD)	Accuracy, mean (SD)	Element recall rate, mean (SD)	IDIAE[c], mean (SD)
AE2IN	0.215 (0.018)	0.484 (0.021)	0.898 (0.022)	0.815 (0.016)	0.91 (0.01)
Only LoRA[d]	0.187 (0.015)	0.412 (0.019)	0.824 (0.020)	0.763 (0.015)	0.83 (0.01)
Only (IA)[3e]	0.176 (0.015)	0.398 (0.017)	0.795 (0.018)	0.741 (0.014)	0.79 (0.01)
Full fine-tuning	0.168 (0.022)	0.388 (0.024)	0.802 (0.028)	0.735 (0.021)	0.78 (0.02)
None	0.137 (0.031)	0.208 (0.025)	0.785 (0.028)	0.691 (0.023)	0.61 (0.02)

[a]BLEU: Bilingual Evaluation Understudy.

[b]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence.

[c]IDIAE: Information Density Index of Adverse Events.

[d]LoRA: low-rank adaptation.

[e](IA)^3: Infused Adapter by Inhibiting and Amplifying Inner Activations.

Similarly, to conduct a more comprehensive analysis of the system model, we selected cases in which regulatory question-answering accuracy was below 0.5 for further study. We found that hallucinations still occurred—for instance, when the model should have referenced article 23 of the "Administrative Measures for the Monitoring and Re-evaluation of Adverse Events of Medical Devices," it instead cited article 24. This occurs because clauses with high similarity but different article numbers exhibit only minor differences in cosine similarity, leading to the retrieval of incorrect clauses and subsequently triggering hallucinations.

Ablation Experiments

To validate the contributions of each component in the dual-adapter joint fine-tuning architecture, this study designed a systematic ablation experiment, maintaining the same testing environment and hyperparameter settings as previously described. The experiment employed a controlled variable approach, sequentially removing key components and observing changes in model performance.

Table 11 shows that the complete model improves feature recall by an average of 9.25% compared with the single-adapter solution, validating the complementarity of low-rank updates and feature scaling in equation (1).

XSL•FO
RenderX

When the LoRA component is removed, BLEU-4 and ROUGE-L decrease notably. This is mainly because LoRA optimizes the long-range dependency modeling capability of the Transformer layer through low-rank decomposition. Its absence makes it difficult for the model to effectively capture the logic in infant incubator failure events. At the same time, the feature recall rate decreases by 10.0%, further validating LoRA's key role in structured information extraction—its trainable parameters are concentrated in the ΔW matrix of the attention layer, enabling more accurate localization of core elements in fault descriptions.

When the $(IA)^3$ component is removed, the IDIAE index decreases notably, due to the weakening of the dynamic feature-scaling function achieved by $(IA)^3$ through the gating mechanism. Specifically, the key-value intervention vector in $(IA)^3$ suppresses the activation of irrelevant features while enhancing the response strength of neurons associated with medical terms.

Following full fine-tuning, metrics exhibited varying degrees of decline compared with AE2IN or single-method fine-tuning. This phenomenon stems from the inevitable overfitting of the 7-billion-parameter model on finite, small samples, validating the necessity of structured parameter fine-tuning for efficiency in small-sample scenarios.

Observation of Table 12 data reveals that, in long-text scenarios, the recall rate of the LoRA configuration alone exhibits a lower decay rate than $(IA)^3$, demonstrating that LoRA's low-rank updates effectively capture long-range semantic associations across clauses. Figure 15 further shows that the attention weights in the LoRA branch decay by 16% for long-distance word pairs, representing a 14-percentage-point reduction compared with the baseline. This indicates that LoRA effectively enhances the interaction strength of distant semantic units. $(IA)^3$ can boost weights over short distances, but long-range dependencies are easily lost, leading to reduced recall rates. This also explains why, under dual-adapter collaboration, the final performance in the infant incubator adverse-event domain improved by approximately 14%, further demonstrating that the 2 adapters can complement each other across different distance scales.

Table 12. AE2IN element recall rate ablation experiment results at different input lengths. This table compares the element recall rate of different model architectures across short and long text adverse event reports for infant incubators.

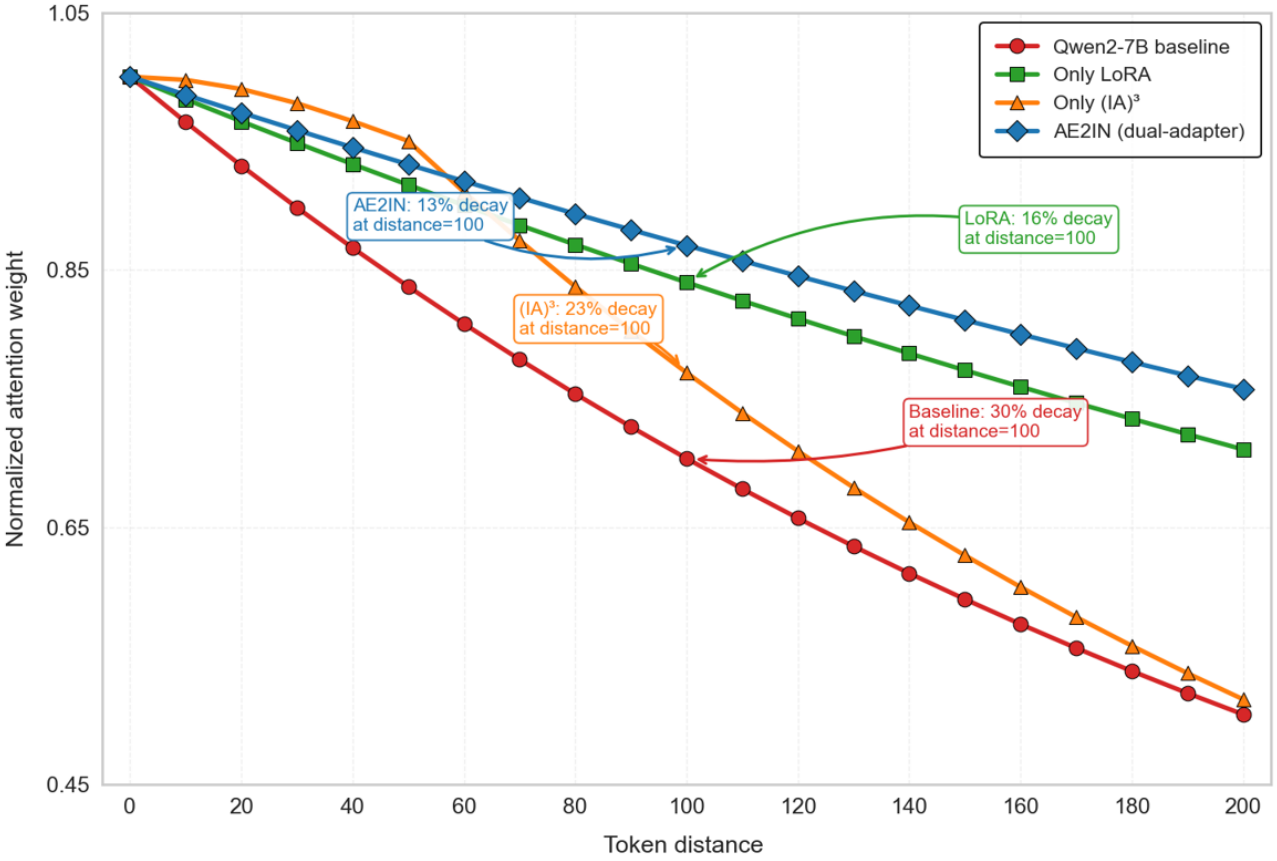
Configuration	Short text (≤64 tokens), mean (SD)	Long text (>128 tokens), mean (SD)	Attenuation level (%)
AE2IN	0.847 (0.012)	0.810 (0.019)	4.4 (1.1)
Only LoRA[a]	0.834 (0.015)	0.766 (0.024)	8.2 (1.8)
Only $(IA)^3$[b]	0.829 (0.018)	0.739 (0.029)	10.9 (2.2)
Full fine-tuning	0.781 (0.021)	0.704 (0.032)	9.9 (2.6)
None	0.733 (0.028)	0.661 (0.035)	9.7 (3.1)

[a]LoRA: low-rank adaptation.

[b]$(IA)^3$: Infused Adapter by Inhibiting and Amplifying Inner Activations.

XSL•FO
RenderX

Figure 15. Dual-adapter semantic modeling capability. The figure demonstrates the model's ability to capture semantics using a long-text test dataset of infant incubator adverse events. It shows the decay of attention weights with increasing token distance: the x-axis represents token distance, and the y-axis shows normalized attention weights, comparing long- and short-range semantic capture across four architectures. (IA)3: Infused Adapter by Inhibiting and Amplifying Inner Activations; LoRA: low-rank adaptation.



Discussion

Principal Findings

This study proposes AE2IN, a large-model framework that integrates RAG with dual-adapter joint fine-tuning for the intelligent analysis of adverse events in infant incubators. Experimental results demonstrate that this framework significantly outperforms baseline models across key metrics, including text generation quality, element recall rate, and regulatory question-and-answer accuracy. Under the same experimental conditions, the element recall rate reaches 0.815, the accuracy of infant incubator adverse event analysis is 0.898, and the accuracy in regulatory clause question-answering tasks attains 0.938. The dual-adapter architecture, that is, LoRA + $(IA)^3$, effectively infuses domain knowledge while preserving model generalization, thereby mitigating catastrophic forgetting. The RAG module, combined with the FINBGE embedding model optimized through supervised contrastive semantic learning, significantly enhances knowledge retrieval accuracy and reduces hallucination.

Comparison With Prior Work

Compared with traditional deep learning models, AE2IN demonstrates superior contextual modeling and causal reasoning capabilities, making it particularly well-suited for complex tasks involving multiple intertwined factors and specialized terminology, such as adverse events in infant incubators. Compared with single fine-tuning approaches, the dual-adapter architecture proposed in this paper achieves a better balance between parameter efficiency and task adaptability. Compared with existing RAG medical applications, AE2IN further introduces a domain-specific embedding model fine-tuning mechanism, enhancing the matching accuracy of Chinese regulatory clauses. When benchmarked against general models such as Qwen and Llama, AE2IN demonstrates higher professional consistency and stronger element-extraction capabilities in specialized tasks, particularly excelling in structured output and terminology standardization.

Limitations

First, the dataset size is limited. The infant incubator adverse-event dataset used in this study comprises 2530 entries. Despite undergoing joint cleaning and enhancement by human experts and GPT-4, the dataset remains insufficient for rare fault types or complex scenarios. To mitigate this issue, we introduced PediaBench general pediatric data for mixed training, employing a 3:1 domain-to-general data sampling ratio to enhance generalization. Future efforts could expand multicenter, multiyear data through collaborations with additional medical institutions to further improve model robustness. Second, the regulatory knowledge base updates lag. The regulatory knowledge base used in this study cannot guarantee real-time coverage of the latest policy revisions. We partially mitigated this knowledge lag by improving the embedding model's semantic understanding of clauses through supervised

XSL•FO
RenderX

contrastive fine-tuning. Future work should consider establishing a dynamic regulatory update mechanism, such as periodically crawling the National Medical Products Administration website or integrating official application programming interfaces to enable automatic knowledge base synchronization. Third, prospective clinical validation was not conducted. This retrospective study did not verify the model's usability and safety within actual clinical workflows. We collaborated with the Henan Drug Administration to implement manual quality control, maintaining a "significant drift" rate ≤5%. However, there remains a lack of assessment regarding clinicians' adoption of model recommendations and the impact of false positives. Future prospective studies should be conducted, inviting hospital equipment departments or adverse-event monitoring personnel to participate in pilot programs and collecting user feedback for iterative optimization.

Future Directions

Current research is limited to adverse events involving infant incubators. Future studies may further explore data-level approaches, such as constructing larger-scale, multicenter, multilingual adverse-event databases covering more high-risk medical device types to enhance model generalizability. Then, at the fine-tuning method level, future work could draw inspiration from distributed entropy regularization [37] to further optimize the learning objective of scaling factors, thereby improving sensitivity to uncommon failure patterns in the training data. At the model level, multimodal fusion could be explored by incorporating visual encoders to process device alarm screenshots or failure photographs, enabling more comprehensive fault identification. Finally, at the system level, an integrated regulatory support system should be developed. This system would integrate the AE2IN model, regulatory retrieval, report generation, and manual review workflows to support multiparty collaboration among hospital equipment departments, regulatory agencies, and manufacturers, thereby advancing the implementation of intelligent regulation.

Conclusions

The safety management of medical devices encompasses all stages of the product life cycle, including research and development, design, manufacturing, clinical application, maintenance, and product recalls, all of which require strict quality control measures. By establishing standardized protocols, refining risk-prevention mechanisms, conducting continuous monitoring, and enhancing professional training, the failure rate of medical devices can be significantly reduced, thereby effectively safeguarding the safety interests of both medical professionals and patients. To advance the management of adverse events related to medical devices and comprehensively support professionals in this field—such as surveillance personnel—this study proposes a large-model solution that integrates RAG technology with dual-adapter fine-tuning. This approach provides an efficient and reliable technical pathway for analyzing adverse events involving infant incubators.

Acknowledgments

We thank the Center for Drug Reevaluation of Henan for providing adverse-event data on infant incubators for this research. Generative artificial intelligence was not used for writing or editing.

Data Availability

These data contain confidential information. To access them, authorization must be requested from the corresponding author (PZ).

Funding

This study was supported by the National Natural Science Foundation of China (grant 12302417), the Shanghai Pujiang Program (grant 23PJ1409200), the Professional Degree Postgraduate Practice Base Project at the University of Shanghai for Science and Technology (grant 2025-009), the Henan Province Artificial Intelligence Medical Devices Supervision Strategy Research Project (grant 262400410407), and the Chenguang Program of the Shanghai Education Development Foundation and Shanghai Municipal Education Commission (grant 24CGA51).

Authors' Contributions

Conceptualization: PZ, WX
Funding acquisition: PZ, WL
Investigation: LW, WK, TL, WZ
Methodology: PZ, WX
Resources: PZ, LW
Supervision: PZ, LW, WL
Writing – original draft: WX
Writing – review & editing: PZ

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Example of raw data for adverse events in infant incubators.
[XLSX File (Microsoft Excel File), 12 KB - medinform_v14i1e83745_app1.xlsx]

References

1. See C, Mooghali M, Dhruva SS, Ross JS, Krumholz HM, Kadakia KT. Class I recalls of cardiovascular devices between 2013 and 2022: a cross-sectional analysis. Ann Intern Med 2024 Nov;177(11):1499-1508. [doi: 10.7326/ANNALS-24-00724] [Medline: 39284187]
2. Sun J, Pan J, Jin Y, Zhang Q, Lv Y, Feng J. Establishment of a medical device adverse event management system for hospitals. BMC Health Serv Res 2022 Nov 24;22(1):1406 [FREE Full text] [doi: 10.1186/s12913-022-08830-5] [Medline: 36419040]
3. Badnjević A, Pokvić LG, Deumić A, Bećirović LS. Post-market surveillance of medical devices: a review. THC 2022 Nov 12;30(6):1315-1329. [doi: 10.3233/thc-220284]
4. Simunovic A, Kranjcec K, Pekas M, Tomic S. Quality analysis of manufacturer's incident reports regarding medical devices. J Pharm Policy Pract 2023 Jan 16;16(1):3 [FREE Full text] [doi: 10.1186/s40545-023-00519-2] [Medline: 36647170]
5. Samore M, Evans RS, Lassen A, Gould P, Lloyd J, Gardner RM, et al. Surveillance of medical device-related hazards and adverse events in hospitalized patients. JAMA 2004 Jan 21;291(3):325-334. [doi: 10.1001/jama.291.3.325] [Medline: 14734595]
6. Ruseva V, Dobrev S, Getova-Kolarova V, Peneva A, Getov I, Dimitrova M, et al. In situ development of an artificial intelligence (AI) model for early detection of adverse drug reactions (ADRs) to ensure drug safety. PHAR 2025 Jul 30;72:1-8. [doi: 10.3897/pharmacia.72.e160997]
7. Hauben M. Artificial intelligence and data mining for the pharmacovigilance of drug-drug interactions. Clin Ther 2023 Feb;45(2):117-133. [doi: 10.1016/j.clinthera.2023.01.002] [Medline: 36732152]
8. Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: a case study of detecting total hip replacement dislocation. Comput Biol Med 2021 Feb;129:104140. [doi: 10.1016/j.compbiomed.2020.104140] [Medline: 33278631]
9. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: The 2015 Conference on Empirical Methods in Natural Language Processing; September 17–21, 2015; Lisbon, Portugal p. 1422-1432. [doi: 10.18653/v1/D15-1167]
10. Pearl J. The seven tools of causal inference, with reflections on machine learning. Commun ACM 2019 Feb 21;62(3):54-60. [doi: 10.1145/3241036]
11. Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. New York, NY: Association for Computational Linguistics; 2012 Presented at: 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012); July 8-14, 2012; Jeju Island, Korea p. 90-94.
12. Lu H, Ehwerhemuepha L, Rakovski C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Med Res Methodol 2022 Jul 02;22(1):181 [FREE Full text] [doi: 10.1186/s12874-022-01665-y] [Medline: 35780100]
13. Naik A, Lehman JF, Rose C. Adapting event extractors to medical data: bridging the covariate shift. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. New York, NY: Association for Computational Linguistics; 2021 Presented at: 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021); April 16, 2021; Online p. 2963-2975. [doi: 10.18653/v1/2021.eacl-main.258]
14. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Red Hook, NY: Curran Associates Inc; 2020 Presented at: NIPS'20: 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Vancouver BC Canada p. 1877-1901.
15. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 2020;21(140):1-67.
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug 12;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
17. Liu Z, Liu P, Sun Y, Nie Z, Zhang X, Zhang Y, et al. DIA-BERT: pre-trained end-to-end transformer models for enhanced DIA proteomics data analysis. Nat Commun 2025 Apr 14;16(1):3530 [FREE Full text] [doi: 10.1038/s41467-025-58866-4] [Medline: 40229248]
18. Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations. 2022 Presented at: International Conference on Learning Representations 2022 (ICLR 2022); April 25-29, 2022; Online p. 3.
19. Wang C, Qian P, Wang Z, Cai W, Yao J, Jiang Y, et al. Multicenter knowledge transfer calibration with rapid zeroth-order TSK fuzzy system for small sample epileptic EEG signals. IEEE Trans Fuzzy Syst 2024 Nov;32(11):6224-6236. [doi: 10.1109/tfuzz.2024.3443878]

XSL•FO
RenderX

20. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM 2021 Nov 19;64(12):86-92. [doi: 10.1145/3458723]
21. Xie Q, Dai Z, Hovy E, Luong T, Le Q. Unsupervised data augmentation for consistency training. 2020 Presented at: Neural Information Processing Systems 2020 (NeurIPS 2020); December 6-12, 2020; Virtual p. 6256-6268.
22. Weinan E. A mathematical perspective of machine learning. In: Proceedings of the International Congress of Mathematicians. Berlin, Germany: The International Mathematical Union; 2022 Presented at: International Congress of Mathematicians 2022 (ICM 2022); July 6-14, 2022; Virtual p. 914-954. [doi: 10.4171/ICM2022/155]
23. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020 Presented at: Neural Information Processing Systems 2020 (NeurIPS 2020); December 6-12, 2020; Virtual p. 9459-9474.
24. Kim Y, Kim JH, Kim YM, Song S, Joo HJ. Predicting medical specialty from text based on a domain-specific pre-trained BERT. Int J Med Inform 2023 Feb;170:104956 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104956] [Medline: 36512987]
25. Zhang X, Zhao G, Ren Y, Wang W, Cai W, Zhao Y, et al. Data augmented large language models for medical record generation. Appl Intell 2024 Dec 06;55(2):88. [doi: 10.1007/s10489-024-05934-9]
26. Tang D, Xi X, Li Y, Hu M. Regulatory approaches towards AI medical devices: a comparative study of the United States, the European Union and China. Health Policy 2025 Mar;153:105260. [doi: 10.1016/j.healthpol.2025.105260] [Medline: 39951854]
27. Kim Y, Kim H, Choi J, Cho K, Yoo D, Lee Y, et al. Early prediction of need for invasive mechanical ventilation in the neonatal intensive care unit using artificial intelligence and electronic health records: a clinical study. BMC Pediatr 2023 Oct 23;23(1):525 [FREE Full text] [doi: 10.1186/s12887-023-04350-1] [Medline: 37872515]
28. Matheny ME, Yang J, Smith JC, Walsh CG, Al-Garadi MA, Davis SE, et al. Enhancing postmarketing surveillance of medical products with large language models. JAMA Netw Open 2024 Aug 01;7(8):e2428276 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.28276] [Medline: 39150707]
29. Jia M, Hu X, Ji L, Wang Y. An algorithm for automatic generation of tags for children's health science resources based on instruction-tuned LLM and similarity retrieval. : IEEE; 2024 Presented at: IEEE International Conference on Bioinformatics and Biomedicine 2024 (BIBM 2024); December 3-6, 2024; Lisbon, Portugal p. 4656-4659. [doi: 10.1109/BIBM62325.2024.10822325]
30. Abbasian M, Yang Z, Khatibi E, Zhang P, Nagesh N, Azimi I, et al. Knowledge-infused LLM-powered conversational health agent: a case study for diabetes patients. New York, NY: IEEE; 2024 Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 15-19, 2024; Orlando, FL p. 1-4. [doi: 10.1109/embc53108.2024.10781547]
31. Hu HW, Lin YC, Chia CH, Chuang E, Ru CY. Leveraging large language models for generating personalized care recommendations in dementia. New York, NY: IEEE; 2024 Presented at: 2024 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM); July 10-12, 2024; Taoyuan County, Taiwan p. 1-4. [doi: 10.1109/iwem59914.2024.10649066]
32. Kulshreshtha A, Choudhary A, Taneja T, Verma S. Enhancing healthcare accessibility: a RAG-based medical chatbot using transformer models. : IEEE; 2025 Presented at: 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD); April 13-15, 2025; Manama, Bahrain p. 1-4. [doi: 10.1109/itikd63574.2025.11005179]
33. Rettig L, Audiffren J, Cudré-Mauroux P. Fusing vector space models for domain-specific applications. New York, NY: IEEE; 2019 Presented at: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI); November 4-6, 2019; Portland, OR p. 1110-1117. [doi: 10.1109/ictai.2019.00155]
34. Yan Q, Huifang N. Risk analysis and recommendations for clinical use of infant incubators based on real-world data. China Medical Devices 2024;39(3):98-102. [doi: 10.3969/j.issn.1674-1633.2024.03.017]
35. Kaur S, Gandhi A, Sandhu SK, Baldi A. Barriers in reporting adverse effects of medical devices: a literature review. Naunyn Schmiedebergs Arch Pharmacol 2025 Feb 11;398(2):1145-1153. [doi: 10.1007/s00210-024-03431-x] [Medline: 39259332]
36. Jiang Y, Gu X, Ji D, Qian P, Xue J, Zhang Y, et al. Smart diagnosis. ACM Trans Multimedia Comput Commun Appl 2020 Jun 21;16(2s):1-21 [FREE Full text] [doi: 10.1145/3340240]
37. Wang C, Ning X, Qian P, Hu W, Yao J, Ng E, et al. Distribution entropy regularized multimodal subspace support vector data description for anomaly detection. Pattern Recognition 2026 Apr;172:112478. [doi: 10.1016/j.patcog.2025.112478]

Abbreviations

(IA)3: Infused Adapter by Inhibiting and Amplifying Inner Activations
AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BLEU: Bilingual Evaluation Understudy
C-MTEB: Chinese Massive Text Embedding Benchmark
FDA: Food and Drug Administration
IDIAE: Information Density Index of Adverse Events

XSL•FO
RenderX

LLM: large language model
LoRA: low-rank adaptation
MAUDE: Manufacturer and User Facility Device Experience
MMLU: Massive Multitask Language Understanding
MTEB: Massive Text Embedding Benchmark
Q/V: query/value
RAG: retrieval-augmented generation
ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence

Edited by A Coristine; submitted 08.Sep.2025; peer-reviewed by E Kim, C Wang, T Yusuff; comments to author 27.Dec.2025; accepted 10.Mar.2026; published 31.Mar.2026.

Please cite as:
Xia W, Zhu W, Li T, Wang L, Li W, Zhang P
Analysis Model for Infant Incubator Adverse Events Using Retrieval-Augmented Generation Combined With Dual-Adapter Fine-Tuning: Development and Evaluation Study
JMIR Med Inform 2026;14:e83745
URL: https://medinform.jmir.org/2026/1/e83745
doi:10.2196/83745
PMID:41915420

©Wenke Xia, Wanting Zhu, Tianchun Li, Li Wang, Weiqi Li, Peiming Zhang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Design Principles for Interactive Dashboards in Drug Safety Surveillance: Design Science Research

Malwina Kotowicz, MSc; Mexhid Ferati, PhD; Soumitra Chowdhury, PhD; Cláudio Martins Pires, BSc

Department of Informatics, Faculty of Technology, Linnaeus University, Växjö, Sweden

Corresponding Author:
Mexhid Ferati, PhD
Department of Informatics, Faculty of Technology, Linnaeus University, Växjö, Sweden

Abstract

Background: Adverse drug reactions pose a serious threat to health care, leading to patient harm and substantial economic burden. Dashboards for drug safety surveillance are a valuable tool to tackle it.

Objective: This qualitative study aims to develop a dashboard for drug safety tracking with active and iterative involvement of end users. To support dashboard development, we formulate and iteratively refine design principles (DPs) for drug safety dashboards using the affordance theory.

Methods: Following a design science research approach, we conducted 3 cycles of iterative design and evaluation involving the end users. Four professional end users (with expertise in drug screening, drug discovery, and data science) and 6 nonprofessional end users (drug consumers) were engaged in the requirements gathering sessions through co-design workshops, usability testing through think-aloud sessions, and heuristic evaluation.

Results: The analysis resulted in a set of 8 DPs refined using the prototype's affordances (ie, actionable properties of the dashboard that guide user interaction and interpretation of data). Following themes emerged in the formulation and refinement of DPs: addressing the bootstrap problem through designing for immediate use (DP1a), allowing identification of patterns through visualizing causality while signalizing uncertainty (DP1b), tracking trends for relevant variables (DP1c), implementing user-controlled views (DP2a) and customizable levels of data granularity (DP2b), guiding user's visual attention through spatial layouts (DP2c), designing for higher public value (DP3a) and providing features to support decision-making for varied stakeholders' groups (DP3b). A high-fidelity dashboard prototype for drug safety surveillance was proposed as a result of applying the final set of DPs. Heuristic evaluation of the prototype revealed an overall usability score of 84%.

Conclusions: Applying DPs rooted in affordance theory led to a purposeful and user-relevant artifact that can improve understanding of drug safety data and potentially guide decision-making processes for professional users. Our theoretical contribution lies in providing refined DPs, while also demonstrating how affordances can aid dashboard development in pharmacovigilance. Our findings may be applicable to similar health information systems in related domains.

(JMIR Med Inform 2026;14:e75936) doi:10.2196/75936

KEYWORDS

drug safety surveillance; adverse drug reactions; pharmacovigilance; dashboard systems; medical informatics applications; design science research; decision support systems; user-centered design; clinical

Introduction

Overview

Adverse drug reactions (ADRs) pose a serious risk, leading to hospitalization of millions of patients every year, and the incidence of ADRs resulting in hospitalization has been consistently increasing in recent years [1-5]. ADRs are not only a major health concern, but also a tremendous economic burden, generating costs of several billion dollars each year [6,7]. Monitoring side effects associated with medicinal products is critical for regulatory bodies, health care providers, pharmacists, and, most importantly, drug consumers.

The rapid spread of health misinformation through online channels poses significant risks to public health [8,9]. Evidence suggests that many individuals, including physicians and medical students, resort to online searches for clinical information (ie, general web and social media queries rather than official pharmacovigilance portals). These channels frequently present user-generated or commercially sponsored content (eg, and influencer posts) that lacks expert review. Only a fraction of users consult credible resources (eg, specialized academic databases) [10,11]. Importantly, more than half of the online searches by health care professionals focus on drug safety and ADRs [12]. Relying on unverified sources for health information can be a major threat to public health, erode public trust, and can lead to severe consequences, including harm or

XSL•FO
RenderX

death [13]. Specialized platforms, such as interactive visualization dashboards for drug safety information at-a-glance, offer drug safety information in a trusted and concise way. Such platforms could be intended for professional users (ie, health care professionals, pharmacists, researchers, pharmaceutical companies, and regulatory agencies concerned with pharmacovigilance), as well as nonprofessional users (ie, patients, medical journalists, or lawyers protecting patient rights).

Dashboards in pharmacovigilance (DiPs), however, are scarce and often have complex interfaces resembling decision support systems rather than accessible designs. Additionally, most of the DiPs are not available online as open-source platforms, but are offered as proprietary software under license [14-17]. As a result, they have a limited audience and lack widespread availability. Moreover, DiP often lacks salient usability features, is poorly designed due to the absence of a structured design approach, or has functional deficiencies [17,18]. Several data mining techniques have been proposed as a valuable tool for early detection of potential safety signals (ie, indications or evidence that suggest a possible new ADR) in pharmacovigilance [19,20]. However, existing DiP rarely integrates data mining methods to enhance raw drug safety information and therefore offers limited insight [21-23]. Given the shortcomings of current platforms for drug safety surveillance and the scarcity of prescriptive knowledge regarding their design, we investigate the idea of finding more fitting designs for handling such data. In search of suitable design principles (DPs) that should lead the development of similar platforms, this study relies on the design science research (DSR) strategy and is guided by the following research question: what are the appropriate DPs for interactive dashboards in drug safety surveillance?

To address this question, we set out to develop and evaluate DPs for a class of dashboards in pharmacovigilance, following the DSR approach. We show that engaging in prototype creation reveals general DPs applicable to similar systems. Consequently, besides the artifact being developed, emergent DPs are proposed as transferable concepts for similar systems within the boundary conditions studied. To this end, we use the notion of affordances—properties of physical things or relationships between actors and these things—to support our research. Affordances help in exploring different designs in this study, since different physical features of the prototype can afford (ie, provide and bring about) the same functionality.

This study consists of the following steps. First, we identify general DPs for visual dashboards of drug safety data. Second, following the DSR methodology, we perform the requirements gathering session with a group of experts in drug screening, bioinformatics, and data science. Third, we create a prototype of an interactive dashboard, where we present drug safety data in a visually friendly manner, with the purpose of improving its general understandability. To this end, we also focus on finding suitable data mining techniques for knowledge extraction from publicly available data in pharmacovigilance. Fourth, we assess the prototype in demonstration and evaluation sessions with users to detect potential usability issues and receive feedback on possible improvements to the design. Altogether, we carry out 3 iterations of the DSR phases.

The aim of this research is thus to enhance the understanding of drug safety data and improve the potential for decision-making among intended users. Specifically, our objectives are to (1) formulate DPs, (2) re-evaluate these in 3 iterations of alternating phases of prototype development, demonstration, and evaluation, and (3) gain insight on how the notion of affordances and DPs can support the development of visualization tools in drug safety surveillance. It is important to mention that, in this study, by “professional users” we refer specifically to researchers in drug screening, bioinformatics, and data science; clinicians and pharmacists were not recruited in our sample. Accordingly, the DPs and evaluation results reported here are grounded in researcher workflows. While some principles are likely cross-cutting across stakeholder groups, we do not evaluate their applicability to clinician- or pharmacist-facing workflows in this study.

Current State of Drug Safety Surveillance Systems and Data Mining Standards

Three major global databases monitor drug safety: Food and Drug Administration (FDA) Adverse Events Reporting System (FAERS) database, VigiBase maintained by the Uppsala Monitoring Center (UMC) and World Health Organization (WHO), and EudraVigilance managed by the European Medicines Agency (EMA). FAERS, the world’s largest repository with more than 24 million reports as of March 2022, serves as a key postmarketing pharmacovigilance tool. EudraVigilance and VigiBase similarly collect ADR reports from the European Economic Area. These reports, submitted by manufacturers (mandatory) and health care professionals or consumers (voluntary), document suspected side effects from one or more drugs per patient case [20,24,25].

Although crucial for surveillance, these reports cannot establish direct causality because of factors such as comorbidities or concurrent medication [25]. To address this, database curators apply data mining to detect safety signals for expert review. The FDA encourages signal detection in drug development, as it positively impacts the process, while the WHO and UMC offer similar services to industry partners [20,26].

Several digital platforms exist for drug safety surveillance, but many, including FAERS, rely solely on reporting frequencies rather than integrated data mining [21]. According to Kumar [17], the FAERS dashboard is affected by limited awareness, unclear purpose, unintuitive navigation, and challenges in interpreting its data. Most reviewed platforms pose accessibility challenges, as they rely on proprietary databases, restricting access, reproducibility, and public transparency [16]. Many are outdated or limited in scope, relying on small datasets or alternative data sources instead of major repositories [14-16,27,28], while others focus only on ADR reporting rather than analysis and visualization [29-31].

Moreover, most platforms were not developed using the DSR process or participatory design methods and lack usability testing. Therefore, their complex interfaces often resemble expert decision support systems rather than accessible designs

XSL•FO
RenderX

[14-16], which can impact usability, hinder tool adoption, and compromise their practical impact [18,32]. These factors emphasize the need for open-source, user-centered drug safety tools that enhance accessibility and stakeholder engagement.

Data mining techniques are routinely used in pharmacovigilance to enhance the understanding of drug and vaccine safety profiles [33-36]. While no universal standard exists, guidelines provide statistical support for interpreting detected signals [37]. Disproportionality analysis (DPA) is currently the most common technique recommended by the FDA, UMC (WHO), and EMA, although the organizations deploy different statistical approaches [38]. Despite its importance, DPA results are seldom integrated into public dashboards, which continue to rely mainly on reporting frequencies [21-23]. This limitation may hinder a thorough understanding of the drug safety profiles and potentially lead to overlooking important trends, which further emphasizes the importance of incorporating safety signal assessments in public dashboards in pharmacovigilance.

New platforms have been developed in recent years primarily leveraging artificial intelligence (AI), machine learning, and real-world data integration for ADR detection in clinical practice. Specific new platforms include the PVClinical system using OMOP-CDM and Observational Health Data Sciences and Informatics (OHDSI) software stack for electronic health record integration [39] and established platforms such as Oracle Argus Safety, repClinical, PvNET, and ArisG for ADR administration and reporting [40]. AI-powered approaches show particular promise, with systematic reviews demonstrating effective natural language processing application across 16 studies for user-generated content analysis [41] and 12 studies showing deep learning models outperforming traditional machine learning for adverse event extraction [42]. However, challenges remain, including data quality, regulatory compliance, and the need for clinical validation of these emerging platforms [43].

Recent research on human-computer interaction (HCI) and user-centered design in pharmacovigilance dashboards is limited but emerging, with only a few directly relevant studies identified in the past few years. The strongest evidence comes from Yu et al [44] who developed a visualization platform for FDA adverse event data with usability demonstration through case studies and validated accuracy against manually processed data. In a study on integrating human-centered design of a public health data dashboard for sexually transmitted infections in the state of New York, Ansari and Martin [45] found that dashboards must reflect actual workflow needs, address data limitations transparently, and use design elements, such as clear notes and estimated data displays, to build user trust and handle common data quality issues. This process resulted in a highly functional design template. Malkani et al [46] identified and ranked key health care dashboard design attributes such as clarity, simplicity, visual hierarchy, responsiveness, and intuitive navigation as most critical for improving usability and user-centered decision-making. Gavriilidis et al [47] applied user-centered DPs to create the PVClinical platform for ADR investigation, identifying “actionability” as a key adoption factor and emphasizing needs for explainable, human-interpretable results. Broader context comes from patient safety dashboard research, where Murphy et al [48] found that among 33 studies, only 4 incorporated informatics or human factors principles in development or evaluation, highlighting a significant gap in rigorous HCI application.

The evidence base remains thin, with most pharmacovigilance visualization research focusing on technical implementation rather than systematic user-centered design evaluation. More rigorous HCI research is needed in this specialized domain [44,47,48]. Table 1 summarizes the limitations of available dashboards and shows the prototype features we explore in response to the identified gaps.

XSL•FO
RenderX

Table . Research gaps were identified through limitations to available dashboards for pharmacovigilance and prototype features were intended to address them.

Platform category (examples)	Key limitations noted in the literature	Intended prototype features
Regulatory public dashboards (FAERS[a]; EudraVigilance; VigiLyze)	Emphasize reporting frequencies; statistical signal detection (eg, DPA)[b] typically not available in the public UI[c]; access and reproducibility constraints (open access but not open source; VigiLyze requires membership). The FAERS public dashboard shows issues with awareness, navigation, and interpretability for nonspecialists [17,21-23]	Integrates DPA outputs in the UI; offers guided interactions; co-designed and usability-tested; open data and code with an open-access dashboard
Enterprise pharmacovigilance dashboards (Oracle Argus Safety; ArisG; repClinical; PvNET)	Proprietary, case-management and regulatory-reporting focus; limited public documentation of integrated analytics, no participatory design or formal usability evaluation; constrained transparency and reproducibility [40]	Offers exploratory analysis; transparent metrics (downloadable DPA files); hybrid-open model enabling inspection, reuse, and migration to a fully open-source stack
Clinical and EHR analytics on OMOP-CDM (PVClinical)	Integrates EHRs[d] via the OHDSI[e] stack and reports user-centered design; scope differs from spontaneous report dashboards; platform availability varies; continued need for clinical validation and regulatory alignment [39,43,47]	Complements PVClinical by focusing on FAERS spontaneous reports; reports outcomes of co-design, think-aloud usability (and its quantitative outcomes), and heuristic evaluation as a result of systematic user involvement in multicycle evaluation process
Academic or research visualization platforms	Often older or limited-scope implementations with small or alternative datasets; few report systematic user-centered design [14-16,28,44]	Built on open FAERS data—the largest, regularly updated public repository with API[f] access; applies recent visualization techniques via implementation of domain-informed design principles
Patient-facing ADR[g] reporting and medication information apps	Designed primarily for ADR submission or mobile information; not intended for exploratory signal analytics or integrated DPA visualization [27,29-31]	Analytic dashboard for exploring suspected drug-event patterns, supporting stakeholder decision contexts

[a]FAERS: Food and Drug Administration Adverse Events Reporting System.

[b]DPA: disproportionality analysis.

[c]UI: user interface.

[d]EHR: electronic health record.

[e]OHDSI: Observational Health Data Sciences and Informatics.

[f]API: application programming interface.

[g]ADR: adverse drug reaction.

Affordance as a Lens to Study DPs

Affordances are the action possibilities an artifact enables through its material features [49-51]. We use this lens operationally: each DP states (1) the intended user activity (affordance), (2) the dashboard features that enable it (material properties), and (3) the boundary conditions in which the principle applies. In this study, the artifact does not itself cause understanding or decisions; it supports them insofar as its features enable the specified affordances [52].

Formulating the DPs

DPs provide prescriptive guidance for constructing an artifact to meet stated objectives [50,53-55]. Drawing on literature on pharmacovigilance data literacy and dashboard design, we formulated DPs using the structure by Chandra et al [53] that ties (1) intended user activity (affordance), (2) enabling features (material properties), and (3) boundary conditions. We use the following template for DPs formulation:

Provide the system with (material property—in terms of form and function) in order for users to (activity of user or group of users—in terms of action) given that (boundary conditions—user group's characteristics or implementation settings) [53].

We consider drug safety surveillance as the boundary condition and analyze 2 user groups for this DSR prototype: (1) nonprofessionals (drug consumers) and (2) professional users as operationalized in this study (researchers in drug screening, bioinformatics, and data science). While health care professionals and pharmacists are central stakeholders for pharmacovigilance dashboards, they were not recruited in the present sample; consequently, findings and DPs are grounded in these boundary conditions.

We categorize the DPs into three classes: (1) DPs related to functionality that address the question of how well the dashboard serves its purpose, (2) DPs related to the mode of display that refer to the specific visualization techniques, and (3) DPs related to added value, which focus on the benefits the artifact provides to the public or end users.

We selected and refined DPs by applying an affordance alignment test: a principle was retained only when a target user

XSL•FO
RenderX

goal could be mapped to a concrete dashboard feature expected to enable the activity in the drug safety context [52,53]. Subsequent evaluation rounds confirmed and sharpened their scope.

DPs Related to Functionality

The value of new information infrastructures (IIs) lies in the users' community. However, since this community is initially small, the benefits of using the system may not be immediately apparent. Several authors emphasized the bootstrap problem that arises when the benefits of new IIs are perceived as indirect, delayed, or uncertain, which can lead to resistance in the system's adoption [56,57]. To address this issue, Hanseth and Lyytinen [57] recommend designing IIs for direct usefulness, that is, prioritizing immediate practical benefits for the users. This approach may enhance the perceived value of the system and help expand the user community. It may also create a positive feedback loop, where the use of the system leads to its further improvements, which in turn increase its usefulness and value. Consequently, we formulated the following DP: DP1a: it provides features to address the bootstrap problem, so that the system affords immediate use of the dashboard in drug safety surveillance.

Recker [58] investigates public dashboards related to the coronavirus pandemic, concluding that many present data through simple attributes (eg, number of cases, number of deaths, and number of recovered). Integrating state-tracking features that enable monitoring changes in states and events that trigger them (such as fluctuations in infection rates, or timelines of lockdowns and restrictions) could be a significant improvement, and the information presented in dashboards should answer the question how, instead of just how many. We formulated two DPs: (1) DP1b: it provides features to describe possible causality between variables, so that the system affords identification of patterns and trends in drug safety surveillance. (2) DP1c: it provides features to track changes in variables, so that the system affords tracking of the progression of relevant states in drug safety surveillance.

DPs Related to the Mode of Display

Matheus et al [59] suggest incorporating customizable views into decision-making dashboards, as this design strategy may help bypass the interpretation bias—"a single view might result in a limited picture of the situation." Expanding views to more than one view may lead to better understanding of the problem. Customized views help users to focus on relevant information and reduce the cognitive load associated with processing large volumes of data. We formulated the following DP: DP2a: it provides features to support customized views, so that the system affords examining various dimensions of data in drug safety surveillance.

Matheus et al [59] also suggest incorporating customizable features into decision-making dashboards, including the ability to toggle between detailed and high-level data visualizations. As these dashboards often contain a large volume of data, such features can help users understand the problem at hand by providing varying levels of detail. We formulated the following DP: DP2b: it provides overview and details, so that the system affords exploring data of different granularity levels in drug safety surveillance.

Toreini et al [60] follow a set of experiments to find ways to decrease task resumption failures. Tracking the real-time eye movements of the participants, they conclude that users start from and focus their attention on the upper left corner of the dashboard. Several authors suggest incorporating attention-augmenting features into dashboard elements (such as varied color, size, shape, and orientation) and visual cues (buttons and dials) to direct attention toward objects of higher importance [60,61] Consequently, DP2c: it provides features that stimulate attention and place important elements on the top-left area of the dashboard, so that the system affords improved attention management while using the dashboard in drug safety surveillance.

DPs Related to Added Value

Matheus et al [59] claim that creating dashboards should go beyond simply visualizing data. Creating public value should be embedded in the design process, as the data on its own is of little use. This way of thinking of the design process—through the prism of public value—transgresses the initial idea of merely finding the ways to present data in the best visual way. Data visualization should be the means, not the goal, and asking the question "what is the added value?" during the design process is necessary. Designers should prioritize generating public values such as engagement, transparency, and accountability, while also upholding public values such as privacy. Consequently: DP3a: it provides features to create public values, so that the system affords users' engagement in dashboard use in drug safety surveillance.

Embedding features that support decision-making is a crucial aspect of dashboard development [59]. Such features can be valuable for professional users, including health care professionals, pharmacists, researchers, and regulatory bodies in pharmacovigilance. We also expect the users would gain greater understanding and take up higher responsibility toward their own health, which would equip them better for decision-making. We formulated the following DP: DP3b: it provides features to support decision-making, so that the system affords users to evaluate what-if scenarios and consider different alternatives in drug safety surveillance. Initial DPs are shown in Table 2.

XSL•FO
RenderX

Table . Overview of initial design principles.

Design principle	Affordance	Material properties
DP1a	Obtaining immediate practical benefits while using the dashboard [56]	Features that respond to users' requirements, features that address actual user problems and offer solutions
DP1b	Identification of patterns and trends through exploration of DPA[a] analysis [58]	Features that enable identification and examination of patterns and trends in drug-ADR[b] evaluations
DP1c	Monitoring the progression of relevant states (how variables change over time) [58]	Features that allow tracking changes in relevant variables
DP2a	Exploring, analyzing, and interpreting data from different angles [59]	Features that enable analysis of data from multiple perspectives
DP2b	Exploring data from a broad and narrow perspective by switching between levels of granularity [59]	Features to adjust the level of detail depending on users' needs and the context of analysis
DP2c	Improved attention management while using the dashboard [60,61]	Features that stimulate attention, visual cues that direct attention toward important elements
DP3a	Active involvement in exploring the data, potentially increasing users' satisfaction [59]	Features with a potential to create public value (engagement, transparency, accountability, and adherence to privacy)
DP3b	Evaluating alternatives of actions, exploring what-if scenarios [59]	Features that support decision-making

[a]DPA: disproportionality analysis.

[b]ADR: adverse drug reaction.

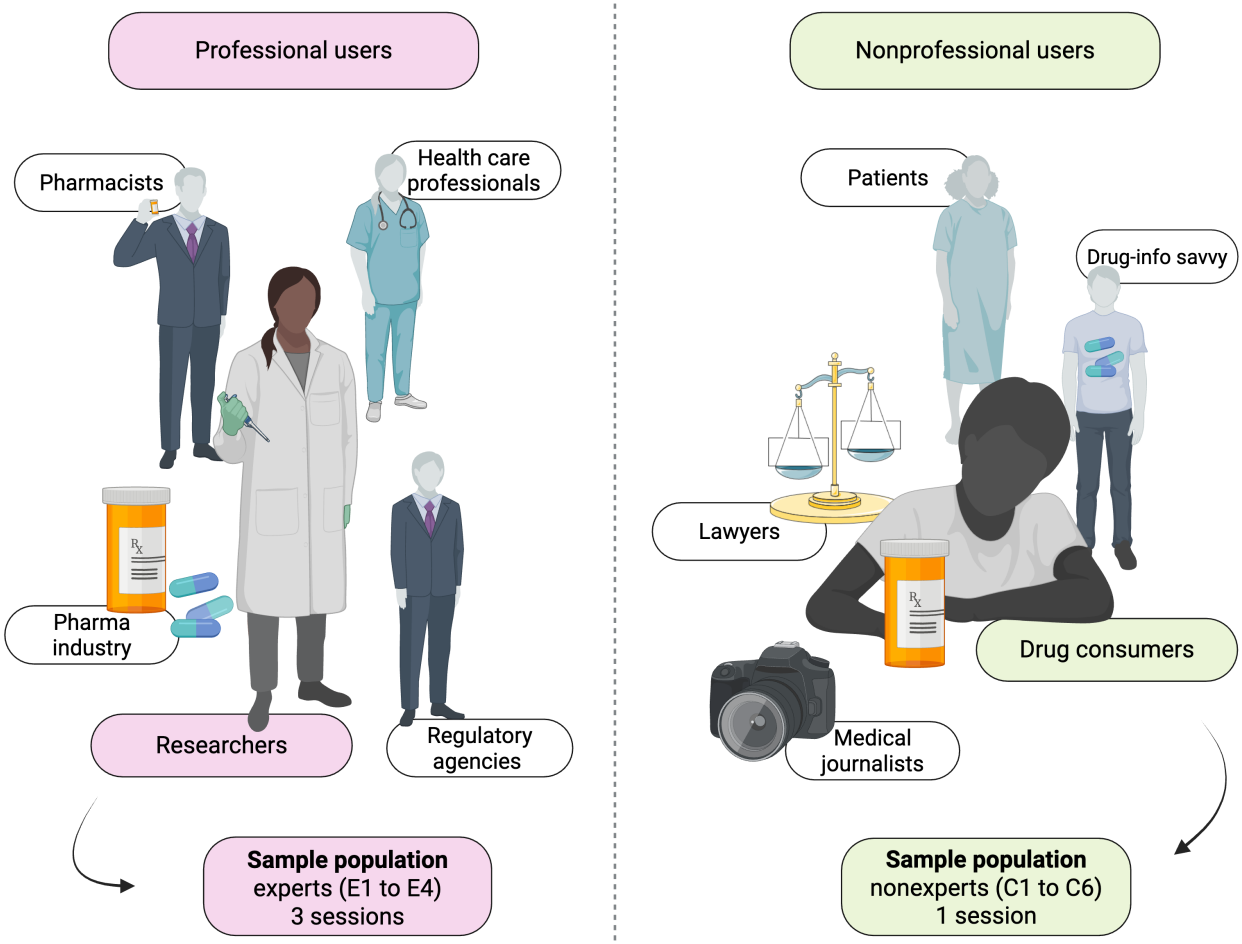
Methods

Data Collection and Target Groups

We have considered 2 target groups for this work: professional users and nonprofessional users. We define professional users as those whose professional activities are concerned with drug safety surveillance and thus have extensive knowledge in the domain. For clarity, throughout the paper, "professional users" refers to this researcher group; the implications for clinician- and pharmacist-facing workflows lie outside this study's boundary conditions. In contrast, nonprofessional users are not professionally concerned by drug safety surveillance and in that sense are laymen in the domain of pharmacovigilance. Figure 1 depicts the user groups.

XSL•FO
RenderX

Figure 1. Professional (left) and nonprofessional (right) user groups. The broader professional stakeholder landscape includes health care professionals, pharmacists, researchers, regulatory agencies, and industry. In this study, the professional participants recruited were researchers in drug screening or drug development (E1-E4). Nonprofessional participants were drug consumers (C1-C6).



We have recruited a group of 4 expert users (referred to as experts, with participant IDs of E1 to E4) for this study. All experts were scientists of different academic degrees and expertise levels and were working on drug development and screening projects. Three out of 4 experts had extensive experience in data science and bioinformatics. In addition, all experts had many years of experience in data visualization and communication to the general and scientific community. Experts participated in the co-design workshop, pilot evaluation, and heuristic evaluation sessions.

For usability testing (think-aloud session), we have recruited a group of 6 nonexpert users (referred to as nonexperts, with participant ID of C1 to C6). All nonexperts were drug consumers and had a history of continuous medication therapy of at least 7 days at least once in the previous year. We assumed these users were potentially also drug safety information-savvy and might provide meaningful insights into which complex analyses are possible. Importantly, nonexperts were novices in the sense that they had no prior exposure to the tool in question or any other similar dashboards [62]. Although some nonexperts shared parts of professional experience with experts (eg, C4 and C5 had expertise in biomedical engineering and data science), they were not working on drug screening and discovery projects.

By design, think-aloud usability testing was conducted solely with nonprofessional “novice” users to evaluate first-time learnability and immediate practical use (DP1a) without prior-exposure effects from experts who had already interacted with the prototype during co-design and/or the pilot evaluation [62].

Overall, all participants were between 28 and 44 years old and were based in Germany, Sweden, France, Finland, and Portugal. A detailed description of participants recruited for this study is shown in Table S1 in Multimedia Appendix 1. Given the age range and the presence of multiple participants with advanced technical or professional backgrounds, this sample likely exhibits above-average digital and data literacy; while valuable for investigating complex interaction issues in early iterations, this may inflate task performance relative to older adults and users with lower digital health literacy.

DSR Approach

We followed a framework proposed by Sauro and Lewis [63] to quantitatively analyze the user experience. Furthermore, in heuristic evaluation, we quantitatively collected and analyzed data in line with the framework proposed by Dowding and Merrill [64]. Additionally, we qualitatively collected and analyzed data in all experimental sessions in this study.

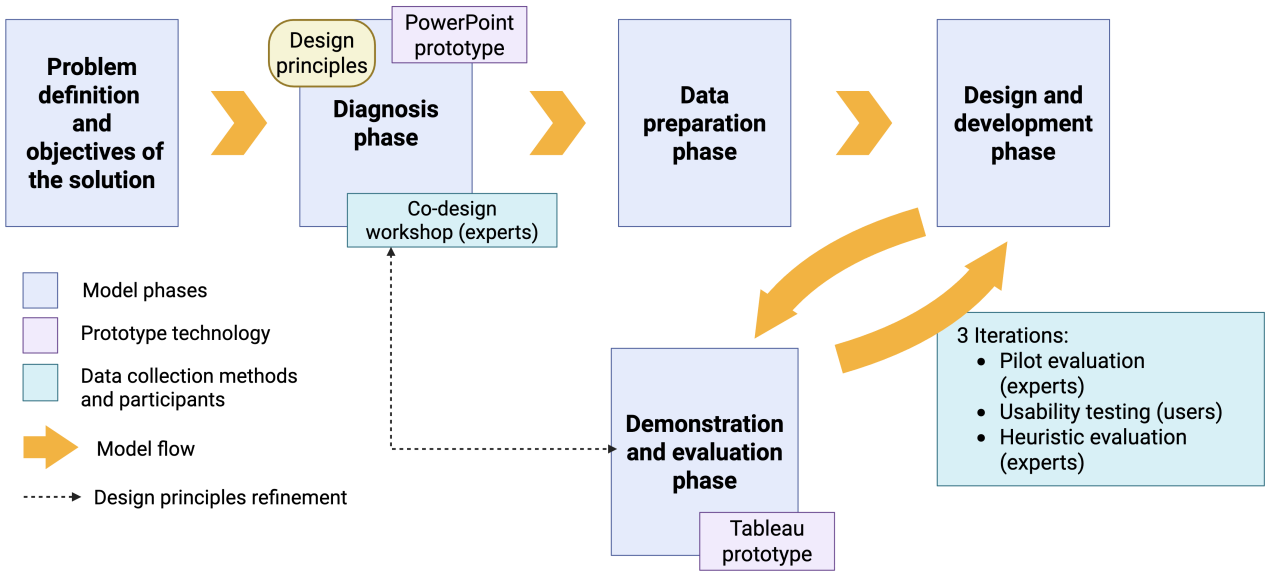
We modified a staged DSR model, initially proposed by Peffers et al [65]. Thus, the DSR framework applied in this study consisted of the following stages: (1) problem definition and

XSL•FO
RenderX

objectives of the solution, (2) diagnosis phase, (3) data preparation phase, (4) design and development phase, and (5) demonstration and evaluation phase. Detailed descriptions of each phase are found in Multimedia Appendix 2.

A depiction of the DSR process is shown in Figure 2. v1-PowerPoint and v2-v4-Tableau prototypes were used in diagnosis and demonstration and evaluation phases, respectively. Blue boxes show individual phases and the order in which they occurred in DSR (yellow arrows). DPs emerged in the diagnosis phase and were then iteratively refined in the demonstration and evaluation phases.

Figure 2. Applied design science research model.



Data Collection in DSR Cycle

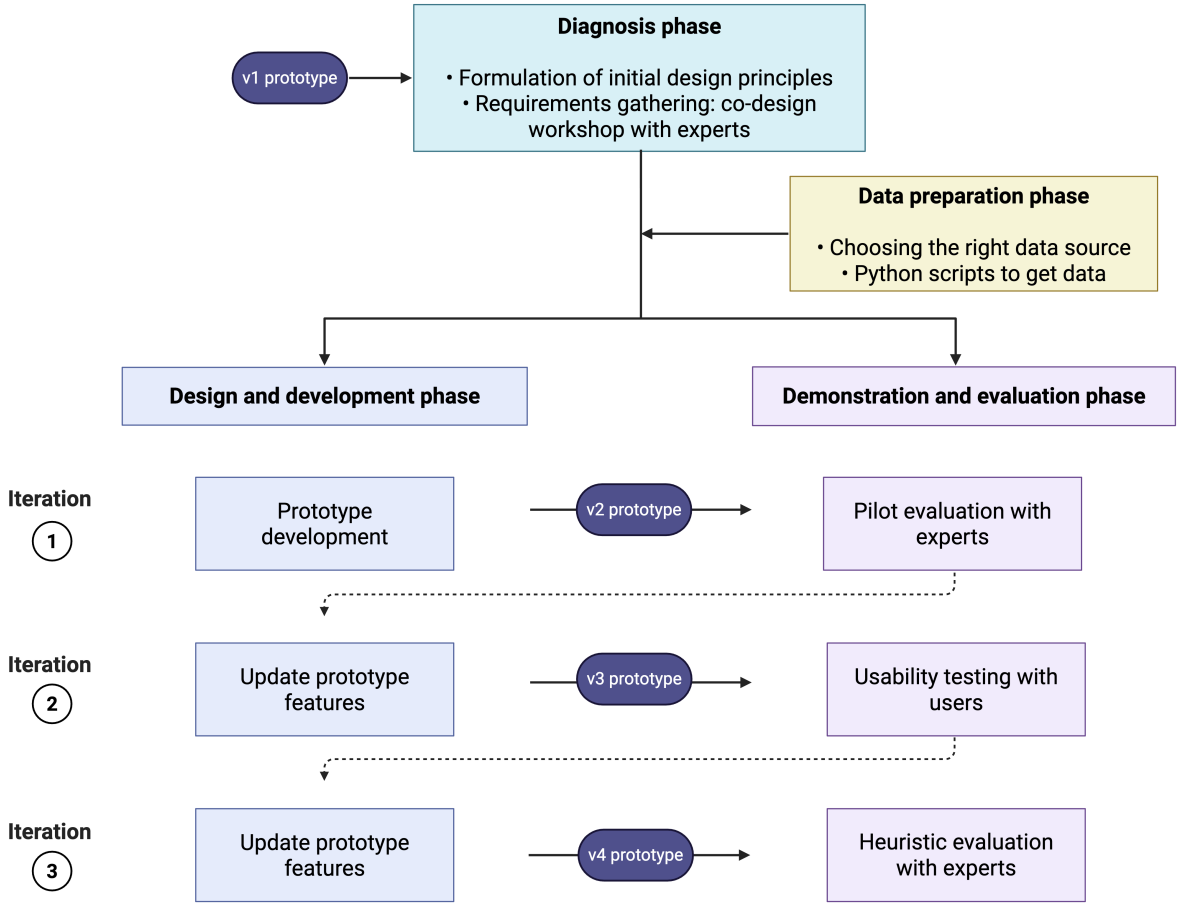
We have performed 3 iterations of the DSR cycle. That is, we started with the diagnosis phase, in which we first identified initial DPs by gathering secondary data available in the literature. Then, we proceeded with one requirement gathering session, in the form of a co-design workshop with experts. Next, we have engaged in the data preparation phase, where—after selecting a suitable drug safety data source—we have written Python scripts to get and process the data.

We then performed 3 iterations of alternating design and development and demonstration and evaluation phases. In the first iteration, we developed a prototype in Tableau software (v2; Tableau Software LLC) [66], and tested it in the pilot evaluation session with experts. The results obtained in this way fed the second DSR iteration. We have updated prototype features (v3) and proceeded with usability testing (in a think-aloud session) with nonexperts. Finally, in the last iteration, upon updating prototype features (v4), we performed heuristic evaluation with experts. A flowchart of research activities is shown in Figure 3.

Further details about the design and development phase are provided in Multimedia Appendix 2. In that appendix, we describe the diagnosis phase, including the development of DPs and requirements gathering. We also provide information on how we used Statista [67] and participant requests to inform the selection process of 5 drugs used for prototype development: ibuprofen, acetaminophen (paracetamol), prednisone, quetiapine, and morphine.

XSL•FO
RenderX

Figure 3. Overview of data collection activities.



Demonstration and Evaluation Phase

Iteration 1: Pilot Evaluation

The researcher was moving through the dashboard by performing a set of predefined activities, while the participants controlled the navigation process by suggesting improvements whenever necessary. The session lasted 1 hour and was recorded with Otter.ai (Otter.ai, Inc) [68].

Iteration 2: Usability Testing

We have performed 6 usability tests with nonexpert users in think-aloud sessions. These included live meetings with 4 participants (C1, C2, C3, and C4) and virtual meetings with 2 participants (C5 and C6). Before the session, we have provided the users with a detailed protocol including introduction of the study, explanation on usability testing in the think-aloud session and a list with predefined tasks. These tasks required participants to do activities on the interface without the researchers' guidance. For the most part, researchers remained silent during users' task performance. Users were encouraged to express their thoughts while conducting the tasks. Additionally, 2 participants (C2 and C3) shared their thoughts on the usefulness of the dashboard and their interest in the tool in general. These short and unstructured discussions unfolded at the end of the sessions and were incorporated into findings. Each session lasted between 1.5 hours and 2.5 hours and was recorded with the same software, Otter.ai [68].

Experts were intentionally not included at this think-aloud stage to avoid learning effects from earlier sessions; their structured feedback was captured in the expert-focused co-design phase, pilot evaluation, and in the subsequent heuristic evaluation with 4 and 3 experts, respectively, which is within the 3 - 5 evaluator range known to uncover the majority of usability problems [69,70].

Iteration 3: Heuristic Evaluation

As originally suggested by Nielsen [71] and confirmed by other authors [62,69,72,73], 85% of usability violations can be detected by 3 to 5 evaluators in heuristic evaluation. We have performed heuristic evaluation with 3 experts (E1, E2, and E3).

Data Analysis

Overview

We used the framework method by Gale et al [74] to qualitatively analyze data in search of codes and themes. Importantly, and due to the specificity of DSR as a methodology (ie, when data are collected in iterations to repeatedly improve an artifact), codes and themes in this work refer to identified usability issues, dashboard features or elements undergoing evaluation and refinement, or any concept related to these (ie, the motivation behind the revision of DPs). Additionally, to prioritize the issues detected in usability testing and select appropriate solutions, we applied a method by Sauro and Lewis [63]. Finally, we analyzed the data from the heuristic evaluation through the assessment of 10 usability factors. These concepts

XSL•FO
RenderX

are discussed further below. Overall, we detected 33 themes (entities) in the co-design workshop session and 20 themes in the pilot evaluation session.

Usability Testing

For detecting usability issues, we applied a quantitative method by Sauro and Lewis [63]. We prioritized the usability issues according to their severity, considering the following factors: how much the issue has impacted the user trying to achieve the task, how critical the task was, and what was the frequency of the issue. Thus, the severity of issues was calculated by multiplying task criticality, impact, and frequency of occurrence. Overall, we detected 69 usability issues.

Each issue was then assigned one or more solutions. To choose the optimal solution to each usability issue, we considered 2 factors: complexity and effectiveness. Finally, a cost-benefit ratio (return on investment [ROI]) was calculated by dividing the effectiveness by the complexity. We decided which solutions to implement based on the following criteria: the solution had an ROI of 5 or more, was technically feasible, and was not solving the issue already covered by another solution with higher ROI. Overall, out of 56 conceptualized solutions, we arrived at 22 final solutions to be implemented. For a detailed description of the applied method, refer to Sauro and Lewis [63] and Rosemberg [75].

Heuristic Evaluation

Applying Dowding and Merrill's [64] heuristic evaluation, we assessed 10 usability factors using 49 close-ended questions. Usability factors had varying question counts; the maximum score per factor was calculated by summing questions assigned to it. Scores obtained per expert were calculated by summing the questions answered yes by each expert, per factor. A mean score was calculated by summing individual expert scores and dividing by the number of experts. The final result was the mean score divided by the maximum score (per factor and cumulatively). Next, overall severity rankings per usability factor were assigned, ranging from 0 (no usability problem) to 5 (usability catastrophe). Finally, qualitative analysis of comments per factor informed suggestions on future implementations. For a detailed description of the applied method, refer to Dowding and Merrill [64] and Nielsen [71].

Ethical Considerations

Informed consent was obtained from all participants prior to inclusion. Participation was voluntary, and participants were informed that they could opt out at any time. They were provided with information on how their data would be processed and that study data would be retained for no longer than 2 years from study start. No ethics or institutional review board approval was required for this work.

According to the Swedish Ethical Review Act (2003:460) [76], this study did not require formal ethical approval because it did not involve sensitive personal data, biological material, or any physical or psychological intervention. Only non-sensitive demographic information (country of residence, age, gender, profession) was collected. All participants were adults and data were pseudonymised using participant ID codes prior to analysis. Participants also provided consent for publication of the study results.

Results

Diagnosis Phase: DPs and Low-Fidelity Prototype

We have implemented DPs into a v1-low-fidelity prototype in PowerPoint. One dashboard element could be assigned to multiple DPs, but each assignment involved different functionalities of the element. Some DPs were action-oriented (DP1a and DP3a), and all material properties were assigned to them. We show how DPs shaped the v1-low-fidelity PowerPoint prototype in Table 3.

To afford exploration of data of varied granularity (DP2b), we created two dashboard sections: (1) an overview of common side effects for the chosen drug (the overview section) and (2) a detailed analysis of specific drug-side effect pairs (the detail section), split in separate pages where different visualization techniques were used (eg, a graph with age-gender distribution of affected populations). Additionally, we added drug and side effect search fields (Figure 4A–C).

To afford identification of patterns and trends (DP1b), the overview section displays a histogram of commonly reported side effects for a queried drug. Users could toggle between positive and negative side effects using buttons (Figure 4A). To afford monitoring the progression of relevant states (DP1c), a line graph with a year filter tracked reports for drug-side effect pairs over time (Figure 4C). Additionally, to afford improved attention management (DP2c), key graphs were positioned top-left, and warning messages were color-coded.

The system provided downloadable evaluation files containing detailed drug safety metrics (DA.csv, Figure 4A–C) to afford evaluating alternatives of actions (DP3b). Warning messages indicated whether an event was statistically significantly related to queried side effects and provided DPA information on mouse hover (Figure 4C). Two affordances, obtaining immediate practical benefits while using the dashboard (DP1a) and active involvement in exploring the data (DP3a) were expected to be enacted across all prototype features, as they aligned with the prototype's purpose.

XSL•FO
RenderX

Table . Instantiation of material properties.

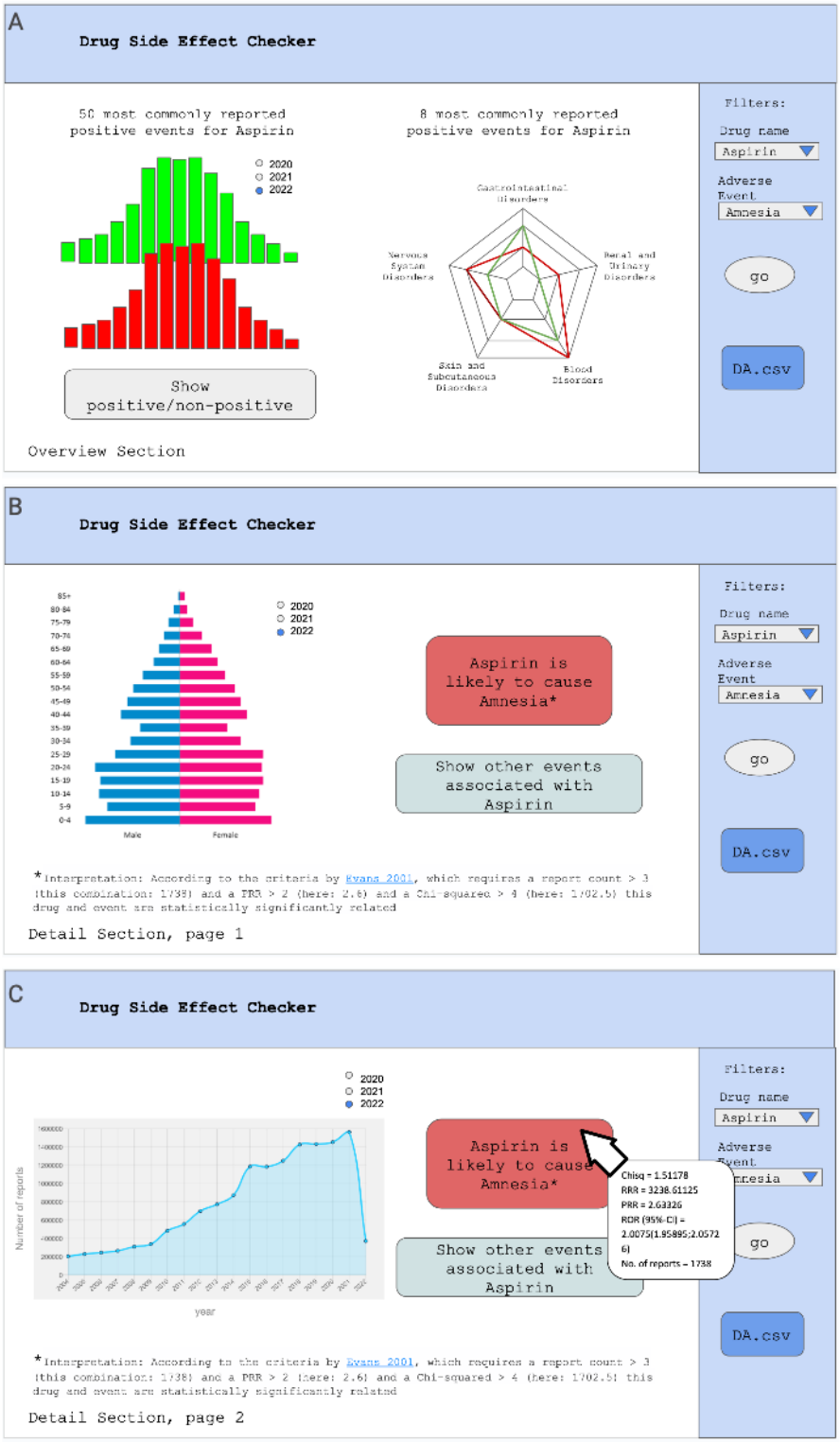
Design principle	Instantiation of prototype features
DP1a	Initiated by all features and iteratively refined in response to users' input
DP1b	Graphs belonging to the overview section, warning messages with DPA[a] metrics on hover
DP1c	Line graph with the number of reports per drug-ADR[b] pairs, with year filters
DP2a	Interactivity of the graphs (eg, buttons to toggle between positive and negative ADRs), with year filters
DP2b	Splitting the dashboard into the overview and detail sections, with year filters
DP2c	Placing the graphs on the top-left, color-coded warning messages
DP3a	Initiated by all features expected to elicit values related to the transparency of drug safety data or the user engagement in health information tracking
DP3b	DPA metrics as downloadable files (DA.csv), warning messages with DPA information on hover

[a]DPA: disproportionality analysis.

[b]ADR: adverse drug reaction.

XSL•FO
RenderX

Figure 4. An excerpt from the v1-low-fidelity PowerPoint prototype. (A) The overview section. (B-C) The detail section.



Diagnosis Phase: Requirements Gathering

Through a collaboration with a group of experts in a co-design workshop, we gathered the requirements for the prototype development. The v1-low-fidelity PowerPoint prototype was used in this session to visualize content of the dashboard, serving as a canvas for design activities. We summarize the principal findings in Table 4.

XSL•FO
RenderX

Table . Principal findings from the co-design workshop.

Feature or element	Related DPs[a]	Feedback
Indication graph	• DP2b (overview and details) • DP3b (decision-making support)	Offers details on common prescriptions for a queried drug, adding an extra variable to the displayed dataset.
Scrolling	• DP2c (attention stimulation)	Use scrolling instead of pagination, it is a preferred way for users to interact with the dashboards.
Drug and event search fields	• DP2b (overview and details)	Users seek information on potential side effects both before and after drug administration. Separate search fields allow users to look up a drug only (before an event occurs) or a drug and ADR[b] (after an event occurs).
DPA[c] metrics files	• DP3b (decision-making support) • DP2b (overview and details)	To influence decision-making of professional users, DPA files should include formulas and information on statistical models used for calculations, as well as the interpretation of calculated indicators.
Removing DPA info from the warning message's hover	• DP2b (overview and details)	While beneficial for researchers, high-level granularity data may confuse nonprofessional users. Access to DPA metrics is provided through downloadable files.
Age-gender distribution graph with modified age ranges, positioned top-left	• DP2b (overview and details) • DP3b (decision-making support) • DP2c (attention stimulation)	The graph provides information on common demographic attributes (age and gender). Using 10-year ranges provides optimal data granularity.
Histogram of most commonly reported side effects, positioned top-left	• DP2b (overview and details) • DP2c (attention stimulation) • DP2a (customized views) • DP1b (causality between variables)	The graph provides an overview of the most commonly reported ADRs, with the option to customize the display using a histogram filter (Related, Unrelated, All events). Limiting the bars to 20 provides optimal data granularity.
Per drug, there is a line graph showing the sum of side effects reported over time (the first line graph) and a second line graph illustrating specific side effects reported over time	• DP1c (state-tracking) • DP2a (customized views)	For researchers, viewing time series of drug-ADR report numbers is crucial. The second line graph helps identify, for example, specific events contributing to sudden increases in ADRs in selected years. The histogram filter is linked with line graphs, enhancing customization options.

[a]DP: design principle.

[b]ADR: adverse drug reaction.

[c]DPA: disproportionality analysis.

Adjusting Data Granularity

We found evidence that the dashboard supported the exploration of data of varied granularity (DP2b). Feedback from experts suggested adjusting the detail level on display. Some elements were considered too detailed, resulting in information overload, especially on the histogram and the age-gender distribution graph. Recommendations included fixing the histogram's displayed bars at 20 and setting age intervals on the age-gender distribution graph to 10 years (Figures 5A and 6A, respectively). Additionally, it was suggested that displaying detailed DPA metrics while hovering the warning message might confuse users unfamiliar with similar statistical methods, as per E1: this proportional reporting ratio and all that (...) Not super easy to interpret, right? This is greater than 2, but like, what is the upper bound? What is the lower bound? Like, how low can it go? (...) These values are not so easy to understand if you're not the professional who uses them.

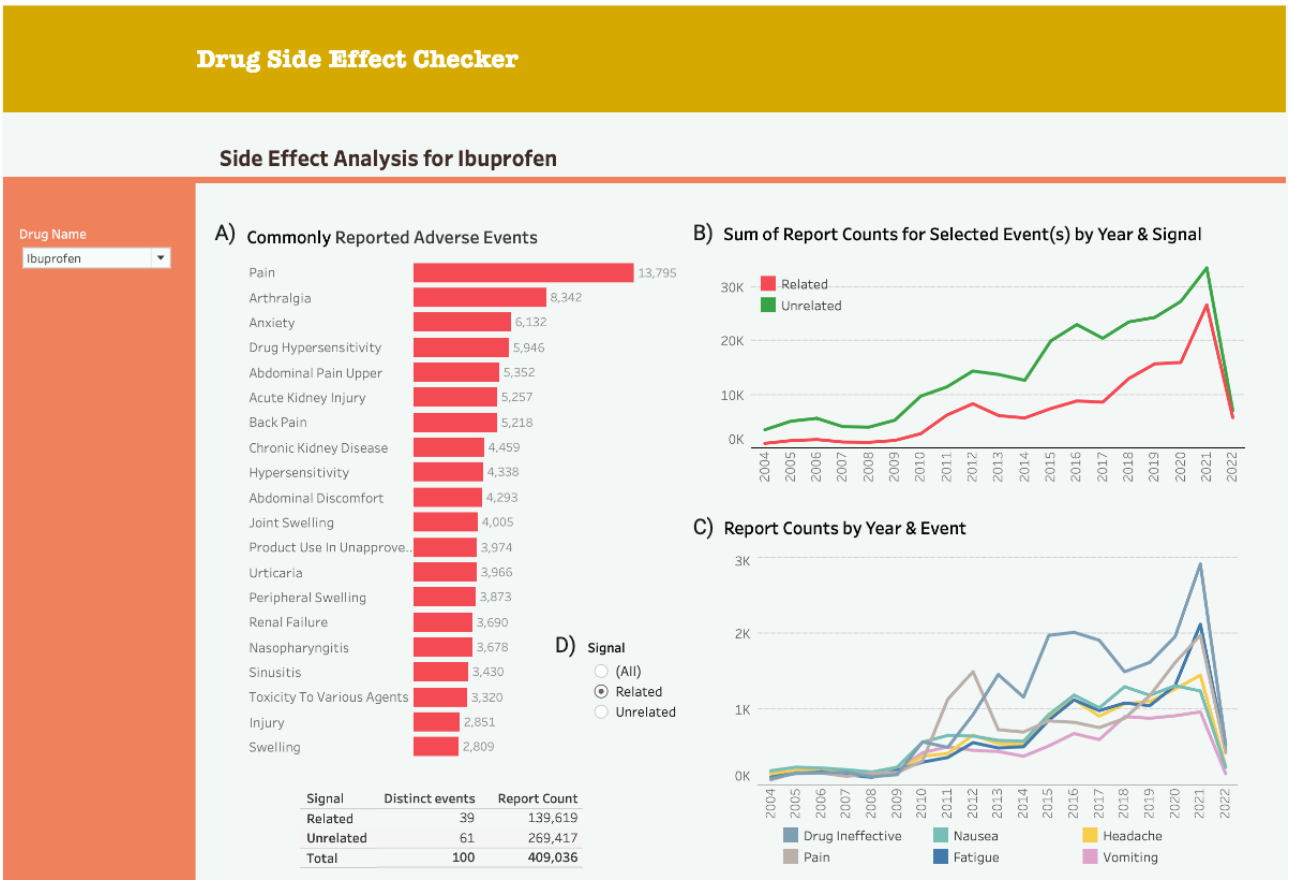
Thus, the DPA information was removed from the warning message and added to downloadable DPA files, together with formulas for score calculation and interpretation of indicators. This was also in line with the suggested increase of data granularity on some occasions. Likewise, it has been proposed to implement a functionality to view the data by the condition for which a drug was prescribed. Finally, while initially considered, the BMI variable was ultimately excluded from the age-gender distribution graph. Although useful for analysis, BMI is rarely provided in the reports. Therefore, it was necessary to adjust the intended data granularity to what was available in the datasets. Considering all the above, we revised the following DP: DP2b: It provides an overview and details, while controlling overview-to-detail ratio, so that the system affords exploring data of different granularity levels in drug safety surveillance.

Accounting for different data granularity interest among various user groups, and considering that customizable level of detail

XSL•FO
RenderX

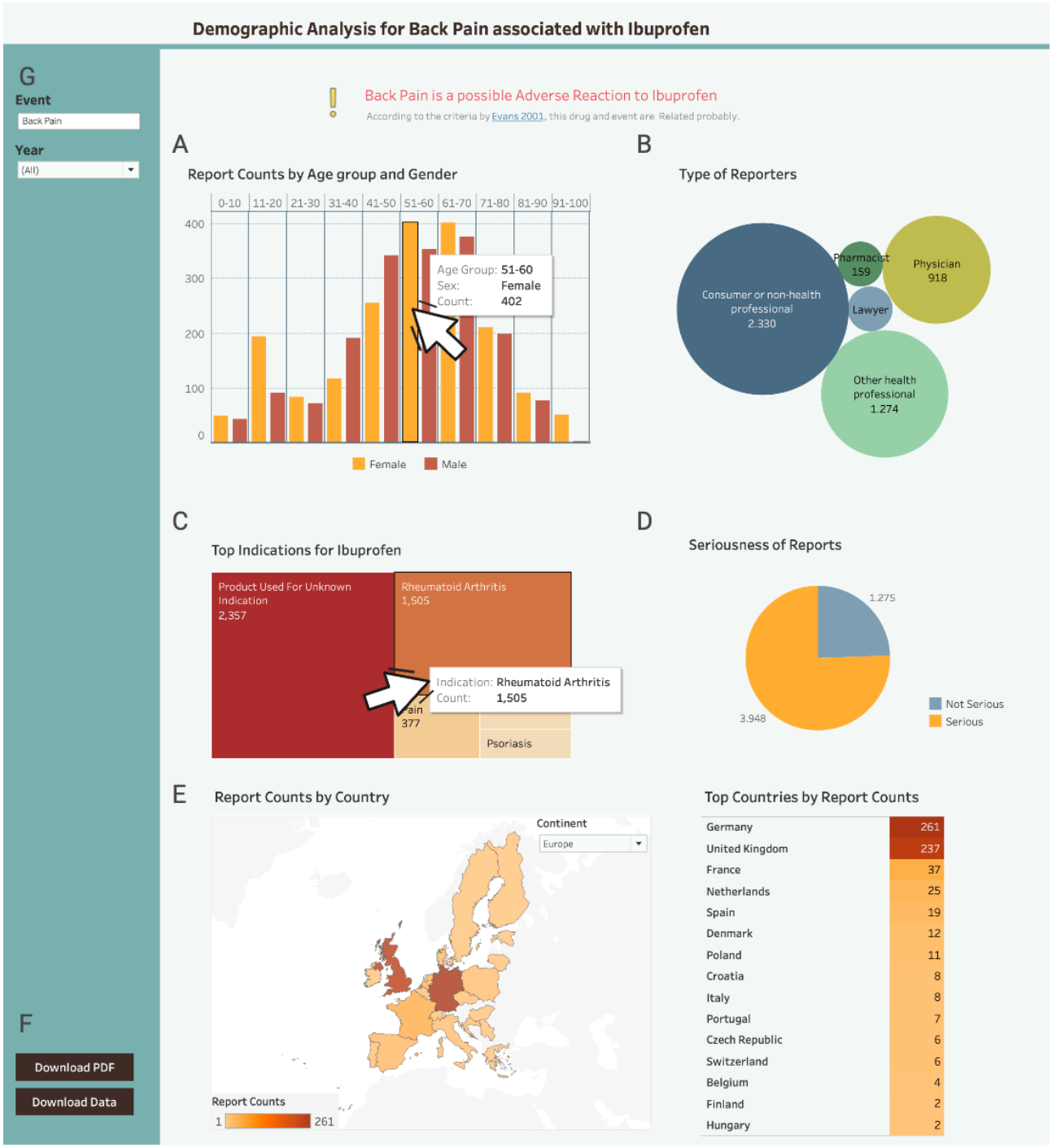
in displayed data can reduce distraction and potentially influence decision-making, we revised another DP: DP3b: It provides features to support decision-making and cluster them by the needs of distinct user groups, so that the system affords users to evaluate what-if scenarios and consider different alternatives in drug safety surveillance.

Figure 5. The first Tableau (v2-high-fidelity) prototype with analysis for ibuprofen (the overview section). (A) Modified histogram. (B) The first line graph. (C) The second line graph. (D) The histogram's filter.



XSL•FO
RenderX

Figure 6. The first Tableau (v2-high-fidelity) prototype with analysis for ibuprofen and back pain (the detail section). (A) Modified age-gender distribution graph. (B) The reporter's type of occupation. (C) The indication graph. (D) The seriousness of the outcome. (E) The map with number of reports per country and the associated table. (F) Download buttons. (G) The event search field.



Scrolling Over Pagination

Feedback suggested changing from pagination to scrolling, as users often prefer to scroll, rather than having to navigate through multiple pages. Pagination is also considered more expensive in terms of browsing and might significantly slow the system down. Thus, we revised the DP: DP2c: it provides features that stimulate attention, place important elements on the top-left area of the dashboard, and promote scrolling over pagination, so that the system affords improved attention management while using the dashboard in drug safety surveillance.

Iteration 1

Design and Development: First Tableau Prototype

We transformed the v1-low-fidelity PowerPoint prototype into a v2-high-fidelity dashboard in Tableau, incorporating refined DPs based on the guidance from the co-design workshop. Figures 5 and 6 show the prototype after the first iteration.

After requirements analysis, to better afford monitoring the progression of relevant states (DP1c), we introduced 2 graphs: a line graph showing all side effect reports per drug per year

XSL•FO
RenderX

(Figure 5B) and a line graph showing specific side effect reports per drug per year (Figure 5C).

For exploration of data of varied granularity (DP2b), we added an indication graph showing the conditions for which the queried drug was prescribed when a side effect occurred (Figure 6C). In line with revised DP2b, we modified the age-gender distribution graph (Figure 6A). We also modified the histogram by reducing the number of bars to better afford identification of patterns and trends (DP1b). A map displayed side effect distribution per country with interactive continent selection and year filtering (Figure 6E).

With regard to improved attention management (DP2c), we positioned the histogram and the age-gender distribution graph on the top-left overview and detail sections, respectively. We replaced pagination with scrolling, aligning with DP2c. To better afford evaluating alternatives of actions (DP3b), we introduced safety signal reports with DPA information, displayed upon clicking the download data button (Figure S1 in Multimedia Appendix 3). To better afford exploring data from different angles (DP2a), all graphs in the overview section were made interactive: for example, selecting up to 6 histogram bars was reflected in the second line chart, where up to 6 individual events were displayed. It was also possible to limit the dataset by year ranges (Figure S2 in Multimedia Appendix 3).

Demonstration and Evaluation: Pilot Evaluation

While conducting a pilot evaluation, we were looking for suggestions on possible improvements to the prototype before embarking on usability testing. Overall, we received advice to implement a getting started tutorial containing screenshots of the dashboard elements with a short explanation on salient functionalities (eg, how to remove the selection and how to select using keyboard or mouse).

Importantly, expert E3 suggested implementing a feature related to selection-keeping—when the user chooses one or more side effects by clicking on the histogram bars, and then changes their drug selection, the dashboard should keep the selection of side effects. This feature can be useful to quickly compare drugs from the same drug classes (eg, ibuprofen and aspirin). According to E3: if you want to compare that one drug to another drug, to see if it (selected side effect) also occurs really often, then this makes sense (...) And you can see it (selected side effect) directly in the ranking.

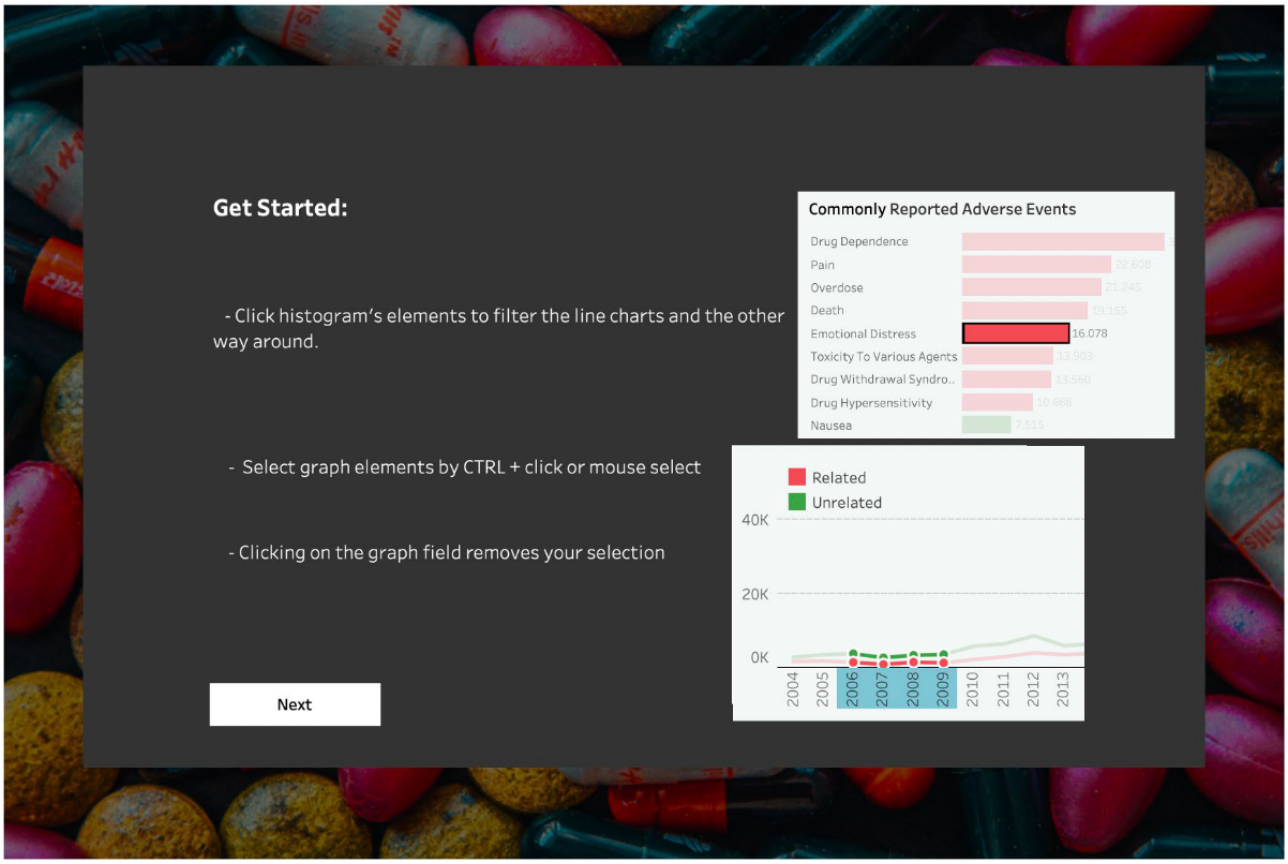
This feature was in line with DP3b, potentially enhancing the decision-making of professional users.

Iteration 2

Design and Development: Updating Prototype Features

We updated the system by implementing selected findings from the pilot evaluation session. Thus, we introduced a getting started tutorial with informative screenshots that guided users through the dashboard functionalities (Figure 7).

Figure 7. An exemplary page from the getting started tutorial in the v3-updated prototype.



XSL•FO
RenderX

To better afford evaluating alternatives of actions (DP3b), we introduced the selection-keeping feature. For a detailed explanation, refer to Figure S3 in Multimedia Appendix 3, where we present a real-world scenario demonstrating the comparison of 2 commonly used pain-relieving drugs: ibuprofen and acetaminophen (also known as paracetamol).

Demonstration and Evaluation: Usability Testing

We conducted usability tests with nonprofessional users (refer to Multimedia Appendix 4 for details). We found several usability issues that could hinder obtaining immediate practical benefits while using the dashboard (DP1a). We grouped the issues according to the affected dashboard features or elements, and we discussed them in the corresponding order.

The getting started tutorial was hard to follow and long. Users struggled to grasp how the individual screenshots relate to the dashboard concepts. This confusion arose because the screenshots presented isolated dashboard elements instead of the dashboard as a whole. The inability to skip or revisit the tutorial was also a major issue. Regarding the histogram and associated filter**,** users had trouble understanding how to select bars by clicking or by dragging sections. Finding the histogram's filter (Figure 5D) took more than 15 seconds, which could hinder identification of patterns and trends (DP1b), as failing to use the filter could impede the thorough exploration of data.

Usability problems also arose for year selection across different dashboard sections. Users didn't fully explore year selection options and found deselecting elements challenging (Multimedia Appendix 4). The absence of a year selection field in the overview section could impede exploration of data of varied granularity (DP2b) and, in consequence, interfere with monitoring the progression of relevant states (DP1c).

Some dashboard elements lacked visibility and clear explanations of their utility. For example, the event search field was not visible enough (Figure 6G) and the utility of the second line graph was not clear to the users. Additionally, the dashboard language was too technical. We also found issues with download buttons, DPA metrics files, and safety signal reports. Multiple download buttons and files made downloads difficult. Moreover, the download dialog box was located beyond eye reach and the system failed to inform about the download status. Finally, safety signal reports were too detailed and specific. According to C2, the report provided "unnecessary information overload" in which the participant was not interested.

End questions revealed users' confusion about the data source and the meaning behind the evaluation of side effects as related or unrelated to a drug by data mining algorithms. This lack of understanding could constrain evaluating alternatives of actions (DP3b). However, users found the system attractive for exploring drug safety information (DP3b). They were often encouraged by its interactivity to explore data from different angles (DP2a) and expressed their interest in using the system as an auxiliary drug information source. C4 suggested that people generally search online for drug safety information, seeking medical advice on websites of compromised quality. The participant considered the dashboard a suitable information source even for users with no pharmacovigilance background, as long as it is supplementary to professional medical advice. Similarly, C3 suggested the system could be useful for direct comparison of drugs with the same active substance produced by different manufacturers. C3 often asks for a cheaper drug alternative at the pharmacy and could use the system to compare the safety information of equivalent drugs.

Iteration 3

Design and Development: Updating Prototype Features

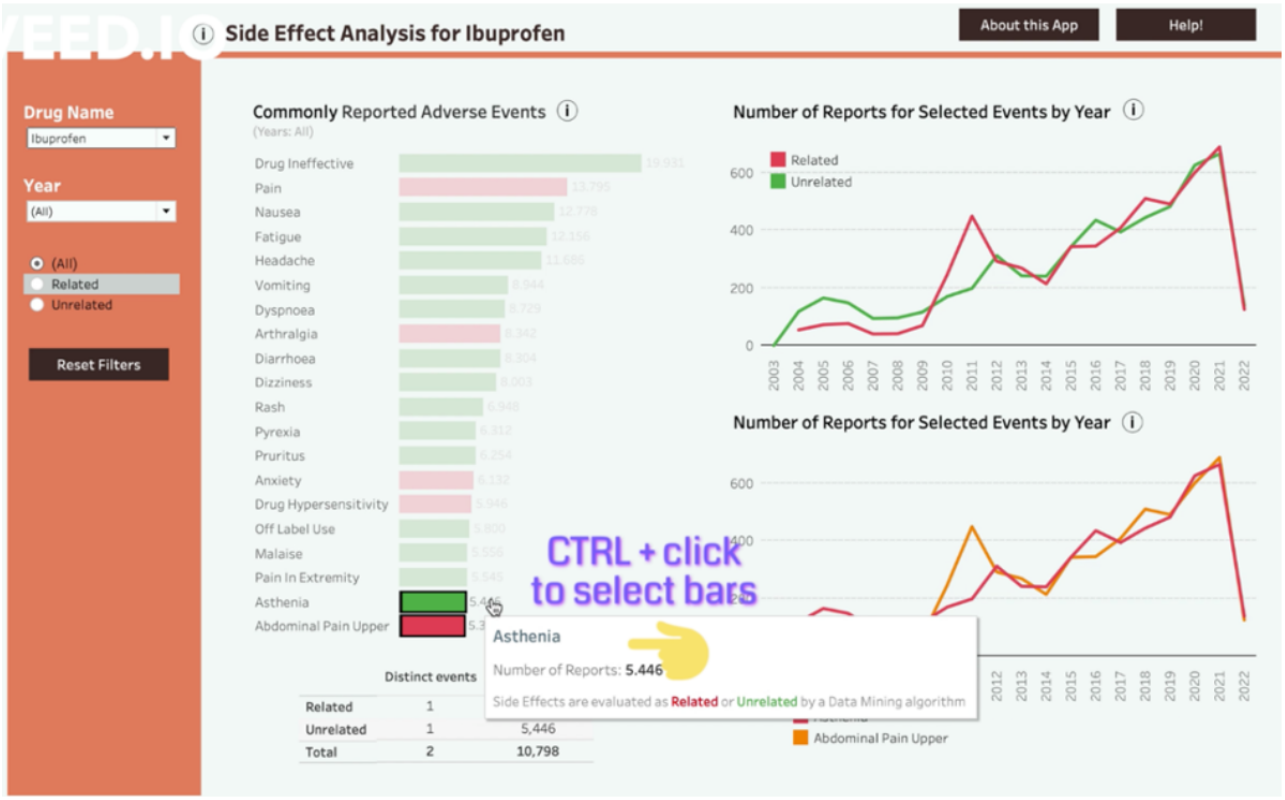
Introducing the Video Tutorial

To better afford obtaining immediate practical benefits while using the dashboard (DP1a), we modified the tutorial. Usability testing revealed that users learn best by doing, but they require basic knowledge to get started. Well-designed tutorials allow users to uncover the true potential of the app and understand the associated benefits [77]. Following the idea that undiscovered features practically do not exist, we replaced the tutorial with the animated walkthrough video.

We recorded key features in a brief video (more than 1 minute) with subtitles explaining dashboard actions. We used icons (eg, pointing finger and large mouse cursor) to guide users and added a skip video button for more experienced users (Figure S5 in Multimedia Appendix 3). Finally, we included a help button allowing users to rewatch the tutorial (Figure 8).

XSL•FO
RenderX

Figure 8. A frame from the getting started video tutorial of the v4-updated prototype.



Emphasizing the Utility of Dashboard Elements

We addressed many usability issues by adding short labels to dashboard elements (Figures 9 and 10).

To better afford monitoring the progression of relevant states (DP1c), we included an info icon next to graphs and section headers. Users could see information about graph functionality and displayed data, as well as the explanation of complex terms. We also included information on the content of the overview and detail sections and explained how graphs interacted with each other.

Similarly, we expected the issue of insufficient visibility and unclear utility of some elements (eg, the event search field and the second line graph) would be addressed by including additional information in the tutorial (Figures S6 and S7 in Multimedia Appendix 3). Finally, we implemented the "About this App" section (Figure 11), where users could find information about the data source, safety signal evaluation, possible limitations to the tool, and external links.

Considering all the above, we refined the DPs: (1) DP1a: it provides features to address the bootstrap problem, add walkthrough videos, and explain dashboard elements**,** so that the system affords immediate use of the dashboard in drug safety surveillance. (2) DP1c: it provides features to track changes in the variables and highlight their utility, so that the system affords tracking of the progression of relevant states in drug safety surveillance.

XSL•FO
RenderX

Figure 9. Histogram with the info icon. Prototype v4.

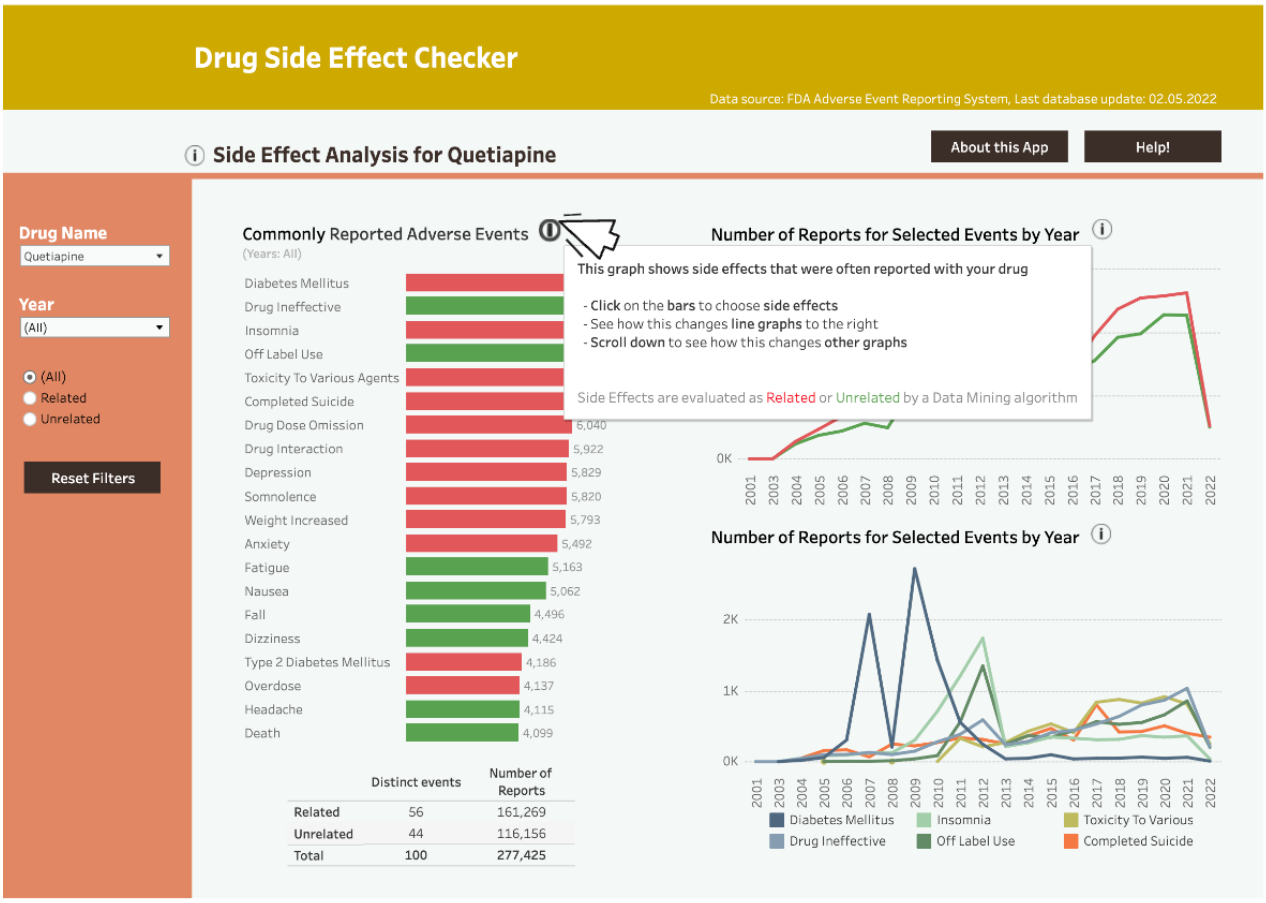
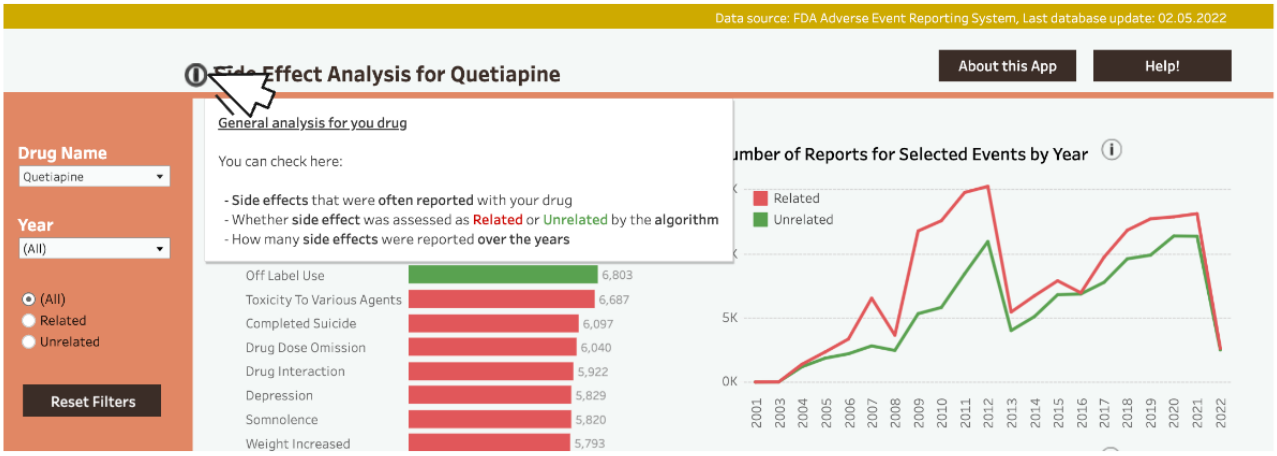
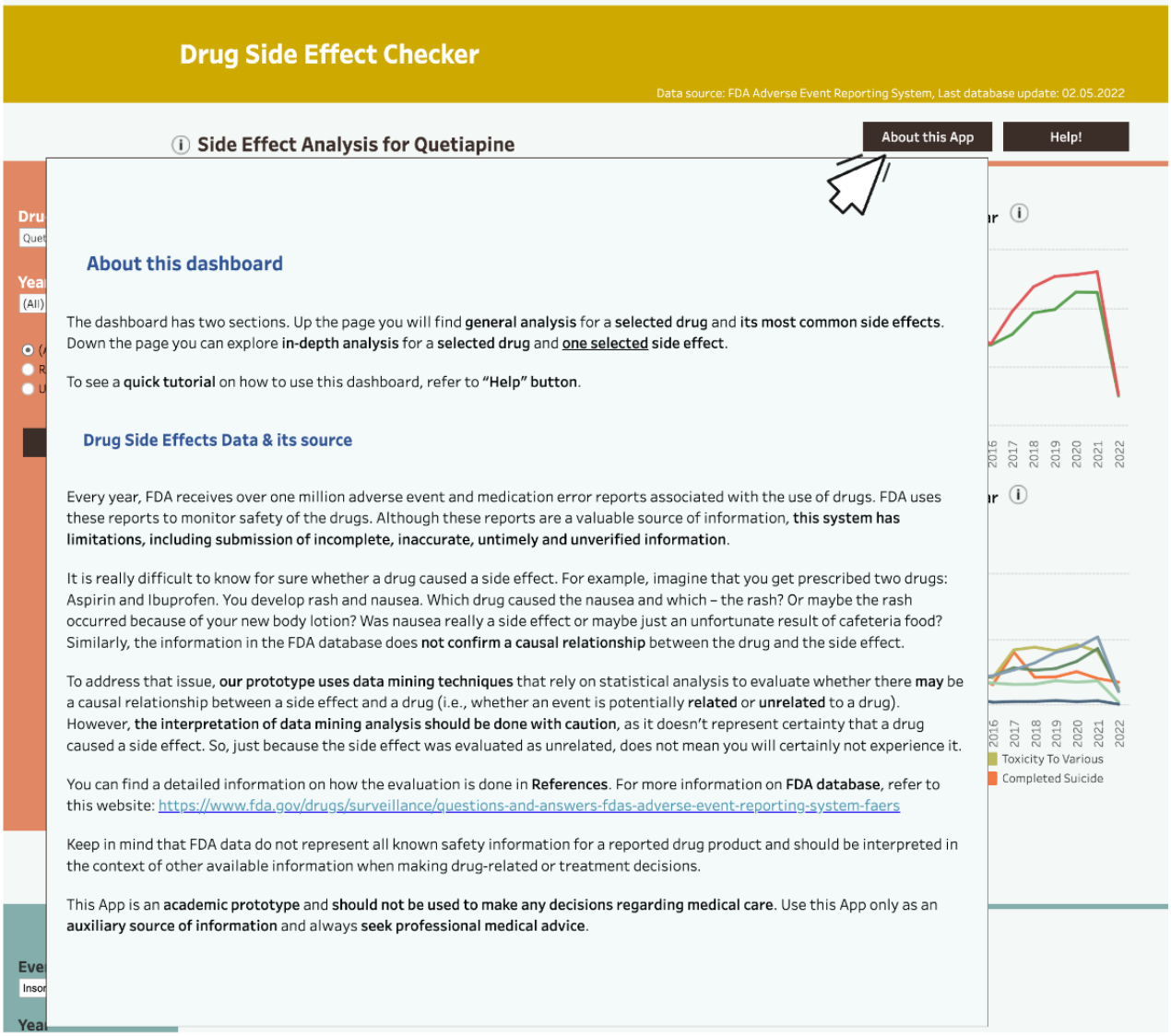


Figure 10. The overview section with the info icon. Prototype v4.



XSL•FO
RenderX

Figure 11. The About this App dashboard section. Prototype v4.



Signalizing the Uncertainty of Data Mining Results

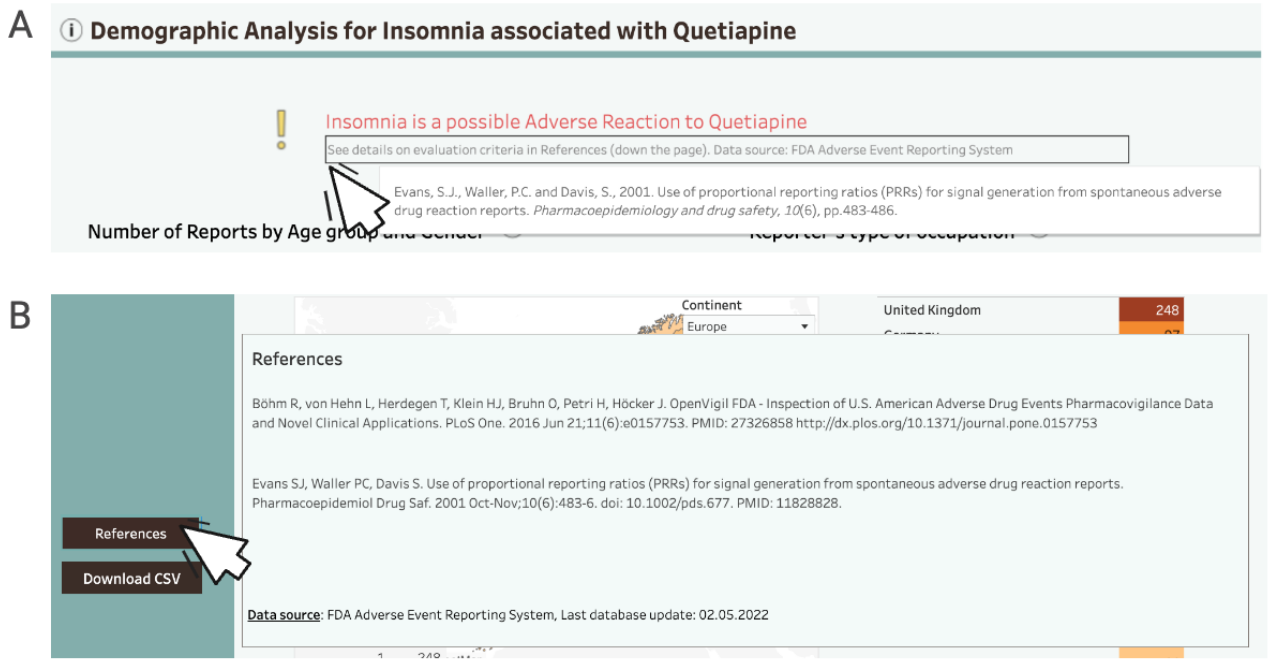
To better afford identification of patterns and trends (DP1b), we added the information on the uncertainty behind DPA evaluation and advised caution when interpreting the results. Since safety signal detection can be used for hypothesis building, but not testing, we found it crucial to highlight these concepts. Safety signal detection is a starting point for further investigations, not a conclusion. Considering ethical and safety concerns, appropriate warnings are necessary.

We addressed all the above in several ways. First, we provided the explanation on the interpretation of DPA in a disclaimer (not shown) and in the About this App section. Accounting for users less familiar with the topic, we provided an illustrative real-world case scenario to explain the concepts. We also implemented the References button, providing the links to evaluation criteria and including similar information in the warning message (Figure 12). Finally, we added links to external resources on many occasions and further enriched DPA files in explanation. In line with the above, we refined the following DP: DP1b: it provides features to describe possible causality between variables and signalize uncertainty behind DPA results in different ways, so that the system affords identification of patterns and trends in drug safety surveillance.

Together, these changes replace specialist terminology with plain-language labels and confine full statistical detail to downloadable files, reducing cognitive load for lay users while preserving transparency for professional review.

XSL•FO
RenderX

Figure 12. (A) The warning message with the information on hover. (B) The References button with the information on hover. Prototype v4.



Improving the Usability of Other Elements

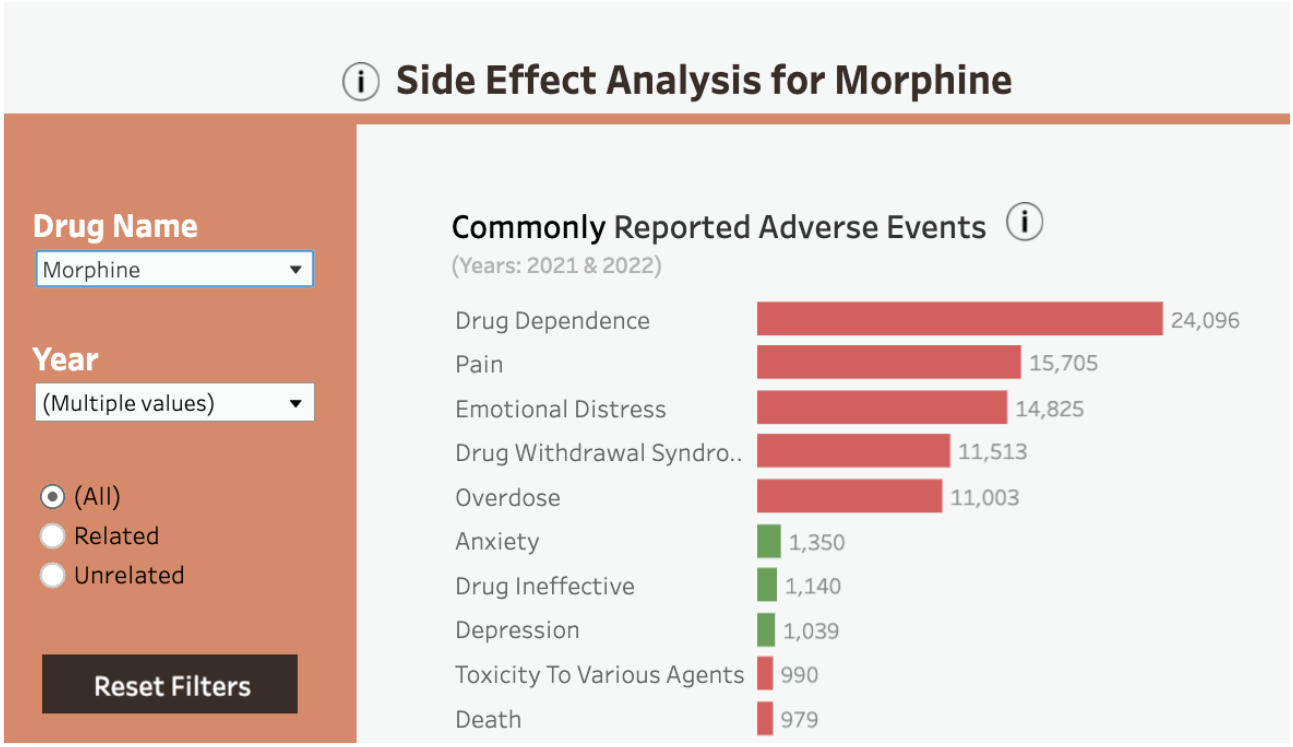
To address the usability issues related to the histogram, and to improve attention management while using the dashboard (DP2c), we relocated the histogram's filter to the upper-left corner, removed the confusing filter's name, and added a histogram's year filter with the information about the current year selection. We also implemented a Reset Filters button to unselect elements in the overview section (Figure 13).

To facilitate obtaining immediate practical benefits while using the dashboard (DP1a), we simplified the language and modified download options, removing redundant download buttons and files. With regard to evaluating alternatives of actions (DP3b), and considering different data granularity needs of users (DP2b), we removed safety signal reports from the background of the download page and simplified the view. Finally, we moved the download dialog box to the eye reach (Figure S8 in Multimedia Appendix 3).

The summary of all solutions implemented after usability testing is shown in Table S2 in Multimedia Appendix 4. The final version (v4) of the prototype was made available online [78]. We recommend using Google Chrome as a browser for a better experience.

XSL•FO
RenderX

Figure 13. Histogram with the side effect filter (All, Related, Unrelated), year filter, the Reset Filters button (bottom left), and information on current year selection (in fine print, gray). Prototype v4.



Demonstration and Evaluation: Heuristic Evaluation

We concluded the DSR cycle by performing heuristic evaluation (Figure 14). We identified 4 out of 10 usability factors with the lowest usability score: flexibility and efficiency of use (52%), user control and freedom (67%), visibility of system status (83%), and consistency and standards (83%). Two usability factors had the usability scores above 90%: recognition rather than recall (92%) and match between system and real world (93%), while 4 usability factors had the usability scores of 100%: aesthetic and minimalist design, orientation, spatial organization, and information coding, where the last 3 factors were related to dashboard visualization. The overall usability score was 84% (Table 5).

XSL•FO
RenderX

Figure 14. Usability scores per usability factor. Total is the overall usability score.

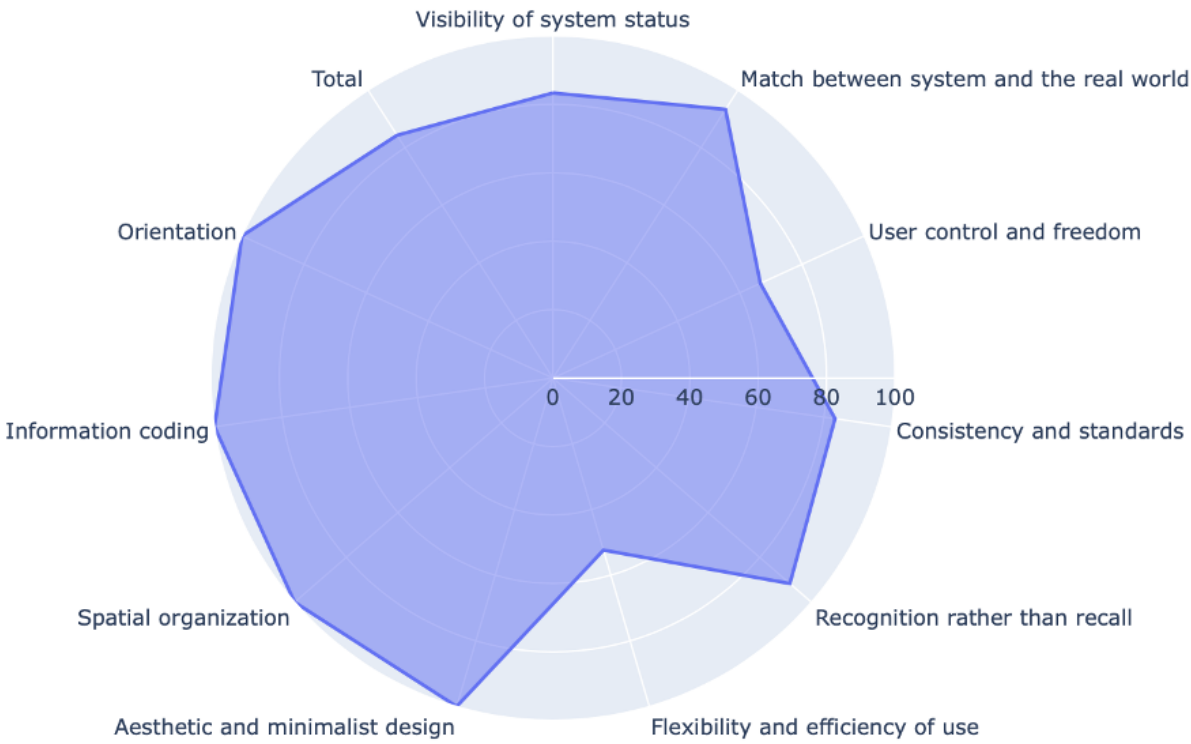


Table . A summary of usability factors with corresponding usability scores. Maximum score is a sum of close-ended questions per usability factor (eg, usability factor #9 had 2 questions, thus its maximum score was 2). E1, E2, and E3 are the scores assigned to each usability factor by each expert (ie, the sum of questions answered as yes by E1, E2, and E3). The result is the mean score divided by the maximum score, displayed as a percentage. Total is the overall usability score.

ID	Usability factor	Maximum score	E1	E2	E3	Mean score	Result (%)
1	Visibility of system status	6	5	5	5	5	83
2	Match between system and the real world	5	5	5	4	4.7	93
3	User control and freedom	5	4	3	3	3.3	67
4	Consistency and standards	6	5	4	6	5	83
5	Recognition rather than recall	4	4	3	4	3.7	92
6	Flexibility and efficiency of use	7	4	4	3	3.7	52
7	Aesthetic and minimalist design	7	7	7	7	7	100
8	Spatial organization	3	3	3	3	3	100
9	Information coding	2	2	2	2	2	100
10	Orientation	4	4	4	4	4	100
	Total	49	43	40	41	41	84

XSL•FO
RenderX

The overall severity ratings (Table 6) were in line with the obtained usability scores in most cases. Usability factor #3—user control and freedom was assigned the severity rating of 2 (minor usability problem), while the usability factor #1—visibility of system status was assigned the severity rating of 1 (cosmetic problem only) by all experts. Surprisingly, the usability factor #6—flexibility and efficiency of use, although it obtained the lowest overall usability score, was assigned the severity rating of 0 (no usability problem) by all experts. Several other usability factors were assigned the severity rating of 1 by some experts: usability factor #2—match between system and the real world (E3), usability factor #4—consistency and standards (E1), and usability factor #5—recognition rather than recall (E2). All remaining usability factors were evaluated as 0 (no usability problem).

The majority of usability issues referred to the flexibility and efficiency of use. We found evidence that the system lacked the ability to control display configurations and to hide unnecessary displays. Interestingly, although all participants noticed the lack of the option to hide unnecessary displays, it was not considered essential. According to E1: no buttons to hide or unhide panels, but not sure if the button is necessary. Information is well spread.

This could explain why this usability factor—although it obtained a relatively low usability score—was assigned an overall severity rating of 0 (no usability problem). Yet, as we saw a clear consensus regarding insufficient control of display configurations, we refined the following DP: DP2a: it provides features to support customized views and to control display configurations, so that the system affords examining different aspects of data in drug safety surveillance.

Table . Overall severity ratings assigned by experts (E1, E2, and E3) in heuristic evaluation.[a]

ID	Usability factor	Overall severity rating		
		E1	E2	E3
1	Visibility of system status	1	1	1
2	Match between system and the real world	0	0	1
3	User control and freedom	2	2	2
4	Consistency and standards	1	0	0
5	Recognition rather than recall	0	1	0
6	Flexibility and efficiency of use	0	0	0
7	Aesthetic and minimalist design	0	0	0
8	Spatial organization	0	0	0
9	Information coding	0	0	0
10	Orientation	0	0	0

[a] A score of 2: minor usability problem, 1: cosmetic problem only, 0: no usability problem.

Feedback also suggested that the system lacked clear exit options and a universal undo function, hindering user control and freedom. We also found support for the utility of the reset button; however, some participants noted that it did not deselect all elements when clicked. Regarding the visibility of system status**,** participants pointed out insufficient system feedback on its status. E3 identified a lack of visual feedback when selecting elements on some graphs. Although this action affected histograms, it was not communicated to the user. Similarly, E2 identified missing information on year selection in the detail section. Finally, concerning the consistency and standards**,** participants noted the absence of visual cues to identify active screens. E2 struggled to identify active screens when navigating between the main and download pages.

As a future agenda, we recommend implementing enhanced user control over display configurations, options to hide unnecessary displays, and adding a back or undo button. The functionality of the reset filters button should be improved to uniformly remove selections. Better feedback on the system status, like displaying selected years in the detail section, is crucial. Prompts and visual cues (eg, the event search box) should be placed where the eye is likely to be looking. Finally, all interactive elements (eg, the histogram and the first line graph) should provide feedback upon selection.

Discussion

DPs and the Value of Affordance Theory in This Study

Following the DSR methodology, we were able to empirically test and refine the initial, theory-driven DPs, by iteratively juxtaposing the results of experimental prototyping with theory-grounded, conceptual design recommendations. Consequently, we prove that the iterative translation of evaluation outcomes into prototype features results in refined design guidelines that could lead other practitioners in similar domains.

XSL•FO
RenderX

Throughout the study, we applied the affordance lens pragmatically: target affordances remained stable while prototype features were iteratively refined to instantiate them, which is in line with the view of Iivari [79] and Chandra et al [53]. Table 7 summarizes each principle in terms of user activity (affordance), enabling features (material properties), and the boundary conditions.

Table . The final list of refined design principles.

Design principle	Revised content
DP1a	Provide features to address the bootstrap problem, add walkthrough videos and explain dashboard elements, so that the system affords immediate use of the dashboard in drug safety surveillance
DP1b	Provide features to describe possible causality between variables and signalize uncertainty behind DPA[a] results in different ways, so that the system affords identification of patterns and trends in drug safety surveillance
DP1c	Provide features to track changes in the variables and highlight their utility, so that the system affords tracking of the progression of relevant states in drug safety surveillance
DP2a	Provide features to support customized views and to control display configurations, so that the system affords examining different aspects of data in drug safety surveillance
DP2b	Provide overview and details, while controlling overview-to-detail ratio, so that the system affords exploring data of different granularity levels in drug safety surveillance
DP2c	Provide features that stimulate attention, place important elements on the top-left area of the dashboard, and promote scrolling over pagination, so that the system affords improved attention management while using the dashboard in drug safety surveillance
DP3a	Provide features to create public values, so that the system affords users' engagement in dashboard use in drug safety surveillance
DP3b	Provide features to support decision-making and cluster them by the needs of distinct user groups, so that the system affords users to evaluate what-if scenarios and consider different alternatives in drug safety surveillance

[a]DPA: disproportionality analysis.

Addressing Existing Research Gaps

Our literature review identified pharmacovigilance systems that did not sufficiently incorporate user requirements, with resulting usability shortcomings [17,18,32]. We respond to this by using the DSR as a methodology. Additionally, this research gap was closely related to the bootstrap problem and DP1a (designing for direct use). We tackled the bootstrap problem in 2 ways: first, by including the requirements gathering phase in the design cycle, we increased the probability of addressing the real needs of users and therefore solving the right problem; second, by giving the users the opportunity to engage with the prototype in the initial stages of development, we initiated its use from the early versions. This positively affects the quality of the prototype and increases the chances of its successful adoption in the future. By iteratively responding to users' needs and targeting usability issues, we built a high-fidelity prototype that is usable for the evaluated tasks and user groups (overall usability score: 84%) and is designed to support sustained use; broader deployment remains future work.

We previously identified the lack of data mining techniques as a limitation of some of the systems for pharmacovigilance [17,21-23]. We suggested this was an important research gap to address, related directly to DP1b (causality between variables). We implemented DPA evaluation in the dashboard by including the information on potential safety signals [80]. We emphasized the uncertainty behind these results in multiple ways to make sure users were aware of their interpretation. Finally, the advantage of our prototype comes with improved visualization techniques. To that end, we followed the guidelines for visual dashboards by Bremser and Wagner [61] when deciding on visualization elements and features.

Consistent with DP3b ("cluster features by the needs of distinct user groups"), we view clinicians and pharmacists, researchers, and public users as distinct personas requiring layered functionality and language; our present DSR cycles cover the latter 2 groups (public users and research professionals), and clinician-facing tailoring should be undertaken in subsequent evaluations by others. In addition, regulatory decision-makers, policy analysts, and global health agencies represent further personas aligned with DP3b; though not evaluated here, they are recognized within our stakeholder scope and remain outside this study's boundary conditions.

Creating Public Value

Here, we articulate how the DPs and their instantiation can create public value when scaled and deployed in appropriate contexts; in this study, we assess feasibility and task-level usability rather than population-level impact.

XSL•FO
RenderX

We previously reviewed systems for pharmacovigilance that use proprietary databases or other data sources limited in diversity of stored information, volume, and availability [14-16,27,28]. To respond to this gap, related directly to DP3a (public value), we based the prototype on rich, open-access datasets with safety reports registered worldwide. In addition, both the code and the prototype are open-source and made available online, potentially contributing to the pharmacovigilance community. Moreover, through improved visualizations, we aim to enhance understanding and encourage users' active participation in interpreting the information.

To counter the misinformation during the COVID-19 pandemic, Murthy [81] recommends addressing information deficits where there is a high public interest but limited quality information available. Additionally, the author advocates for the modernization of public health communications by developing new, credible, and evidence-based online tools to effectively convey information to communities. By providing transparent and accurate drug safety data from reliable sources, our system may contribute to fighting misleading information and negative attitudes toward medication.

Our system is also a step forward in a higher degree of patient empowerment (the restoration of patients' access to public health data) and patient participation (the ability of patients to receive drug safety information and act upon it). In line with our findings, drug consumers could use the prototype for direct comparison of drugs with the same active compound, but produced by different manufacturers. Generic drugs (drug equivalents), although chemically the same as brand-name drugs (drugs produced by the pharmaceutical company that discovered them, protected by the patent), are often negatively perceived by drug consumers [82-84]. This is due to misbeliefs about brand-name drugs being better medicines per se.

There is strong evidence that, through a variety of incentives, medical doctors are influenced by pharmaceutical companies on which drugs to prescribe [85-88]. Physicians often enter contractual relationships with vendors, the latter being mostly representatives of brand-name drug companies, rather than cheaper generic drugs [89]. We suggest that by having a tool to directly compare primary safety profiles of drugs and their alternatives at hand, doctors may be encouraged to make more ethical decisions on their choice of collaboration with pharmaceutical companies.

Enhancing the Decision-Making

We integrated exportable DPA files into the prototype, considering them the main enabler of decision-making. When combined with data visualizations, this feature can significantly benefit pharmacovigilance professionals, including those affiliated with pharmaceutical industry or research institutes. These same affordances are pertinent to policy analysts and global health agencies, who require transparent, auditable summaries and cross-jurisdictional trend overviews. Detecting unusual or previously unknown drug-ADR associations is a first step in drug safety investigations [23,90]. Identified signals prompt further research, influencing actions such as product labelingchanges or potential drug withdrawal [91].

Regulatory bodies such as FDA, WHO, and EMA systematically monitor drug safety signals. Unusual associations undergo thorough scrutiny, involving literature reviews, clinical study revisions, and exploration of biochemical mechanisms. Results are compared with product information, and pharmaceutical companies are asked to provide relevant data, promoting higher industry accountability [91,92].

Contributions

Theoretical Contribution

Our work contributes to the prescriptive knowledge [93] by providing a list of refined, empirically-driven DPs. We suggest that, by delineating DPs, our work puts forward general design recommendations in similar solution domains. These can be extrapolated to a class of similar problems concerned with data visualization in pharmacovigilance. We do not claim the DPs are completely generalizable to other settings. Instead, they are put forward with the ambition to lead research on similar problems in wider contexts.

To our knowledge, there is no previous research that uses DSR methodology in designing dashboards for pharmacovigilance. By engaging in the DSR cycle, we were able to define affordances for similar systems, contributing to better understanding of the problem domain. These were (1) prompting users to engage with the dashboard (enabled by improved usability and attention management), (2) aiding analysis of what-if scenarios to support decision-making processes, (3) presenting data from different angles and at multiple granularity levels to examine its different facets, (4) showing changes in relevant variables over time, and (5) creating public value through generated insights.

Finally, we contribute to DSR by providing insights into how different requirements gathering and evaluation methods can support the development of visualization dashboards. While no claims can be made about the validity of these findings to other systems, we hope the techniques proposed here can be a useful instrument for identifying design issues early in the prototype development.

Practical Contribution

The practical contribution of our study lies in the developed artifact, publicly available online. Our system can be used by researchers and practitioners in pharmacovigilance, upon refinements. The code is open-source [94] and others can repurpose or reuse the scripts as partial solutions or inspiration for their projects. Though we confined our prototype to 5 drugs and 500 secondary events, the code can be adapted to an infinite number of drugs or secondary events to suit others' needs. For larger multisource deployments, the ETL and DPA steps can be executed in batches without altering the front end. Because the user interface consumes flat exports, the prototype can be reproduced in fully open-source dashboards or a custom front end.

The prototype is intended to demonstrate and operationalize the DPs. While usable for the studied tasks, broader public deployment requires the scale-up steps outlined in the next section. Any public-facing release by adopters would also

XSL•FO
RenderX

require clear plain-language disclaimers and risk communication for lay users, basic accessibility compliance, and a documented support pathway; these deployment activities are outside the scope of this study.

Our project however can be considered a hybrid open-source considering our reliance on the Tableau, which is a proprietary software. Table 8 provides an accurate description of the hybrid nature of our project with its respective components.

Finally, we do not claim full generalizability. The DPs are intended as reusable, prescriptive heuristics for similar dashboards within the stated boundary conditions; adaptation to clinician and pharmacist-facing workflows and additional user groups should be iteratively refined in those settings.

Table . Open-source status of project components.

Component	Status	Implication
Data	Open source	Fully accessible and usable by anyone
Code	Open source	All transformation and analysis code (eg, Python or R scripts) is on GitHub under an Open Source Initiative–approved license, allowing others to reuse and modify it
Dashboard prototype	Proprietary/Open access	The dashboard itself is viewable and accessible to anyone via Tableau Public (making it "open access"), but its dependency on proprietary Tableau software means others cannot easily reproduce, fork, or host a modified version without their own Tableau software or an equivalent proprietary viewer

Limitations and Future Research Agenda

This study has some limitations. First, drug safety surveillance can be viewed differently, leading to different affordances than those outlined here. Second, the DSR cycle could be extended to yet another phase. Iivari et al [95] proposes a framework for DPs evaluation in terms of their reusability: through the criteria of accessibility, importance, novelty, actability and effectiveness. Similarly, empirical testing of how revised DPs apply to other settings could be done through their repeated application in different boundary conditions [52]. Future research could address these issues to test the quality of the outcomes of this DSR project. Third, to support broader use, the prototype should be expanded to all FDA drugs and events via direct application programming interface calls [21].

Establishing readiness for broad public deployment of the prototype will require expanded datasets and multisite evaluations beyond the scope of this study. Moreover, including data from other databases (eg, EudraVigilance and VigiBase) would improve DPA evaluation, but requires (1) source-specific mapping into the common schema, (2) robust cross-repository deduplication, and (3) compliance with licensing and governance constraints. If further deployments are undertaken, evaluation can focus on user engagement, operational reliability, and decision support alignment with expert assessments, as these activities are beyond this study.

Relatedly, in the final prototype version, we keep shortcuts and advanced controls to a minimum to maintain clarity for nonprofessional users. This design choice explains the lower expert scores on flexibility, efficiency of use, and user control and freedom, which were nevertheless judged as low-severity issues. For expert-only deployments, future implementers can add more controls and configuration.

Importantly, our participant sample represented young and digitally proficient (late 20s to mid 40s), and several nonexpert participants held technical or professional roles (eg, data science, biomedical engineering, and humanitarian logistics). This sampling may limit generalizability to older adults and individuals with lower digital health literacy [96]. Older adults are more likely to experience polypharmacy [97] and may have distinct interaction needs (eg, larger text, higher contrast, simplified and error-tolerant interactions) [98], which we did not assess in this study. Future independent evaluations should specifically include older adults and lower-literacy users and consider accessibility guidance (eg, larger typographic scales, high-contrast palettes, and keyboard navigability) when adapting the dashboard for these groups.

Relatedly, some nonprofessional users found data mining terms challenging even when aided by tooltips and tutorials. The present prototype mitigates this by using plain-language labels, info icons, an "About this App" section, and by keeping advanced formulas in downloadable files; however, residual risk of misinterpretation remains. We therefore reiterate that DPA supports hypothesis generation rather than confirmation and should be interpreted cautiously by lay audiences. To reduce potential user anxiety, we emphasize this uncertainty in plain language and signpost users to seek professional advice for individual decisions. However, we did not evaluate affective outcomes (eg, worry and decisional conflict) among lay users; given the potential for spontaneous report dashboards to provoke anxiety, such outcomes should be assessed in independent evaluations by others.

Furthermore, it would be beneficial to implement a higher variety of data sources: scientific literature, clinical studies or data from text mining [20]. Also, the prototype would benefit from comprehensive benchmarking, such as head-to-head comparisons with existing pharmacovigilance tools on performance, cost, or user satisfaction. Future work could also

XSL•FO
RenderX

expand DPA methods to improve the validity of detected safety signals. Integrating and systematically evaluating additional machine learning approaches for ADR signal detection could enhance analytical breadth and validate performance across settings.

Finally, integration into health care IIs (eg, electronic health records) was outside the scope of this DSR prototype. The artifact targets public FDA data and does not process identifiable patient records. Any operational deployment by adopters would require institutional governance (eg, data use agreements, access control, auditability, and security increase) and, where applicable, compliance with jurisdictional regulations. The use of standard terminologies, decoupled Python scripts, and portability to open source (Multimedia Appendix 2) is intended to ease such translation, but we do not evaluate these deployment aspects here.

Conclusion

In this study, we applied the DSR cycle to investigate the design of visual dashboards to support activities within pharmacovigilance. To that end, we identified required affordances, developed corresponding DPs that satisfied them, and demonstrated their implementation, showing that they brought about expected outcomes. The developed DPs are anticipated to be applicable in various pharmacovigilance settings. Finally, we hope our research will be a catalyst of further research in similar domains.

Acknowledgments

All figures were created with BioRender [98].

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available in the AdverseDrugEventPipeline repository [93].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics of the target groups.
[DOCX File, 15 KB - medinform_v14i1e75936_app1.docx]

Multimedia Appendix 2
Design science research approach details.
[DOCX File, 20 KB - medinform_v14i1e75936_app2.docx]

Multimedia Appendix 3
Prototype in Tableau.
[DOCX File, 6699 KB - medinform_v14i1e75936_app3.docx]

Multimedia Appendix 4
Usability testing and an overview of solutions implemented.
[DOCX File, 22 KB - medinform_v14i1e75936_app4.docx]

References

1. Das S. Predicting unknown adverse drug reactions using an unsupervised node embedding algorithm. : Massachusetts Institute of Technology; 2019 URL: https://dspace.mit.edu/handle/1721.1/124239 [accessed 2026-01-21]
2. Adedapo ADA, Adedeji WA, Adedapo IA, Adedapo KS. Cohort study on adverse drug reactions in adults admitted to the medical wards of a tertiary hospital in Nigeria: prevalence, incidence, risk factors and fatality. Br J Clin Pharmacol 2021 Apr;87(4):1878-1889. [doi: 10.1111/bcp.14577] [Medline: 32991765]
3. Giardina C, Cutroneo PM, Mocciaro E, et al. Adverse drug reactions in hospitalized patients: results of the FORWARD (Facilitation of Reporting in Hospital Ward) study. Front Pharmacol 2018;9:350. [doi: 10.3389/fphar.2018.00350] [Medline: 29695966]

XSL•FO
RenderX

4. Zaidi AS, Peterson GM, Bereznicki LRE, Curtain CM, Salahudeen MS. Ten-year trends in adverse drug reaction-related hospitalizations among people with dementia. Ther Adv Drug Saf 2022;13:20420986221080796. [doi: 10.1177/20420986221080796] [Medline: 35295667]
5. World Health Organization. In: Takanashi F, editor. WHO Pharmaceuticals Newsletter - No 1, 2022: World Health Organization; 2022:978-992.
6. Slight SP, Seger DL, Franz C, Wong A, Bates DW. The national cost of adverse drug events resulting from inappropriate medication-related alert overrides in the United States. J Am Med Inform Assoc 2018 Sep 1;25(9):1183-1188. [doi: 10.1093/jamia/ocy066] [Medline: 29939271]
7. Field TS, Gilman BH, Subramanian S, Fuller JC, Bates DW, Gurwitz JH. The costs associated with adverse drug events among older adults in the ambulatory setting. Med Care 2005 Dec;43(12):1171-1176. [doi: 10.1097/01.mlr.0000185690.10336.70] [Medline: 16299427]
8. Al Khaja KAJ, AlKhaja AK, Sequeira RP. Drug information, misinformation, and disinformation on social media: a content analysis study. J Public Health Policy 2018 Aug;39(3):343-357. [doi: 10.1057/s41271-018-0131-2] [Medline: 29795521]
9. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
10. Daei A, Soleymani MR, Ashrafi-Rizi H, Zargham-Boroujeni A, Kelishadi R. Clinical information seeking behavior of physicians: a systematic review. Int J Med Inform 2020 Jul;139:104144. [doi: 10.1016/j.ijmedinf.2020.104144] [Medline: 32334400]
11. O'Mathúna DP. How should clinicians engage with online health information? AMA J Ethics 2018 Nov 1;20(11):E1059-E1066. [doi: 10.1001/amajethics.2018.1059] [Medline: 30499435]
12. Dorfman S. What physicians want and need: insights from CMI/compas media vitals research. PM360. 2019. URL: https://www.pm360online.com/what-physicians-want-and-need-insights-from-cmi-compas-media-vitals-research/ [accessed 2026-01-21]
13. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 2;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
14. Duke J, Faiola A, Kharrazi H. A novel visualization tool for evaluating medication side-effects in multi-drug regimens. In: Jacko JA, editor. Presented at: Human-Computer Interaction Interacting in Various Application Domains: 13th International Conference, HCI International 2009; Jul 19-24, 2009. [doi: 10.1007/978-3-642-02583-9_52]
15. Wei JX, Zhu YX, Sun J, Xu HM, Li M, Sun YH. ADRVis: an information visualization platform for adverse drug reactions. IJUNESST 2015 Oct 31;8(10):139-150. [doi: 10.14257/ijunesst.2015.8.10.14]
16. Shoshi A, Hoppe T, Kormeier B, Ogultarhan V, Hofestädt R. GraphSAW: a web-based system for graphical analysis of drug interactions and side effects using pharmaceutical and molecular data. BMC Med Inform Decis Mak 2015 Feb 28;15:15. [doi: 10.1186/s12911-015-0139-5] [Medline: 25881043]
17. Kumar A. The newly available FAERS public dashboard: implications for health care professionals. Hosp Pharm 2019 Apr;54(2):75-77. [doi: 10.1177/0018578718795271] [Medline: 30923396]
18. Vingen D, Andrews EJ, Ferati M. Usability in patient-oriented drug interaction checkers—a Scandinavian sampling and heuristic evaluation. Informatics (MDPI) 2020;7(4):42. [doi: 10.3390/informatics7040042]
19. Fukazawa C, Hinomura Y, Kaneko M, Narukawa M. Significance of data mining in routine signal detection: analysis based on the safety signals identified by the FDA. Pharmacoepidemiol Drug Saf 2018 Dec;27(12):1402-1408. [doi: 10.1002/pds.4672] [Medline: 30324671]
20. Duggirala HJ, Tonning JM, Smith E. Data mining at FDA – white paper. US Food and Drug Administration. 2018. URL: https://www.fda.gov/media/91848/download [accessed 2026-01-21]
21. FDA aes reporting system (FAERS) public dashboard. Food and Drug Administration. 2017 Sep. URL: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm [accessed 2022-01-16]
22. VigiBase - the WHO global database. Uppsala Monitoring Centre. 2021 Dec. URL: https://who-umc.org/vigibase/vigibase-services/ [accessed 2022-08-07]
23. EudraVigilance. European Medicines Agency. 2022 Jun. URL: https://www.ema.europa.eu/en/human-regulatory/research-development/pharmacovigilance/eudravigilance [accessed 2022-08-08]
24. Duggirala HJ, Tonning JM, Smith E, et al. Use of data mining at the Food and Drug Administration. J Am Med Inform Assoc 2016 Mar;23(2):428-434. [doi: 10.1093/jamia/ocv063] [Medline: 26209436]
25. Guide on the interpretation of spontaneous case reports of suspected adverse reactions to medicines. inspections, human medicines pharmacovigilance and committees. European Medicines Agency. 2011. URL: https://www.ema.europa.eu/en/documents/report/guide-interpretation-spontaneous-case-reports-suspected-adverse-reactions-medicines_en.pdf [accessed 2026-01-21]
26. Signal management at UMC. Uppsala Monitoring Centre. 2022 Mar 23. URL: https://who-umc.org/signal-work/signal-detection/ [accessed 2022-03-23]
27. Roosan D, Li Y, Law A, et al. Improving medication information presentation through interactive visualization in mobile apps: human factors design. JMIR Mhealth Uhealth 2019 Nov 25;7(11):e15940. [doi: 10.2196/15940] [Medline: 31763991]

XSL•FO
RenderX

28. Sorbello A, Ripple A, Tonning J, et al. Harnessing scientific literature reports for pharmacovigilance. Prototype software analytical tool development and usability testing. Appl Clin Inform 2017 Mar 22;8(1):291-305. [doi: 10.4338/ACI-2016-11-RA-0188] [Medline: 28326432]
29. Defer G, Le Caignec F, Fedrizzi S, et al. Dedicated mobile application for drug adverse reaction reporting by patients with relapsing remitting multiple sclerosis (Vigip-SEP study): study protocol for a randomized controlled trial. Trials 2018 Mar 9;19(1):174. [doi: 10.1186/s13063-018-2560-4] [Medline: 29523169]
30. Delio J, Catalanotti JS, Marko K, et al. Integrating adverse event reporting into a free-text mobile application used in daily workflow increases adverse event reporting by physicians. Am J Med Qual 2020;35(5):374-379. [doi: 10.1177/1062860619891995] [Medline: 31795736]
31. Rubin DS, Pesyna C, Jakubczyk S, Liao C, Tung A. Introduction of a mobile adverse event reporting system is associated with participation in adverse event reporting. Am J Med Qual 2019;34(1):30-35. [doi: 10.1177/1062860618781920] [Medline: 29938520]
32. Corrao NJ, Robinson AG, Swiernik MA, Naeim A. Importance of testing for usability when selecting and implementing an electronic health or medical record system. J Oncol Pract 2010 May;6(3):120-124. [doi: 10.1200/JOP.200017] [Medline: 20808553]
33. Bae JH, Baek YH, Lee JE, Song I, Lee JH, Shin JY. Machine learning for detection of safety signals from spontaneous reporting system data: example of nivolumab and docetaxel. Front Pharmacol 2020;11:602365. [doi: 10.3389/fphar.2020.602365] [Medline: 33628176]
34. Lee H, Kim JH, Choe YJ, Shin JY. Safety surveillance of pneumococcal vaccine using three algorithms: disproportionality methods, empirical Bayes geometric mean, and tree-based scan statistic. Vaccines (Basel) 2020 May 22;8(2):242. [doi: 10.3390/vaccines8020242] [Medline: 32456068]
35. Omar NE, Fahmy Soliman AI, Eshra M, Saeed T, Hamad A, Abou-Ali A. Postmarketing safety of anaplastic lymphoma kinase (ALK) inhibitors: an analysis of the FDA Adverse Event Reporting System (FAERS). ESMO Open 2021 Dec;6(6):100315. [doi: 10.1016/j.esmoop.2021.100315] [Medline: 34864500]
36. Xiao C, Li Y, Baytas IM, Zhou J, Wang F. An MCEM framework for drug safety signal detection and combination from heterogeneous real world evidence. Sci Rep 2018 Jan 29;8(1):1806. [doi: 10.1038/s41598-018-19979-7] [Medline: 29379048]
37. Böhm R. Primer on disproportionality analysis. OpenVigil. 2018 Oct. URL: http://openvigil.sourceforge.net/doc/DPA.pdf [accessed 2022-02-07]
38. Caster O, Sandberg L, Bergvall T, Watson S, Norén GN. vigiRank for statistical signal detection in pharmacovigilance: first results from prospective real‐world use. Pharmacoepidemiol Drug Saf 2017 Aug;26(8):1006-1010 [FREE Full text] [doi: 10.1002/pds.4247]
39. Dimitriadis VK, Gavriilidis GI, Natsiavas P. Pharmacovigilance and clinical environment: utilizing OMOP-CDM and OHDSI software stack to integrate EHR data. Stud Health Technol Inform 2021 May 27;281:555-559. [doi: 10.3233/SHTI210232] [Medline: 34042637]
40. Budha MS, Khan J, Mujawar DT. Review on pharmacovigilance method, recent developments, future perspective and software. IJRASET 2024;12(2):251-255 [FREE Full text] [doi: 10.22214/ijraset.2024.58231]
41. Pilipiec P, Liwicki M, Bota A. Using machine learning for pharmacovigilance: a systematic review. Pharmaceutics 2022 Jan 23;14(2):266. [doi: 10.3390/pharmaceutics14020266] [Medline: 35213998]
42. Li Y, Tao W, Li Z, et al. Artificial intelligence-powered pharmacovigilance: a review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. J Biomed Inform 2024 Apr;152:104621. [doi: 10.1016/j.jbi.2024.104621] [Medline: 38447600]
43. R P, Andrews S D, GS N. Advancing drug safety surveillance: a review of AI techniques and strategies for improving pharmacovigilance and adverse event reporting. Int J Res Publ Rev 2024;5(5):13348-13351. [doi: 10.55248/gengpi.5.0524.1468]
44. Yu HM, Huang YM, Xiao J, et al. Establishment of a visualization platform for ADR query and analysis: an example of severe skin adverse reactions caused by sulfonylureas. Eur J Clin Pharmacol 2025 Jul;81(7):1055-1067. [doi: 10.1007/s00228-025-03842-6]
45. Ansari B, Martin EG. Integrating human-centered design in public health data dashboards: lessons from the development of a data dashboard of sexually transmitted infections in New York State. J Am Med Inform Assoc 2024 Jan 18;31(2):298-305. [doi: 10.1093/jamia/ocad102] [Medline: 37330670]
46. Malkani M, Madan E, Malkani D, et al. Rank ordered design attributes for health care dashboards including artificial intelligence: usability study. Online J Public Health Inform 2024 Nov 20;16(1):e58277. [doi: 10.2196/58277] [Medline: 39566038]
47. Gavriilidis GI, Dimitriadis VK, Jaulent MC, Natsiavas P. Identifying actionability as a key factor for the adoption of “Intelligent” systems for drug safety: lessons learned from a user-centred design approach. Drug Saf 2021 Nov;44(11):1165-1178. [doi: 10.1007/s40264-021-01103-w] [Medline: 34674190]
48. Murphy DR, Savoy A, Satterly T, Sittig DF, Singh H. Dashboards for visual display of patient safety data: a systematic review. BMJ Health Care Inform 2021 Oct;28(1):e100437. [doi: 10.1136/bmjhci-2021-100437] [Medline: 34615664]

XSL•FO
RenderX

49. Gibson JJ. The theory of affordances. In: Shaw R, Bransford J, editors. Perceiving, Acting, and Knowing: Toward an Ecological Psychology: Lawrence Erlbaum Associates; 1977:67-82.
50. Pan SL, Li M, Pee LG, Sandeep MS. Sustainability design principles for a wildlife management analytics system: an action design research. Eur J Inf Syst 2021 Jul 4;30(4):452-473. [doi: 10.1080/0960085X.2020.1811786]
51. Maier JRA, Fadel GM. Affordance based design: a relational theory for design. Res Eng Design 2009 Mar;20(1):13-27. [doi: 10.1007/s00163-008-0060-3]
52. Seidel S, Chandra Kruse L, Székely N, Gau M, Stieger D. Design principles for sensemaking support systems in environmental sustainability transformations. Eur J Inf Syst 2018 Mar 4;27(2):221-247. [doi: 10.1057/s41303-017-0039-0]
53. Chandra L, Seidel S, Gregor S. Prescriptive knowledge in IS research: conceptualizing design principles in terms of materiality, action, and boundary conditions. Presented at: 2015 48th Hawaii International Conference on System Sciences (HICSS); HI, USA p. 4039-4048. [doi: 10.1109/HICSS.2015.485]
54. Sein MK, Henfridsson O, Purao S, Rossi M, Lindgren R. Action Design Research1. MIS Q 2011 Mar 1;35(1):37-56. [doi: 10.2307/23043488]
55. Baskerville R, Baiyere A, Gergor S, et al. Design science research contributions: finding a balance between artifact and theory. J Assoc Iinform Syst 2018;19(5):358-376. [doi: 10.17705/1jais.00495]
56. Aanestad M, Jensen TB. Building nation-wide information infrastructures in healthcare through modular implementation strategies. J Strateg Inf Syst 2011 Jun;20(2):161-176. [doi: 10.1016/j.jsis.2011.03.006]
57. Hanseth O, Lyytinen K. Design theory for dynamic complexity in information infrastructures: the case of building internet. J Inf Technol 2010 Mar;25(1):1-19. [doi: 10.1057/jit.2009.19]
58. Recker J. Improving the state-tracking ability of corona dashboards. Eur J Inf Syst 2021 Sep 3;30(5):476-495. [doi: 10.1080/0960085X.2021.1907235]
59. Matheus R, Janssen M, Maheshwari D. Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. Gov Inf Q 2020 Jul;37(3):101284. [doi: 10.1016/j.giq.2018.01.006]
60. Toreini P, Langner M, Maedche A, et al. Designing attentive information dashboards. J Assoc Inform Syst 2022;22(2):521-552. [doi: 10.17705/1jais.00732]
61. Bremser W, Wagner WP. Developing dashboards for performance management. CPA J 2013;83(7):62-67.
62. Thyvalikakath TP, Monaco V, Thambuganipalle H, Schleyer T. Comparative study of heuristic evaluation and usability testing methods. Stud Health Technol Inform 2009;143:322-327. [Medline: 19380955]
63. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research, 2nd edition: Morgan Kaufmann; 2016:356.
64. Dowding D, Merrill JA. The development of heuristics for evaluation of dashboard visualizations. Appl Clin Inform 2018 Jul;9(3):511-518. [doi: 10.1055/s-0038-1666842] [Medline: 29998455]
65. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst 2007 Dec;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
66. Beard L, Aghassibake N. Tableau (version 2020.3). J Med Libr Assoc 2021;109(1). [doi: 10.5195/jmla.2021.1135]
67. Statista OTC pharmaceuticals - worldwide. Statista. 2022. URL: https://www.statista.com/outlook/hmo/otc-pharmaceuticals/worldwide?srsltid=AfmBOoqOJV3dvb5VZaEczqUObuIS58T7fc9X1lQ43jIXDhkHj97tbJcK [accessed 2026-01-22]
68. Liang S, Fu Y. Voice meeting notes & real-time transcription. Otter.ai. 2016. URL: Otter.ai [accessed 2026-01-22]
69. Khanbhai M, Symons J, Flott K, et al. Enriching the value of patient experience feedback: web-based dashboard development using co-design and heuristic evaluation. JMIR Hum Factors 2022 Feb 3;9(1):e27887. [doi: 10.2196/27887] [Medline: 35113022]
70. Nielsen J. Finding usability problems through heuristic evaluation. 1992 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92. [doi: 10.1145/142750.142834]
71. Nielsen J. Usability Inspection Methods: John Wiley and Sons; 1994.
72. Georgsson M, Staggers N, Weir C. A modified user-oriented heuristic evaluation of a mobile health system for diabetes self-management support. Comput Inform Nurs 2016 Feb;34(2):77-84. [doi: 10.1097/CIN.0000000000000209] [Medline: 26657618]
73. Snyder ME, Jaynes H, Gernant SA, et al. Alerts for community pharmacist-provided medication therapy management: recommendations from a heuristic evaluation. BMC Med Inform Decis Mak 2019 Jul 16;19(1):135. [doi: 10.1186/s12911-019-0866-0] [Medline: 31311532]
74. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
75. Rosemberg C. Turning usability testing data into action. Toptal. 2017 Apr 20. URL: https://www.toptal.com/designers/usability-testing/turning-usability-testing-data-into-action [accessed 2022-09-10]
76. Sveriges riksdag. Lag (2003:460) om etikprövning av forskning som avser människor [swedish ethical review act]. SFS 2003:460. 2003. URL: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/lag-2003460-om-etikprovning-av-forskning-som_sfs-2003-460/ [accessed 2026-02-27]

XSL•FO
RenderX

77. Balboni K. Creating app walkthroughs (+ 4 walkthrough examples). Appcues. 2022 Apr 20. URL: https://www.appcues.com/blog/the-5-best-walkthrough-examples [accessed 2022-09-05]
78. Drug side effect checker. Tableau Public. 2022. URL: https://public.tableau.com/app/profile/cl.udio.pires4914/viz/DrugSideEffectChecker/DrugSideEffectChecker [accessed 2026-01-22]
79. Iivari J. A paradigmatic analysis of information systems as a design science. Scand J Inf Syst 2007;19:39-64 [FREE Full text]
80. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001;10(6):483-486. [doi: 10.1002/pds.677] [Medline: 11828828]
81. Murthy V. Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment. : Office of the Surgeon General (U.S.); 2021. [Medline: 34283416]
82. Sanchez CK. Patient perceptions of generic drugs: dispelling misconceptions. US Pharm 2016;41(6):36-41 [FREE Full text]
83. Dunne SS, Dunne CP. What do people really think of generic medicines? A systematic review and critical appraisal of literature on stakeholder perceptions of generic drugs. BMC Med 2015 Jul 29;13:173. [doi: 10.1186/s12916-015-0415-3] [Medline: 26224091]
84. Aivalli PK, Elias MA, Pati MK, et al. Perceptions of the quality of generic medicines: implications for trust in public services within the local health system in Tumkur, India. BMJ Glob Health 2017;2(Suppl 3):e000644. [doi: 10.1136/bmjgh-2017-000644] [Medline: 29531844]
85. Fickweiler F, Fickweiler W, Urbach E. Interactions between physicians and the pharmaceutical industry generally and sales representatives specifically and their association with physicians' attitudes and prescribing habits: a systematic review. BMJ Open 2017 Sep 27;7(9):e016408. [doi: 10.1136/bmjopen-2017-016408] [Medline: 28963287]
86. DeJong C, Aguilar T, Tseng CW, Lin GA, Boscardin WJ, Dudley RA. Pharmaceutical industry-sponsored meals and physician prescribing patterns for medicare beneficiaries. JAMA Intern Med 2016 Aug 1;176(8):1114-1122. [doi: 10.1001/jamainternmed.2016.2765] [Medline: 27322350]
87. Wazana A. Physicians and the pharmaceutical industry: is a gift ever just a gift? JAMA 2000 Jan 19;283(3):373-380. [doi: 10.1001/jama.283.3.373] [Medline: 10647801]
88. Mitchell AP, Trivedi NU, Gennarelli RL, et al. Are financial payments from the pharmaceutical industry associated with physician prescribing?: a systematic review. Ann Intern Med 2021 Mar;174(3):353-361. [doi: 10.7326/M20-5665] [Medline: 33226858]
89. Lieb K, Scheurich A. Contact between doctors and the pharmaceutical industry, their perceptions, and the effects on prescribing habits. PLoS One 2014;9(10):e110130. [doi: 10.1371/journal.pone.0110130] [Medline: 25330392]
90. Analyse your data in Vigilyze. Uppsala Monitoring Centre. 2022 May 13. URL: https://who-umc.org/pv-products/vigilyze/ [accessed 2022-08-08]
91. Regulation and prequalification. World Health Organization. 2021 Dec 21. URL: https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance [accessed 2022-08-07]
92. Ventola CL. Big data and pharmacovigilance: data mining for adverse drug events and interactions. P T 2018 Jun;43(6):340-351. [Medline: 29896033]
93. Gregor S. The nature of theory in information systems1. MIS Q 2006 Sep 1;30(3):611-642. [doi: 10.2307/25148742]
94. AdverseDrugEventPipeline. GitHub. 2022. URL: https://github.com/Malwoiniak/AdverseDrugEventPipeline [accessed 2025-11-15]
95. Iivari J, Rotvit Perlt Hansen M, Haj-Bolouri A. A proposal for minimum reusability evaluation of design principles. Eur J Inf Syst 2021 May 4;30(3):286-303. [doi: 10.1080/0960085X.2020.1793697]
96. European Commission. Digital Decade: Report on the State of the Digital Decade 2023: Publications Office of the European Union; 2023. [doi: 10.2759/318547]
97. OECD. Health at a Glance 2023: OECD Indicators: OECD Publishing; 2023. [doi: 10.1787/7a7afb35-en]
98. Budiu R, Pernice K. UX guidelines for older adults (ages 65 and older). : Nielsen Norman Group; 2024 URL: https://www.nngroup.com/reports/senior-citizens-on-the-web/ [accessed 2026-01-31]
99. BioRender. BioRender. URL: https://www.biorender.com [accessed 2026-02-18]

Abbreviations

ADR: adverse drug reaction
AI: artificial intelligence
DiP: dashboard in pharmacovigilance
DP: design principle
DPA: disproportionality analysis
DSR: design science research
EMA: European Medicines Agency
FAERS: Food and Drug Administration Adverse Events Reporting System

XSL•FO
RenderX

FDA: Food and Drug Administration
HCI: human-computer interaction
II: information infrastructure
OHDSI: Observational Health Data Sciences and Informatics
ROI: return on investment
UMC: Uppsala Monitoring Centre
WHO: World Health Organization

Edited by C Perrin; submitted 13.Apr.2025; peer-reviewed by A Azizi, D Helminski, T Lima; accepted 03.Jan.2026; published 27.Feb.2026.

Please cite as:
Kotowicz M, Ferati M, Chowdhury S, Martins Pires C
Design Principles for Interactive Dashboards in Drug Safety Surveillance: Design Science Research
JMIR Med Inform 2026;14:e75936
URL: https://medinform.jmir.org/2026/1/e75936
doi:10.2196/75936

© Malwina Kotowicz, Mexhid Ferati, Soumitra Chowdhury, Cláudio Martins Pires. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Data-Centric Approach for Health Care and Research in a Health Knowledge Management Platform: Implementation and Requirement-Based Evaluation Study

Björn Schreiweis[1,2*], Prof Dr; Benjamin Kinast[1,2*], MA; Hannes Ulrich[1,2], Dr rer nat; Tobias Bronsch[1,2], MSc; Ann-Kristin Kock-Schoppenhauer[2,3], Dr hum biol; Björn Bergh[1,2], Prof Dr Med

[1]Institute for Medical Informatics and Artificial Intelligence, Kiel University and University Hospital Schleswig-Holstein, Arnold-Heller-Straße 3, Kiel, Germany

[2]Medical Data Integration Center, University Hospital Schleswig-Holstein, Kiel/Lübeck, Germany

[3]Institute for Medical Biometric and Statistics, Section for Clinical Research IT, Universität zu Lübeck and University Hospital Schleswig-Holstein, Lübeck, Germany

[*]these authors contributed equally

Corresponding Author:
Benjamin Kinast, MA
Institute for Medical Informatics and Artificial Intelligence, Kiel University and University Hospital Schleswig-Holstein, Arnold-Heller-Straße 3, Kiel, Germany

Abstract

Background: In the evolving landscape of health care, data use plays an ever-increasing role in health care IT. However, data are often siloed and uncoded free text distributed across several IT systems. This paper introduces a health knowledge management platform, designed to integrate, harmonize, and enable reuse of health care and medical research data. The platform aims to bridge the gap between research and patient care, showcased through real-world scenarios, emphasizing data harmonization and knowledge management within a health care institution. The study is based at the University Hospital Schleswig-Holstein.

Objective: The main objective of this project is to design, implement, and evaluate a knowledge management platform that integrates health care and biomedical research to support use cases in both domains.

Methods: The study describes the “health knowledge management platform” designed to access and gain knowledge from health care and medical research data. We performed several rounds of focus groups with stakeholders to elicit the platform requirements. In the process, we identified key aspects of the platform. From the functional requirements, we designed an architectural concept. The platform evaluation follows the Framework for Evaluation in Design Science Research and International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 25010 standard with a focus on key aspects identified and real-world scenarios. Two application scenarios, cardiology and radiology, are selected for a requirement-based, qualitative evaluation.

Results: We show that our health knowledge management platform is capable of integrating diverse data formats like Health Level 7 Version 2 messages, CSV exports, and Digital Imaging and Communications in Medicine. It currently integrates over 46 million admit, discharge, transfer messages, 38 million imaging studies, and structured clinical data for approximately 1.5 million patients. The platform supports different scenarios based on its 5-layer architecture, including a clinical data repository and services like Master Patient Index and Consent Management. The evaluation against 39 predefined functional requirements showed our platform’s capability in certain real-world scenarios of cardiology and radiology. Our evaluation demonstrates that the platform covers the majority of the identified requirements to support knowledge management in health care institutions.

Conclusions: Our requirement-based evaluation of the health knowledge management platform at University Hospital Schleswig-Holstein reveals its capabilities, which is possibly leading to better knowledge transfer between patient care and research. The platform’s architecture and standardized data improve the quality of data and facilitate access to knowledge. Ongoing development and potential quantitative measures will further enhance its applicability in dynamic health care landscapes.

(JMIR Med Inform 2026;14:e83608) doi:10.2196/83608

KEYWORDS

clinical data repository; data integration; knowledge management; interoperability; secondary use; requirements engineering

XSL•FO
RenderX

Introduction

In modern health care, data play a crucial role in supporting clinical decisions, improving treatment strategies, and enabling clinical research. Clinicians and researchers require access to reliable, structured, and machine-readable data for applications such as clinical decision support systems (CDSS), artificial intelligence (AI), and federated learning models [1-4]. However, in many hospitals, data remain siloed and poorly coded and exist often in free-text formats within the hospital information systems. These silos limit data reuse, complicate integration, and limit the deployment of intelligent systems that depend on structured and consistent inputs [5-7]. Additionally, the coexistence of multiple interoperability standards adds to the complexity of data integration and management [8,9].

However, instead of focusing only on technical interoperability between systems, hospitals need platforms that structure, harmonize, standardize, and make data usable across various scenarios [10]. Structured and centrally available data are a central prerequisite for building learning health care systems [11], which aim to improve patient care and research at the same time. Such systems rely on data from diverse sources, such as clinical information systems (CIS), picture archiving and communication systems (PACS), and laboratory information systems (LIS), or data from medical devices, wearable sensors, and mobile apps [12,13]. It is not sufficient to exchange data between systems. Data must be transformed into usable knowledge. This highlights the importance of knowledge management in health care. It shifts the focus from isolated data handling toward creating value from data by organizing, linking, and making it accessible for different stakeholders via an integrated platform. “Knowledge Management is therefore a conscious strategy of getting the right knowledge to the right people at the right time and helping people share and put information into action in ways that strive to improve organizational performance” [14]. This definition shows that the value of data lies not only in its availability but also in its structured, contextualized, and accessible form. There are different kinds of knowledge, but for now, we mainly focus on empirical knowledge. To implement knowledge management in health care settings, a technical infrastructure is needed that ensures interoperability, data harmonization, and support for various applications on a consistent data basis.

The Medical Informatics Initiative funded by the German Federal Ministry of Research, Technology and Space aims toward reliable and fast access to data-centric platforms for biomedical research by setting up (medical) data integration centers (Medical Data Integration Center or Data Integration Center) at the German university hospitals [15,16]. Our approach goes beyond pure secondary use of data by aiming to support both research and immediate patient care. To achieve this, we iteratively develop a health knowledge management platform architecture that not only facilitates structured data integration but also provides interfaces and functionalities, which allow the reuse of data for, for example, clinical decision support and cohort exploration. In doing so, it contributes to a future-ready infrastructure for emerging applications such as AI-driven analytics, personalized medicine, or federated learning.

The main objective of this project is to design, implement, and evaluate a knowledge management platform that integrates health care and biomedical research to support use cases in both domains. Therefore, the three subobjectives are (1) to identify the requirements of such a knowledge management platform, (2) to design an architectural concept for the knowledge management platform based on the requirements, and (3) to implement and qualitatively evaluate the platform against the requirements.

Methods

Ethical Considerations

The referenced studies were carried out in adherence with the Declaration of Helsinki and approved by the local medical ethics committees of Kiel University (approval B279/18 and D596/22). Informed consent was obtained from all of the participants in the studies.

Overview

Our methodology is based on the Framework for Evaluation in Design Science Research (FEDS) by Venable et al [17], which describes the process from defining the problem, through deriving a solution based on requirement engineering (RE), evaluating the platform, and communicating the outcomes.

Framework for Evaluation in Design Science Research

The FEDS consists of 4 components: problem, solution, evaluation, and communication.

Problem

The first step involves identifying and understanding the challenges related to siloed and insufficiently structured and coded health care data in a university hospital setting. Much of the clinical information is documented in unstructured formats such as free text or PDF documents. These are often embedded in hospital information systems, which limits accessibility and reuse [17]. Additionally, relevant data are distributed across various primary source systems, complicating semantic interoperability and secondary (research) use. Moreover, structured and centrally accessible data are essential prerequisites to enable future applications such as AI-based decision support or federated learning approaches [18].

Solution (via RE)

Overview

The second step of the FEDS focuses on designing and developing an effective artifact to address the identified problems. In the FEDS context, an artifact can refer to a tangible or intangible entity—such as software applications, system architectures, databases, protocols, guidelines, or frameworks—created or modified as part of the solution development process [17]. In this study, the solution is our health knowledge management platform, designed to integrate structured and unstructured, alphanumeric, and multimedia health care data while harmonizing them. The platform is further designed to improve interoperability, thereby creating a unified data management environment for both health care and research-related applications. Therefore, the development was

XSL•FO
RenderX

preceded by a systematic RE process to capture and formalize the platforms' functional needs.

RE Process

We used a structured RE approach to define the platform architecture [19,20]. From 2019 to 2024, we conducted multiple focus groups with a total of 27 stakeholders, covering a broad spectrum of roles and expertise:

- Strategic and technical leadership: including the chief digital officer and the heads of the Medical Data Integration Center.
- Technical and domain experts: such as specialists in metadata management, data integration, interoperability standards (eg, Health Level 7 [HL7] Fast Healthcare Interoperability Resources [FHIR], openEHR, and Digital Imaging and Communications in Medicine [DICOM]), and terminologies (eg, Systematized Nomenclature of Medicine—Clinical Terms, *International Classification of Diseases and Related Health Problems, 10th Revision, German Modification* [*ICD-10-GM*], Anatomical-Therapeutic-Chemical, and Logical Observation Identifiers Names and Codes [LOINC]).
- User representatives: including pharmacologists, an anesthesiologist, biobank directors, and data privacy specialists.
- Scenario-specific stakeholders: such as cardiologists, radiologists, a neurologist, a nephrologist, and a sleep medicine specialist who acted as product owners for the real-world evaluation.

The interdisciplinary composition of the focus groups ensured alignment with strategic goals and operational needs. Iterative discussions and repeated validation reasonably support the assumption that thematic saturation is achieved, in line with the approach described by Vasileiou et al [21].

The discussions were guided by the central question: "Which functions should the knowledge management platform provide to support health care and biomedical research?" [22,23]. Session results were summarized in a mind map (Multimedia Appendix 1) and clustered into thematic areas: data management and persistence, architecture, Findable, Accessible, Interoperable, Reusable principles [24], AI, data and information extraction, health telematics and regional integration, integration of research and patient care data, general knowledge integration, and integrated frontends.

Using this structure, we mapped each of the identified key aspects to functional requirements, which we derived from a previously created requirements catalog [25]. Additional requirements were specified using International Organization for Standardization (ISO) 29148 syntax ([Subject] [Action] [Constraint of Action]) [19], for example, "The platform must integrate data." Possible functionalities or solutions are annotated in the respective requirements notes (Multimedia Appendix 1). Disagreements during the requirements elicitation process were resolved through moderated discussions with the RE expert.

Subsequently, we mapped these requirements to the HiGHmed architecture [15] and extended the existing layer model to define our five-layer architecture: (1) source layer, (2) staging layer, (3) normalization layer, (4) data access layer, and (5) application layer, in line with the principle of separation of concerns [26].

Evaluation

The evaluation component involves a requirement-based artifact evaluation of the platform against the identified requirements. The primary focus lies on evaluating the functional aspects identified in the RE process. While detailed benchmarking of nonfunctional qualities (eg, performance or scalability) is beyond the current scope, quantitative insights (eg, volume of integrated laboratory data) are included. We use a technical artifact evaluation strategy guided by the quality model defined in International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 25010 [27], part of the ISO/IEC 25000 SQuaRE framework. This allows us to focus on relevant quality attributes such as interoperability, reusability, data consistency, and usability (see Key Aspects of the Platform section). The evaluation is accompanied by 2 real-world scenario-based assessments, each chosen for its distinct technical requirements:

- Cardiology: This scenario evaluates the platform's ability to integrate and harmonize structured and unstructured clinical routine data from hospital IT systems (eg, CIS, PACS, LIS, Electronic Prescribing and Medicine Administration [ePMA], and patient-generated health data [PGHD] from mobile apps and wearables for patients with heart failure patients, with a focus on technical integration, interoperability, and traceable data reuse).
- Radiology: This scenario evaluates the platform's functionality in processing imaging metadata and integrating demographic data to enable consent-compliant cohort identification and follow-up planning, focusing on system-level integration and traceability.

To operationalize the qualitative scenario evaluation, we applied different evaluation activities depending on the nature of the requirement and its implementation status. These activities were selected to provide traceable evidence for requirement fulfillment, in line with the ISO/IEC 25000 standard framework. Specifically, the following evaluation activities were used:

- Scenario walkthroughs: used to evaluate end-to-end functionality under real-world conditions.
- System and architecture inspections: used to verify data flows, persistence architecture, and interoperability aspects.
- Log inspections and data audits: applied to confirm the integration of structured and unstructured data (eg, Health Level 7 Version 2 [HL7 V2] and DICOM).
- Conceptual design reviews: used for requirements that have been specified, modeled, or architecturally prepared but are not yet in operational use.

These activities support a systematic evaluation of how well the platform meets the functional requirements derived from the RE process. All evaluation activities were conducted by the institution's Data Integration Unit, a dedicated team of medical informatics experts, data engineers, and platform architects. An additional evaluation scenario is provided in Multimedia Appendix 2.

XSL•FO
RenderX

Communication

The final step involves disseminating results to relevant stakeholders. In our case, we communicate findings through this study and use them internally to inform researchers and clinicians to promote the platform and its functionalities.

Following these structured phases ensures that the platform's development and evaluation align with its goal of enabling interoperability and the effective health care data use across multiple domains by leveraging data integration as a central strategy.

Results

Problem Statement

With more than 500,000 encounters (approximately 100,000 inpatients and approximately 400,000 outpatients) annually and about 17,000 employees, University Hospital Schleswig-Holstein (UKSH) is one of the largest university hospitals in Germany and the only hospital providing maximum medical care in the state of Schleswig-Holstein (Northern Germany) [28]. It operates across 2 locations, Kiel and Lübeck, in partnership with the Medical Faculty of Kiel University and the Medical Section of the Universität zu Lübeck, forming a central hub for academic medicine in Northern Germany.

The current IT system landscape at UKSH consists of different types of information systems. The CIS builds the core of the IT system landscape, provides a general overview of the patient's history, and serves as a basis for documentation and communication across health care domains. In addition, there are specialized department-specific systems, such as radiology information systems (RIS) or echocardiography systems in cardiology. Finally, there are the function-specific specialized systems like LIS and ePMA that support specific workflows or tasks independent of the respective department. The existence of these various systems challenges data integration, harmonization, and coding and requires careful attention to ensure that all necessary information is captured and managed appropriately. The IT system landscape and its limited semantic interoperability between the different systems often result in patient data spreading across multiple systems, sometimes even redundantly. This fragmentation complicates timely access to required information and poses challenges for researchers who require timely access to relevant data, ultimately hindering effective data use and analysis.

Moreover, there is a client separation in health care, with patient care and research operating independently from an IT and data perspective. Typically, clinicians are responsible for treating patients and have access to patient-level data for that purpose only. The processes and IT systems used in patient care are often disjoint from those used for clinical and biomedical research. Accordingly, it is inherent in the nature of these IT systems that they are primarily designed for patient care purposes rather than research, that is, supporting patient-level data access, but not cohort-level access. The organizational separation between patient care and research is partly also mandated by regulation, requiring researchers to obtain patient consent and adhere to additional processes to access the necessary data. As a result, clinicians researching on patient data either request a data export from the clinical IT service provider or manually extract the data via the applications' user interfaces (ie, redundant data entry). Both are time-consuming, highly inefficient, and prone to transmission errors, leading to potential inaccuracies in the data and generating irreproducible research findings. Data exports are often raw CSV files per IT system, which requires manual data linkage and transformations to ensure consistency in units of measurement, as well as annotations with appropriate terminologies before allowing effective data analysis. In addition to the client separation, there is also a significant lack of automatic knowledge transfer between patient care and research. However, there is a demand in patient care for access to the latest scientific discoveries and insights to inform and improve clinical decision-making.

Solution

Key Aspects of the Platform

The focus groups conducted during the requirements elicitation process identified key aspects mandatory for a knowledge management platform. The first set of key aspects relates to the scenarios the knowledge management platform needs to support, which span 3 domains: health care, research enablement, and research or secondary use. For health care, participants emphasized the need to visualize heterogeneous data types (eg, laboratory data, medication data, vital parameters, reports, images, and biosignals). Requested features included patient trajectories and the comparison of patients with similar characteristics. The integration of CDSS and AI, in general, for functionalities like smart alerting was also considered essential. A further requirement, for both scenarios, health care and research, was the inclusion of PGHD, patient-reported outcome measures, and patient-reported experience measures, with patients also having access to their own health data.

For research enablement, three key aspects were mentioned: (1) self-service feasibility analysis based on cohort definitions, (2) prescreening support for clinical trial recruitment, and (3) cohort exploration tools.

In terms of secondary use, the platform should enable care-connected research, such as research registries, by reducing redundant documentation and enabling seamless reuse of health care data. Researchers should be able to access data directly for analysis purposes or receive data exports.

After defining scenario-based needs, the view of the focus groups shifted to the data perspective. The platform needs to support various data types including structured and unstructured text, streaming data, omics data, and multimedia objects (DICOM and non-DICOM). In addition to clinical IT systems, other data sources must be integrated, including electronic health records, biomaterial information, and research data (eg, Clinical Trial Management System [CTMS]). Prefilling electronic case report forms in the CTMS with care data and integrating research data back into care are desirable. Furthermore, access to data from public sources (eg, literature databases) is requested. The platform should persist raw data on premises to serve as a single point of truth, equipped with a searchable index for data usability from the beginning.

XSL•FO
RenderX

Finally, the security and functional optimization were discussed. A clear separation of concerns is required—distinguishing between fully identifiable data, pseudonymized data, and restricted-use datasets. Raw data must be securely stored and accessible only to a small administrative group. Semantically enriched and pseudonymized datasets free from direct identifiers should be made available for health care applications and research queries. Exported data should be pseudonymized or, if possible, anonymized per project or cohort, enabling external access, for example, via national portals like the German Portal for Medical Research Data [29].

Finally, focus group participants emphasized that end users, for example, health care professionals, researchers, and lecturers, would not prefer to access data directly, but through scenario-specific applications that provide tailored visualizations and functionalities.

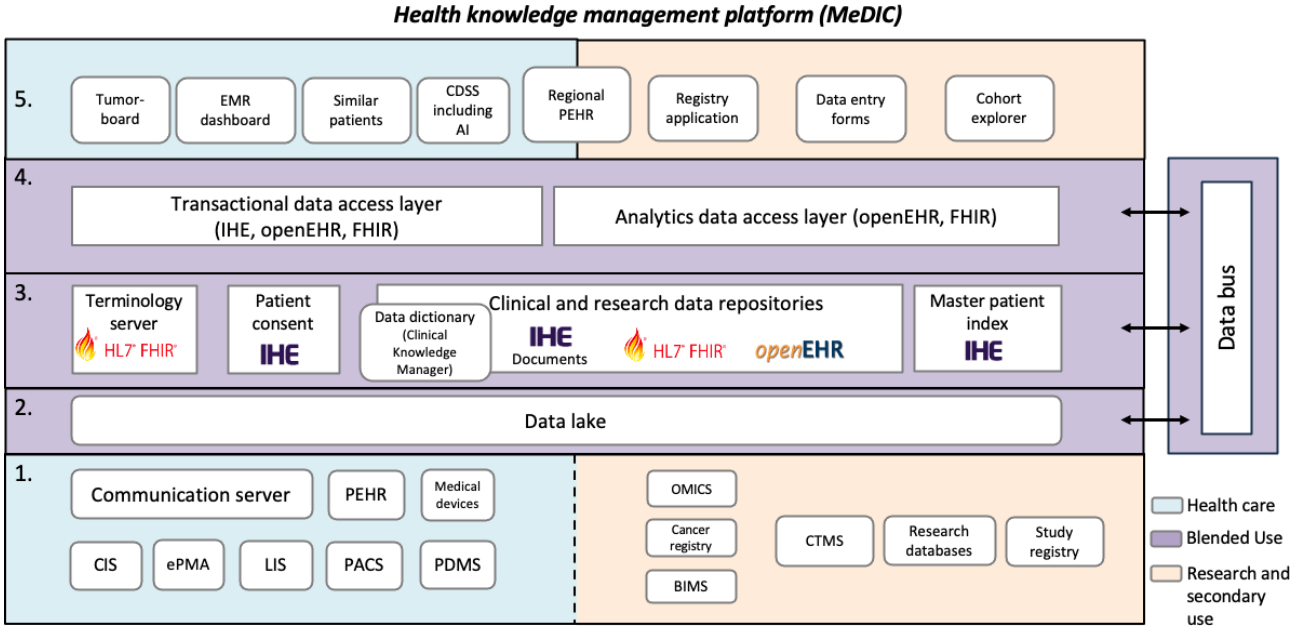
Requirements

A total of 39 requirements were organized into 5 distinct layers for the envisioned platform architecture to ensure a structured and modular approach. Specifically, 9 (23%) requirements were assigned to the source layer, 4 (10%) to the staging layer, 15 (39%) to the normalization layer, 2 (5%) to the data access layer, and 9 (23%) newly identified requirements to the application layer. A detailed overview of all 39 requirements and their assignment to the architectural layers is provided in Multimedia Appendix 2.

Description of the Knowledge Management Platform

The knowledge management platform consists of 5 layers and a data bus: source, staging, normalization, data access, and application layer (Figure 1).

Figure 1. Architecture of the health knowledge management platform of UKSH MeDIC with its five layers: (1) source layer, (2) staging layer, (3) normalization layer, (4) data access layer, and (5) application layer. BIMS: Biomaterial Information Management System; CIS: clinical information system; CTMS: Clinical Trial Management System; EMR: electronic medical record; ePMA: Electronic Prescribing and Medicine Administration; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7; IHE: Integrating the Healthcare Enterprise; LIS: laboratory information system; MeDIC: Medical Data Integration Center; PACS: picture archiving and communication system; PDMS: patient data management system; PEHR: personal cross-enterprise health record; UKSH: University Hospital Schleswig-Holstein.



Source Layer

The source layer contains all health care and research IT systems from which data are integrated (DA-01; see Key Aspects of the Platform section). Health care systems include CIS, LIS, ePMA, PACS (including multimedia data from radiology, cardiology, neurology, and gynecology; DA-03), patient data management system, personal cross-enterprise health record (PEHR; DA-07 and DA-08-3), and medical devices. Research systems include a clinical cancer registry and tumor documentation system, a Biomaterial Information Management System, a CTMS, systems for generating and analyzing OMICS data, research databases, and a local study registry. The CTMS supports the data entry process in clinical trials, but also the management of study appointments, participants, and electronic case report forms. The local study registry contains all studies running at USKH including eligibility criteria to support the prescreening for clinical trials in the future. These studies can be integrated from national and international study registries like ClinicalTrials.gov or drks.de (DA-27) [30]. Data are transferred via interfaces (eg, HL7 V2) and a communication server to the data lake (see Staging Layer section; DA-02) or exported and stored consistently into the data lake (DA-01-4). Data types range from structured (eg, laboratory reports; DA-01-4) to semistructured and free-text (eg, discharge letters; DA-01-3) and multimedia data (eg, computed tomography [CT] images; DA-03).

Staging Layer

The staging layer consists of a central data lake with analytical processing capabilities (DP-02) [31], supporting both batch and message-based (streaming) data (DP-07). Data flows are orchestrated using Apache NiFi. Incoming data are first persisted in S3 buckets (DS-03) to ensure reliability and data provenance (DS-02-2). A Master Patient Index Identifier is requested from

XSL•FO
RenderX

the Master Patient Index (MPI) in the normalization layer and replaces the hospital's patient identifier in all processed records or data (DP-20-3, DP-20-5, and DP-21). Subsequently, data are indexed in Elasticsearch to enable searchability and discovery (DS-13) and forwarded via the data bus. Unlike the source layer, the staging layer processes both health care and research domains (blended use).

Data Bus

The data bus is built on Apache Kafka and enables scalable, topic-based data distribution for both health care and research data [32]. Topics may contain either full data records or references (eg, images or electrocardiography [ECG] paths) for downstream retrieval.

Normalization Layer

The normalization layer provides harmonization and standardization of data from internal (eg, CIS) and external sources (eg, PEHR; DP-19). It includes the MPI, Patient Consent Manager (PCM), clinical and research data repositories, and a terminology server, based on interoperability standards, including HL7 FHIR, openEHR, DICOM (DP-23), and Integrating the Healthcare Enterprise (IHE). Patient identity management is handled via the MPI (DP-20-3 and DP-20-5) compliant with IHE Patient Identifier Cross-Referencing [33] and IHE Patient Demographics Query (DSC-04) [34]. The PCM manages and distributes versioned broad consents to authorized downstream systems (DAS-13) [35]. Patient-referenced metadata (DP-21) are stored in an openEHR-based clinical data repository (DS-02-2, DS-05, DS-08, and DP-33-1), while nonpatient-referenced data, such as study metadata or data about drugs, are maintained in a FHIR-based repository (DS-02-2, DS-05, DP-06, and DP-33-1). OpenEHR compositions may additionally be persisted as IHE Cross-Enterprise Document Sharing (XDS) documents (DP-33-1). The CDR is queried via openEHR Archetype Query Language (AQL), and its data models are openly accessible and managed in a clinical knowledge management tool (MDM-01) [36], while FHIR models are provided via Simplifier [37].

The terminology server supports mapping of local codes to standardized terminologies such as LOINC (eg, for laboratory data), *ICD-10-GM*, and Systematized Nomenclature of Medicine—Clinical Terms (DP-01). It provides versioned terminologies and FHIR ConceptMaps to ensure semantic interoperability across systems (DP-01) [38].

Data Access Layer

The data access layer includes transactional and analytical interfaces. The transactional data access layer supports persistence and retrieval via openEHR, FHIR, and IHE-based application programming interfaces, while the analytical data access layer supports cohort-based queries and research-driven analysis (DAS-03). Data exchange between the data lake, CDR, and project-specific data marts is handled via the data bus (DAS-12).

Application Layer

The application layer provides end-user applications that access clinical and research data via the data access layers. In the health care domain, this includes a Molecular Tumor Board (MTB) application that supports preparation for MTB sessions by providing access to patient-level data including imaging, clinical, and OMICS data (DR-01 and DR-03-3) [39]. An Electronic Medical Record Dashboard offering unified patient-level views (DR-01, DR-03, and DR-04) and a "Patients like mine" similarity application (DR-03-1, DR-03-2, and DR-03-3) [40,41] are planned, as well as various CDSS including AI-based tools. A regional PEHR application (DA-07) to integrate PGHD from smartphones, wearables, body sensors, and self-reported data via FHIR (DA-18) was tested and evaluated in a proof-of-concept. Research-oriented applications include structured data entry forms (DR-01), registry integration (DR-01) [42], and cohort explorer for feasibility analysis (DR-02) and data requests (DR-01 and DR-05). Access to data based on use requests is governed via consent management (DAS-13 and DSC-04). A centralized user management system is integrated with the hospital identity management system (DSC-12).

Evaluation

Overview

The evaluation assesses the platform against the previously identified functional requirements. We conducted scenario-based evaluations in 2 distinct medical domains to cover a broad range of technical and research-related requirements. The primary results from the cardiology and radiology scenarios are presented below, with an additional neurology scenario provided in Multimedia Appendix 2.

Clinical Routine Data in a Cardiology Scenario

The cardiology scenario evaluates the platform's technical integration, traceability, and reuse of structured clinical data for patients with heart failure, integrating routine care data [15]. The evaluation focuses on system-level interoperability and requirements fulfillment rather than on clinical effectiveness, patient outcomes, or therapeutic impact.

Patient data were processed under a study-specific consent, covering electronic medical record reuse (ethics approval obtained at Kiel University B279/18). Clinical data from CIS, LIS, and ePMA were integrated (DA-01). Demographic, laboratory, and anamnesis data (DA-01-2 and DA-01-3) were ingested as HL7 V2 admit, discharge, transfer (ADT) and observation result unsolicited (ORU) messages from the hospital's communication server (DA-01 and DA-02), while medication data were imported via CSV exports. Data were persisted in the staging layer (DS-13) and semantically harmonized using standardized terminologies such as LOINC, Anatomical-Therapeutic-Chemical, and *ICD-10-GM* (DP-01). Follow-up and echocardiography data were captured via openEHR-based data entry forms (DR-01) [43,44]. Cross-system record linkage and separation of identifying and medical data were ensured via the MPI (DP-20-3, DP-20-5, and DSC-04), with metadata stored for traceability in the openEHR-based CDR (DS-02-2 and DS-03), including the technically implemented ingestion of PEHR-derived structured data (DA-08-3) as a proof-of-concept. In addition, sensor-derived PGHD, including heart rate and activity measurements recorded via wearable devices, were transmitted as FHIR resources from

XSL•FO
RenderX

a mobile app to the PEHR and ingested into the platform (DA-18, DA-07, and DA-08 - 3) and linked via the MPI (DP-21 and DP-19) as a technical proof-of-concept.

Finally, structured data corresponding to metadata from 710 patients were transformed into validated openEHR templates (DP-06, DP-23, and DP-33-1) and persisted in the openEHR-based CDR (DA-01-4, DS-05, DS-08, and DP-21) [43-47], derived from more than 38,000 HL7 messages (DA-01, DA-02, and DP-07). All openEHR templates were validated with clinicians and published (DP-23) [36]. The resulting dataset was queried using AQL and exported as a CSV file for analysis (DAS-12 and DR-05).

Thus, the scenario meets 26 of 39 (67%) platform requirements. The evaluation activities and corresponding evidence at the requirements level are detailed in Table 1. A consolidated assessment of the resulting quality attributes according to ISO/IEC 25010 across both scenarios is summarized in Table 2.

XSL•FO
RenderX

Table . Alignment of platform requirements with scenarios.

Requirement ID	Layer	Description	Implemented in platform	Evaluation activity	Evidence or outcome
DA-01	SrcL[a]	The platform must acquire data from different heterogeneous source systems.	✓[b]	Scenario walkthrough (cardiology, radiology)	Successful ingestion of HL7 V2[c] (ADT[d]/ORU[e]) DICOM[f], and CSV data from CIS[g], LIS[h], and PACS[i] persisted as records in the data lake
DA-01 - 2	SrcL	The platform must acquire structured alphanumeric EHR[j] data.	✓	Log inspection (cardiology)	2902 laboratory reports acquired from HL7 V2
DA-01 - 3	SrcL	The platform must acquire unstructured alphanumeric EHR data.	✓	Cardiology scenario	1016 anamnesis forms acquired
DA-01 - 4	SrcL	The platform must acquire data from heterogeneous sources in a consistent manner.	✓	Technical architecture inspection	HL7 V2, CSV, FHIR[k], and DICOM processed through standardized ETL[l] pipelines and stored in uniform data structures within the data lake and CDR[m] (verified via ETL configuration and storage inspection)
DA-02	SrcL	The platform must acquire structured data based on interoperability standards from the IT and health care domains.	✓	Scenario walkthrough	Incoming data received via HL7 V2 (ADT, ORU, BAR[n]) messages from, eg, CIS and LIS
DA-03	SrcL	The platform must acquire multimedia data from various domains.	✓	Radiology scenario; ingestion test	CT[o] DICOM successfully retrieved from PACS (including data from, eg, radiology, cardiology, neurology, and gynecology) and stored in the platform [48]
DA-07	SrcL	The platform must acquire data from cross-institutional sources.	✓	Cardiology PGHD[p] ingestion in PEHR[q] (technically implemented)	FHIR-based PGHD integration pipeline from external medPower to PEHR to the platform, verified by successful ETL execution and persisted openEHR compositions in proof-of-concept
DA-08 - 3	SrcL	The platform must acquire data from personal health records with all of the following nature: * structured * unstructured.	✓	Cardiology PGHD workflow	Structured FHIR resources generated from Apple HealthKit; transmitted via PEHR, transformed into openEHR compositions, and stored in CDR (verified by stored compositions in proof-of-concept)

XSL•FO
RenderX

Requirement ID	Layer	Description	Implemented in platform	Evaluation activity	Evidence or outcome
DA-18	AppL[r]	The platform must acquire data from (FHIR-based) apps.	✓	Cardiology scenario	Successful ingestion of FHIR data from HealthKit (Apple Watch) via medPower and PEHR; ETL implemented to map to openEHR in proof-of-concept
DA-27	SrcL	The platform must integrate data from public sources.	TBD[s]	Conceptual design review	Functionality not yet implemented due to regulatory prerequisites
DP-07	SL[t]	The platform must process data from data streams.	✓	Cardiology scenario	HL7 V2 data streams (ORU, ADT, and BAR) processed weekly
DS-02 - 2	SL	The platform must store both data and metadata from multiple sources in a single storage.	✓	System inspection	HL7[u], DICOM, and CSV data including metadata, persisted in S3 storage and indexed via Elasticsearch (verified by storage inspection)
DS-03	SL	The platform must store the original data.	✓	System inspection	Original HL7 and CSV files stored in S3 buckets for traceability (verified by storage inspection)
DS-05	NL[v]	The platform must store data in a data repository.	✓	System inspection	Structured data stored in openEHR-based CDR; ~1.5 million patient records integrated (verified by storage inspection)
DS-08	NL	The platform must store the data in a way that it is available in a clinical data repository.	✓	Scenario execution (all)	AQL[w] queries executed on CDR returned structured longitudinal patient data, enabling cohort identification in both scenarios (verified via query results)
DS-13	SL	The platform must store data in all of the following ways: * full-text indexed * searchable.	✓	Technical inspection	Full-text indexing via Elasticsearch; supports full-text search (verified via query results)
DP-01	NL	The platform must process data by mapping it to multiple terminology standards.	✓	Cardiology scenario	LOINC[x], ATC[y], and ICD-10-GM[z] applied during ETL for laboratory, medication, and diagnosis data (verified via query results)
DP-02	NL	The platform must process unstructured data by using natural language processing (NLP) methods.	TBD	Conceptual design review	NLP functionality conceptually designed but not implemented

XSL•FO
RenderX

Requirement ID	Layer	Description	Implemented in platform	Evaluation activity	Evidence or outcome
DP-06	NL	The platform must process all of the following: * data * metadata.	✓	System design inspection	Metadata extraction from HL7, DICOM, CSV; structured storage of both in openEHR CDR (verified via query results)
MDM-01	NL	The platform must provide metadata management (tools).	✓	Scenario execution (radiology)	Imaging metadata stored, searchable via DICOM tag filters (verified via query results)
DP-19	NL	The platform must link external data with local data.	(✓)[aa]	Conceptual design review	PGHD from patients linked via MPI to clinical data in proof-of-concept
DP-20 - 3	NL	The platform must process data by linking it to at least 1 of the following Master Patient Indexes: * institutional * cross-institutional.	✓	Scenario execution (all)	MPI links data across systems; contains 2.3 million patient identities
DP-20 - 5	NL	The platform must process data by using at least 1 of the following IDs: * locally unique record ID * locally unique case IDs.	✓	Scenario execution (all)	Local patient IDs used in MPI, mapped to MPI-ID for further identification
DP-21	NL	The platform must process data to link records from various source systems.	✓	Scenario execution (all)	For example, record linkage across CIS, PACS, LIS, ePMA[ab]
DP-23	NL	The platform must process data according to a specified semantic data model.	✓	Technical inspection	Structured data instances validated against openEHR archetypes and templates; conformance verified through successful persistence in the CDR without schema violations
DP-33 - 1	NL	The platform must process data by transforming it into an open standardized format.	✓	Scenario execution (all)	HL7, CSV, and FHIR inputs transformed into openEHR compositions and stored in CDR; successful transformation confirmed by absence of processing errors and validated template bindings
DSC-04	NL	The platform must provide a pseudonymization ID (PID) to disconnect medical data from identifying data.	✓	Scenario execution (all)	PID service integrated; separates identifying (MPI) from clinical data (CDR)
DAS-03	AL[ac]	The platform must provide various data analysis techniques.	TBD	Conceptual design review	Framework supports integration of external analysis tools; not yet implemented

XSL•FO
RenderX

Requirement ID	Layer	Description	Implemented in platform	Evaluation activity	Evidence or outcome
DAS-12	AL	The platform must extract datamarts for research-related analysis.	✓	Scenario execution (all)	Structured data exports to CSV; used for analytics in cardiology and radiology scenarios
DAS-13	NL	The platform must only analyze (patient-) data for which consent has been obtained.	✓	Scenario execution (radiology)	Consent status checked in PCM[ad] before recontact; enforced before analysis
DR-01	AppL	The platform must provide user group–specific interfaces.	✓	Scenario execution (cardiology)	Data entry tool tailored for study nurses
DR-02	AppL	The platform must provide a user interface for cohort exploration.	✓	Scenario execution (cardiology)	Cohort builder enables data transfer unit to search by, eg, diagnosis, laboratory, procedure, and consent data
DR-03	AppL	The platform must provide enhanced visualization for holistic multimodal care information.	TBD	Conceptual design review	Not implemented
DR-03 - 1	AppL	The platform must provide user group–specific interfaces for predictive data analysis.	TBD	Conceptual design review	Not implemented; planned for future artificial intelligence module integration
DR-03 - 2	AppL	The platform must provide a user interface for the comparison of patients with similar medical histories.	TBD	Conceptual design review	Not implemented; AQL queries for patient grouping prepared
DR-03 - 3	AppL	The platform must provide trajectory data representation.	✓	Conceptual design review	Longitudinal patient trajectories rendered in the Molecular Tumor Board application, displaying multimodal data on timeline (verified via system demonstration)
DR-04	AppL	The platform must provide an interface for context-related external calls from the clinical information system (CIS/EMR[ae]).	TBD	Conceptual design review	Not implemented
DR-05	AppL	The platform must allow querying qualified search results.	✓	Scenario execution (cardiology)	AQL used for structured queries across CDR; results exported for analysis

XSL•FO
RenderX

Requirement ID	Layer	Description	Implemented in platform	Evaluation activity	Evidence or outcome
DSC-12	AppL	The platform must enable single-sign-on.	(✓)	Conceptual design review	Designed but not yet fully enabled in the live system

[a]SrcL: source layer.
[b]✓: fulfilled.
[c]HL7 V2: Health Level 7 Version 2.
[d]ADT: admit, discharge, transfer.
[e]ORU: observation result unsolicited.
[f]DICOM: Digital Imaging and Communications in Medicine.
[g]CIS: clinical information system.
[h]LIS: laboratory information system.
[i]PACS: picture archiving and communication system.
[j]EHR: electronic health record.
[k]FHIR: Fast Healthcare Interoperability Resources.
[l]ETL: extract-transform-load.
[m]CDR: clinical data repository.
[n]BAR: billing account record.
[o]CT: computed tomography.
[p]PGHD: patient-generated health data.
[q]PEHR: patient cross-enterprise health record.
[r]AppL: application layer.
[s]TBD: to be done.
[t]SL: staging layer.
[u]HL7: Health Level 7.
[v]NL: normalization layer.
[w]AQL: Archetype Query Language.
[x]LOINC: Logical Observation Identifiers Names and Codes.
[y]ATC: Anatomical-Therapeutic-Chemical.
[z]ICD-10-GM: International Classification of Diseases and Related Health Problems, 10th Revision, German Modification.
[aa](✓): partially fulfilled.
[ab]ePMA: Electronic Prescribing and Medicine Administration.
[ac]AL: Access layer
[ad]PCM: Patient Consent Manager.
[ae]EMR: electronic medical record.

XSL•FO
RenderX

Table . Consolidated quality attribute evaluation.

Quality attribute	Cardiology scenario findings	Radiology scenario findings
Interoperability	Demonstrated integration of structured and unstructured data from 4 source systems (HL7 V2[a] and FHIR[b]).	Integrated demographic and imaging metadata from PACS[c] and CIS[d] via CSV-based ETL[e] processes and MPI[f] linkage.
Data accuracy and consistency	Achieved through the use of standardized terminologies (*ICD-10-GM*[g] and LOINC[h]) and validation of templates by clinical experts.	Ensured by automated metadata extraction from radiological systems and consistent use of patient IDs via MPI ensured reliable linkage and cohort formation.
Reusability	Extraction logic and openEHR templates were reused from more than 38,000 HL7[i] messages (anamnesis forms: n=1016, echocardiography entries: n=616, laboratory reports: n=2902, and medication entries: n=6231).	The cohort identification process (unique patients: n=1744 and unique CT[j] studies: n=4146) and corresponding query logic can be reused for other imaging-based follow-up studies with similar criteria.
Traceability	Enabled by storing metadata and source system information alongside clinical data, ensuring transparent data provenance.	Full provenance tracking of imaging data was achieved by retaining source system identifiers and metadata (eg, modality and scan date).
User or acceptance or usability	Manual data entry interfaces (eg, echocardiography) were cocreated with clinicians and study nurses, and the final dataset was validated with domain experts from external study sites (eg, other HiGHmed sites).	Automated generation and delivery of final patient lists to the radiology department replaced manual chart review and simplified recruitment workflows.

[a]HL7 V2: Health Level 7 Version 2.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]PACS: picture archiving and communication system.

[d]CIS: clinical information system.

[e]ETL: extract-transform-load.

[f]MPI: Master Patient Index.

[g]*ICD-10-GM*: *International Classification of Diseases and Related Health Problems, 10th Revision, German Modification*.

[h]LOINC: Logical Observation Identifiers Names and Codes.

[i]HL7: Health Level 7.

[j]CT: computed tomography.

Recontacting Patients in a Radiology Scenario

The radiology scenario evaluates the platform's technical integration, traceability, and reuse of CT imaging metadata for cohort identification of patients who underwent hip or spine CT examinations (ethics approval obtained at Kiel University [D596/22]). The evaluation focuses on data integration, metadata processing, and consent-compliant record linkage (DAS-13), while diagnostic accuracy or clinical effectiveness are out of scope [49]. The platform supported cohort preparation by combining pseudonymized CT imaging data from PACS (DA-03) with demographic data from the CIS (DA-02 and DP-21) using record linkage. Eligibility criteria (age, broad consent permitting data use and recontact, CT scan type, and CT scan age) were applied via structured queries (DP-21 and DR-05). Imaging metadata (DP-06) stored in the data lake were processed within the platform (MDM-01, DS-05, and DS-02 - 2) [48]. Imaging metadata and radiological reports were ingested as daily CSV reports into the S3 bucket in the staging layer (DA-01, DA-01-2, DA-01-4, DS-03, and DP-07), persisted in the data lake, and linked via the MPI (DP-20-3 and DP-20-5) [48]. Patient demographics were accessed via the MPI, and consent information was verified through the PCM prior to analysis (DAS-13, DP-20-3, and DSC-04). The CDR provided structured *ICD-10-GM* diagnoses (DS-08, DP-01, DP-23, and DP-33 - 1) for eligibility filtering and querying (DR-05 and DS-13). Query results were exported as CSV datasets (MPI Identifiers, patient addresses, and vital status; DAS-12), containing 1744 unique eligible patients for follow-up processing. Subsequent patient communication was outside the platform's scope.

The structured integration of imaging metadata and patient demographics enabled automated cohort identification and facilitated follow-up planning. Beyond the immediate gains for recruitment and research workflows, the setup also provides long-term efficiency: once implemented, additional cohorts can now be identified based on the continuously updated data within the platform. Thus, the radiology scenario fulfills 23 of 39 (59%) identified platform requirements. The evaluation activities and corresponding evidence at the requirements level are detailed in Table 1. A consolidated assessment of the resulting quality attributes according to ISO/IEC 25010 across both scenarios is summarized in Table 2.

Operational Assessment

In addition to the scenario-based requirement evaluation, we collected operational data metrics to provide quantitative

XSL•FO
RenderX

evidence of the platform's real-world data handling capabilities. The platform is currently connected to 10 heterogeneous clinical source systems, including ORBIS (CIS with RIS module), OPUS-L (LIS), MEONA (medication documentation), MUSE (ECG), Deep Unity (PACS), Viewpoint (gynecology and obstetrics), ODSEasy (clinical cancer registry and tumor documentation system), NEXUS (pathology information system), COPRA (patient data management system), and CentraXX (CTMS).

On a weekly basis, the platform processes approximately 180,000 HL7 ADT, 120,000 HL7 ORU, and 120,000 HL7 billing account record messages, in addition to around 30,000 DICOM studies, 12,000 RIS reports, 2500 ECG datasets, and 7000 discharge letters. The total alphanumeric data volume amounts to approximately 1 to 1.5 GB per week. As of early 2026, the platform has processed approximately 46.5 million ADT, 39.8 million ORU, 28.1 million billing account record messages, 38.3 million DICOM studies, 144,000 ECG datasets, and 4.1 million discharge letters. The MPI contains demographic records for over 2.3 million patients (since 2000), and the openEHR-based CDR stores structured data from approximately 1.5 million patients in more than 85 million compositions (including ~12 million diagnoses, ~13 million procedures, ~10 million medication requests, ~2.8 million scores from intensive care units, ~340,000 UKSH broad consents, and information on ~58,000 biospecimens). This data volume results from the platform's integration strategy: when specific data elements are requested, they are not only integrated for the selected cohort but also systematically incorporated for all patients across legacy and ongoing data. For example, a request for anesthetics data in a specific intensive care unit in 2020 leads to the integration of anesthetic data from the introduction of the respective IT system to the present, including future records.

Discussion

Principal Findings

This study presents the implementation and requirement-based evaluation of a health knowledge management platform at UKSH, assessed against 39 functional requirements and 2 real-world scenarios. To overcome the traditional separation between patient care and biomedical research, the platform integrates heterogeneous clinical data (eg, CIS, PACS, and LIS) into a structured, interoperable environment [50]. Requirements were derived through stakeholder focus groups, building upon the developed requirements catalog by Kinast et al (30/39, 77%) [25]. The catalog provided a comprehensive and systematically defined foundation for key interoperability and data integration requirements, enabling efficient requirements formalization and traceable coverage. In total, 9 of 39 (23%) additional requirements were specified using ISO 29148-compliant syntax to address context-specific needs. Overall, 30 of 39 (77%) requirements are fully implemented, while the cardiology scenario fulfills 26 of 39 (67%) and the radiology scenario fulfills 23 of 39 (59%).

However, its functional scope extends well beyond the functionalities necessary to support the selected scenarios. For example, the platform integrates radiological images (eg, x-ray, CT, and magnetic resonance imaging) together with corresponding free-text reports (DA-01-3) [48], although structured extraction and storage of report content in the CDR are not yet implemented. Two requirements stay partially or not yet fulfilled: single sign-on, which is currently limited to selected applications (eg, MTB; DSC-12), and linkage of external public data sources (DP-19). The latter, including integration of regional death registries, is on hold due to legal and organizational constraints.

The integration of patient profile-based information from public sources (DP-19), for example, to support advanced decision support use cases, remains conceptually defined but not yet operational due to legal and technical constraints. Together with the incomplete single-sign-on integration (DSC-12), these gaps reflect the current implementation status rather than architectural limitations.

Our platform supports a range of data-driven applications via the data access layer, although user group-specific interfaces for predictive data analysis (DR-03-1) have yet to be developed. A key challenge is the prevalence of unstructured clinical documentation, which limits structured reuse and data quality [51]. To address this, the normalization layer is being extended with natural language processing capabilities, and an initial pipeline for allergy information extraction from discharge letters (DP-02) has been implemented [31].

An MTB application has been deployed, enabling visualization of oncological histories, patient comparison and cohort identification, and analytical functionalities (DAS-03, DR-03, and DR-03-3) [39,52]. It is also in active use to support patient care as part of the MTB meetings. Additional functionalities include end-user cohort exploration (DR-01) and data entry (DR-05) that are planned but currently restricted to specific user groups, that is, system administrators or study nurses, respectively. Our platform shares its openEHR-based foundation with the system described by Biermann et al [53], whose work on patient trajectory analysis and predictive interfaces illustrates the extensibility of this architecture. However, some applications envisioned remain on the roadmap (eg, patients like mine and end user–enabled cohort exploration). The identified gaps highlight the need for further iterative development to realize the platform's full potential. Its design was guided by defined key aspects (see Key Aspects of the Platform section), aiming to ensure relevant quality attributes such as interoperability, reusability, data consistency, and usability aligned with the specific requirements of the UKSH environment.

A key architectural decision was the adoption of an openEHR-based CDR instead of a FHIR-based repository. While FHIR primarily supports real-time data exchange, openEHR provides a semantically rich, longitudinal data model suitable for persistent storage and consistent querying across both patient and cohort levels, posing an essential requirement for secondary use and advanced analytics [54]. Its archetype-based approach enables explicit clinical content modeling, ensuring semantic consistency and modular reuse across use cases and front ends while maintaining interoperability via FHIR interfaces. This decision is supported by the growing adoption of openEHR across European health care systems, including countries such

XSL•FO
RenderX

as Slovenia [55], Sweden [56], Norway, and Finland [57] as well as major academic hospitals such as Karolinska University Hospital and University Hospital Basel [58], and the region of Catalonia [59]. The approach also aligns with the open platform principles defined by the Aperta Foundation [60], which promote openness, semantic interoperability, and modularity.

To complement this foundation, FHIR-based applications could be integrated into the application layer, enabling patients to contribute personal health data (PGHD). Conceptually, such applications could also be considered part of the PEHR within the source layer, with data integrated bottom-up via the data lake. The platform architecture and interfaces are designed to support both integration options, allowing source systems either to maintain independent persistence or to rely on the platform's normalization layer as a shared persistence layer. We did not consider integrating the knowledge management platform directly into the CIS, as CIS platforms are primarily designed for billing, individual patient-level care, and documentation rather than cohort-level analysis or research data export. Establishing a separate platform increases architectural flexibility, without interfering with clinical workflows or billing-related functionality constraints. This also ensures that patient care remains uncompromised, as CIS and corresponding systems are not burdened with additional loads from research-related queries. Furthermore, it aligns with Findable, Accessible, Interoperable, Reusable principles by enhancing data accessibility and promoting reuse in research contexts [24].

In the platform's initial architectural design, IHE profiles played a significant role [15]. Currently, Patient Identifier Cross-Referencing [33] and Patient Demographics Query [34] are used within the MPI, along with Advanced Patient Privacy Consents [61] for consent management. Many openEHR compositions are additionally persisted as IHE XDS-compliant documents. However, as XDS-based access retrieval is not required for current scenarios, data retrieval is performed via openEHR AQL. The architectural separation of patient identifiers, consent data, and medical content supports data protection and enables modular governance. In alignment with IHE principles, our platform manages an MPI, a consent repository, and a CDR to persist the respective data types. This modular structure enables precise permission management via fine-grained consent enforcement [35]. Currently, the platform supports consents for the research-related use of patient care data, leftover biomaterial, recontact in case of incidental findings, and recontact for future research projects (ie, recruitment and follow-up) [35]. Because the platform processes both consented and nonconsented patient data, it can also support legally authorized research models beyond informed consent, such as opt-out–based research based on a specific regulation.

Limitations

One limitation concerns the stakeholder selection in the requirements elicitation process, which primarily emphasized functional requirements. Nonfunctional aspects such as scalability, security, performance, and usability were considered at the architectural level but not systematically elicited as explicit requirements. The platform's modular and layered architecture follows separation of concerns principles, allowing nonfunctional requirements to be addressed within specific components. However, these aspects were not formally evaluated in this study.

Although the requirements catalog by Kinast et al [25] facilitated the formalization and data integration requirements, it did not address user interface aspects. Consequently, 9 additional requirements had to be specified separately, expanding the scope of the RE process. The RE process did not include a dedicated study-specific stakeholder analysis. However, participation in the HiGHmed consortium ensured broad clinical domain representation across multiple use cases [15]. Additional requirements were elicited through engagement with clinicians involved in the evaluated scenarios.

The platform builds on an openEHR CDR, a technology traditionally applied to electronic health records in institutional or regional infrastructures [62-64]. Our application as a knowledge management platform integrating both research and patient care contexts represents an extended use of this technology.

Structured data are a prerequisite not only for research scenarios and applications but also for decision support. However, clinical documentation still heavily relies on unstructured free-text, either entered directly into the CIS or recorded through dictation and transcription. As a result, requirements such as DR-03-1 to DR-03-3 remain a challenge, as the respective functionalities depend on structured data. This highlights the need to establish structured data as a foundation before enabling advanced care functionalities. To address this, we plan to expand the platform's normalization layer to further include natural language processing and information extraction features, enabling the processing and structuring of the large volume of free-text reports from the CIS and its subsystems [31].

Finally, the evaluation has primarily focused on research-oriented scenarios. Broader deployment in routine care, especially for real-time decision support, remains to be assessed. Although the platform provides the necessary technical interfaces, additional application development is required to evaluate clinical integration at scale, fully demonstrating the platform's potential in both biomedical research and patient care settings.

Risk Assessment

Every modification of an IT system poses risks to data consistency, availability, and quality, particularly when integrating heterogeneous source systems via different interfaces [65]. Each data transformation step introduces the potential for errors, and repeated transformation increases the risk of inconsistencies between datasets. Our approach of acquiring raw data and transforming them centrally only once reduces repeated transformations but still introduces transformation risks. To mitigate these, we use automated processing, code reviews, and structured testing procedures [66]. Nevertheless, occasional errors occur. Once identified, the corresponding extract-transform-load process is corrected, and the affected data are reintegrated from the original S3 storage. Although

XSL•FO
RenderX

issues in the source data itself cannot be resolved, detected anomalies can be flagged within the integrated data view.

Data reliability varies by type and depends on the original purpose of documentation. For example, laboratory measurements are typically generated through standardized processes, whereas billing-related data such as diagnoses or procedures follow regulatory and reimbursement-driven documentation logic. Platform transparency regarding data provenance and quality is therefore essential. Each newly integrated data source typically includes the complete dataset of the corresponding system, including live and legacy data, rather than cohort-specific subsets. This "integrate once, use many" strategy promotes long-term sustainability and data reuse.

Beyond technical risks, governance and long-term maintenance represent additional challenges. Sustainable operation depends on strong organizational support and compliance with data privacy regulations. At UKSH, the platform is institutionally anchored through a board resolution and supported by internal funding as well as national initiatives (Network University Medicine), which adds to its long-term stability.

Comparison With Prior Work

Our health knowledge management platform integrates both health care and research functionalities within a unified architecture. This contrasts with many existing initiatives that focus either on specific use cases or predominantly on research-oriented data integration [67-69].

In Germany, initiatives such as the Medical Informatics Initiative and Network University Medicine [16] have significantly advanced health care data integration. As part of the HiGHmed consortium, we set the foundation for our platform by introducing open platforms and interoperability principles to enhance health care and research across institutional boundaries [15]. Building on this foundation, our platform extends these concepts by integrating both research and health care functionalities within a single integrated platform. Similarly, MIRACUM and DIFUTURE focus on cross-institutional data integration and secondary analyses to inform clinical practice and medical innovation [70,71]. While these initiatives emphasize research data use, our platform bridges the gap by unifying research and health care workflows, enabling bidirectional data reuse.

Internationally, comparable efforts exist. For instance, the Mayo Clinic Platform facilitates secure access to deidentified clinical data for research and innovation [72]. Unlike Mayo's Enterprise Data Trust, which primarily functions as a nontransactional analytics repository, our platform additionally supports real-time clinical workflows (eg, MTB) and research data capture directly from patient care processes, enabling immediate translational use within UKSH. Similarly, the World Health Organization's Integrated Data Platform supports large-scale population health monitoring [73] but does not integrate institutional patient care and research workflows within a unified platform.

Overall, while prior initiatives have significantly contributed to secondary use of health care data, our platform emphasizes the operational integration of health care and research through centralized consent management and a layered architecture based on separation of concerns principles.

Conclusions

This study describes and evaluates the UKSH health knowledge management platform, demonstrating how its multistandards approach effectively addresses interoperability challenges while ensuring data consistency. The scenario-based evaluation illustrates the platform's capability to support research based on real-world health care data. Its layered architecture and harmonization processes provide a structured data access layer that enables both clinical and research applications.

Future developments will focus on (1) expanding the application layer with cohort exploration tools and AI-supported data analysis tools [18] and (2) developing user-specific interfaces to enhance clinical usability. Hereby, we expect benefits through a reduction of manual data preparation for cohort identification, enabling more efficient study planning and data-driven health care. These applications will be developed to create a direct impact on patient care based on the partially and unfulfilled requirements, which will undergo further evaluation. Inspired by the "Ten Topics to Get Started in Medical Informatics Research" [74], our platform has the potential to improve several key areas in medical informatics, pushing the boundaries of both clinical practice and research capabilities.

Acknowledgments

The authors thank all members of the Medical Data Integration Center and HiGHmed team at both University Hospital Schleswig-Holstein campus—Lübeck and Kiel. Furthermore, the authors thank the members of IMPETUS and NEMO at the University Hospital Schleswig-Holstein campus Kiel.

Funding

This work was funded by the German Federal Ministry of Education and Research (grants 01ZZ1802T, 01ZZ2302F, 01KX2121, 01ZZ2011, 01KX2524, and 16KISA063).

Data Availability

All openEHR templates used can be accessed via the Clinical Knowledge Manager [75].

XSL•FO
RenderX

Authors' Contributions

BK selected the method for the requirements elicitation and evaluation, chose the scenarios, conducted the evaluation, and led the drafting of the manuscript, including the initial and final versions. BS derived the initial architecture based on the requirements. BK provided expertise on the cardiology scenario, organized and moderated the expert discussions, and incorporated their results into the manuscript draft. HU conducted the radiology and neurology scenarios and contributed to expert insights. TB conceptualized the integration of patient-generated health data into the platform. BS and BB refined the platform's architecture, actively participated in expert discussions, and provided critical input to the manuscript. AKK-S contributed to the requirements for biomaterial information and study registry. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mind map of identified functional aspects. This appendix contains the structured mind map generated during the stakeholder focus group sessions. The visualization summarizes and clusters the identified functional aspects of the envisioned knowledge management platform and documents the conceptual basis for the derivation of the 39 functional requirements.
[PDF File, 148 KB - medinform_v14i1e83608_app1.pdf]

Multimedia Appendix 2
Additional scenario implemented and evaluated with the platform. This appendix contains the detailed description of the neurology evaluation scenario. It further provides the complete requirements mapping table (39 functional requirements) including architectural layer assignment and fulfillment across the cardiology, radiology, and neurology scenarios, as well as the corresponding literature references from Kinast et al.
[DOCX File, 91 KB - medinform_v14i1e83608_app2.docx]

References

1. Sa S. Big data in healthcare management: a review of literature. Am J Theor Appl Bus 2018;4(2):57. [doi: 10.11648/j.ajtab.20180402.14]
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
3. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219. [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
4. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018;1(1):5. [doi: 10.1038/s41746-017-0012-2] [Medline: 31304291]
5. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2(1):77. [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
6. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc 2012 Jun;19(e1):e119-e124. [doi: 10.1136/amiajnl-2011-000508] [Medline: 22437072]
7. Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform 2001 Aug;34(4):285-298. [doi: 10.1006/jbin.2001.1024] [Medline: 11977810]
8. LeSueur D. 5 reasons healthcare data is unique and difficult to measure. Health Catalyst. 2014. URL: https://www.healthcatalyst.com/insights/5-reasons-healthcare-data-is-difficult-to-measure [accessed 2023-10-13]
9. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79. [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
10. Schubert A, Truxillo TM, Guthrie R. The power of the driver diagram: a conceptual approach. In: Schubert A, Kemmerly SA, editors. Optimizing Widely Reported Hospital Quality and Safety Grades, An Ochsner Quality and Value Playbook: Springer International Publishing; 2022:21-29. [doi: 10.1007/978-3-031-04141-9_4]
11. Ethier JF, McGilchrist M, Barton A, et al. The TRANSFoRm project: experience and lessons learned regarding functional and interoperability requirements to support primary care. Learn Health Syst 2018 Apr;2(2):e10037. [doi: 10.1002/lrh2.10037] [Medline: 31245579]
12. Kock-Schoppenhauer AK, Schreiweis B, Ulrich H, et al. Communications in Computer and Information Science: Springer Science and Business Media Deutschland GmbH; 2021:269-284. [doi: 10.1007/978-3-030-87657-9_21]
13. Dhayne H, Haque R, Kilany R, Taher Y. In search of big medical data integration solutions—a comprehensive survey. IEEE Access 2019;7:91265-91290. [doi: 10.1109/ACCESS.2019.2927491]
14. O'Dell C, Grayson CJ. If Only We Knew What We Know: Identification and Transfer of Internal Best Practices: Free Press; 1998. [doi: 10.2307/41165948]

XSL•FO
RenderX

15. Haarbrandt B, Schreiweis B, Rey S, et al. HiGHmed—an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018 Jul;57(S 01):e66-e81. [doi: 10.3414/ME18-02-0002] [Medline: 30016813]
16. Albashiti F, Thasler R, Wendt T, Bathelt F, Reinecke I, Schreiweis B. Data integration centers-from a concept in the Medical Informatics Initiative to its local implementation in the Network of University Medicine. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Jun;67(6):629-636. [doi: 10.1007/s00103-024-03879-5] [Medline: 38662020]
17. Venable J, Pries-Heje J, Baskerville R. FEDS: a Framework for Evaluation in Design Science Research. Eur J Inf Syst 2016 Jan;25(1):77-89. [doi: 10.1057/ejis.2014.36]
18. Arshad K, Ardalan S, Schreiweis B, Bergh B. Integrating an AI platform into clinical IT: BPMN processes for clinical AI model development. BMC Med Inform Decis Mak 2025 Jul 2;25(1):243. [doi: 10.1186/s12911-025-03087-4] [Medline: 40604977]
19. ISO/IEC/IEEE 29148:2018—systems and software engineering—life cycle processes—requirements engineering. International Standardization of Organization. 2018. URL: https://www.iso.org/standard/72089.html [accessed 2026-04-23]
20. Bühne S, Glinz M, van Loenhoud H, Staal S. Certified professional for requirements engineering—foundation level syllabus. CPRE. Version 3.3.0. IREB. 2026. URL: https://cockpit-v1.ireb.org/media/pages/downloads/cpre-foundation-level-syllabus/f305d53dab-1775054523/cpre_foundationlevel_syllabus_en_v.3.3.0.pdf [accessed 2026-04-22]
21. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148. [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
22. Maiden N, Gizikis A. Where do requirements come from? IEEE Softw 2001;18(5):10-12. [doi: 10.1109/52.951486]
23. Pohl K. Requirements Engineering: Springer; 2010.
24. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
25. Kinast B, Ulrich H, Bergh B, Schreiweis B. Functional requirements for medical data integration into knowledge management environments: requirements elicitation approach based on systematic literature analysis. J Med Internet Res 2023 Feb 9;25:e41344. [doi: 10.2196/41344] [Medline: 36757764]
26. Hafedh M, Elkharraz A, Mcheick H. Understanding separation of concerns. Presented at: Early Aspects: Aspect-Oriented Requirements Engineering and Architecture Design; Mar 21, 2004 URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4b53c4af6254e7530fa4652d6fb0013680835ab1#page=76 [accessed 2026-04-23]
27. ISO/IEC 25010:2023–systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—product quality model. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/78176.html [accessed 2025-12-18]
28. Jahresbericht 2024 interaktiv. Universitätsklinikum Schleswig-Holstein. 2024. URL: https://user-4lj71tt.cld.bz/Jahresbericht-2024-interaktiv/14-15/ [accessed 2026-01-05]
29. Kleinert P, Gebhardt M, Buckow K, et al. Germany's approach for the secondary use of federated real-world data in the German Portal for Medical Research Data. Stud Health Technol Inform 2024 Aug 22;316:1704-1708. [doi: 10.3233/SHTI240751] [Medline: 39176538]
30. Ablaß T, Simon F, Schwitlick C, et al. Development, involvement and use of an overarching in-house registry for clinical trials. Stud Health Technol Inform 2024 Aug 30;317:123-128. [doi: 10.3233/SHTI240846] [Medline: 39234714]
31. Macedo M, Wiedekopf J, Hillmer T, Schreiweis B, Saalfeld S, Ulrich H. From text to knowledge: an end-to-end extraction pipeline for clinical information. Stud Health Technol Inform 2025 Aug 7;329:1074-1078. [doi: 10.3233/SHTI251004] [Medline: 40776022]
32. Kreps J, Narkhede N, Rao J. Kafka: a distributed messaging system for log processing. Presented at: NetDB 2011; Jun 12, 2011 URL: https://pages.cs.wisc.edu/~akella/CS744/F17/838-CloudPapers/Kafka.pdf [accessed 2025-07-07]
33. ITI technical framework, volume 1: integration profiles—Section 5: Patient Identifier Cross-Referencing (PIX). IHE International. 2020. URL: https://profiles.ihe.net/ITI/TF/Volume1/ch-5.html [accessed 2025-07-07]
34. ITI technical framework, volume 1: integration profiles—Section 8: Patient Demographics Query (PDQ). IHE International. 2020. URL: https://profiles.ihe.net/ITI/TF/Volume1/ch-8.html [accessed 2025-07-07]
35. Bronsch T, Anywar M, Kinast B, et al. Implementation of a Patient Consent Management at a German Medical Data Integration Center. Stud Health Technol Inform 2025 Aug 7;329:1626-1627. [doi: 10.3233/SHTI251134] [Medline: 40776151]
36. Incubator: UKSH Universitätsklinikum Schleswig-Holstein. OpenEHR Clinical Knowledge Manager. 2024. URL: https://ckm.highmed.org/ckm/incubators/1246.152.60 [accessed 2026-04-08]
37. Ammon D, Kurscheidt M, Buckow K, et al. Interoperability Working Group: core dataset and information systems for data integration and data exchange in the Medical Informatics Initiative. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Jun;67(6):656-667. [doi: 10.1007/s00103-024-03888-4] [Medline: 38753022]
38. Kinast B, Wiedekopf J, Ulrich H, Schreiweis B. Extracting LOINC codes from a laboratory information system's index: addressing semantic interoperability with web scraping. Stud Health Technol Inform 2025 Apr 24;324:234-239. [doi: 10.3233/SHTI250194] [Medline: 40270418]

XSL•FO
RenderX

39. Reimer N, Unberath P, Busch H, et al. Challenges and experiences extending the cBioPortal for cancer genomics to a Molecular Tumor Board platform. Stud Health Technol Inform 2021 Nov 18;287:139-143. [doi: 10.3233/SHTI210833] [Medline: 34795098]
40. Rosati RA, McNeer JF, Starmer CF, Mittler BS, Morris JJJ, Wallace AG. A new information system for medical practice. Arch Intern Med 1975 Aug;135(8):1017-1024. [doi: 10.1001/archinte.1975.00330080019003] [Medline: 1156062]
41. Gombar S, Callahan A, Califf R, Harrington R, Shah NH. It is time to learn from patients like mine. NPJ Digit Med 2019;2(1):16. [doi: 10.1038/s41746-019-0091-3] [Medline: 31304364]
42. Kinast B, Rohde H, Anywar M, et al. Implementation of an atrioventricular valve intervention registry: comparative study of REDCap vs. CDR-based openEHR registry. Stud Health Technol Inform 2024 Aug 22;316:1069-1073. [doi: 10.3233/SHTI240595] [Medline: 39176974]
43. Follow-up, published template. openEHR Clinical Knowledge Manager HiGHmed, Project UKSH. 2022. URL: https://ckm.highmed.org/ckm/templates/1246.169.189 [accessed 2026-04-08]
44. Echokardiographie, published template. openEHR Clinical Knowledge Manager HiGHmed, Project UKSH. 2021. URL: https://ckm.highmed.org/ckm/templates/1246.169.85 [accessed 2026-04-08]
45. Anamnesis-ext, published template. openEHR Clinical Knowledge Manager HiGHmed, Project UKSH. 2023. URL: https://ckm.highmed.org/ckm/templates/1246.169.3576 [accessed 2026-04-08]
46. KDS_Laborbericht_UKSH, draft template. openEHR Clinical Knowledge Manager HiGHmed, Project UKSH. URL: https://ckm.highmed.org/ckm/templates/1246.169.3582 [accessed 2026-04-08]
47. KDS_Medikationseintrag, published template. openEHR Clinical Knowledge Manager HiGHmed, Project UKSH. 2021. URL: https://ckm.highmed.org/ckm/templates/1246.169.3568 [accessed 2026-04-08]
48. Ulrich H, Anywar M, Kinast B, Schreiweis B. Large-scale standardized image integration for secondary use research projects. Stud Health Technol Inform 2024 Jan 25;310:174-178. [doi: 10.3233/SHTI230950] [Medline: 38269788]
49. Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care–embedded biobanking: understanding and reasons to donate in a large patient sample. Genet Med 2018 Jan;20(1):76-82. [doi: 10.1038/gim.2017.82]
50. Beyer M, Lenz R, Kuhn KA. Health information systems (Informationssysteme im Gesundheitswesen). it - Information Technology 2006 Jan 1;48(1):6-11. [doi: 10.1524/itit.2006.48.1.6]
51. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;17(1):128-144. [doi: 10.1055/s-0038-1638592] [Medline: 18660887]
52. Bossenz J, Manuilova I, Weise AB, et al. Prototypical visualization of patient similarities in cBioPortal to enhance decision-making in Molecular Tumor Boards. Stud Health Technol Inform 2025 May 15;327:487-491. [doi: 10.3233/SHTI250385] [Medline: 40380495]
53. Biermann P, Baier C, Vietor AC, et al. An openEHR based infection control system to support monitoring of nosocomial bacterial clusters and contacts. npj Digit Med ;8(1). [doi: 10.1038/s41746-025-01795-9]
54. Pedrera-Jiménez M, García-Barrio N, Frid S, et al. Can OpenEHR, ISO 13606, and HL7 FHIR work together? An agnostic approach for the selection and application of electronic health record standards to the next-generation health data spaces. J Med Internet Res 2023 Dec 28;25(1):e48702. [doi: 10.2196/48702] [Medline: 38153779]
55. Bajrić S. Building a sustainable ecosystem for eHealth in Slovenia: opportunities, challenges, and strategies. Digit Health 2023;9:20552076231205743. [doi: 10.1177/20552076231205743] [Medline: 37808243]
56. Karolinska Uuniversity Hospital and Tieto Caretech drive healthcare data integration. Tieto. 2025. URL: https://www.tietoevry.com/en/success-stories/2025/karolinska-university-hospital-and-tietoevry-care-drive-healthcare-data-integration/ [accessed 2026-01-05]
57. Norway PH. Norway, Sweden and Finland as forerunners in open ecosystems and openEHR. In: Roadmap to Successful Digital Health Ecosystems: Elsevier; 2022:457-471. [doi: 10.1016/B978-0-12-823413-6.00011-2]
58. OWT to advance digital health at University Hospital Basel with openEHR standard. OWT (Open Web Technology). 2025. URL: https://www.owt.swiss/en/news/owt-to-advance-digital-health-at-university-hospital-basel-with-openehr-standard/ [accessed 2026-01-06]
59. Piera-Jiménez J, Carot-Sans G. Catalonia's journey to the open platform paradigm in healthcare. In: Scheuer A, Studzinski J, editors. Digital Maturity in Hospitals: Springer Nature Switzerland; 2025:111-127. [doi: 10.1007/978-3-031-80704-6_8]
60. Defining an open platform. Apperta Foundation. 2018. URL: https://apperta.org/assets/Apperta_Defining_an_Open_Platform.pdf [accessed 2026-01-06]
61. IHE IT infrastructure—Technical Framework supplement—Advanced Patient Privacy Consents (APPC). Integrating the Healthcare Enterprise. 2016. URL: https://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Suppl_APPC.pdf [accessed 2026-04-23]
62. Christensen B, Ellingsen G. Evaluating model-driven development for large-scale EHRs through the openEHR approach. Int J Med Inform 2016 May;89:43-54. [doi: 10.1016/j.ijmedinf.2016.02.004] [Medline: 26980358]
63. Min L, Wang L, Lu X, Duan H. Case study: applying OpenEHR archetypes to a clinical data repository in a Chinese hospital. Stud Health Technol Inform 2015;216:207-211. [Medline: 26262040]

XSL•FO
RenderX

64. Bode L, Schamer S, Böhnke J, et al. Tracing the progression of sepsis in critically ill children: clinical decision support for detection of hematologic dysfunction. Appl Clin Inform 2022 Oct;13(5):1002-1014. [doi: 10.1055/a-1950-9637] [Medline: 36162433]
65. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. CODATA 2015 May 22;14:2. [doi: 10.5334/dsj-2015-002]
66. Förstel S, Förstel M, Gallistl M, Zanca D, Eskofier BM, Rothgang EM. Data quality in hospital information systems: lessons learned from analyzing 30 years of patient data in a regional German hospital. Int J Med Inform 2024 Dec;192:105636. [doi: 10.1016/j.ijmedinf.2024.105636] [Medline: 39357217]
67. Ozaydin B, Zengul F, Oner N, Feldman SS. Healthcare research and analytics data infrastructure solution: a data warehouse for health services research. J Med Internet Res 2020 Jun 4;22(6):e18579. [doi: 10.2196/18579] [Medline: 32496199]
68. Rance B, Canuel V, Countouris H, Laurent-Puig P, Burgun A. Integrating heterogeneous biomedical data for cancer research: the CARPEM infrastructure. Appl Clin Inform 2016;7(2):260-274. [doi: 10.4338/ACI-2015-09-RA-0125] [Medline: 27437039]
69. Doods J, Bache R, McGilchrist M, et al. Piloting the EHR4CR feasibility platform across Europe. Methods Inf Med 2014;53(4):264-268. [doi: 10.3414/ME13-01-0134] [Medline: 24954881]
70. Prokosch HU, Acker T, Bernarding J, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91. [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
71. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data Integration for Future Medicine (DIFUTURE): an architectural and methodological overview. Methods Inf Med 2018 Jul;57:e57-e65. [doi: 10.3414/ME17-02-0022] [Medline: 30016812]
72. Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17(2):131-135. [doi: 10.1136/jamia.2009.002691] [Medline: 20190054]
73. Nadal S, Jovanovic P, Bilalli B, Romero O. Operationalizing and automating data governance. J Big Data 2022;9(1):117. [doi: 10.1186/s40537-022-00673-5] [Medline: 36532842]
74. Wolfien M, Ahmadi N, Fitzer K, et al. Ten topics to get started in medical informatics research. J Med Internet Res 2023 Jul 24;25:e45948. [doi: 10.2196/45948] [Medline: 37486754]
75. Clinical knowledge manager. HIGHmed. URL: https://ckm.highmed.org/ckm/projects/1246.152.61 [accessed 2026-04-27]

Abbreviations

ADT: admit, discharge, transfer
AI: artificial intelligence
AQL: Archetype Query Language
CDR: clinical data repository
CDSS: clinical decision support system
CIS: clinical information system
CT: computed tomography
CTMS: Clinical Trial Management System
DICOM: Digital Imaging and Communications in Medicine
ECG: electrocardiography
ePMA: Electronic Prescribing and Medicine Administration
FEDS: Framework for Evaluation in Design Science Research
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
HL7 V2: Health Level 7 Version 2
***ICD-10-GM*:** *International Classification of Diseases and Related Health Problems, 10th Revision, German Modification*
IHE: Integrating the Healthcare Enterprise
ISO: International Organization for Standardization
ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
LIS: laboratory information system
LOINC: Logical Observation Identifiers Names and Codes
MPI: Master Patient Index
MTB: Molecular Tumor Board
ORU: observation result unsolicited
PACS: picture archiving and communication system
PCM: Patient Consent Manager
PEHR: patient cross-enterprise health record
PGHD: patient-generated health data

XSL•FO
RenderX

RE: requirement engineering
RIS: radiology information system
UKSH: University Hospital Schleswig-Holstein
XDS: Cross-Enterprise Document Sharing

Edited by C Perrin; submitted 05.Sep.2025; peer-reviewed by A Gupta, B Senst, M Meleka; revised version received 02.Mar.2026; accepted 02.Mar.2026; published 30.Apr.2026.

Please cite as:
Schreiweis B, Kinast B, Ulrich H, Bronsch T, Kock-Schoppenhauer AK, Bergh B
A Data-Centric Approach for Health Care and Research in a Health Knowledge Management Platform: Implementation and Requirement-Based Evaluation Study
JMIR Med Inform 2026;14:e83608
URL: https://medinform.jmir.org/2026/1/e83608
doi:10.2196/83608

© Björn Schreiweis, Benjamin Kinast, Hannes Ulrich, Tobias Bronsch, Ann-Kristin Kock-Schoppenhauer, Björn Bergh. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Modeling Diabetes Risk and Progression With Public Health Data: Ontology-Guided, Simulation-Capable Digital Twin Study

Qingrui Li[1]; Kapileshwor Ray Amat[1]; Eric L Johnson[2]; Juan Li[1]

[1]Department of Computer Science, North Dakota State University, PO Box 6050, NDSU Dept #2455, Fargo, ND, United States

[2]Department of Family and Community Medicine, School of Medicine & Health Sciences, University of North Dakota, Grand Forks, ND, United States

Corresponding Author:
Juan Li
Department of Computer Science, North Dakota State University, PO Box 6050, NDSU Dept #2455, Fargo, ND, United States

Abstract

Background: Digital twins (DTs) offer a paradigm for health care by enabling data-driven, simulation-capable representations of individual health trajectories. However, DT development remains limited by the scarcity of standardized, temporally structured, and multidomain data suitable for modeling chronic disease progression. Most existing DT studies rely on narrowly scoped or proprietary datasets, restricting generalizability. Public health datasets, such as the Midlife in the United States study, provide rich biopsychosocial information but are underused due to structural complexity and lack of semantic integration frameworks.

Objective: This study aimed to develop and evaluate an ontology-guided, agent-orchestrated framework for constructing offline, simulation-capable, and progression-aware DTs from public health datasets. Using diabetes as a case study, the framework integrates agent-based orchestration, medical ontologies, and large language model (LLM)–assisted semantic reasoning with machine learning to support explainable feature structuring, risk prediction, and predictive "what-if" progression analysis.

Methods: A 6-stage DT framework was developed and applied to Midlife in the United States wave 2 (baseline) and wave 3 (follow-up) data. Ontology- and LLM-assisted feature selection identified predictors across biological, behavioral, psychosocial, and socioeconomic domains. Cleaned and harmonized data were used to train predictive models (random forest, eXtreme gradient boosting, and logistic regression) to estimate diabetes onset at follow-up. A state-transition simulator was implemented to model between-wave progression dynamics, quantify transitions across low-, medium-, and high-risk states, and evaluate predictive "what-if" scenarios such as weight reduction and lifestyle improvement. Model performance was assessed using accuracy, F_1 score, area under the receiver operating characteristic curve (AUC), and calibration metrics.

Results: From 9976 candidate variables, ontology- and LLM-guided selection retained the top 200 relevant predictors spanning biological, behavioral, psychosocial, and socioeconomic domains. Predictive modeling achieved strong discrimination, with random forest (AUC=0.82, accuracy=0.76) and eXtreme gradient boosting (AUC=0.81, accuracy=0.75) outperforming logistic regression (AUC=0.78). The state-transition simulator reproduced realistic progression patterns: 33.9% (1414/4174) of participants changed risk states between waves, and the high-risk group increased from 10.8% (451/4174) to 32.2% (1344/4174). Next-state prediction accuracy reached 92.5%. Predictive "what-if" analyses showed that with a simulated 10% weight reduction, model-estimated diabetes cases decreased by 98 (from 576 to 478). A placebo test (0% weight change) produced less than 0.3% difference in risk distribution, confirming model stability.

Conclusions: This study presents a foundational, ontology-guided, and agent-orchestrated framework for constructing offline, simulation-capable, and progression-aware DTs from public datasets. By combining semantic reasoning, multidomain predictors, and predictive "what-if" progression simulation, the framework transforms static population data into longitudinal, interpretable representations of individual health trajectories. The proof-of-concept application to diabetes demonstrates that public health data can support robust and explainable DT models for exploratory risk analysis and hypothesis generation, without implying causal intervention effects or direct clinical decision support.

(JMIR Med Inform 2026;14:e87374) doi:10.2196/87374

KEYWORDS

digital twin; diabetes; public health datasets; Midlife in the United States; MIDUS; multiagent AI; ontologies; large language models; risk prediction; simulation; chronic disease

XSL•FO
RenderX

Introduction

Background

Digital twins (DTs) are virtual representations of physical entities that are designed to represent and simulate real-world systems to predict and optimize outcomes [1]. In health care, DTs can serve as individualized, data-driven models of patients, enabling proactive monitoring, personalized intervention, and scenario-based simulation [2]. By integrating longitudinal biomedical, behavioral, psychosocial, and environmental data, DTs offer the potential to transform chronic disease management from reactive treatment to anticipatory care.

Challenges in Health Care DTs

Despite this promise, the development of health care DTs is constrained by several challenges. First, there is a scarcity of high-quality, structured datasets that support temporal modeling and personalization, especially in chronic disease contexts [3]. Second, there is a lack of standardized frameworks for organizing heterogeneous health variables across domains and time points [4,5]. Finally, feature selection and semantic integration remain complex tasks, requiring both domain expertise and explainability [6]. These limitations are particularly acute for type 2 diabetes, where long-term behavioral, physiological, and psychosocial factors interact in complex, nonlinear ways. Current prototypes typically rely on narrowly scoped, proprietary datasets collected in controlled clinical settings, which limits generalizability and reuse and constrains their applicability beyond specific populations or case settings.

Opportunity in Public Health Datasets

In contrast, large-scale public health datasets, such as the MIDUS (Midlife in the United States) study [7], the National Health and Nutrition Examination Survey [8], and the UK Biobank [9], contain rich biopsychosocial profiles of individuals over time. These resources offer an underused opportunity for constructing inclusive and population-grounded DT representations.

However, they present structural and semantic challenges, including inconsistent variable definitions, fragmented modules, missing values, and a lack of standardized feature hierarchies. To date, most uses of these datasets have focused on regression or correlation analyses rather than longitudinal, simulation-capable modeling of individual risk trajectories.

Objective

This study aimed to develop a foundational, ontology-guided, and agent-orchestrated framework for constructing offline, simulation-capable DTs from public health datasets. In this work, we adopt a restricted and pragmatic definition of a health care DT, referring to an offline, simulation-capable virtual representation derived from population-level data, rather than a continuously synchronized or clinically validated dynamic twin.

Using diabetes risk prediction as a demonstrative case study, we sought to do the following:

1. Transform longitudinal public health data into temporally structured, semantically enriched feature sets suitable for DT modeling;
2. Integrate medical ontologies and large language model (LLM)–assisted semantic reasoning to support intelligent and explainable feature selection; and
3. Develop prediction and simulation models that support risk estimation and predictive "what-if" scenario analysis rather than causal intervention inference.

By addressing data reuse, semantic organization, and interpretable predictive simulation, this work establishes a new methodological pathway for constructing progression-aware and explainable DTs from population-level data, with a focus on predictive modeling and exploratory analysis rather than real-time operation, causal inference, or direct clinical decision-making.

Methods

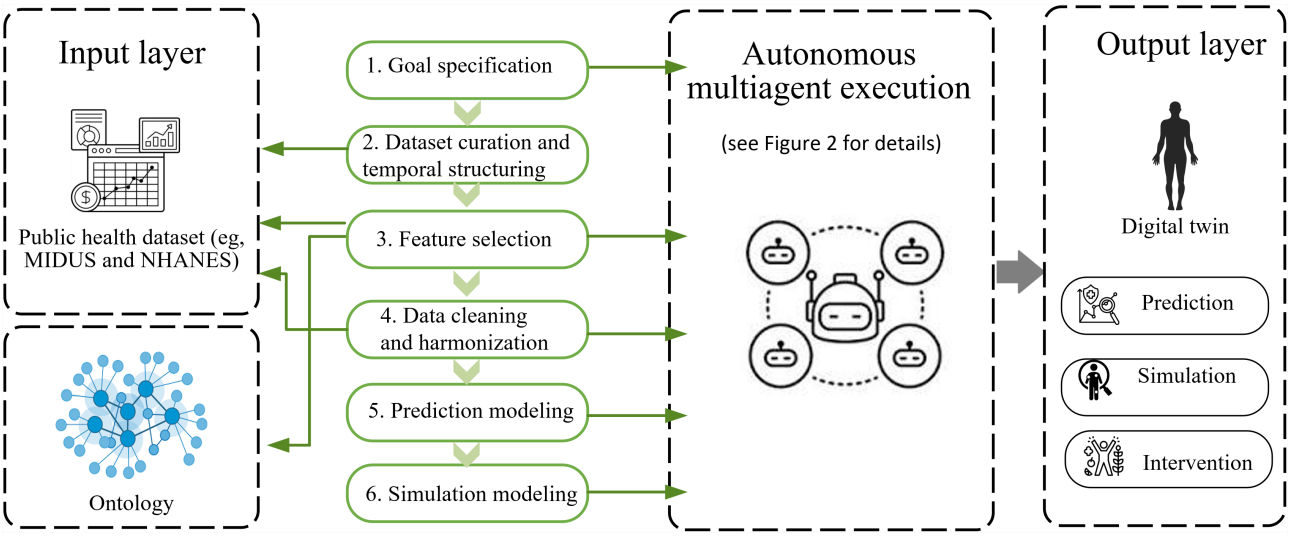
Framework Design

Overview

We developed a foundational, modular framework for constructing offline, simulation-capable DTs from public health data, illustrated using diabetes risk prediction in the MIDUS cohort. As illustrated in Figure 1, the framework is organized into six stages: (1) goal specification, (2) dataset curation and temporal structuring, (3) ontology- and LLM-assisted feature selection, (4) data cleaning and harmonization, (5) prediction modeling, and (6) simulation modeling.

XSL•FO
RenderX

Figure 1. Multiagent, ontology-guided framework for constructing digital twins from public health data. MIDUS: Midlife in the United States; NHANES: National Health and Nutrition Examination Survey.



These stages are coordinated through an agent-orchestrated architecture, in which specialized agents are assigned distinct roles, such as ontology mapping, LLM-based reasoning, data harmonization, prediction, and simulation. Agents operate as LLM-driven reasoning components with bounded autonomy, enabling semantic interpretation, ambiguity resolution, and iterative refinement rather than rigid rule-based execution.

This design supports scalability, transparency, and modularity, enabling each component to function independently while contributing to the overall DT framework without implying real-time synchronization, causal inference, or clinical decision-making capability.

Scope of the DT Implementation

The DT instantiated in this study is offline and population-trained, using paired observations from MIDUS waves 2 and 3. Models are trained once and applied deterministically during simulation. The system is intended for progression-aware research and hypothesis exploration, rather than real-time operation, causal inference, or clinical decision support.

Agent Autonomy and Interagent Reasoning

While the overall workflow of the framework follows a logical progression from data curation to prediction and simulation, the internal execution relies on agentic reasoning rather than rigid pipeline heuristics. Unlike traditional software modules that execute predefined rules (eg, “if X then Y”), the agents in this framework leverage LLMs to perform probabilistic reasoning, semantic interpretation, and adaptive decision-making.

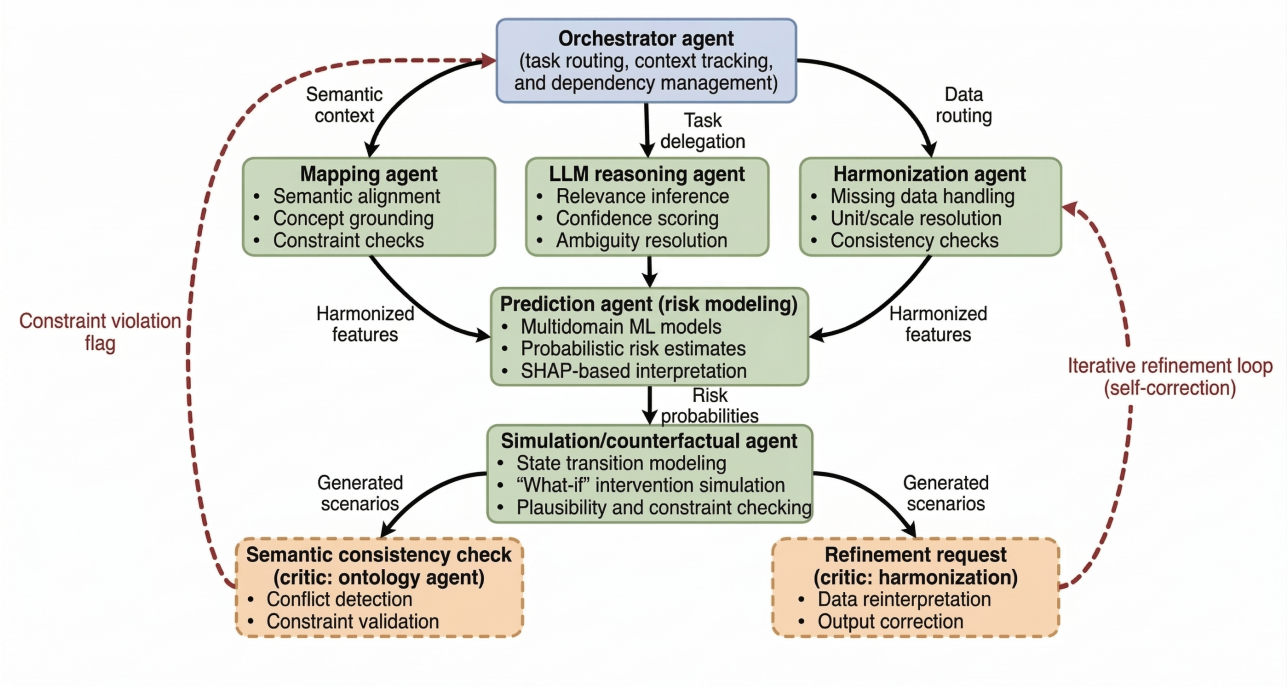
As illustrated in Figure 2, the framework departs from a static pipeline by employing a “Hub-and-Spoke” agentic topology centered on a dynamic orchestrator agent. This architecture enables 3 distinct forms of autonomy not present in the standard modular software:

1. Semantic adaptability: Agents resolve ambiguity in MIDUS variable definitions by inferring meaning from metadata, ontology context, and prior knowledge rather than requiring explicit hard-coded mappings.
2. Contextual interaction: Outputs produced by one agent condition the reasoning of downstream agents, and simulation outputs are checked against ontology constraints, enabling iterative refinement rather than one-pass execution.
3. Reflective evaluation: The simulation agent evaluates generated counterfactual scenarios for plausibility and internal consistency and regenerates outputs when conflicts are detected.

These behaviors distinguish the framework from a static processing pipeline and motivate the use of an agentic abstraction, while remaining intentionally lightweight and constrained compared to fully autonomous multiagent systems.

XSL•FO
RenderX

Figure 2. The multiagent collaboration architecture. Unlike a linear software pipeline, the orchestrator agent dynamically manages task execution based on semantic context. The architecture features explicit iterative refinement loops (red-dashed lines), where "Critic" functions (eg, semantic consistency check) autonomously flag errors or implausible simulations, triggering replanning and data reinterpretation by the functional agents. LLM: large language model; ML: machine learning; SHAP: Shapley Additive Explanations.



To complement the conceptual overview in Figure 2 and Table 1, we provide 2 brief execution traces (Textboxes 1 and 2) illustrating (1) semantic ambiguity resolution during feature selection and (2) critic-triggered refinement during counterfactual simulation.

Table . Comparison between a traditional pipeline and the proposed agentic framework.

Aspect	Traditional pipeline	Proposed agentic framework
Logic	Deterministic rules	LLM[a]-based probabilistic reasoning
Ambiguity handling	Manual preprocessing	Semantic inference and confidence scoring
Interaction	Unidirectional	Iterative, context-conditioned refinement
Adaptation	None	Output regeneration under constraints
Failure mode	Error or crash	Reasoned approximation

[a]LLM: large language model.

Textbox 1. Example trace: resolving semantic ambiguity in a MIDUS variable.

Input (ambiguous item). The orchestrator agent receives the MIDUS item "Falling or staying sleep frequency - 30 days", which conflates sleep-onset and sleep-maintenance difficulty and thus lacks a single unambiguous semantic label.

Ontology-grounded context. The orchestrator agent delegates grounding to the ontology mapping agent, which retrieves compact ontology evidence (top-matched concepts and a few relations) consistent with a broader sleep disturbance symptom frequency construct and packages it as a structured context for downstream inference.

Structured decision (LLM reasoning). Conditioned on the MIDUS metadata and retrieved ontology context, the LLM reasoning agent outputs a constrained JSON decision, eg,

{"classification":"Indirect","relevance_score":0.6,"reasoning":"Sleep disturbance is associated with metabolic dysregulation and increased diabetes risk."}

The orchestrator agent records a canonical label (eg, sleep_disturbance_symptom_frequency_30d) for harmonization and retains the feature in the ranked set used for modeling.

Reflective check (bounded loop). A Critic function validates schema/score range and checks semantic consistency with the retrieved evidence; if the score is below a predefined threshold, the orchestrator triggers a single re-query, otherwise the decision is accepted and logged.

XSL•FO
RenderX

Textbox 2. Example trace: critic-flagged implausible counterfactual and orchestrated refinement.

Input (candidate scenario). The simulation agent proposes a candidate "what-if" perturbation that attempts to vary a prescription-medicine checklist variable to reduce predicted diabetes risk; this proposal is subsequently screened by the Critic before any simulation is accepted.

Critic flag (implausible/nonactionable). A Critic function flags the scenario as nonactionable and semantically inconsistent with the DT's intervention scope: prescription history is a downstream proxy/outcome, not an upstream modifiable determinant for counterfactual exploration.

Orchestrator replanning (context-conditioned refinement). The orchestrator agent removes the nonmodifiable proxy variable and replans the scenario using upstream, modifiable factors already supported by the simulator (eg, body weight and physical activity). The simulator is rerun under these perturbations.

Refined output (accepted). The refined scenarios yield stable, monotonic risk changes consistent with the simulator's sensitivity design (eg, 5% and 15% weight reductions decrease average predicted risk), and the accepted trace is logged.

Goal Specification Using the Population, Intervention, Comparison, and Outcome Approach

Defining a clear objective is essential for aligning data engineering, modeling, and evaluation with actionable health care needs. In this framework, the objective is formalized through the Population, Intervention, Comparison, and Outcome (PICO) approach [10], which provides a structured foundation for prediction and simulation.

For this study, the following points should be noted:

- Population (P): adults aged ≥40 years in MIDUS wave 2, a group at elevated risk of type 2 diabetes due to age-related metabolic changes and cumulative lifestyle factors
- Intervention (I): a data-driven DT integrating multidomain predictors (eg, BMI, physical activity, sleep quality, psychosocial stress, and family history) derived from surveys and biomarkers to simulate diabetes risk
- Comparison (C): benchmark models using only biomedical or clinical variables, assessing the added value of including psychosocial and behavioral factors
- Outcome (O): diabetes onset at wave 3, defined through self-reported diagnosis, medication use, or biomarkers (fasting glucose and HbA_{1c})

This specification clarifies the DT's purpose, guides variable selection, and defines evaluation benchmarks. The PICO structure also generalizes easily to other chronic diseases or datasets.

Dataset Curation and Temporal Structuring

High-quality DTs require longitudinal data that capture changes in individual health states. Although MIDUS contains only 2 major waves, it provides sufficient longitudinal contrast for modeling transitions between baseline (wave 2) and follow-up (Wave 3).

The following procedures were applied:

1. Wave alignment and participant tracking: wave 2 served as baseline and wave 3 as follow-up. Participants with complete paired data were retained, yielding 4173 individuals with consistent identifiers and matched features.
2. Outcome labeling: diabetes onset was defined using wave 3 data (self-report, medication use, and biomarkers such as fasting glucose and HbA_{1c}) to provide ground truth for supervised learning.
3. Feature harmonization: variables across biological, behavioral, psychosocial, and socioeconomic domains were standardized and harmonized. Comparable constructs (eg, BMI, activity, stress, and income) were aligned for consistency across waves.
4. Temporal pairing for transition modeling: instead of computing explicit Δ features, health progression was represented implicitly through paired records (wave 2 → wave 3). Each baseline profile served as model input and its follow-up state as the target.

This process transformed MIDUS from a general-purpose survey into a longitudinal, simulation-ready dataset consistent with DT principles. Although limited to 2 waves, the paired design enabled population-level transition modeling and individualized progression analysis over nearly a decade.

Ontology-Assisted and LLM-Driven Feature Selection

Feature-Selection Overview

Public health datasets, such as MIDUS, include thousands of variables spanning biomedical, psychological, behavioral, and social domains. Selecting the most relevant features for DT construction requires scalable, explainable reasoning beyond traditional statistical filters. Conventional feature selection techniques, such as association tests [11], correlation thresholds [12], or heuristics [13], often overlook semantically subtle but influential variables (eg, stress or coping behavior).

To overcome this limitation, we developed a hybrid feature selection pipeline that integrates domain ontologies with LLM reasoning in a retrieval-augmented setting [14]. Feature selection is treated as a structured semantic reasoning task rather than a purely statistical filter [14].

Semantic Enrichment With Ontologies

Each MIDUS variable was mapped to a curated set of diabetes-related and study-design ontologies, enabling the LLM to reason across biomedical and psychosocial dimensions. The following ontologies were integrated:

- The Diabetes Mellitus Treatment Ontology (DMTO) [15] capturing treatment strategies, complications, and management protocols
- The Common Diabetes Medications Ontology [16] linking medication variables to pharmacological mechanisms (of action)
- The FHIR (Fast Healthcare Interoperability Resources) And SSN (Semantic Sensor Network)-based Type 1 diabetes Ontology (FASTO) [17] providing real-time monitoring and device-related concepts

XSL•FO
RenderX

- The BioMedBridges Diabetes Ontology [18] defining diagnostic and phenotypic concepts for translational research

Crucially, to contextualize the role of each variable within the MIDUS longitudinal design, we incorporated the Study Cohort Ontology [19,20]. This allowed the agent to understand the structural purpose of a variable (eg, as a baseline demographic, psychosocial instrument, or clinical outcome), which is essential for distinguishing direct from indirect relevance.

Variable metadata (labels, descriptions, and coding) were mapped using keyword matching, semantic similarity, and vector embeddings. The result was a structured, ontology-aligned representation that enabled the retrieval-augmented generation (RAG) agent to reason over deep conceptual relationships rather than surface-level text similarity.

In this framework, the ontology layer serves primarily as a semantic and organizational scaffold rather than as a direct performance-optimizing component. Its role is to (1) impose a consistent hierarchical structure over heterogeneous public health variables, (2) guide LLM-assisted feature selection toward domain-relevant predictors, and (3) enable interpretable grouping and reuse of features across modeling stages. Accordingly, ontology integration is intended to improve transparency, explainability, and reproducibility, rather than to guarantee higher predictive accuracy.

Relevance Classification With LLMs

After enrichment, the RAG agent evaluated each variable's relevance to diabetes risk prediction using structured prompts containing ontology context and metadata. Variables were classified as directly relevant, indirectly relevant, or not relevant, each with a confidence score and explanatory note. The agent operated in an iterative loop: (1) extract variable metadata and ontology matches; (2) construct prompts integrating semantic context; (3) generate classification, justification, and confidence; (4) requery or reformulate if confidence is low; and (5) log outputs for transparency and reproducibility.

This process yielded a feature set annotated with ontology links, metadata, and reasoning traces, ensuring transparency, auditability, and explainability for downstream modeling.

The confidence scores generated by the LLM represent the internal consistency of the relevance classification given the ontology-grounded prompt context and were used to control iterative requerying rather than to impose a hard inclusion threshold. To assess face validity, we manually audited the resulting feature set with respect to established diabetes risk factors. Core predictors, such as age, BMI, blood pressure, family history, physical activity, smoking, mental health indicators, and medication use, were consistently retained, and we did not observe systematic exclusion of known clinical predictors. Disagreements primarily occurred for borderline psychosocial or survey-specific variables rather than for established biomedical risk factors.

The LLM-assisted feature selection is not intended to replace statistical feature selection or expert curation, nor is it strictly required to achieve high discrimination with flexible tree-based models. Its primary contribution lies in semantic coverage, explainability, and scalability when working with large, heterogeneous public health datasets, where manual curation is impractical and purely statistical filters risk omitting contextually relevant variables.

Data Cleaning and Harmonization

Following feature selection, a standardized cleaning pipeline ensured data quality and consistency:

- Variable-type normalization: unified categorical, ordinal, and Boolean variables
- Missing-data handling: mean/mode imputation for low-missing features; k-nearest neighbor (kNN) imputation for complex patterns
- Scale and unit standardization: normalized continuous features and harmonized measurement units
- Semantic harmonization: merged equivalent constructs across waves via ontology tags
- Outlier detection: applied statistical thresholds and domain-specific plausibility rules

MIDUS contains both item-level missingness and wave-level attrition, which may correlate with health status and socioeconomic factors. We therefore do not assume that data are missing completely at random. We selected kNN imputation because it preserves multivariate relationships among mixed-type variables and avoids distributional assumptions required by parametric multiple imputation models, which are difficult to validate in high-dimensional survey data with thousands of heterogeneous variables. kNN uses observed similarity across individuals to infer plausible values and is widely used in biomedical machine-learning pipelines where predictive performance and internal consistency are the primary goals. Nevertheless, kNN does not correct for missing-not-at-random mechanisms, and any bias arising from differential dropout or nonresponse may propagate into downstream prediction and simulation.

Prediction and Simulation Modeling

The final stage of the framework focuses on developing models that support both risk prediction and scenario-based simulation, which are 2 core capabilities of a DT.

Prediction Modeling

Prediction models estimate the likelihood of diabetes onset at follow-up using multidomain features. This component extends beyond traditional biomedical risk scores by incorporating behavioral, psychosocial, and socioeconomic predictors.

Model objectives emphasize personalization, interpretability, and temporal continuity, ensuring that predictions remain clinically meaningful and adaptable as health states evolve.

Simulation Modeling

Given the availability of only 2 MIDUS waves, the simulator implements a single-step, between-wave state-transition model (wave 2 → wave 3). Therefore, it supports coarse-grained progression modeling and scenario-based sensitivity analysis but not continuous, multistep temporal dynamics or online updating. The counterfactual simulations are predictive rather

XSL•FO
RenderX

than causal: they quantify how model-estimated diabetes risk changes when selected inputs are perturbed, holding the learned model structure fixed. They do not represent real-world intervention effects. The simulation module operationalizes the DT's capacity to model how an individual's predicted diabetes risk evolves under changing behavioral and psychosocial conditions. Conceptually, the simulator is designed as a coarse-grained state-transition system, where each participant's health state at a future timepoint is estimated as a function of their baseline characteristics and current states. These states summarize multidomain attributes, including biological (eg, BMI and hypertension), behavioral (eg, physical activity, smoking, and alcohol use), and psychosocial (eg, anxiety and depression), into interpretable clusters that represent distinct risk levels.

To model transitions, a supervised learning function approximates P(St+1 ∣ St,Xt), where St and St+1 denote latent health states at consecutive waves and Xt represents observed features. Given the availability of only 2 MIDUS waves, this formulation supports single-step, between-wave transition modeling rather than fine-grained or continuous temporal dynamics. This enables the simulator to reproduce observed between-wave progression patterns progression patterns and estimate next-state probabilities under hypothetical scenarios.

Counterfactual reasoning is implemented through predictive scenario simulation, in which modifiable factors (eg, body weight, physical activity, smoking, alcohol use, stress, and depression scores) are systematically perturbed to assess their influence on predicted transition probabilities and future diabetes risk. The simulator then recomputes the resulting state distributions and outcome probabilities using the trained transition and outcome models.

By comparing such simulated trajectories, the DT supports multifactor "what-if" analysis at both individual and population levels, enabling exploration of model sensitivity to changes in key risk factors. Methodologically, this simulation framework complements the predictive component: while prediction estimates a static risk at a given timepoint, the simulator enables exploratory analysis of coarse-grained health evolution across survey waves.

The counterfactual simulations implemented in this framework are predictive rather than causal. They quantify how model-estimated risk changes when selected input variables are perturbed and do not represent real-world intervention effects. Accordingly, these simulations should be interpreted as model-based sensitivity analyses and hypothesis-generation tools, rather than estimates of causal impact or decision-grade clinical recommendations.

Integration

Together, prediction and simulation form the methodological foundation of the DT. Prediction provides individualized risk estimates, whereas simulation allows for the exploration of dynamic scenarios and potential interventions. This dual design ensures that the DT framework not only forecasts outcomes but also offers actionable insights for prevention and management in diabetes and other chronic diseases.

Agentic LLM Workflow and Reproducibility

The feature selection component is implemented as a retrieval-augmented, LLM-driven agent. We used OpenAI GPT-3.5 turbo (OpenAI Python SDK, openai>=1.0.0) with a temperature of 0.3 to reduce stochasticity. For each MIDUS variable, the agent receives the variable's name and metadata together with ontology-grounded context retrieved from 6 diabetes-related ontologies: Diabetes Mellitus Diagnosis and Support Ontology, DMTO, BioMedBridges Diabetes Ontology, Common Diabetes Medications Ontology, FASTO, and the Study Cohort Ontology.

Additional details are provided here:

- Ontology RAG. Feature names and ontology concept labels are embedded using sentence transformers (allMiniLML6v2). For each feature, the topk =5 ontology concepts above a similarity threshold of 0.3 are retrieved. For each retrieved concept, ontology relationships are extracted as triples (eg, Type2Diabetes → hasRiskFactor → Obesity), and short concept explanations are generated and filtered. These labels, Internationalized Resource Identifiers, triples, and explanations form a structured context block injected into the LLM prompt.
- Prompt and output schema. The LLM is prompted with a fixed system instruction and a user template containing (1) MIDUS variable metadata, (2) retrieved ontology context, and (3) an explicit scoring rubric. The model must return valid JSON with fields {relevance_score ∈ [0,1], classification (Direct/Indirect/Unrelated), reasoning}.
- Iteration and thresholds. Features are processed in batches of 40. If a feature's relevance score is below 0.3, it is requeried once with expanded ontology context; if it is still below the threshold, it is excluded. This bounded control flow ensures deterministic termination. Final selection is performed by ranking features by relevance_score and retaining the top 200.
- Reproducibility. We fix the feature-processing order, ontology-retrieval parameters, and LLM temperature and record all configuration parameters, data and ontology versions, and LLM outputs in run logs. Baseline selectors (L1-regularized logistic regression [LASSO] and MI) use random_state =42 to ensure deterministic comparisons.
- Agent orchestration layer. The orchestration layer acts as a central controller that sequentially invokes specialized agents, including ontology mapping, feature selection, data harmonization, prediction, and simulation. Agents do not communicate peer-to-peer; all interactions are mediated by the orchestrator. Feedback loops are implemented, wherein simulation outputs are checked for semantic consistency against ontology constraints and refined when inconsistencies are detected.

Ethical Considerations

This study used publicly available, deidentified data and did not involve interaction with individuals or the use of identifiable private information; therefore, it did not involve a human subject as defined under 45 CFR 46.102 [21], and institutional review board review was not sought. Informed consent for this secondary analysis was not required. Data were accessed and

XSL•FO
RenderX

analyzed in accordance with the applicable MIDUS and Inter-university Consortium for Political and Social Research data use terms [22]. To ensure compliance with these terms, the LLM was strictly provided with variable names, feature descriptions, and other documentation-level metadata; no participant-level or individual MIDUS data was entered into the LLM.

Results

Cohort Characteristics

After applying inclusion criteria, the analytic cohort consisted of 4174 participants with complete records in both MIDUS wave 2 (baseline) and wave 3 (follow-up), allowing individual trajectories to be tracked longitudinally. The sample size, therefore, remained constant across waves (n=4174), although the number of participants with diabetes increased over time. The mean age at baseline was 56.0 (SD 12.4) years; 54.8% (n=2288) were female. By wave 3615 (14.7%), participants met the diabetes definition, indicating a moderate class imbalance.

As shown in Table 2, across both waves, participants who had or developed diabetes were older, more often male, had lower education and household income, and showed higher cardiometabolic burden (eg, BMI and hypertension) than those without diabetes. Behavioral differences were smaller, but people with diabetes reported slightly higher levels of light physical activity at follow-up, possibly reflecting postdiagnosis behavior changes. By wave 3, depressive symptoms were also higher in the diabetes group.

Table . Descriptive statistics of the study cohort (MIDUS waves 2 and 3), stratified by diabetes status.

Variable	MIDUS[a] 2 overall (n=4174)	No diabetes (n=3730)	Diabetes (n=444)	*P* value	MIDUS 3 overall (n=4174)	No diabetes (n=3559)	Diabetes (n=615)	*P* value
Demographics								
Age (y), mean (SD)	56.0 (12.4)	55.4 (12.3)	61.6 (11.2)	<.001	64.0 (11.4)	63.3 (11.4)	68.5 (10.7)	<.001
Male, n (%)	1886 (45.2)	1643 (44.0)	243 (54.7)	<.001	1886 (45.2)	1561 (43.9)	325 (52.8)	.004
Female, n (%)	2288 (54.8)	2087 (56.0)	201 (45.3)	<.001	2288 (54.8)	1998 (56.1)	290 (47.2)	.004
Education (y), mean (SD)	7.3 (2.5)	7.3 (2.5)	6.6 (2.6)	<.001	7.6 (2.5)	7.7 (2.5)	7.0 (2.6)	<.001
Household income, median (IQR), $	57,500 (29,000-95,272)	59,500 (30,250-97,500)	46,250 (22,000-75,024)	<.001	68,875 (34,250-119,938)	72,500 (36,938-123,750)	55,250 (25,000-95,250)	<.001
Biological factors								
BMI, mean (SD)	27.9 (5.7)	27.4 (5.4)	31.9 (6.9)	<.001	28.2 (6.1)	27.6 (5.6)	31.5 (7.6)	<.001
Hypertension, n (%)	1263 (30.3)	982 (26.3)	281 (63.3)	<.001	1049 (38.9)	786 (34.6)	263 (61.6)	<.001
Behavioral factors								
Ever smoked, n (%)	2022 (65.1)	1782 (64.5)	240 (69.4)	.09	1321 (62.4)	1118 (62.1)	203 (64.2)	.51
Light activity (h/wk), mean (SD)	2.0 (1.5)	2.0 (1.5)	2.3 (1.7)	<.001	2.0 (1.5)	1.9 (1.5)	2.5 (1.7)	<.001
Psychosocial factors								
Depression score, mean (SD)	0.6 (1.7)	0.6 (1.7)	0.6 (1.7)	.91	0.6 (1.7)	0.5 (1.6)	0.7 (1.9)	.03

[a]MIDUS: Midlife in the United States.

Feature-Selection Outcomes

Ontology- and LLM-assisted feature selection yielded a curated set of predictors spanning biological, behavioral, psychosocial, and socioeconomic domains. A representative subset is shown in Table 3. As expected, classic biomedical risk factors—BMI, hypertension, waist-to-hip ratio, and current weight—were classified as directly relevant. The pipeline also surfaced function- and lifestyle-related variables (eg, basic activities of daily living, shortness of breath while walking, and health limits walking >1 mile), which capture mobility and cardiopulmonary reserve and are not always selected by purely statistical filters. In addition, socioeconomic (household income) and psychosocial (anxiety symptoms) variables were retained as

XSL•FO
RenderX

indirectly relevant, extending the model beyond a biomedical-only view of diabetes risk.

Table . Representative variables selected by ontology- and LLM[a]-assisted feature selection for diabetes risk prediction.

Variable	Domain	Relevance classification	Ontology alignment (examples)	Notes/explanation
Body Mass Index - Metric	Biological	0.8 (direct)	DMTO[b]: 0000362 (overweight/obesity)	BMI is a well-established indicator of obesity and a major risk factor for diabetes.
Age	Biological	0.8 (direct)	DMDSONT[c]: Patient_hasAge; SNOMED CT[d]: chronological age	Age is a strong predictor due to accumulated lifestyle exposures and metabolic changes.
Hypertension (ever diagnosed, 12 mo)	Biological	0.8 (direct)	DMDSONT: Hypertension_b	Hypertension is a known comorbidity and a direct risk factor for type 2 diabetes.
Any chronic conditions (12 mo)	Biological	0.8 (indirect)	DMTO: DDO_0000201 (chronic disease)	Multimorbidity contributes to systemic inflammation and increases diabetes risk.
Weight compared to 5 years ago	Biological	0.8 (indirect)	DMTO: DDO_0000130 (weight trajectory)	Weight gain over time is closely associated with higher diabetes risk.
Current weight (pounds)	Biological	0.8 (direct)	DMTO: DDO_0000130 (weight)	Absolute weight is a central predictor of diabetes risk.
Waist-to-hip ratio	Biological	0.8 (direct)	DMTO: DDO_0000320 (waist circumference)	Central obesity predicts diabetes more strongly than BMI alone.
Basic activities of daily living (3-item)	Behavioral	0.7 (indirect)	DMTO: DDO_0000127 (activity level)	Functional limitations are correlated with higher chronic disease risk.
Shortness of breath while walking	Behavioral	0.7 (indirect)	DMTO: DDO_0009259 (respiratory function)	Reduced capacity reflects poorer cardiometabolic health.
Health limits walking >1 mile	Behavioral	0.7 (indirect)	DMTO: DMTO_0001830 (walk)	Mobility restrictions often signal underlying cardiometabolic burden.
Household income	Socioeconomic	0.6 (indirect)	FASTO[e]: socioeconomic factor	Lower income is associated with limited health care access and higher diabetes risk.
Anxiety symptoms	Psychosocial	0.6 (indirect)	DMTO: DDO_0000483 (anxiety)	Anxiety can indirectly influence diabetes risk via stress responses and lifestyle.

[a]LLM: large language model.

[b]DMTO: Diabetes Mellitus Treatment Ontology.

[c]DMDSONT: Diabetes Mellitus Diagnosis and Support Ontology.

[d]SNOMED_CT: Systematized Nomenclature of Medicine—Clinical Terms.

[e]FASTO: FHIR (Fast Healthcare Interoperability Resources) And SSN (Semantic Sensor Network)-based Type 1 diabetes Ontology.

To support transparency, the full ranked list of the top 200 features, including ontology alignment, relevance level, and notes, is provided in Multimedia Appendix 1. These results confirm that the ontology-guided, multiagent pipeline can identify multidomain predictors consistent with a biopsychosocial model of diabetes.

To evaluate the specific contribution of the ontology-guided RAG agent, we conducted a comparative ablation study against 2 established statistical baselines:

1. Lasso: a standard method for sparse feature selection that penalizes nonzero coefficients.
2. Mutual information (MI): a univariate filter that selects features based on their statistical dependency with the target.

All methods were restricted to a top-K setting (K=200) to ensure a fair comparison of information density. The complete list of 200 features can be found in Multimedia Appendices 1-3. Selection quality was evaluated based on downstream predictive

XSL•FO
RenderX

performance (area under the receiver operating characteristic curve [AUC]) and a manual semantic audit of feature relevance.

As shown in Table 4, the ontology-guided RAG agent achieved an AUC of 0.82, outperforming Lasso (AUC=0.73) and MI (AUC=0.78). This indicates that the agent identifies predictive signals that purely statistical filters overlook.

Table . Ablation study—feature-selection method comparison.

Feature-selection method	Number of features selected	AUC[a] (prediction)	"Noise" features included[b]	Clinical relevance score[c]
Baseline 1: mutual information	200	0.78	Medium (110)	Low
Baseline 2: Lasso[d] (L1)	200	0.73	High (135)	Medium
Ours (ontology+LLM[e] agent)	200	0.82	Low (70)	High

[a]AUC: area under the receiver operating characteristic curve.

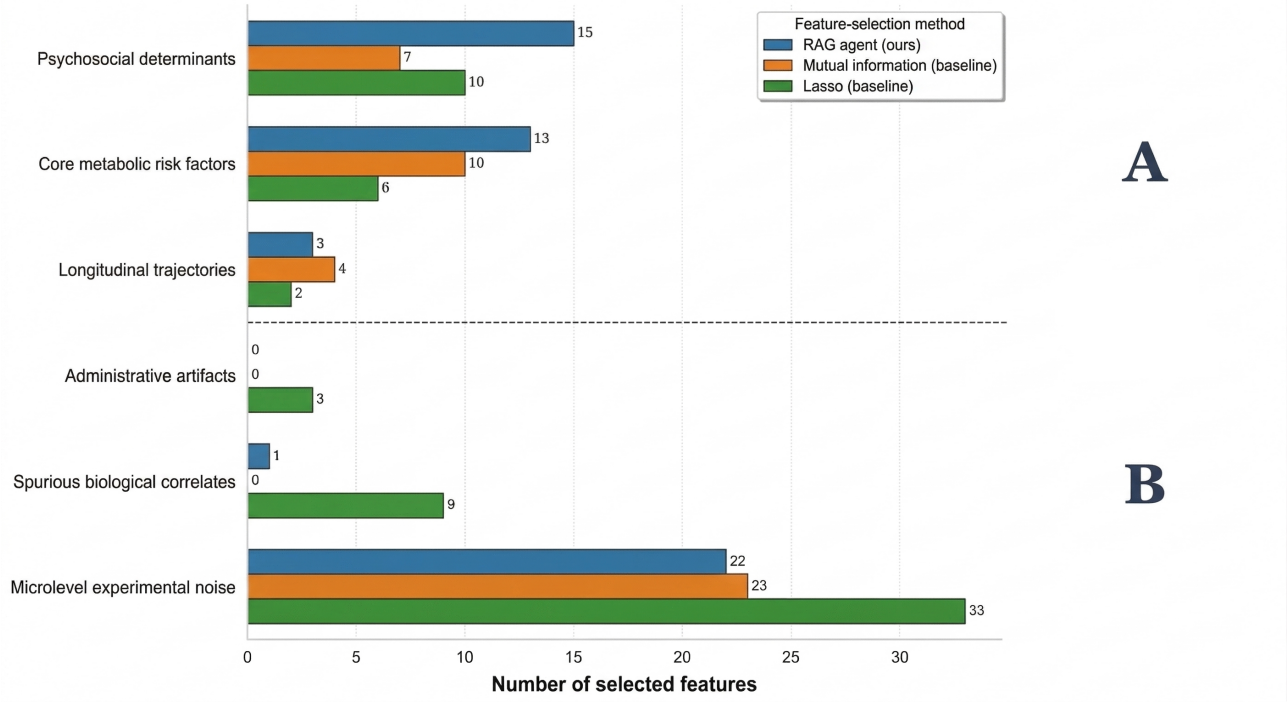
[b]"Noise" features refer to variables classified through manual semantic audit as artifacts or administratively correlated variables without clear causal or actionable relevance to diabetes risk.

[c]Clinical relevance score indicates the overall qualitative relevance of the selected feature set to established diabetes risk factors, based on manual semantic audit.

[d]Lasso: L1-regularized logistic regression.

[e]LLM: large language model.

Figure 3. Comparative semantic audit of feature selection. (A) Target-relevant features (higher count is better) and (B) artifactual and noisy features (lower count is better). The RAG agent (blue) selects significantly more target-relevant features (eg, psychosocial determinants and metabolic risk factors) while suppressing artifactual noise compared to L1-regularized logistic regression (Lasso, green) and mutual information (orange). RAG: retrieval-augmented generation.



Beyond raw accuracy, the agentic approach demonstrated superior semantic precision. Statistical methods frequently selected "noisy" artifacts: variables that correlate with the target but lack causal validity.

- Reduction of artifacts: Lasso selected 135 features classified as noise or administrative artifacts (eg, "Interviewer ID" and "Data Collection Site"), whereas the RAG agent reduced this to 70, effectively filtering out spurious correlations.
- Actionable versus proxy features: as detailed in Figure 3, the RAG agent prioritized modifiable lifestyle constructs essential for DT simulation, selecting 21 physical activity items compared to only 4 for Lasso. Conversely, Lasso heavily relied on "Prescription Medicine" checklists (31 items vs 7 for RAG). While prescription history is a strong statistical proxy for disease, it represents a downstream outcome rather than an upstream, modifiable risk factor. By favoring upstream determinants (eg, "Lose 10 pounds by diet"), the agentic framework ensures the resulting DT is actionable.

XSL•FO
RenderX

Figure S1 and Tables S1–S5 (Multimedia Appendix 4) provide a detailed ablation and semantic analysis of feature-selection methods. These results show that ontology-guided RAG selects a substantially different and nonredundant feature set relative to Lasso and MI, with greater emphasis on modifiable lifestyle and psychosocial constructs and fewer spurious or administrative artifacts, thereby supporting its suitability for interpretable, intervention-oriented DT modeling.

Prediction Performance

Prediction models were trained on wave 2 features to predict diabetes onset in wave 3. Of the 4174 participants, 615 (14.7%) developed diabetes, so the synthetic minority oversampling technique was applied during training to mitigate class imbalance. Three models were evaluated: random forest, eXtreme gradient boosting, and logistic regression. A 5-fold cross-validation with stratified splits was used.

All models showed good discrimination, with tree-based models performing the best:

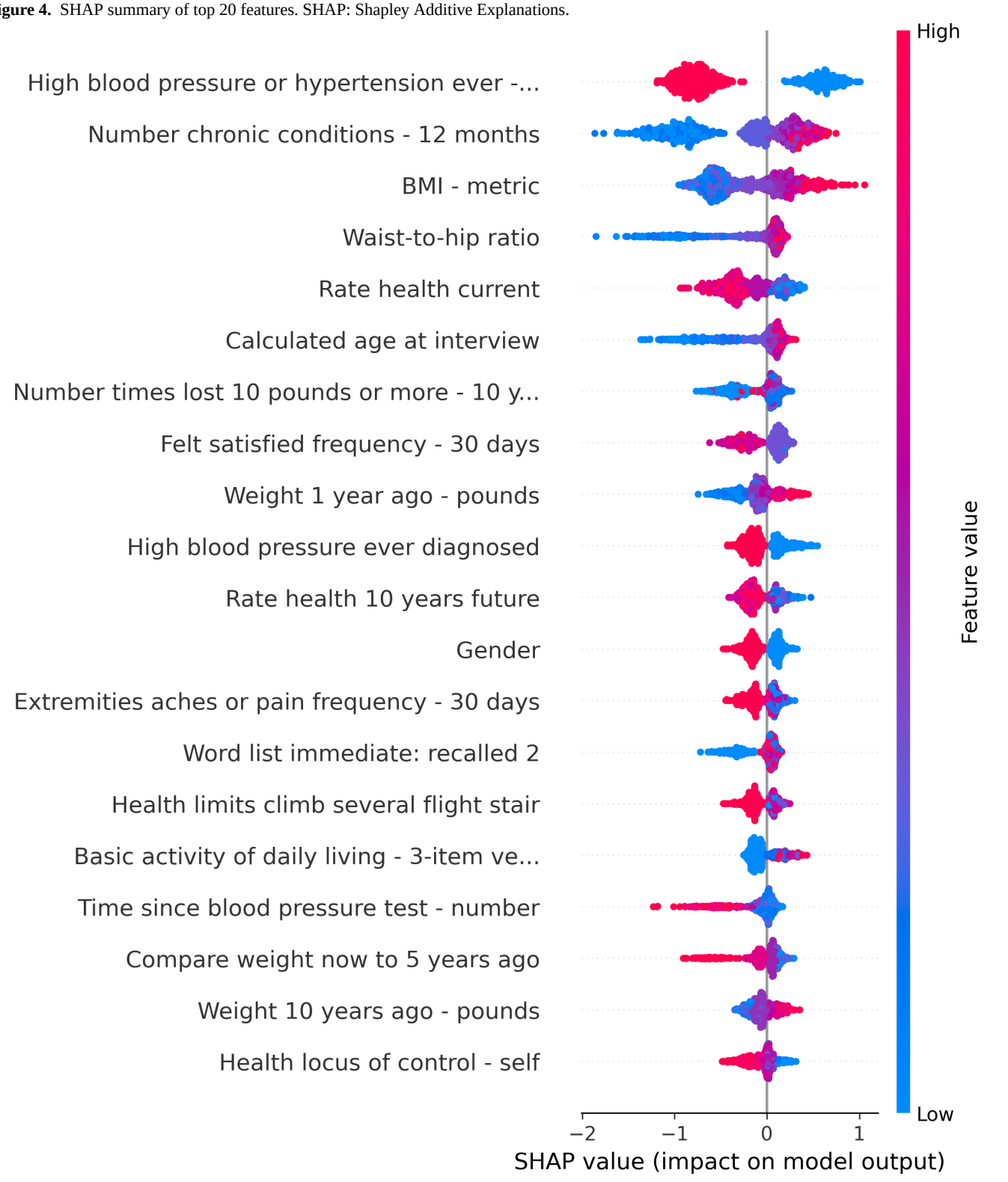
- Random forest: accuracy 0.856, AUC 0.826
- eXtreme gradient boosting: accuracy 0.868, AUC 0.829
- Logistic regression: accuracy 0.807, AUC 0.81

Performance improved steadily as additional domains were added. Models restricted to biological variables reached only about AUC ≈ 0.75; adding behavioral variables increased performance to about AUC ≈ 0.79; and using the full multidomain set (biological + behavioral + psychosocial + socioeconomic) yielded the best performance (AUC up to 0.82). This pattern underscores that psychosocial and functional factors in MIDUS carry incremental signals for diabetes onset.

Model interpretability was examined with the SHAP (Shapley Additive Explanations) analysis [23]. As shown in Figure 4, the highest impact features included hypertension status, number of chronic conditions, BMI, waist-to-hip ratio, several functional-limitation items, and indicators of health monitoring (eg, time since the last blood pressure test). Together, these suggest that the learned model is capturing not only metabolic risk but also the accumulation of chronic disease burden and health care–engagement patterns.

XSL•FO
RenderX

Figure 4. SHAP summary of top 20 features. SHAP: Shapley Additive Explanations.



Simulation Findings

State Construction and Risk Stratification

Discrete risk states were introduced to support interpretable, coarse-grained transition modeling, given the sparse temporal structure of the data. Clustering is used here only as a discretization of a continuous predicted-risk score into interpretable strata. We selected K=3 to align with standard clinical stratification into low-, medium-, and high-risk categories. When K>3, the clustering yields finer-grained strata of the same risk distribution rather than a different disease model. To preserve semantic comparability across K, we define the high-risk population as the union of the highest-risk clusters whose combined size matches the K=3 high-risk group.

Alternative state definitions could be used, including quantile-based binning of predicted risk, latent class or hidden Markov models, or continuous-state formulations that avoid discretization altogether. We selected k-means for this study

XSL•FO
RenderX

because it provides a transparent and data-driven partition of the risk distribution that is compatible with sparse 2-wave data. Exploring continuous-time and latent-state models is an important direction for future work as higher-frequency data become available.

We fixed the centroids learned at wave 2 and applied them to wave 3, so that states were compared between waves.

- Wave 2: low 79.1% with mean risk 3.03% (SD 3.35%), medium 10.1% with mean risk 27.8% (SD 10.77%), and high 10.8% with mean risk 93.5% (SD 10.84%)
- Wave 3: low 62.4% with mean risk 1.20% (SD 5.84%), medium 5.4% with mean risk 49.7% (SD 7.75%), and high 32.2% with mean risk 90.8% (SD 9.77%)
- Participants changing states: 1414/4174 (33.9%)

The substantial increase in the high-risk group (10.8% → 32.2%) and one-third of participants who changed states indicate meaningful between-wave shifts in modeled diabetes risk, supporting the use of a transition-based, coarse-grained simulator.

Transition Model (Wave 2 → Wave 3)

We trained a state-transition model to estimate

P(SW3 | SW2,xW2) where SW2 and SW3 represent cluster-based risk states and xW2 is the participant's wave 2 feature vector. Given the availability of only 2 time points, this model captures single-step, between-wave transitions rather than continuous temporal dynamics. Five-fold cross-validation with stratified splits was used. Inputs included demographics, biological, behavioral, psychosocial, and socioeconomic features. State-transition model performance is summarized in Table 5.

Table . State-transition model performance.

Model	Accuracy (SD), %	Macro F_1, %	Average log-likelihood
XGBoost[a]	92.54 (1.39)	91.64	–0.2572
Random forest	88.57 (1.31)	79.42	–0.4089
Logistic regression	60.42 (2.31)	50.22	–0.9475

[a]XGBoost: eXtreme gradient boosting.

Tree-based models substantially outperformed the linear baseline, capturing nonlinear interactions among M2 features and baseline states. Model calibration was further assessed with Brier scores (≈0.07) and expected calibration error (≈0.04), indicating well-calibrated predictive transition probabilities.

Outcome Model (Diabetes at Wave 3)

To link state evolution to clinical outcomes, we trained a second model to estimate

P(YW3=1 | SW2,SW3,xW2), where YW3 denotes diabetes onset at follow-up. Inputs included baseline features, wave 2 state probabilities, and predicted wave 3 state probabilities (soft outputs) from the transition model.

All 3 models achieved strong discrimination; tree-based models again delivered slightly better calibration, as shown in Table 6. This 2-stage design (state transition → outcome) supports progression-aware predictive modeling, consistent with the framework's DT abstraction.

Table . Diabetes outcome model performance.

Model	AUC[a] (SD), %	PR-AUC[b], %	Brier score
Random forest	90.89 (2.90)	75.60	0.0640
XGBoost[c]	90.34 (2.80)	75.10	0.0638
Logistic regression	88.69 (1.91)	67.45	0.1089

[a]AUC: area under the receiver operating characteristic curve.

[b]PR-AUC: precision-recall area under the curve.

[c]XGBoost: eXtreme gradient boosting.

Predictive "What-If" Simulator

We then used the trained transition and outcome models as a predictive scenario-analysis engine. Modifiable wave 2 covariates (weight/BMI, physical activity, smoking, alcohol consumption, anxiety, and depression) were perturbed, and the models were rerun to obtain new wave 3 state distributions and predicted diabetes risk predictions. All simulation runs were deterministic and logged for reproducibility.

Scenario 1: 10% Weight Reduction (Predictive Sensitivity Analysis)

A uniform 10% decrease in weight (and hence BMI, assuming a fixed height) was applied to all 4174 participants (Table 7).

- Participants with any state change: 280/4174 (6.7%)
- Largest single improvement: 53 transitions from high → low risk
- The model-estimated number of diabetes cases decreased from 576 to 478 (−98) under this perturbation
- Individuals with reduced predicted risk: 2417/4174 (57.9%)

XSL•FO
RenderX

These changes reflect the model's sensitivity to weight-related inputs rather than the projected real-world intervention effects.

Table . Effect of 10% weight/BMI reduction (predictive simulation).

Outcome	Preintervention	Postintervention	Change
Predicted diabetic (count)	576	478	–98
Any state change, n (%)	—[a]	280 (6.7)	—
High → low-risk transitions	—	53	—
Individuals with reduced risk, n (%)	—	2417 (57.9)	—

[a]Not applicable.

Scenario 2: Composite Lifestyle Scenario (Predictive Sensitivity Analysis)

A composite lifestyle scenario was simulated by increasing physical activity by 1 ordinal category (≈20% of the MIDUS 1 - 6 scale), switching current smokers to nonsmokers, reducing alcohol consumption by 2 units, and reducing anxiety and depression scores by 20%.

Model-estimated response includes the following:

- Participants with state change: 300/4174 (7.2%)
- High → low improvement: 49 participants
- The model-estimated number of diabetes cases decreased from 576 to 506 (−70)
- Individuals with lower diabetes risk: 1417/4174 (33.9%)

Because MIDUS stores activity on an ordinal scale rather than as minutes per week, the "20% increase" was operationalized as a +1 shift in the activity category, which preserves the semantic meaning of the item.

Illustrative Case Studies

To demonstrate individual-level exploratory use, we simulated 2 patients:

- Case 1 (10% weight loss): 50-year-old male, BMI 29.6 kg/m², baseline predicted diabetes risk 35.4%. After perturbing weight-related inputs to reflect a 10% reduction (BMI 26.6 kg/m²), the model-estimated risk decreased to 22.1% (−13.3% points).
- Case 2 (composite lifestyle scenario): A 57-year-old male with a baseline risk of 57.6%. After increased activity, smoking cessation, reduced alcohol intake, and a 20% reduction in lower anxiety/depression scores, the predicted risk dropped to 36.0% (−21.4% points).

These vignettes show that the same population-trained simulator can support individualized, predictive "what-if" reasoning, rather than personalized intervention recommendations.

Sensitivity and Robustness Checks

To evaluate the stability of the proposed framework and the robustness of the latent state definitions, we conducted a series of sensitivity analyses examining alternative perturbation magnitudes, state-definition choices, placebo perturbation, and subgroup calibration:

- Alternative perturbation magnitudes: Weight reductions of 5% and 15% yielded monotonic decreases in average predicted diabetes risk (−6% and −14%, respectively).
- State-definition robustness: For K>3, the high-risk state was defined as the union of the highest risk clusters whose combined size matched the K=3 high-risk group. Under this definition, the composition and clinical profile of the high-risk population remained stable for K=3 - 5 (>88% Jaccard overlap; Table S7 in Multimedia Appendix 4), and transition accuracy varied by <2%.
- Placebo test: A 0% weight change produced <0.3% difference in risk distribution, confirming model stability.
- Subgroup calibration: AUCs were consistent across age, sex, and income tertiles (variation <3%), indicating no major bias within the modeled population.

To assess the robustness of the latent state definitions, we performed a sensitivity analysis by varying K from 3 to 5. The "High Risk" state demonstrated high stability, maintaining a Jaccard membership overlap of >0.90 across configurations (see Table S7 in Multimedia Appendix 4). Increasing K primarily resulted in the granular subdivision of the "Healthy" state (eg, distinguishing normative vs optimal profiles) rather than altering the boundaries of the pathological state, confirming that the diabetic phenotype represents a distinct manifold in the data.

Discussion

Principal Results

This study demonstrates the feasibility of constructing a simulation-capable DT from a public, survey-based dataset using MIDUS waves 2 and 3. By integrating multidomain predictors, ontology-guided semantic structuring, and a 2-stage prediction–simulation design, the framework achieved strong diabetes risk prediction and reproduced meaningful between-wave risk transitions.

Our results highlight a critical limitation of traditional statistical feature selection in health informatics: the inability to distinguish causality from correlation. Although statistically rigorous, Lasso identified "Prescription Medicine" checklists as top predictors. In a static risk model, these are useful proxies; however, for a DT simulation, they are dead ends: one cannot "simulate" removing a history of hernia repair to cure diabetes. In contrast, the agentic framework successfully retrieved upstream determinants, such as dietary behavior and physical activity frequency. This shift from proxy-based prediction toward

XSL•FO
RenderX

modifiable-input simulation justifies the computational complexity of the LLM layer, as it enables structured predictive sensitivity analysis (eg, "What if the patient increases vigorous activity?") that statistical baselines cannot support. Across scenarios, observed changes should be interpreted as responses of the trained model to input perturbations rather than estimates of real-world intervention effects.

Generalizability and Portability to Other Diseases and Datasets

Although diabetes is used as a case study, the proposed framework is modular and largely reusable across chronic conditions and public-health datasets. Disease-agnostic components include (1) the data harmonization and temporal alignment pipeline, (2) the agent-orchestration and logging layer (bounded autonomy, traceable message passing), (3) the general ontology-RAG retrieval mechanism (embedding-based concept retrieval + structured context injection), (4) the downstream prediction stack (tree-based and linear models with calibration/SHAP), and (5) the sensitivity-analysis interface that perturbs modifiable inputs and re-evaluates model outputs. Disease-specific components primarily include (1) the ontology set used for grounding retrieval, (2) the prompt rubric and relevance criteria used by the feature-selection agent (what constitutes "direct/indirect relevance" for a target disease), (3) the outcome definition (case ascertainment and temporal window), and (4) the selection of modifiable factors emphasized in the scenario analysis.

Porting the framework to a different chronic disease (eg, cardiovascular disease [CVD] or chronic obstructive pulmonary disease [COPD]) would involve four practical steps: (1) define the target end point and prediction horizon consistent with the dataset (eg, incident CVD event, COPD exacerbation, and hospitalization), (2) swap or extend the ontology layer using disease-appropriate ontologies/terminologies (eg, condition-specific risk-factor and phenotype ontologies, and mappings to the Systematized Nomenclature of Medicine [SNOMED] and International Classification of Diseases [ICD], where applicable), (3) update the agent prompt templates and scoring rubric so that relevance classification and confidence scoring reflect domain-specific pathways (eg, lipids/blood pressure for CVD; smoking history, spirometry proxies, and symptom burden for COPD), and (4) retune the state definition and progression model to match disease dynamics and data cadence. In datasets with sparse waves (as in MIDUS), coarse-grained discretization and single-step transitions can provide an interpretable progression abstraction; however, for conditions with faster dynamics or richer longitudinal sampling, the same pipeline could replace k-means discretization with quantile bins or latent-state models (eg, hidden Markov models) and replace between-wave transitions with continuous-time or event-based progression models.

Finally, we note that portability also depends on dataset structure. Survey-based cohorts provide multidomain exposures but limited clinical end points, whereas electronic health record (EHR)–based cohorts provide richer outcomes but noisier measurement and coding artifacts. The framework's ontology-grounded RAG and semantic auditing steps are intended to mitigate such heterogeneity by making feature selection and simulation inputs more interpretable and traceable, while still requiring disease- and dataset-specific end point definitions and validation protocols.

Limitations

Data Quality, Missingness, and Measurement Error

This study relies on the MIDUS dataset, which consists largely of self-reported behavioral, psychosocial, and health measures. Such data are subject to recall bias, social desirability bias, and reporting inaccuracies, particularly for diet, physical activity, mental health, and substance use. These errors may attenuate or distort learned associations and propagate uncertainty into both risk predictions and counterfactual simulations. Although large sample sizes and robust machine-learning models can mitigate random noise, systematic reporting bias remains an inherent limitation of survey-based DT construction.

MIDUS also exhibits item nonresponse and wave-level attrition that are likely to be partially missing not at random. For example, participants with poorer health may be less likely to complete follow-up surveys, and sensitive behaviors may be systematically underreported. While kNN preserves multivariate structure and reduces information loss, it cannot recover information that is systematically missing. Consequently, both predicted risks and simulated outcomes may be biased if missingness is correlated with unobserved health status. Future work should explore explicit missingness indicators, sensitivity analyses, or joint models of outcome and dropout when higher frequency or linked EHR data become available.

Limited Temporal Resolution

The framework is trained on only 2 observation points (MIDUS waves 2 and 3), separated by approximately a decade. This supports single-step, between-wave transition modeling but does not capture short- or medium-term disease dynamics, nor does it allow continuous updating or multistep forward simulation. As a result, the proposed DT should be viewed as a coarse-grained, long horizon progression model rather than a high-resolution temporal simulator. Additional waves or higher-frequency data (eg, EHRs or wearables) would be required to support continuous or multistep DT updating.

Predictive (Not Causal) Nature of Counterfactuals

The counterfactual "what-if" simulations implemented in this framework are predictive rather than causal. They assume that the statistical associations learned by the transition and outcome models remain stable under hypothetical interventions (eg, weight loss, increased physical activity, and smoking cessation). Accordingly, quantities such as "98 fewer diabetes cases" represent model-estimated or projected changes in risk under perturbed inputs, not real-world treatment effects.

While placebo and monotonicity tests demonstrate internal numerical stability of the simulator, they do not establish causal efficacy. Therefore, the simulations reflect model sensitivity rather than physiologically or clinically validated intervention outcomes and should be interpreted as sensitivity analyses and hypothesis generation tools rather than the estimates of clinical impact. Although the magnitude and direction of predicted

XSL•FO
RenderX

effects are broadly consistent with major lifestyle intervention trials (eg, the Diabetes Prevention Program), such comparisons are illustrative only and do not constitute external validation or causal inference.

Clinical Applicability

The current framework is not intended for direct clinical deployment. Outputs such as transitions between cluster-based risk states or changes in population-level predicted incidence are not directly actionable at the individual patient level. Instead, the system is designed to support exploratory analysis, hypothesis generation, and population-level planning rather than point-of-care decision-making. Substantial additional work would be required for clinical translation, including calibration to clinical end points, alignment with guideline-based thresholds, and integration with EHR workflows.

Model-based simulations of hypothetical lifestyle changes also raise ethical and communication concerns. Without causal validation, simulated risk reductions could be misinterpreted as guaranteed outcomes, potentially leading to false reassurance or undue anxiety. Any future clinical deployment would require careful attention to explainability, uncertainty communication, and clinician oversight.

State Definition and Discretization

Health states were defined using k-means clustering for its simplicity, transparency, and compatibility with sparse 2-wave data. However, alternative discretization strategies—such as quantile-based risk bins, latent class models, hidden Markov models, or continuous-state formulations—could be used to define health states without hard clustering. Exploring such alternatives, particularly in datasets with higher temporal resolution, represents an important direction for future work.

Scope of the DT

The proposed framework does not implement several features of mature health care DTs, including continuous synchronization with real-world data, online learning, bidirectional patient feedback, or causal intervention modeling. It should therefore be interpreted as a foundational, offline, simulation-capable DT suitable for population-level exploration and hypothesis generation rather than a clinically deployed or continuously updating patient-specific twin.

Comparison With Prior Work

Most existing health care DT prototypes have relied on narrowly scoped, proprietary datasets, often collected in clinical settings with limited generalizability. Prior studies in diabetes modeling [24-26] have focused primarily on short-term glucose forecasting or clinical risk scoring, with relatively little emphasis on integrating psychosocial and behavioral variables. Our approach advances this field in two ways: (1) by leveraging publicly available, longitudinal health datasets to promote inclusivity and scalability, and (2) by combining multiagent orchestration, ontologies, and LLM reasoning to enable principled, explainable feature selection. These contributions distinguish our framework from traditional statistical or machine-learning models and align with emerging calls for DTs that are both interpretable and adaptable to real-world complexity.

It is important to note the distinction between the "simulation-centric" twin presented here and "operational" twins used in acute care. Our system lacks the continuous, real-time data synchronization typical of industrial DTs [1]. Instead, it functions as a progression-aware simulator, a foundational prototype that establishes the semantic and computational architecture (agents, ontologies, and state transitions) required for a DT. While currently powered by static survey waves, this same architecture is designed to ingest real-time streams (eg, wearables) in future iterations.

Conclusions

This study presents a foundational, progression-aware DT framework for predictive risk projection using public health data, rather than a validated dynamic or clinical DT. We show how large, heterogeneous cohort datasets can be transformed into simulation-capable, longitudinal digital patient representations using ontology-guided feature selection, agent-orchestrated LLM reasoning, and machine learning–based prediction and transition modeling. While the current implementation is offline and limited to 2 survey waves, it establishes a methodological foundation for DT construction as richer and higher-frequency data become available.

A key contribution of this work is demonstrating that public health data, when semantically aligned and processed through an LLM-assisted, agent-orchestrated pipeline, can support not only risk prediction but also predictive simulation. The resulting DT enables the exploration of how model-estimated risk trajectories change under hypothetical modifications to modifiable factors, such as body weight, physical activity, smoking, or psychosocial stress. These simulations reflect model sensitivity rather than real-world intervention effects, but they provide a practical and interpretable way to explore risk dynamics and generate hypotheses in chronic disease research.

For public health planners and clinical researchers, this framework enables population-level "what-if" analyses and supports exploratory individualized risk profiling and scenario analysis grounded in real cohort data. For medical informatics researchers, the proposed architecture, which combines ontologies, LLM-based agents, and transition-aware prediction models, provides a reusable blueprint for other slowly progressing, multidomain conditions, such as CVD, frailty, and COPD.

Future work will focus on the following 4 directions:

1. Extending single-step progression to multistep and higher-frequency updating through integration with EHR, wearable, and mobile-health data;
2. Incorporating decision-analytic and reinforcement-learning components to identify which modifiable factors are most influential for reducing individual risk;
3. Performing external and prospective validation to move from predictive projections toward decision-grade DTs; and
4. Integrating causal-inference and trial-linked data to transition from predictive counterfactuals to causally grounded DT simulation.

XSL•FO
RenderX

Funding

The authors express their gratitude to the funding provided to support this study from the National Science Foundation (NSF) with award number 2218046.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 200 features selected by ontology-assisted large language model.
[XLSX File, 198 KB - medinform_v14i1e87374_app1.xlsx]

Multimedia Appendix 2
Top 200 features selected with Lasso (L1-regularized logistic regression).
[XLSX File, 22 KB - medinform_v14i1e87374_app2.xlsx]

Multimedia Appendix 3
Top 200 features selected with mutual information.
[XLSX File, 22 KB - medinform_v14i1e87374_app3.xlsx]

Multimedia Appendix 4
Supplementary analyses of feature selection and latent-state stability.
[DOCX File, 125 KB - medinform_v14i1e87374_app4.docx]

References

1. Papachristou K, Katsakiori PF, Papadimitroulas P, Strigari L, Kagadis GC. Digital twins' advancements and applications in healthcare, towards precision medicine. J Pers Med 2024 Nov 11;14(11):1101. [doi: 10.3390/jpm14111101] [Medline: 39590593]
2. Zhang K, Zhou HY, Baptista-Hon DT, et al. Concepts and applications of digital twins in healthcare and medicine. Patterns (N Y) 2024 Aug 9;5(8):101028. [doi: 10.1016/j.patter.2024.101028] [Medline: 39233690]
3. Turner K, Hohman KH. Demonstrated progress and future promise of chronic disease data modernization. Prev Chronic Dis 2024 Oct 31;21:E86. [doi: 10.5888/pcd21.240396] [Medline: 39481011]
4. Dang Y, Li F, Hu X, et al. Systematic design and data-driven evaluation of social determinants of health ontology (SDoHO). J Am Med Inform Assoc 2023 Aug 18;30(9):1465-1473. [doi: 10.1093/jamia/ocad096] [Medline: 37301740]
5. Wang S, McDermott MBA, Chauhan G, Hughes MC, Naumann T, Ghassemi M. MIMIC-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. Presented at: ACM CHIL '20: ACM Conference on Health, Inference, and Learning; Apr 2-4, 2020; Toronto, Ontario, Canada p. 222-235. [doi: 10.1145/3368555.3384469]
6. Orhan F, Kurutkan MN. Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors. BMC Health Serv Res 2025 Mar 12;25(1):366. [doi: 10.1186/s12913-025-12502-5] [Medline: 40075408]
7. Radler BT. The Midlife in the United States (MIDUS) series: a national longitudinal study of health and well-being. Open Health Data 2014;2(1):e3. [doi: 10.5334/ohd.ai] [Medline: 25558376]
8. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/nchs/nhanes/index.html [accessed 2025-10-30]
9. Health research data for the world. UK Biobank. URL: https://www.ukbiobank.ac.uk/ [accessed 2025-10-30]
10. Schiavenato M, Chu F. PICO: what it is and what it is not. Nurse Educ Pract 2021 Oct;56:103194. [doi: 10.1016/j.nepr.2021.103194] [Medline: 34534728]
11. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinform 2022;2:927312. [doi: 10.3389/fbinf.2022.927312] [Medline: 36304293]
12. Van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R. Threshold-based feature selection techniques for high-dimensional bioinformatics data. Netw Model Anal Health Inform Bioinforma 2012 Jun;1(1-2):47-61. [doi: 10.1007/s13721-012-0006-6]
13. Sadeghian Z, Akbari E, Nematzadeh H, Motameni H. A review of feature selection methods based on meta-heuristic algorithms. J Exp Theor Artif Intell 2025;37(1):1-51. [doi: 10.1080/0952813X.2023.2183267]
14. Gao Y, Xiong Y, Jia K, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]

XSL•FO
RenderX

15. El-Sappagh S, Kwak D, Ali F, Kwak KS. DMTO: a realistic ontology for standard diabetes mellitus treatment. J Biomed Semantics 2018 Feb 6;9(1):8. [doi: 10.1186/s13326-018-0176-y] [Medline: 29409535]
16. Common diabetes medications ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/CDMONTO [accessed 2025-11-03]
17. HL7 FHIR and SSN ontology based type 1 diabetes mellitus ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/FASTO [accessed 2026-01-13]
18. BioMedBridges diabetes ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/DIAB [accessed 2025-11-03]
19. Study cohort ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/SCO [accessed 2025-11-03]
20. Chari S, et al. Making study populations visible through knowledge graphs. Presented at: 18th International Semantic Web Conference (ISWC 2019); Oct 26-30, 2019; Auckland, New Zealand p. 53-68. [doi: 10.1007/978-3-030-30796-7_4]
21. 45 CFR 46.102. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-04-16]
22. Terms of use. ICPSR. URL: https://www.icpsr.umich.edu/sites/icpsr/about/policies/terms-of-use [accessed 2026-04-16]
23. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed 2022 Feb;214:106584. [doi: 10.1016/j.cmpb.2021.106584] [Medline: 34942412]
24. Yang X, Li J. A clustering-based federated deep learning approach for enhancing diabetes management with privacy-preserving edge artificial intelligence. Healthc Anal 2025 Jun;7:100392. [doi: 10.1016/j.health.2025.100392]
25. Yang X, Li J. Edge AI empowered personalized privacy-preserving glucose prediction with federated deep learning. Presented at: 2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom); Dec 15-17, 2023. [doi: 10.1109/Healthcom56612.2023.10472368]
26. Li Q, Amat KR, Li J. LLM-powered personalized glucose prediction in type 1 diabetes. Comput Struct Biotechnol Rep 2025;2:100068. [doi: 10.1016/j.csbr.2025.100068]

Abbreviations

AUC: area under the receiver operating characteristic curve
CVD: cardiovascular disease
DMTO: Diabetes Mellitus Treatment Ontology
DT: digital twin
EHR: electronic health record
FASTO: FHIR And SSN-based Type 1 diabetes Ontology
kNN: k-nearest neighbor
Lasso: L1-regularized logistic regression
LLM: large language model
MI: mutual information
MIDUS: Midlife in the United States
PICO: population, intervention, comparison, and outcome
PR-AUC: precision-recall area under the curve
RAG: retrieval-augmented generation
SHAP: Shapley Additive Explanations

Edited by A Benis; submitted 14.Nov.2025; peer-reviewed by J Guo, T Peng, Z Liu; revised version received 06.Mar.2026; accepted 06.Mar.2026; published 21.Apr.2026.

Please cite as:
Li Q, Amat KR, Johnson EL, Li J
Modeling Diabetes Risk and Progression With Public Health Data: Ontology-Guided, Simulation-Capable Digital Twin Study
JMIR Med Inform 2026;14:e87374
URL: https://medinform.jmir.org/2026/1/e87374
doi:10.2196/87374

© Qingrui Li, Kapileshwor Ray Amat, Eric L Johnson, Juan Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making

Adriano Fabi[1,2*], MD; Caroline E Egli[1,2*], MD; Séverin R Wendelspiess[1,2], MD; Sebastian Griewing[3,4], MD, PhD; Yvonne Haas[1,2], MD; Laura De Pellegrin[5], MD; Dirk J Schaefer[1,2], MD; Shan S Qiu[6], MD; Yves Harder[7,8], MD; Elisabeth A Kappos[1,2,9], MD

[1]Department of Plastic, Reconstructive, Aesthetic and Hand Surgery, University Hospital of Basel, Spitalstrasse 21, Basel, Switzerland

[2]Faculty of Medicine, University of Basel, Spitalstrasse 21, Basel, Switzerland

[3]Institute for Digital Medicine, Philipps University of Marburg, Marburg, Germany

[4]Department of Gynecology and Obstetrics, Philipps University of Marburg, Marburg, Germany

[5]Department of Plastic and Hand Surgery, University Hospital of Bern, Bern, Switzerland

[6]Department of Plastic and Reconstructive Surgery, Maastricht University Medical Centre, Maastricht, The Netherlands

[7]Department of Plastic, Reconstructive and Aesthetic Surgery and Hand Surgery, University Hospital of Lausanne (CHUV), Lausanne, Switzerland

[8]Faculty of Biology and Medicine, University of Lausanne (UNIL), Lausanne, Switzerland

[9]Breast Center, University Hospital of Basel, Basel, Switzerland

[*]these authors contributed equally

Corresponding Author:
Elisabeth A Kappos, MD
Department of Plastic, Reconstructive, Aesthetic and Hand Surgery, University Hospital of Basel, Spitalstrasse 21, Basel, Switzerland

Abstract

Background: Upper and lower extremity lymphedema is a chronic, progressive condition that significantly impairs the quality of life of affected patients. Despite the recently established effectiveness of physical therapy and supermicrosurgical interventions, current guidelines frequently lag behind emerging evidence and commonly do not offer stage-specific treatment algorithms. This gap in evidence-based guidance may prompt clinicians with limited experience to seek support from large language models such as ChatGPT.

Objective: Given the potential of artificial intelligence to rapidly integrate emerging research, this study evaluated how clinicians from different professional backgrounds rate the quality and reliability of personalized lymphedema management recommendations generated by ChatGPT.

Methods: In this exploratory cross-sectional study, ChatGPT generated treatment recommendations for 6 standardized lymphedema case scenarios. An international panel of 67 participants (resident doctors, board-certified specialists, physiotherapists, and advanced practice nurses) from 34 institutions across 11 countries assessed the recommendations using a modified DISCERN questionnaire with a 9-point agreement scale ranging from 1 (completely disagree) to 9 (completely agree). Ratings were summarized as pooled means with variability measures and compared across clinician groups (residents vs board-certified physicians vs physiotherapists or advanced practice nurses) using group comparison testing.

Results: ChatGPT was rated most favorably for diagnostic accuracy and treatment relevance, with higher ratings among residents than board-certified physicians. Residents assigned significantly lower scores for source indication, source currency, and communication of uncertainty. Between-group differences were observed across multiple DISCERN items, consistent with systematically more critical appraisal by experienced specialists. Participants reported moderate to high trust and willingness to consider ChatGPT as a supplementary resource, with more favorable perceptions among younger respondents.

Conclusions: Clinicians perceived ChatGPT as potentially useful for preliminary orientation and educational support in lymphedema management, especially for less experienced users. Despite not being blinded, lower ratings in evidence transparency and uncertainty communication, particularly among experienced specialists, suggest that current artificial intelligence outputs should not be used as stand-alone guidance. Future work should test clinically integrated, citation-grounded workflows in prospective settings and evaluate whether they improve decision quality and efficiency.

(JMIR Med Inform 2026;14:e80553) doi:10.2196/80553

XSL•FO
RenderX

KEYWORDS

artificial intelligence; AI; ChatGPT; decision-making; digital health; large language models; lymphedema; personalized medicine

Introduction

Lymphedema is a chronic condition characterized by the accumulation of interstitial fluid, typically resulting from damage or disruption of lymphatic pathways [1]. In the Western world, lymphedema is commonly caused by oncologic surgery or radiotherapy and significantly reduces quality of life due to extremity swelling, limited mobility, and psychosocial distress [1-5].

Lymphedema treatment is highly individualized, influenced by age, preexisting comorbidities, and prior interventions [1]. Conservative treatment includes complex decongestive therapy; however, recent literature has demonstrated promising long-term outcomes after supermicrosurgical interventions, such as lymphovenous anastomosis (LVA) and vascularized lymph node transfer (VLNT) [6-9]. Due to the personalized nature of lymphedema treatment, following the current guidelines of the AWMF (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, S2k) or the International Society of Lymphology (ISL) may result in difficulties providing universally applicable, stage-specific recommendations [10,11]. Furthermore, the guidelines may include outdated information, primarily due to the time-intensive nature of expert consensus and validation processes [10,11]. Given the rapid advancements in lymphatic surgery, the incorporation of novel treatment approaches into official guidelines may take years, as this process requires a predefined and structured consensus. Consequently, there is a possibility that patients may receive care that is no longer up to date in a clinical context.

Large language models (LLMs) such as ChatGPT (OpenAI) represent a new approach to the processing of knowledge, which also applies to medical information [12,13]. These technologies have the potential to reshape medical decision-making by accelerating the integration of novel research into clinical routines [14-17]. Thus, unlike traditional guidelines, artificial intelligence (AI) is hypothesized to accelerate the transition of new findings into personalized treatment recommendations. However, whether this potential is truly realized in practice strongly depends on data implementation and prompting [18].

Therefore, the goal of this study was to assess ChatGPT's ability to define and generate treatment recommendations for lymphedema by evaluating the quality, accuracy, and perceived utility of its recommendations across an interdisciplinary panel of relevant medical specialties. Additionally, the differences in acceptance were compared between resident doctors in training and board-certified plastic surgeons with experience in either lymphatic surgery or lymphedema management.

Methods

Study Design

The study was designed as a multidisciplinary assessment of ChatGPT-generated treatment recommendations of 6 cases of chronic lymphedema, ranging from ISL stage I to ISL stage 3 of the upper and lower limb (Multimedia Appendix 1). To ensure consistency, all cases were submitted to GPT-4o (GPT-4 Omni) on December 18, 2024, at noon Central European Time by the first authors (AF and CEE). Each case was entered into a distinct, separate chat session to limit session history bias. Only the initial response was considered, and the option to regenerate the response was not used. The responses obtained from GPT-4o were compiled into a PDF and were subsequently evaluated by a panel of resident doctors (considered inexperienced), board-certified doctors (considered experienced), and physiotherapists or advanced practice nurses (APNs) using a 9-point agreement scale [19].

Clinical Vignettes and Treatment Recommendations

The clinical vignettes were designed with the intention of incorporating the underlying etiology of lymphedema through past medical history, current patient symptoms, and findings from clinical examination. Cases were prepared to be as realistic as possible and to reflect real clinical scenarios. Patients were clearly categorized into the 3 ISL stages, with corresponding indocyanine green studies illustrating the lymphatic patterns consistent with the stage described in each clinical vignette [10]. Furthermore, the cases were illustrated with images adapted from Principles and Practice of Lymphedema Surgery by Cheng et al [20]. A standardized prompt was used for each case and is provided in Multimedia Appendix 1.

Evaluation

The evaluators had access to the original prompts, all 6 clinical cases, and the corresponding treatment recommendations generated by ChatGPT (Multimedia Appendix 1). The web-based questionnaire for subsequent evaluation was developed and distributed using the Research Electronic Data Capture (REDCap; Vanderbilt University) platform [21]. The DISCERN tool, a validated questionnaire for assessing the quality of written treatment information, was adapted to a 9-point agreement scale and then used for the quantitative evaluation of each generated response [19].

In accordance with the instructions provided in the DISCERN handbook [19], questions 1 and 2 ("Are the aims clear?" and "Does it achieve its aims?") were excluded from the evaluation as they were found to be irrelevant to the defined research question of this study. Two additional questions were included in the evaluation of each case: one concerning correct staging (question 1) and the other regarding overall final agreement with the treatment option recommended by ChatGPT (question 16). In summary, each response was evaluated using 16 questions rated on a 9-point agreement scale, where 1 represents complete disagreement and 9 represents complete agreement. Finally, 3 concluding questions were added to evaluate the overall agreement regarding the treatment recommendation by ChatGPT and the attitude regarding AI use as a clinical decision tool (Multimedia Appendix 2). In addition, each participant was given the option to offer feedback. These open-text comments were reviewed descriptively by the first authors to provide contextual insight alongside the quantitative results. No formal

XSL•FO
RenderX

qualitative analysis or coding approach was applied. All available comments were read in full, and relevant observations are summarized in Multimedia Appendix 3.

Survey Distribution

The survey was conducted from January 7 to March 18, 2025. Resident doctors without prior experience in lymphedema management were recruited internationally through the same broader professional networks as the expert panel, primarily via participating centers in the LYMPH trial, as well as international resident and student networks. Experienced, board-certified doctors specializing in lymphedema treatment were identified and contacted, most of whom had previously participated in the international lymphedema consortium through existing collaborations [22]. A total of 64 board-certified doctors were invited via email to take part in the study, receiving a survey link along with a comprehensive explanation of its objectives and requirements (response rate 24/64, 37.5%). Additionally, 30 physiotherapists and lymphedema APNs specializing in lymphedema treatment were invited (response rate 10/30, 33.3%) through the same consortium and a contact list provided by the Lymphödem Vereinigung Schweiz (a Swiss patient organization) [22,23]. As resident recruitment was conducted via mass email and a preexisting student group, the response rate of invited resident doctors could not be calculated.

Statistical Analysis

Statistical analysis was conducted using R software (version 4.4.3; R Foundation for Statistical Computing) [24]. The primary end point was the scoring of agreement and reliability of ChatGPT-generated treatment recommendations stratified by experience groups. Consequently, participants were categorized into 3 groups: resident doctors, board-certified doctors, and physiotherapists or APNs. For each question, group-wise means and SDs were calculated. The Levene test was used to confirm homogeneity of variances prior to conducting 1-way ANOVA. Depending on the result of the Levene test, either standard ANOVA (for equal variances) or Welch ANOVA (for unequal variances) was used to test for differences between groups. Where applicable, an “overall” group comprising all participants was included for descriptive purposes only and was excluded from inferential statistical testing. A 2-sided P value of ≤.05 was considered statistically significant.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. On the basis of the fictional nature of the cases and the absence of human subject data, the local ethics committee (Ethikkommission Nordwest- und Zentralschweiz) deemed that no formal ethics approval was required. Participation in the web-based survey was voluntary. All participants were informed about the purpose of the study, and informed consent was implied through completion of the survey. Participants had the option to discontinue the survey at any time without any consequences. All data were collected anonymously. Data protection and confidentiality were ensured in accordance with applicable regulations. No compensation was provided. No identifiable individuals are shown in any figures or supplementary materials.

Results

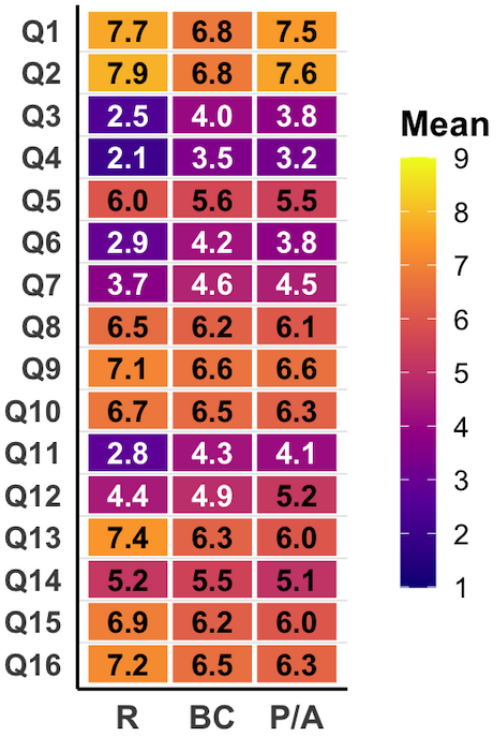
A total of 67 participants from 34 different institutions across 11 countries completed the survey, including 33 (49.3%) resident doctors, 24 (35.8%) board-certified doctors, and 10 (14.9%) lymphedema physiotherapists or APNs. The mean age was 35.6 (SD 14.1) years, with resident doctors averaging 28.0 (SD 6.0) years, board-certified doctors 45.0 (SD 14.0) years, and physiotherapists or APNs 42.0 (SD 19.9) years.

Diagnostic Accuracy and Relevance (Q1 and Q2)

Staging and diagnostic accuracy (question 1; Q1) were generally rated high (mean 7.3, SD 1.5), with board-certified doctors assigning significantly lower scores than resident doctors (P<.001; Table 1). Similarly, the relevance of the recommendations (Q2) was rated highly (mean 7.5, SD 1.4), with board-certified doctors rating significantly lower compared to resident doctors or physiotherapists or APNs (P<.001; Figure 1; Table 1). A detailed overview of scoring per case is provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 1. Heat map of pooled mean item ratings across 6 case scenarios by clinician group. Tiles are color-coded by the pooled mean score for each question (Q1-Q16) within each group; numeric values within tiles indicate the corresponding mean. Higher scores indicate more favorable ratings on a 9-point agreement scale (1=completely disagree and 9=completely agree); variability measures (SD or CI) are not shown in this figure. BC: board-certified; R: resident; P/A: physiotherapist/advanced practice nurse.



XSL•FO
RenderX

Table . Mean scores across all 6 scenarios[a].

Question[b]	Overall (n=67), mean (SD)	Resident doctors (n=33), mean (SD)	Board-certified physicians (n=24), mean (SD)	Physiotherapist or APN[c] (n=10), mean (SD)	*P* value
Q1—diagnostic accuracy	7.3 (1.5)	7.7 (1.2)	6.8 (1.6)	7.5 (1.9)	*<.001*[d]
Q2—relevance	7.5 (1.4)	7.9 (1.0)	6.8 (1.6)	7.6 (1.1)	*<.001*
Q3—indication of sources used	3.2 (2.5)	2.5 (2.1)	4.0 (2.4)	3.8 (2.9)	*<.001*
Q4—currency of sources used	2.7 (2.1)	2.1 (1.6)	3.5 (2.1)	3.2 (2.9)	*<.001*
Q5—balanced and unbiased information	5.8 (1.8)	6.0 (1.7)	5.6 (1.6)	5.5 (2.2)	.07
Q6—additional supporting information	3.5 (2.3)	2.9 (2.1)	4.2 (2.0)	3.8 (2.7)	*<.001*
Q7—communication of areas of uncertainty	4.1 (2.1)	3.7 (2.0)	4.6 (1.8)	4.5 (2.5)	*<.001*
Q8—explanation of how the treatment works	6.3 (1.8)	6.5 (2.0)	6.2 (1.5)	6.1 (2.0)	.23
Q9—benefits of treatment	6.8 (1.6)	7.1 (1.8)	6.6 (1.3)	6.6 (1.7)	*.04*
Q10—risks of treatment	6.6 (1.9)	6.7 (2.0)	6.5 (1.5)	6.3 (2.0)	.22
Q11—consequences of treatment refusal	3.5 (2.2)	2.8 (2.0)	4.3 (1.9)	4.1 (2.7)	*<.001*
Q12—impact on quality of life	4.7 (2.4)	4.4 (2.3)	4.9 (2.1)	5.2 (2.8)	*.03*
Q13—availability of multiple treatment options	6.8 (1.9)	7.4 (1.7)	6.3 (1.8)	6.0 (2.2)	*<.001*
Q14—support for shared decision-making	5.3 (2.3)	5.2 (2.4)	5.5 (2.1)	5.1 (2.3)	.37
Q15—overall quality of the recommendation	6.5 (1.6)	6.9 (1.4)	6.2 (1.8)	6.0 (1.8)	*<.001*
Q16—overall agreement	6.8 (1.8)	7.2 (1.6)	6.5 (1.9)	6.3 (2.3)	*<.001*

[a]Values represent the pooled mean (SD) across the 6 standardized lymphedema case scenarios from an exploratory cross-sectional survey of 67 clinicians (resident doctors, board-certified physicians, and physiotherapists or advanced practice nurses) from 34 institutions in 11 countries (January-March 2025).

[b]Q: question.

[c]APN: advanced practice nurse.

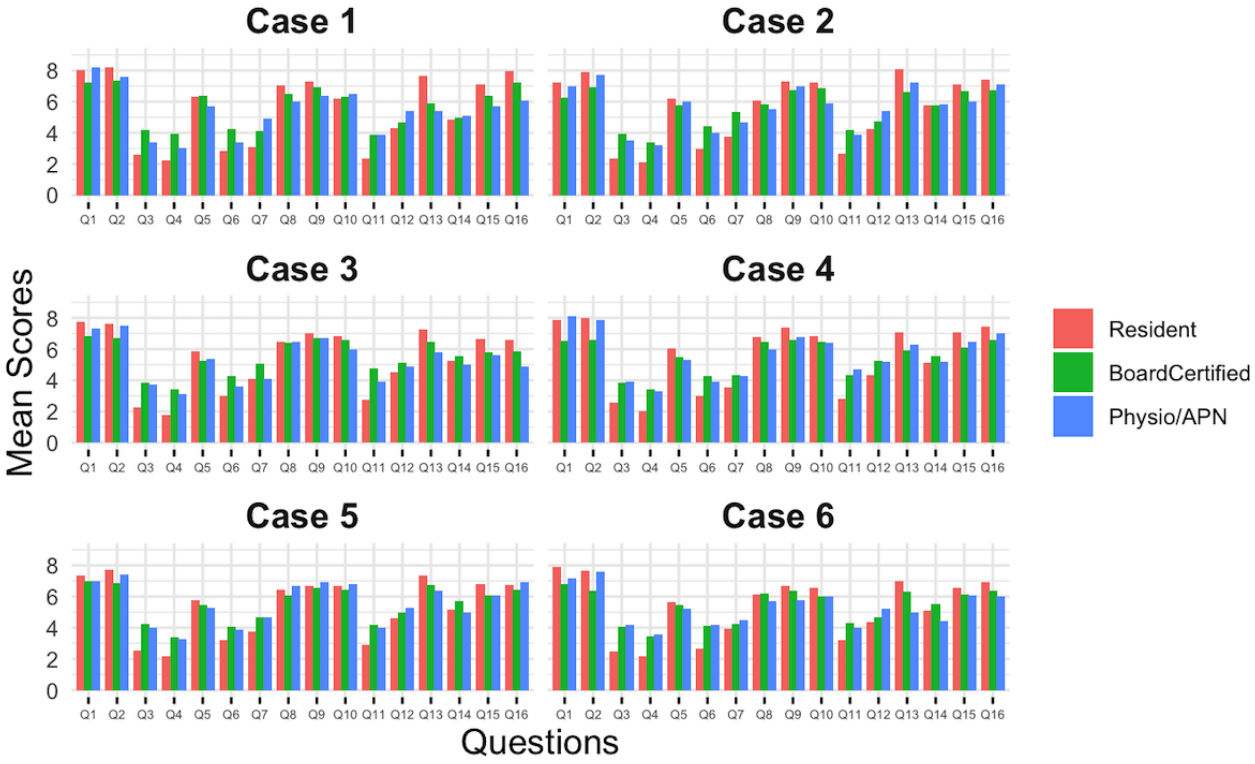
[d]Values in italic demonstrate statistical significance ($P<.05$).

Source Transparency and Objectivity (Q3-Q5)

The clarity of source citation (Q3) and the dating of sources (Q4) were rated very low across all groups, with mean scores of 3.2 (SD 2.5) and 2.7 (SD 2.1), respectively (Figure 2). Interestingly, resident doctors gave both questions significantly lower ratings than the other groups ($P<.001$). Interestingly, Q5 (objectivity or neutrality of ChatGPT) received an average score of 5.8 (SD 1.8), with no significant differences between groups.

XSL•FO
RenderX

Figure 2. Mean scores by professional group for all cases. Overview of mean scores across 6 lymphedema case scenarios by professional group. Bars represent mean scores. APN: advanced practice nurse; Q: question. Exact values can be found in Multimedia Appendix 4.



Supportive Information and Areas of Uncertainty (Q6 and Q7)

The quality of additional resources (Q6) was rated badly, with a mean score of 3.5 (SD 2.3) across all cohorts. Similarly, the communication of uncertainty (Q7) received moderate scores of 4.1 (SD 2.1), with residents giving the lowest scores across both questions (P<.001).

Treatment Explanation and Options, Risk Communication, and Quality of Life Considerations (Q8-Q13)

ChatGPT's performance in communicating treatment options and possible risks was generally rated positively, with Q8 (how the treatment would work), Q9 (benefits of treatment), and Q10 (potential risks) all receiving favorable scores (mean 6.3, SD 1.8; mean 6.8, SD 1.6; and mean 6.6, SD 1.9, respectively). In contrast, Q11 (consequence of no treatment) consistently scored low (mean 3.5, SD 2.2). Q12 assessed whether the provided information addressed the impact of the treatment on patients' quality of life. Ratings were moderate (mean 4.7, SD 2.4), with residents rating the responses significantly lower than other groups. Q13 (ChatGPT mentioning that there may be more than one treatment choice) was generally rated positively (mean 6.8, SD 1.9).

Shared Decision-Making and Overall Quality (Q14-Q16)

Support for shared decision-making (Q14) was rated moderately (mean 5.3, SD 2.3). The overall impression of the quality of responses (Q15) and overall agreement with treatment recommendations (Q16) both received high ratings (mean 6.5, SD 1.6 and mean 6.8, SD 1.8, respectively), with physiotherapists or APNs rating the answers the lowest (P<.001).

Concluding Questions—Trust and Future Use

Concluding question (CQ) 1 assessed trust in ChatGPT's responses, CQ2 assessed willingness to use ChatGPT in future clinical work, and CQ3 assessed whether participants would recommend it for clinical decision-making. Resident doctors rated all 3 concluding questions (CQ1-CQ3) significantly higher than board-certified doctors and physiotherapists or APNs. CQ2 showed a similar pattern, with resident doctors at a mean of 6.5 (SD 1.4), board-certified doctors at 5.3 (SD 2.2), and physiotherapists or APNs at 5.3 (SD 2.3). Finally, for CQ3, which assessed whether participants would recommend AI as a clinical decision-making tool, residents again scored highest, followed by board-certified doctors and physiotherapists or APNs (Figure 3; Table 2). All survey comments are listed in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 3. Concluding questions (CQs) by professional groups. Boxplot overview of concluding questions (trust, willingness to use, and recommendation of ChatGPT) by professional group. APN: advanced practice nurse.

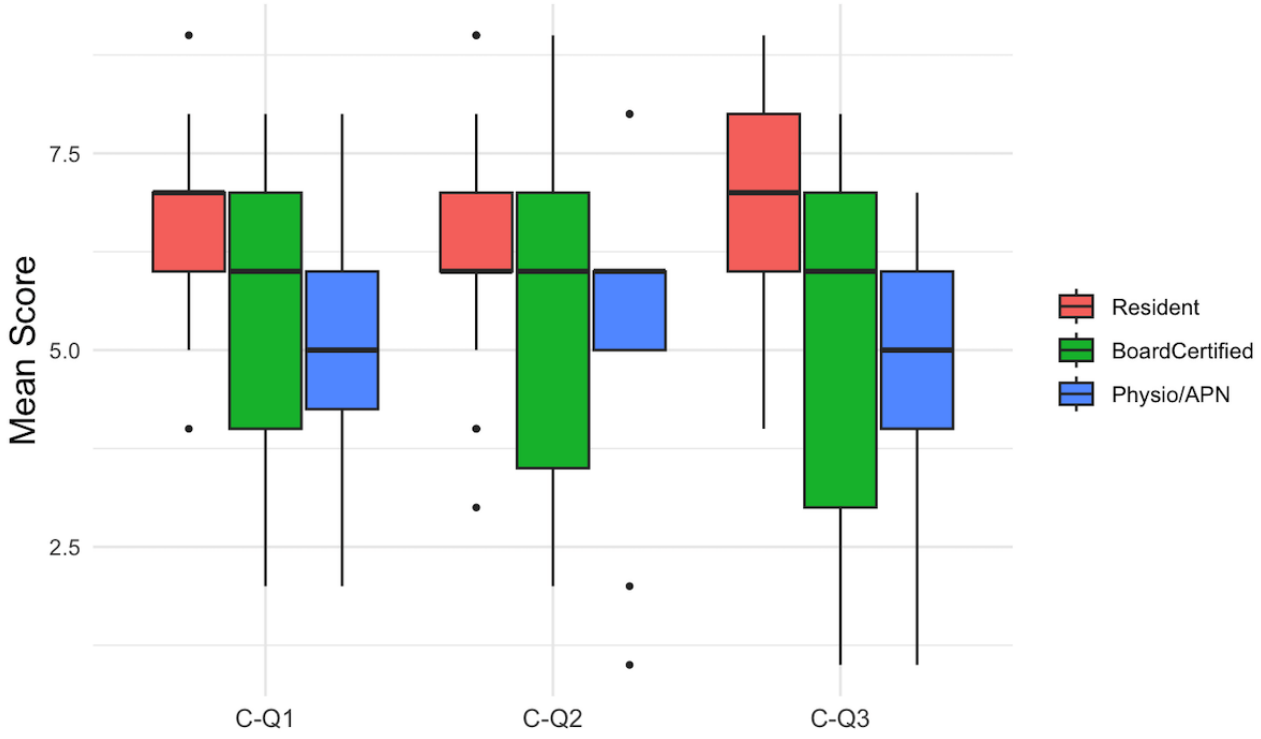


Table . Scores by professional groups across all concluding questions (CQs)[a].

Question	Overall (n=67), mean (SD)	Resident doctors (n=33), mean (SD)	Board-certified physicians (n=24), mean (SD)	Physiotherapist or APN[b] (n=10), mean (SD)	*F* test (*df*)	Levene test for equality of variables	ANOVA *P* value
CQ1—trust in AI[c]	5.9 (1.5)	6.5 (1.0)	5.4 (1.9)	5.1 (1.8)	5.21 (2, 20.58)	0.027	*.007*[d]
CQ2—willingness to use AI	5.9 (1.9)	6.5 (1.4)	5.3 (2.2)	5.3 (2.3)	3.20 (2, 63.00)	0.119	*.04*
CQ3—recommendation of AI	6.1 (2.1)	7.0 (1.4)	5.4 (2.4)	4.7 (2.0)	8.13 (2, 21.55)	0.015	*<.001*

[a]Mean scores for the CQS (trust, willingness to use, and recommendation of ChatGPT), stratified by professional group.

[b]APN: advanced practice nurse.

[c]AI: artificial intelligence.

[d]Values in italics were statistically significant ($P<.05$).

Discussion

In this exploratory, cross-sectional study, a total of 67 participants, comprising 33 resident doctors still in training, 24 board-certified doctors, and 10 lymphedema physiotherapists or APNs, from 34 different institutions across 11 countries evaluated ChatGPT-generated treatment recommendations for 6 fictional lymphedema cases. This is the first study to evaluate ChatGPT-generated treatment recommendations for lymphedema.

Notably Higher Acceptance of ChatGPT Among Younger Doctors

The overall high level of agreement with the ChatGPT-generated treatment recommendations (Q16) indicates broad acceptance across all cases and groups (Table 1). Resident doctors expressed a more positive opinion toward ChatGPT, a tendency that is presumably attributable to their lesser clinical experience and greater openness to digital technologies in medical practice, indicating that the evaluation of ChatGPT may also be influenced by generational factors (Figure 3) [25,26]. Importantly, however, this favorable perception was not uncritical. Residents assigned significantly lower scores than board-certified physicians for source transparency (Q3-Q4;

XSL•FO
RenderX

Figure 1). The lower ratings may reflect stricter expectations regarding explicit source citation, differences in grading thresholds, or a reduced ability to recognize implicit sources that experienced clinicians may infer based on prior knowledge [27]. As the study design does not allow discrimination between these explanations, no conclusions can be drawn regarding the underlying cause of this difference. Nevertheless, several resident doctors reported that ChatGPT already serves them as a useful orientation tool, particularly in situations with limited prior familiarity with a condition or treatment options. This underscores the tool's potential as a rapid decision support or educational aid, as it only uses a few seconds to generate an extensive response [18]. It also highlights the need to integrate the critical use of AI tools into medical education. As residents showed greater willingness to use ChatGPT, targeted training could help prevent misconceptions and ensure these tools complement, rather than replace, clinical reasoning.

However, board-certified doctors demonstrated a lower level of agreement with treatment recommendations provided by ChatGPT, possibly due to their greater clinical experience and higher expectations regarding therapeutic nuances. This may, in part, stem from the perception that AI systems are not yet capable of providing fully personalized treatment algorithms. This viewpoint was supported in the open comments section (Multimedia Appendix 3), where experienced clinicians repeatedly pointed out that well-established surgical interventions, such as LVA and VLNT, were insufficiently addressed. In this study, LVA and VLNT were only briefly mentioned in cases 2, 3, 4, and 6 and only as potential options if compression therapy and liposuction were deemed insufficient. Reconstructive procedures such as LVA and VLNT are mentioned in both the German AWMF (S2k) guidelines and the ISL consensus document [10,11]. While the AWMF (S2k) guidelines suggest considering surgical options only after 6 months of conservative treatment without any lasting beneficial effect, the ISL guidelines refer to microsurgical interventions without specifying any time frame regarding previous complex decongestive therapy [10,11]. However, recent literature suggests that these procedures may also be used earlier or even as a preventative measure, thus challenging the traditional time threshold [28,29]. Given that key papers and guideline discussions on LVA and VLNT were already available well before October 2023, their limited presence in the responses is unlikely to be explained by the model's knowledge cutoff alone [30]. A more plausible explanation is that ChatGPT defaults to what it has seen most often: conservative management pathways that are widely taught and described, and a generally "play it safe" tendency when asked broad clinical questions [31]. In other words, the model may not be unaware of surgery; rather, it may just not surface it unless the prompt explicitly asks for surgical options or provides more detail about operative candidacy [12]. Nevertheless, newer ideas such as prophylactic LVA may still be underrepresented due to the recency of the evidence. Future work should test whether prompting specifically for surgical options changes the balance and completeness of the recommendations. These findings also emphasize the necessity of incorporating temporal limitations when using AI tools in clinical decision-making processes [7,8,20,28-30,32,33]. Furthermore, the strong emphasis on lymphedema management using conservative treatment options may result from the broader and more established evidence base in this area, which has developed over decades and is therefore more prominently represented in ChatGPT's training data.

Physiotherapists or APNs expressed greater skepticism regarding ChatGPT's treatment recommendations, as highlighted in the open comments section. The recommendations were regarded as being too generalized and lacking personalization. It was emphasized that effective lymphedema care necessitates ongoing, individualized assessment, a component that the AI recommendations did not adequately address.

Prompt Quality and Lack of Sources Limit ChatGPT's Clinical Reliability

One of the key characteristics of ChatGPT is its reliance on user input—responses are generated solely based on the prompt it receives [12]. As a result, the specificity and clarity of the input significantly affect the quality, depth, and relevance of the output [12]. Shorter prompts tend to produce shorter, less detailed responses. Citations were completely absent in ChatGPT's responses and consistently rated poorly by participants (Q3; Table 1), demonstrating that the absence of sources significantly limited the reliability of AI-generated treatment recommendation and, consequently, did not earn the physicians' trust. Furthermore, the absence of information regarding additional sources (Q6), the lack of mention of areas of uncertainty (Q7), and the omission of potential consequences of forgoing treatment (Q11) were also rated negatively (Table 1).

As an LLM, ChatGPT does not inherently verify the information it provides, but instead generates content based on linguistic patterns, which may influence the accuracy of treatment recommendations [12]. Notably, previous studies emphasized the need to explicitly ask for citations, as the model may otherwise omit references or produce so-called "hallucinations," meaning fabricated or incorrect references [34,35]. As our prompt did not explicitly request sources, thus reflecting the expected real-world use of ChatGPT, the absence of citations was expected. However, this highlights 2 key limitations. First, when users lack the background knowledge to formulate precise prompts, the model's effectiveness is significantly reduced. Second, users currently have no possibilities to verify the origin or quality of the information provided. Previous studies have shown that such uncritical reliance on unverifiable information has the potential to result in inefficient or inappropriate treatment decisions and, consequently, potentially harm the patient [36]. It is also important to emphasize that ChatGPT does not *collect* information but rather generates responses based on statistical patterns in its training data, which further limit its interpretive reliability.

ChatGPT may also pose security risks [12]. Of particular concern is the use of false or deliberately misleading prompts, the aim of which is to manipulate the model into generating inaccurate responses [12]. To ensure patient safety, any clinical implementation would require strict regulatory oversight. A more realistic, and potentially safer, alternative could involve embedding ChatGPT within a clinical decision support system using prevalidated clinical data [37]. In this scenario, ChatGPT would function as a predictive model, assisting in

XSL•FO
RenderX

decision-making within a strictly regulated and supervised data environment.

Study Strengths and Limitations

This is the first exploratory, cross-sectional study evaluating the agreement, reliability, and decision-making value of AI using LLMs in generating treatment recommendations for lymphedema management in a global, multidisciplinary, and interprofessional panel. Strengths of this study include the sizeable expert panel of 67 doctors and lymphedema therapists from 34 different institutions across 11 countries. A potential limitation of this study is that evaluators were aware that responses were generated by ChatGPT, which may have introduced either positive bias (novelty or expectation effect) or negative bias (preexisting skepticism toward AI). The results should therefore be interpreted as clinicians' perceptions of AI-generated recommendations rather than a purely objective measure of model performance. The lack of human-generated reference standard represents an important limitation, as it prevents the interpretation of absolute DISCERN scores; these should instead be understood in a relative context across professional groups. Although the 6 vignettes were deliberately designed to reflect common clinical presentations across the ISL stage spectrum across both extremities, they cannot capture the full heterogeneity of lymphedema. Consequently, generalizability to all clinical contexts is limited, and future studies should evaluate a larger and more diverse set of cases. Another limitation of this study is the use of an adapted assessment tool based on the DISCERN criteria. Although the underlying framework is validated, modifications to the item set and scoring scale may have altered the original psychometric properties of the instrument. As resident recruitment was conducted via large email and international student group, the response rate of invited resident doctors could not be calculated. Moreover, all analyses were based on a single, time-stamped response per case and were not controlled for random seed values or multirun variability, as this would have substantially increased survey burden. Furthermore, because the study was conducted using the standard ChatGPT web interface rather than the application programming interface, we were unable to access or control key inference parameters such as temperature, top-p, or system prompts. As a result, these factors were not reported. Finally, the study did not assess how variations in prompt design might affect the consistency or accuracy of the generated recommendations. Additionally, participants did not generate the prompts themselves, which may limit user satisfaction to the generated responses.

Conclusions

ChatGPT has the capacity to provide relevant basic knowledge for lymphedema management and may be particularly useful for less experienced clinicians. However, it has not yet succeeded in replacing expert consultation and offers only limited personalization. Additionally, its constraints regarding source transparency and prompt dependency highlight the need for cautious and context-aware integration into clinical practice.

Acknowledgments

During the revision of this manuscript, the authors used GPT-5.2 to improve readability. The authors subsequently reviewed and edited the content as needed and take full responsibility for the final version of the publication.

Funding

This study was funded by the Freiwillige Akademische Gesellschaft Basel, the Swiss Cancer Research Foundation, Rising Tide Foundation, the Swiss National Science Foundation, and the Theodor und Ida Herzog-Egli Stiftung.

Data Availability

The datasets generated and analyzed during this study are stored on the secure Research Electronic Data Capture platform hosted by the Department of Clinical Research at the University of Basel and can be provided by the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. Clinical vignettes were prepared by AF, CEE, and EAK. Data collection was conducted by AF, CEE, and EAK. Statistical analysis and figure preparation were performed by AF and SRW. Data interpretation and drafting of the manuscript were carried out by AF and CEE. The study was supervised by EAK. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompt, cases, and ChatGPT treatment recommendations.
[DOCX File, 36 KB - medinform_v14i1e80553_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Research Electronic Data Capture questionnaire with the validated DISCERN tool.
[DOCX File, 18 KB - medinform_v14i1e80553_app2.docx]

Multimedia Appendix 3
Survey comments per subgroup (anonymous).
[DOCX File, 25 KB - medinform_v14i1e80553_app3.docx]

Multimedia Appendix 4
Group comparisons for questionnaire responses divided by case.
[DOCX File, 64 KB - medinform_v14i1e80553_app4.docx]

References

1. Lin WC, Safa B, Buntic RF. Approach to lymphedema management. Semin Plast Surg 2022 Nov 16;36(4):260-273. [doi: 10.1055/s-0042-1758691] [Medline: 36561430]
2. Grada AA, Phillips TJ. Lymphedema: pathophysiology and clinical manifestations. J Am Acad Dermatol 2017 Dec;77(6):1009-1020. [doi: 10.1016/j.jaad.2017.03.022] [Medline: 29132848]
3. Ridner SH, Bonner CM, Deng J, Sinclair VG. Voices from the shadows: living with lymphedema. Cancer Nurs 2012;35(1):E18-E26. [doi: 10.1097/NCC.0b013e31821404c0] [Medline: 21558848]
4. Vassard D, Olsen MH, Zinckernagel L, Vibe-Petersen J, Dalton SO, Johansen C. Psychological consequences of lymphoedema associated with breast cancer: a prospective cohort study. Eur J Cancer 2010 Dec;46(18):3211-3218. [doi: 10.1016/j.ejca.2010.07.041] [Medline: 20797846]
5. Bowman C, Piedalue KA, Baydoun M, Carlson LE. The quality of life and psychosocial implications of cancer-related lower-extremity lymphedema: a systematic review of the literature. J Clin Med 2020 Oct 2;9(10):3200. [doi: 10.3390/jcm9103200] [Medline: 33023211]
6. Chang DW, Dayan J, Greene AK, et al. Surgical treatment of lymphedema: a systematic review and meta-analysis of controlled trials. Results of a consensus conference. Plast Reconstr Surg 2021 Apr 1;147(4):975-993. [doi: 10.1097/PRS.0000000000007783] [Medline: 33761519]
7. Seidenstuecker K, Fertsch S, Ghazaleh AA, et al. Improving quality of life after breast cancer: a comparison of two microsurgical treatment options for breast cancer-related lymphedema (BCRL). Clin Exp Med 2024 Apr 23;24(1):82. [doi: 10.1007/s10238-024-01344-w] [Medline: 38653874]
8. Kappos EA, Fabi A, Halbeisen FS, et al. Vascularized lymph node transfer (VLNT) versus lymphaticovenous anastomosis (LVA) for chronic breast cancer-related lymphedema (BCRL): a retrospective cohort study of effectiveness over time. Breast Cancer Res Treat 2025 Apr;210(2):319-327. [doi: 10.1007/s10549-024-07567-5] [Medline: 39653884]
9. Yang JCS, Wang YM, Wu SC, et al. Lymphaticovenous anastomosis for treating secondary lower limb lymphedema in older patients-a retrospective cohort study. J Clin Med 2022 May 30;11(11):3089. [doi: 10.3390/jcm11113089] [Medline: 35683479]
10. Executive Committee of the International Society of Lymphology. The diagnosis and treatment of peripheral lymphedema: 2020 Consensus Document of the International Society of Lymphology. Lymphology 2020;53(1):3-19. [Medline: 32521126]
11. Wilting J, Bartkowski R, Baumeister R, Földi E, Stöhr S, Strubel G, et al. S2k Leitlinie: diagnostik und therapie der lymphödeme. : Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften; 2017 URL: http://register.awmf.org/assets/guidelines/058-001l_S2k_Diagnostik_und_Therapie_der_Lymphoedeme_2019-07-abgelaufen.pdf [accessed 2026-03-19]
12. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst 2023;2(2):81-83. [doi: 10.54097/fcis.v2i2.4465]
13. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High Confid Comput 2024 Jun;4(2):100211. [doi: 10.1016/j.hcc.2024.100211]
14. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
15. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. J Pers Med 2023 Oct 16;13(10):1502. [doi: 10.3390/jpm13101502] [Medline: 37888113]
16. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
17. Nietsch KS, Shrestha N, Mazudie Ndjonko LC, et al. Can large language models (LLMs) predict the appropriate treatment of acute hip fractures in older adults? Comparing appropriate use criteria with recommendations from ChatGPT. J Am Acad Orthop Surg Glob Res Rev 2024 Aug 9;8(8):e24.00206. [doi: 10.5435/JAAOSGlobal-D-24-00206] [Medline: 39137403]

XSL•FO
RenderX

18. Trinkley KE, An R, Maw AM, Glasgow RE, Brownson RC. Leveraging artificial intelligence to advance implementation science: potential opportunities and cautions. Implement Sci 2024 Feb 21;19(1):17. [doi: 10.1186/s13012-024-01346-y] [Medline: 38383393]
19. Charnock D. The DISCERN Handbook: Quality Criteria for Consumer Health Information on Treatment Choices: Radcliffe Medical Press; 1998.
20. Cheng MH, Chang DW, Patel KM. Principles and Practice of Lymphedema Surgery, 2nd edition: Elsevier Health Sciences; 2021.
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Kappos EA, Haas Y, Schulz A, et al. The LYMPH trial: comparing microsurgical with conservative treatment for chronic breast cancer-associated lymphoedema - study protocol of a pragmatic randomised international multicentre superiority trial. BMJ Open 2025 Feb 17;15(2):e090662. [doi: 10.1136/bmjopen-2024-090662] [Medline: 39961719]
23. Lymphödem Vereinigung Schweiz. 2025. URL: https://www.lv-schweiz.ch [accessed 2026-03-19]
24. RStudio: integrated development environment for R. Posit. URL: https://posit.co [accessed 2026-03-20]
25. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ 2024 Apr 10;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
26. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? Digit Health 2023 Jan 26;9:20552076231152167. [doi: 10.1177/20552076231152167] [Medline: 36762024]
27. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
28. Sun JM, Yamamoto T. Primary surgical prevention of lymphedema. J Chin Med Assoc 2024 Jun 1;87(6):567-571. [doi: 10.1097/JCMA.0000000000001101] [Medline: 38666773]
29. Wan R, Hussain A, Kuruoglu D, Houdek MT, Moran SL. Prophylactic lymphaticovenous anastomosis (LVA) for preventing lymphedema after sarcoma resection in the lower limb: a report of three cases and literature review. Microsurgery 2023 Mar;43(3):273-280. [doi: 10.1002/micr.30975] [Medline: 36226524]
30. OpenAI. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
31. Saban M, Dubovi I. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. J Adv Nurs 2025 Nov;81(11):7489-7499. [doi: 10.1111/jan.16101] [Medline: 38366690]
32. Cheng MH, Loh CYY, Lin CY. Outcomes of vascularized lymph node transfer and lymphovenous anastomosis for treatment of primary lymphedema. Plast Reconstr Surg Glob Open 2018 Dec 20;6(12):e2056. [doi: 10.1097/GOX.0000000000002056] [Medline: 30656125]
33. Scaglioni MF, Fontein DB, Arvanitakis M, Giovanoli P. Systematic review of lymphovenous anastomosis (LVA) for the treatment of lymphedema. Microsurgery 2017 Nov;37(8):947-953. [doi: 10.1002/micr.30246] [Medline: 28972280]
34. Athaluri SA, Manthena SV, Kesapragada VS, Yarlagadda V, Dave T, Duddumpudi RT. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023 Apr 11;15(4):e37432. [doi: 10.7759/cureus.37432] [Medline: 37182055]
35. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
36. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv. Preprint posted online on Aug 11, 2024. [doi: 10.1056/AIoa2300015]
37. Liao Z, Wang J, Shi Z, Lu L, Tabata H. Revolutionary potential of ChatGPT in constructing intelligent clinical decision support systems. Ann Biomed Eng 2024 Feb;52(2):125-129. [doi: 10.1007/s10439-023-03288-w] [Medline: 37332008]

Abbreviations

AI: artificial intelligence
APN: advanced practice nurse
AWMF: Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften
ISL: International Society of Lymphology
LLM: large language model
LVA: lymphovenous anastomosis
REDCap: Research Electronic Data Capture
VLNT: vascularized lymph node transfer

XSL•FO
RenderX

Edited by A Coristine, A Benis; submitted 12.Jul.2025; peer-reviewed by E Dylke, N Bievre, T Lefèvre; revised version received 02.Feb.2026; accepted 02.Feb.2026; published 08.Apr.2026.

Please cite as:
Fabi A, Egli CE, Wendelspiess SR, Griewing S, Haas Y, De Pellegrin L, Schaefer DJ, Qiu SS, Harder Y, Kappos EA
Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making
JMIR Med Inform 2026;14:e80553
URL: https://medinform.jmir.org/2026/1/e80553
doi:10.2196/80553

© Adriano Fabi, Caroline E Egli, Séverin R Wendelspiess, Sebastian Griewing, Yvonne Haas, Laura De Pellegrin, Dirk J Schaefer, Shan S Qiu, Yves Harder, Elisabeth A Kappos. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning Prediction of Progression to Dialysis in Patients With Polycystic Kidney Disease: Population-Based Retrospective Cohort Study

Cheng-Hao Chang[1*], MD; Mingchih Chen[2,3*], PhD; Ming-Hsien Tsai[1,4], MD, PhD; Yen-Chun Huang[5], PhD; Hung-Hsiang Liou[6], MD; Ben-Chang Shia[2,3], MD, PhD; Chingying Liang[2,3], MD; Yu-Wei Fang[1,4], MD, PhD

[1]Division of Nephrology, Department of Internal Medicine, Shin-Kong Wu Ho-Su Memorial Hospital, No.95, Wen-Chang Road, Shih-lin District, Taipei, Taiwan

[2]Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City, Taiwan

[3]AI Development Center, Fu Jen Catholic University, New Taipei City, Taiwan

[4]Department of Medicine, Fu Jen Catholic University School of Medicine, New Taipei City, Taiwan

[5]Department of Artificial Intelligence, Tamkang University, New Taipei City, Taiwan

[6]Division of Nephrology, Department of Internal Medicine, Hsin-Jen Hospital, New Taipei City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Yu-Wei Fang, MD, PhD
Division of Nephrology, Department of Internal Medicine, Shin-Kong Wu Ho-Su Memorial Hospital, No.95, Wen-Chang Road, Shih-lin District, Taipei, Taiwan

Abstract

Background: Autosomal dominant polycystic kidney disease (ADPKD), characterized by progressive cyst growth and renal decline, is the leading genetic cause of end‐stage renal disease.

Objective: This study aims to develop and validate machine learning (ML) models for predicting the risk of progression to dialysis in patients with ADPKD using a nationwide administrative database. Early identification of high-risk patients is critical for timely monitoring.

Methods: This retrospective cohort study used data from Taiwan's National Health Insurance Research Database (2007‐2018) to identify newly diagnosed patients with ADPKD. Six ML algorithms, including logistic regression, random forest, and extreme gradient boosting (XGBoost), were employed to predict progression to dialysis. Models were developed using 10-fold cross-validation, with the Synthetic Minority Oversampling Technique applied within training folds to address class imbalance. An ensemble-based feature selection strategy was implemented to identify the most robust predictors and optimize final model performance. Model evaluation was conducted using a strict temporal split.

Results: The study included 1856 patients with ADPKD, of whom 302 (16.27%) progressed to dialysis. Multivariable Cox regression identified several significant risk factors, including age 66 years and older (hazard ratio [HR] 4.63, 95% CI 2.71‐7.92; P<.001), anemia (HR 4.33, 95% CI 3.25‐5.78; P<.001), congestive heart failure (HR 1.81, 95% CI 1.29‐2.54; P<.001), and acute kidney injury (HR 1.69, 95% CI 1.19‐2.41; P=.003). Among the ML models, the XGBoost model, using an optimized set of 27 features, demonstrated the highest predictive performance on the held-out temporal test set (accuracy 98.3%; area under the curve 0.955; F_1-score 0.800; Brier score 0.022). The top predictors in the XGBoost model largely aligned with age, comorbidity burden, anemia, and cardiovascular disease markers. Medication use (eg, anticoagulants, loop diuretics, febuxostat) was also among the most influential predictors; however, medication-related predictors should be interpreted as proxies for disease complexity rather than direct risk modulators.

Conclusions: ML models can predict dialysis risk in patients with ADPKD using administrative data with temporal validation. This approach may support risk stratification by helping identify individuals at higher predicted risk who may warrant closer monitoring and further specialist evaluation.

(JMIR Med Inform 2026;14:e80343) doi:10.2196/80343

KEYWORDS

autosomal dominant polycystic kidney disease; ADPKD; artificial intelligence; machine learning; dialysis risk; predictive modeling

XSL•FO
RenderX

Introduction

Background

Polycystic kidney disease (PKD) is the leading genetic cause of end-stage renal disease (ESRD). The most common mutations in PKD occur in the *PKD1* or *PKD2* gene, which encode the proteins polycystin-1 and polycystin-2, respectively. These mutations result in abnormal cell proliferation, fluid secretion, and differentiation, leading to the formation and growth of cysts [1]. The progressive nature of renal cysts may cause structural and functional changes in the kidney. According to the latest guidelines, the estimated prevalence of PKD varies widely, ranging from approximately 2 to 14 per 10,000 individuals, depending on the study methodology and population characteristics [2].

Contemporary management of autosomal dominant polycystic kidney disease (ADPKD) has evolved to include both the management of associated complications and direct interventions aimed at the disease mechanism. A cornerstone of care remains the control of hypertension, a common and impactful complication. Angiotensin-converting enzyme inhibitors (ACEIs) and angiotensin II receptor blockers (ARBs) are widely recommended for their established efficacy in blood pressure control and proteinuria mitigation [3,4]. In addition, lifestyle interventions, such as dietary sodium restriction and weight management, play a crucial role in preserving renal function [5,6]. In parallel, disease-modifying therapies, such as the vasopressin V2 receptor antagonist tolvaptan, have been developed to slow cyst growth in select patients with rapidly progressing disease, who are often identified using prognostic markers such as genetic status or total kidney volume (TKV) [7]. Furthermore, other medications, such as statins and metformin, are frequently prescribed to manage concurrent conditions in this population, though their role as primary treatment options for ADPKD itself remains under investigation [8,9].

Study Objective

While risk stratification using advanced tools is crucial for guiding specialized treatments, a significant need remains for risk assessment methods based on accessible, routinely collected data. Currently, the gold-standard models for prognostication in ADPKD include the Mayo Imaging Classification, which utilizes height-adjusted TKV and patient age to predict the rate of estimated glomerular filtration rate (eGFR) decline, and the PROPKD score, which integrates genetic information (*PKD1* vs *PKD2* mutation type) with early clinical manifestations [2,10,11]. Although these models are invaluable, their broad implementation is often hindered by their dependence on specialized and resource-intensive evaluations. For instance, the Mayo Classification requires magnetic resonance imaging (MRI) or computed tomography to measure TKV, while the PROPKD score necessitates comprehensive genetic sequencing. These resources can be costly and are not universally available in all clinical settings, creating a gap for more scalable screening tools.

In parallel with these clinical tools, artificial intelligence (AI) has been increasingly used in ADPKD management. AI models have been developed to predict glomerular filtration rate decline, aiding in the early identification of high-risk patients and facilitating proactive management [12]. In imaging-based classifications, deep learning algorithms have been applied to automate ADPKD severity assessment, enhancing objectivity and reducing interobserver variability [13]. AI has also been instrumental in TKV quantification, a crucial prognostic marker in ADPKD. Deep learning-based segmentation models have demonstrated high accuracy in automated kidney segmentation from MRI and computed tomography images, significantly reducing manual workload [14,15]. While these advancements improve disease monitoring, they often still rely on the availability of imaging or detailed clinical data, leaving a gap for predictive models that can leverage more widely accessible administrative data to forecast hard clinical end points.

Therefore, this study aimed to develop and validate an AI-assisted model to identify significant predictors of progression to dialysis in patients with ADPKD using a nationwide administrative database. We evaluated multiple machine learning (ML) algorithms using a prespecified temporal split for model development and evaluation and summarized predictors to enhance interpretability and clinical usability. Our goal was to create a complementary tool for risk stratification, potentially aiding clinicians in monitoring patients and identifying those who may warrant more specialized evaluation.

Methods

Data Sources

This retrospective cohort study was based on the National Health Insurance Research Database (NHIRD), which is derived from Taiwan's universal compulsory health insurance system covering nearly 99% of the 23 million individuals in Taiwan since 1998 [16]. The database includes patients' hospitalization and outpatient visit records, and each individual was continuously followed. Disease diagnoses were coded according to the International Classification of Diseases, Ninth Revision, Clinical Modification (*ICD-9*-CM). The NHI Administration fully adopted *ICD-10*-CM for clinical records in 2016.

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki and was reviewed and approved by the Ethics Review Board of Shin-Kong Wu Ho-Su Memorial Hospital (approval number 20211008R). The requirement for informed consent was waived by the review board because this study involved a secondary analysis of routinely collected administrative data, and no direct contact with individual participants occurred. The original data collection procedures and this secondary analysis were approved without the need for additional consent.

All data obtained from the NHIRD were fully deidentified prior to analysis. Personal identifiers were removed by the data holder, and the authors had no access to information that could be used to identify individual patients. Data access and analysis were conducted in compliance with relevant data protection regulations.

XSL•FO
RenderX

No financial or other compensation was provided to participants, as this study did not involve direct recruitment or interaction with human subjects.

The manuscript and all supplementary materials did not include any images or information that could lead to the identification of individual participants.

Study Design and Population

We identified new-onset ADPKD cases from the Catastrophic Illness Patient Registry between 2007 and 2018. To minimize potential misclassification bias from ICD coding, we implemented stringent validation criteria for both the primary diagnosis and comorbidities. The identification of patients with ADPKD followed a 2-step process: subjects were first identified using ICD diagnostic codes, and their diagnosis was then validated against Taiwan's Catastrophic Illness Patient Registry. A catastrophic illness certificate requires formal review and approval by physicians, ensuring a high degree of diagnostic accuracy for our study cohort [17-19].

The index date was defined as the first date on which a patient met the ADPKD case definition and had validation in the Catastrophic Illness Patient Registry. Baseline predictors were ascertained during the 365-day lookback period preceding the index date.

To construct a clean incident cohort and reduce reverse causation, we excluded individuals aged below 19 years or above 85 years, those with missing core administrative information (eg, sex, age, or index date), and those with evidence of ADPKD during the 2005 to 2006 washout period. We also excluded patients with any evidence of dialysis (hemodialysis or peritoneal dialysis) before the index date and those with kidney transplantation before the index date. Patients were followed from the index date until initiation of dialysis, death, or December 31, 2019, whichever occurred first.

Baseline Variables

The dataset includes a variety of demographic, clinical, and medication-related variables for individuals with ADPKD. The key demographic variables include sex and age. Clinical variables encompass a variety of diseases grouped as cerebrovascular (intracranial aneurysms, ischemic stroke, and hemorrhagic stroke), cardiovascular (hypertension, atrial fibrillation, congestive heart failure [CHF], peripheral vascular disease, arrhythmia, and ischemic heart disease), gastrointestinal (liver cirrhosis, peptic ulcer, diverticulosis, cholangitis, and acute pancreatitis), renal and urological (urinary tract infection [UTI] and acute kidney injury [AKI]), pulmonary (chronic obstructive pulmonary disease [COPD], pneumonia, and asthma), metabolic (dyslipidemia, diabetes mellitus, and gout), mental (dementia, anxiety, and depression), and other (glaucoma and anemia). Each comorbidity was confirmed if the corresponding diagnosis code appeared in at least 1 hospitalization record or at least 3 outpatient visit records within the 365-day lookback period preceding the index date.

Medication variables cover a broad spectrum, including antihypertensives (beta-blockers, alpha-blockers, ACEIs/ARBs, calcium channel blockers [CCBs], methyldopa, hydralazine, minoxidil, clonidine, potassium-sparing diuretics, thiazide diuretics, and loop diuretics), antidiabetic drugs (metformin, sulfonylureas, dipeptidyl peptidase-4 inhibitors, thiazolidinediones, other oral hypoglycemic agents, rapid-acting insulins, and long-acting insulins), antimetabolic agents (statins, fenofibrate, and other lipid-lowering agents), urate-lowering agents (benzbromarone, allopurinol, febuxostat), and other medications (nonsteroidal anti-inflammatory drugs [NSAIDs], sedative hypnotics, tranexamic acid, mammalian target of rapamycin inhibitors, sodium bicarbonate, anticoagulants, antiplatelet agents, and vitamin K). As with the definition of comorbidities, we adopted a widely used and validated approach to enhance specificity. Medication variables were quantified as the cumulative number of prescription days during the 365-day lookback period preceding the index date. This approach was used to capture exposure intensity beyond a binary indicator. The specific codes used to define all comorbidities and medications are detailed in Tables S1 and S2 in Multimedia Appendix 1, respectively.

Outcome Definition and Follow-Up

The primary end point was ESRD, defined as initiation of dialysis or receipt of kidney transplantation without prior dialysis. Patients were followed from the index date until ESRD, death, or December 31, 2019, whichever occurred first. For the ML classification models, the binary outcome was ESRD by the end of follow-up (event=1). In the ML analysis, the label reflects whether ESRD was observed by December 31, 2019, and does not represent a fixed-horizon risk. Patients without ESRD were labeled as nonevents (event=0), including those who died before reaching ESRD.

Data Preprocessing and Handling of Missing Data

Our strategy for handling missing data involved 2 stages. First, case-wise deletion was performed to exclude any patient records with missing core administrative data (eg, sex, age, or date of diagnosis), as indicated in the study flowchart. For the final cohort, predictor variables were constructed such that no missing values remained. The absence of a record for a specific variable was treated as meaningful data rather than as a missing value. For instance, if a patient had no prescriptions for a given medication, the cumulative exposure days were coded as 0. Similarly, if a patient did not meet the predefined criteria for a comorbidity, the variable was coded as 0, indicating absence.

Statistical Analysis

Categorical variables were presented as frequencies with percentages and compared using the chi-square test. Continuous variables were expressed as means with SDs and compared using the *t* test. Multivariable Cox proportional hazards regression models were used to estimate adjusted hazard ratios (HRs) and 95% CIs for the risk of dialysis initiation. A 2-tailed *P* value of <.05 was considered statistically significant. All statistical analyses were performed using SAS version 9.4 (SAS Institute Inc).

ML Model Development and Feature Selection

Given the extended accrual period, we used a temporal split to preserve chronological separation between model development and evaluation. Records from 2007 to 2016 (80%) were used

XSL•FO
RenderX

for model training, and records from 2017 to 2018 (20%) were reserved as an independent held-out test set for performance assessment.

Within the training set, we performed 10-fold cross-validation to improve robustness and reduce overfitting. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied only within the training folds of each cross-validation iteration to avoid information leakage into the validation folds. Model performance was evaluated on the held-out test set using accuracy, Cohen κ, sensitivity, specificity, area under the curve (AUC), F_1 score, and Brier score, recognizing that multiple complementary metrics are needed under imbalanced outcome distributions.

The 6 ML algorithms were chosen to encompass a broad spectrum of modeling techniques, ranging from interpretable linear models to complex, high-performance ensembles. Logistic regression was selected as a conventional statistical method widely used in clinical research, serving as a robust and interpretable baseline for comparison. To capture complex nonlinear relationships, we included 2 distinct models: classification and regression trees, a foundational decision tree algorithm, and multivariate adaptive regression splines, a flexible model adept at identifying interactions between predictors. Finally, to leverage the superior predictive power of ensemble learning, 3 state-of-the-art algorithms were selected [20]. Random forest represents the bagging method, which builds independent trees to reduce variance, whereas CatBoost and eXtreme Gradient Boosting (XGBoost) are representatives of boosting methods that construct models sequentially to correct the errors of their predecessors and often yield high accuracy. This comprehensive selection allowed for a thorough evaluation of predictive performance across different algorithmic families, and these methods have been successfully applied in similar medical informatics contexts [21-23] (see Supplementary Methods in Multimedia Appendix 1 for detailed model descriptions). Hyperparameters for each model were optimized using random search within the training set under 10-fold cross-validation. The held-out temporal test set was used only for final performance evaluation. The specific hyperparameters for all models are detailed in Table S3 in Multimedia Appendix 1. All model construction was conducted using Python (version 3.9; Python Software Foundation).

We employed a 2-pronged feature selection approach to identify the most significant predictors. In the first approach, each model performed feature selection based on its own internal mechanism, and its performance was evaluated using this algorithm-specific feature set. In the second approach, we constructed an overall consensus ranking of predictors using rank aggregation across the 6 algorithms. We prespecified that the best-performing algorithm would be selected based on performance on the held-out temporal test set, considering discrimination, class-imbalance–sensitive metrics, and calibration. For the selected best-performing model, we additionally evaluated parsimony by sequentially adding predictors according to the consensus ordering and selecting the feature count that maximized test-set performance.

Results

Characteristics of the Study Population

From the NHIRD, 2677 incident ADPKD cases were identified between 2007 and 2018. After applying exclusion criteria (n=821), 1856 patients were included in the final cohort (Figure 1).

XSL•FO
RenderX

Figure 1. Study flowchart of patient selection and machine learning model development. The diagram shows the patient selection process from the NHIRD for the period 2007 - 2018 and the subsequent workflow for model development. Patients were split temporally into a training set (2007 - 2016) and a held-out test set (2017 - 2018). ESRD was ascertained through December 31, 2019. ADPKD: autosomal dominant polycystic kidney disease; CART: classification and regression trees; ESRD: end-stage renal disease; eXGB: extreme gradient boosting; LGR: logistic regression; MARS: multivariate adaptive regression splines; ML: machine learning; RF: random forest.

XSL•FO
RenderX

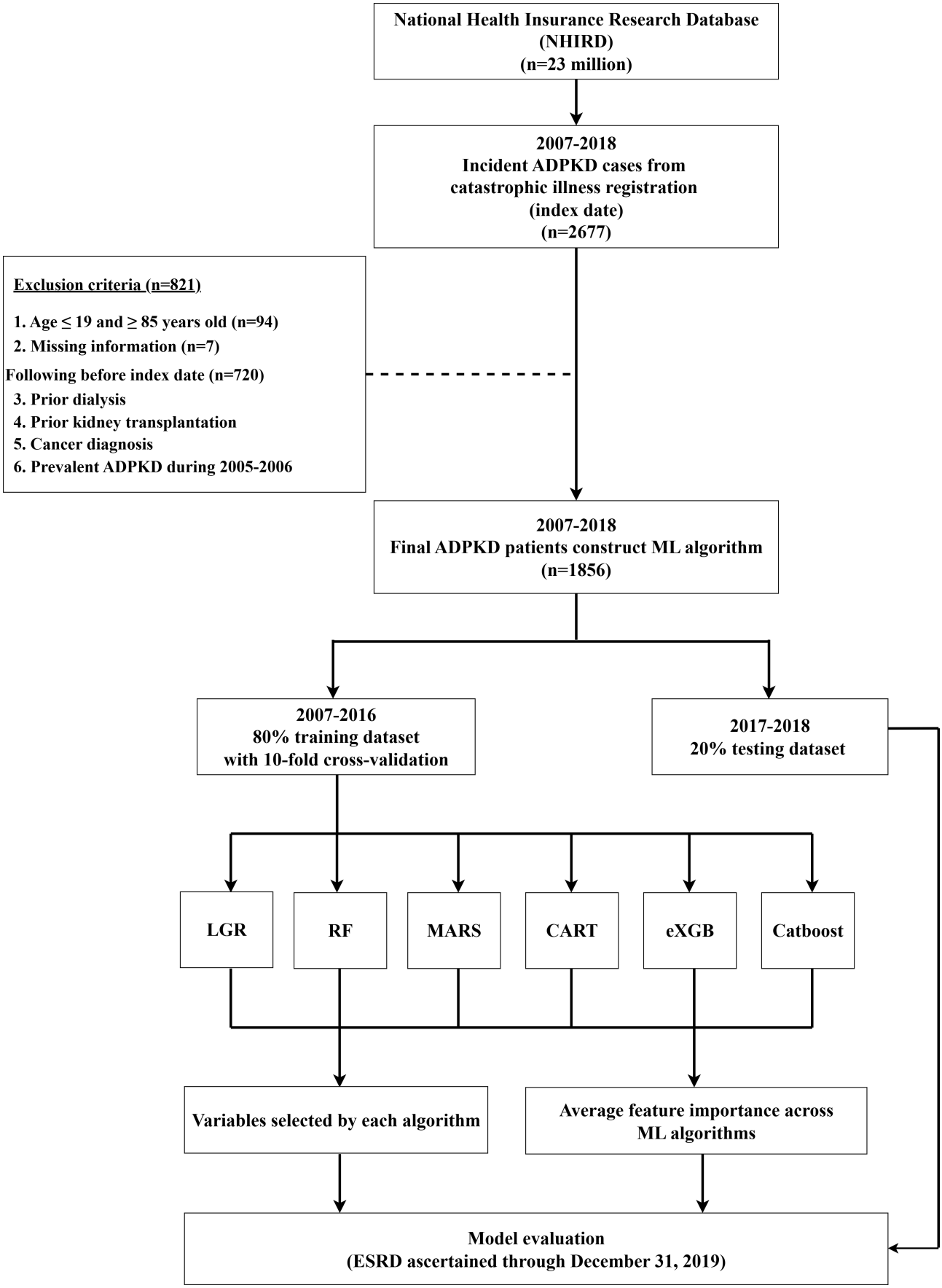


Table 1 summarizes baseline characteristics assessed during the 365-day lookback period preceding the index date. During follow-up through December 31, 2019, 302 patients (302/1856, 16.27%) initiated dialysis. Kidney transplantation without prior dialysis was observed in only 2 patients. Due to small cell size restrictions intended to protect patient privacy in the NHIRD

XSL•FO
RenderX

data release process, transplantation events could not be analyzed separately; therefore, ESRD outcomes are presented primarily in terms of dialysis initiation. The remaining 1554 patients (1554/1856, 83.73%) did not initiate dialysis and were censored at death or end of follow-up.

XSL•FO
RenderX

Table . Baseline characteristics of the cohort stratified by dialysis initiation during follow-up.

Baseline characteristic	Autosomal dominant PKD[a] (n=1856)	Dialysis initiation during follow-up	
		Yes	No
Sex , n (%)			
Male	1006 (54.20)	168 (55.63)	838 (53.93)
Female	850 (45.80)	134 (44.37)	716 (46.07)
Age group (y), n (%)			
≤40	687 (37.02)	56 (18.54)	631 (40.60)
41 - 55	754 (40.63)	152 (50.33)	602 (38.74)
56 - 65	265 (14.28)	56 (18.54)	209 (13.45)
≥66	150 (8.08)	38 (12.58)	112 (7.21)
Age, mean (SD)	45.85 (13.35)	50.94 (11.94)	44.86 (13.39)
CCI[b] group, n (%)			
1 - 2	1014 (54.63)	72 (3.88)	942 (50.75)
3 - 5	660 (35.56)	168 (9.05)	492 (26.51)
6+	182 (9.81)	62 (3.34)	120 (6.47)
CCI scores, mean (SD)	2.58 (2.34)	4.18 (2.27)	2.27 (2.22)
Underlying disease, n (%)			
Cerebrovascular diseases			
Intracranial aneurysms	20 (1.08)	4 (1.32)	16 (1.03)
Ischemic stroke	51 (2.75)	11 (3.64)	40 (2.57)
Hemorrhagic stroke	91 (4.90)	27 (8.94)	64 (4.12)
Cardiovascular diseases			
Hypertension	1164 (62.72)	203 (67.22)	961 (61.84)
AF[c]	34 (1.83)	15 (4.97)	19 (1.22)
CHF[d]	178 (9.59)	77 (25.50)	101 (6.50)
PVD[e]	71 (3.83)	26 (8.61)	45 (2.90)
Arrhythmia	214 (11.53)	69 (22.85)	145 (9.33)
Ischemic heart disease	344 (18.53)	105 (34.77)	239 (15.38)
Gastrointestinal diseases			
Liver cirrhosis	33 (1.78)	14 (4.64)	19 (1.22)
Peptic ulcer bleeding	84 (4.53)	25 (8.28)	59 (3.80)
Diverticulosis	32 (1.72)	10 (3.31)	22 (1.42)
Cholangitis	39 (2.10)	16 (5.30)	23 (1.48)
Acute pancreatitis	41 (2.21)	20 (6.62)	21 (1.35)
Renal and urological diseases			
UTI[f]	938 (50.54)	196 (64.90)	742 (47.57)
AKI[g]	173 (9.32)	71 (23.51)	102 (6.56)
Lung diseases			
COPD[h]	319 (17.19)	71 (23.51)	248 (15.96)
Pneumonia	543 (29.26)	121 (40.07)	422 (27.16)
Asthma	207 (11.15)	31 (10.26)	176 (11.33)

XSL•FO
RenderX

Baseline characteristic	Autosomal dominant PKD[a] (n=1856)	Dialysis initiation during follow-up	
		Yes	No
Metabolic diseases			
Dyslipidemia	706 (38.04)	124 (41.06)	582 (37.45)
DM[i]	277 (14.92)	73 (24.17)	204 (13.13)
Gout	427 (23.01)	122 (40.40)	305 (19.63)
Mental diseases			
Dementia	47 (2.53)	15 (4.97)	32 (2.06)
Anxiety	321 (17.30)	62 (20.53)	259 (16.67)
Depression	250 (13.47)	46 (15.23)	204 (13.13)
Others			
Glaucoma	88 (4.74)	18 (5.96)	70 (4.50)
Anemia	461 (24.84)	199 (65.89)	262 (16.86)
Medication, mean days of use during the 365-day pre-index period (SD)			
Antihypertension			
Beta-blockers	95.85 (107.05)	92.24 (91.14)	97.15 (112.30)
Alpha-blockers	94.40 (106.66)	101.10 (107.30)	90.89 (106.30)
ACEI[j]/ARB[k]	160.11 (132.18)	130.10 (102.30)	167.30 (137.50)
CCB[l]	157.80 (136.36)	162.20 (122)	156.50 (140.40)
Methyldopa	26.79 (37.29)	17.97 (28.55)	31.61 (41.78)
Hydralazine	51.68 (65.70)	46.13 (60.99)	56.27 (69.50)
Minoxidil	53.04 (70.99)	43.11 (51.21)	68.30 (94.18)
Clonidine	41.69 (49.51)	42.35 (38)	40.92 (61.88)
Potassium-sparing diuretics	54.94 (83.58)	31.52 (68.85)	60.38 (85.97)
Thiazide diuretics	66.84 (83.42)	45.18 (67.33)	73.30 (86.68)
Loops diuretics	59.35 (94.95)	60.11 (84.12)	58.75 (102.70)
Antidiabetic			
Metformin	122.29 (122.92)	44.17 (43.31)	130.20 (125.70)
Sulfonylureas	130.89 (120.83)	103.20 (117.20)	137.30 (121.40)
DPP4i[m]	120.03 (101.88)	100.40 (103)	125 (101.50)
Other OHAs	92.69 (102.37)	113.50 (126.70)	82.84 (88.87)
Rapid-acting insulins	14.71 (49.95)	2.12 (8.94)	27.31 (68.23)
Long-acting insulins	100.09 (88.53)	130 (109.90)	84.26 (73.73)
Antimetabolic			
Statin	121.32 (111.87)	124.60 (118.10)	120.60 (110.50)
Fenofibrate	57.96 (72.62)	56.79 (81.12)	58.42 (69.38)
Other lipid-lowering agents	75.24 (90.44)	6.19 (2.38)	92.50 (93.65)
Urate-lowering agents			
Benzbromarone	65.45 (79.37)	41.72 (49.69)	71.38 (84.22)
Allopurinol	82.52 (97.02)	78.39 (84.91)	85.45 (105)
Febuxostat	104.04 (101.03)	92.67 (79.03)	108.20 (107.80)
Other			

XSL•FO
RenderX

Baseline characteristic	Autosomal dominant PKD[a] (n=1856)	Dialysis initiation during follow-up	
		Yes	No
NSAID[n]	15.76 (34.22)	15.69 (29.67)	15.78 (35.12)
Sedative hypnotics	92.03 (158.68)	118.70 (188.30)	83.81 (147.50)
Tranexamic acid	5.05 (16.89)	6.59 (11.14)	4.59 (18.24)
mTOR[o] inhibitors	135.14 (148.87)	127.80 (140.60)	144 (174.70)
Sodium bicarbonate	9.50 (42.61)	20.52 (46.30)	7.25 (41.48)
Anticoagulants	110.46 (155.24)	113.10 (190.60)	109 (132.20)
Antiplatelet agents	211.22 (251.64)	204.20 (228.70)	214.30 (261.10)
Vitamin K	57.41 (91.58)	19.86 (43.90)	107.50 (114.30)

[a]PKD: polycystic kidney disease.
[b]CCI: Charlson Comorbidity Index.
[c]AF: atrial fibrillation.
[d]CHF: congestive heart failure.
[e]PVD: peripheral vascular disease.
[f]UTI: urinary tract infection.
[g]AKI: acute kidney injury.
[h]COPD: chronic obstructive pulmonary disease.
[i]DM: diabetes mellitus.
[j]ACEI: angiotensin-converting enzyme inhibitor.
[k]ARB: angiotensin II receptor blocker.
[l]CCB: calcium channel blocker.
[m]DPP4i: dipeptidyl peptidase-4 inhibitor.
[n]NSAID: nonsteroidal anti-inflammatory drug.
[o]mTOR: mechanistic target of rapamycin.

At baseline, the cohort had substantial cardiorenal comorbidity and medication use. The most prevalent comorbidities were hypertension (86.58%, 1606/1856) and UTI (50.38%, 935/1856), with pneumonia (29.26%, 543/1856), anemia (24.83%, 461/1856), and gout (23.01%, 427/1856) also common. Antihypertensive exposure was widespread (ACEI/ARB 161.30, SD 127.50 d; CCB 160.90, SD 133.70 d), and metabolic therapy was frequent (metformin 116.20, SD 137.10 d; long-acting insulin 91.71, SD 139.10 d).

Stratification by subsequent dialysis initiation highlighted clinically meaningful differences in baseline disease severity and management. Patients who later initiated dialysis were older (50.94, SD 11.94 vs 44.86, SD 13.39 y) and had higher comorbidity burden (Charlson Comorbidity Index [CCI] 4.18, SD 2.27 vs 2.27, SD 2.22; CCI≥6: 20.53% vs 7.72%), with the largest separations for anemia (65.89% vs 16.86%) and AKI (23.51% vs 6.56%). Baseline medication patterns were consistent with greater clinical complexity in the dialysis group, including lower metformin exposure (44.17, SD 95.41 vs 130.20, SD 148 d) and higher sodium bicarbonate exposure (20.52, SD 68.46 vs 7.25, SD 44.30 d).

Risk Analysis for Progression to Dialysis

Table 2 presents the multivariable Cox regression results. After adjustment for sex, age category, CCI category, and comorbidities, male sex was associated with a higher risk of dialysis initiation (aHR 1.55, 95% CI 1.13 - 2.11; *P*=.006). Age remained a strong predictor: compared with age younger than 55 years, the risk increased across categories. Among comorbidities, anemia showed the strongest association with dialysis initiation (aHR 4.33, 95% CI 3.25 - 5.78; *P*<.001). Congestive heart failure, AKI, and gout were also associated with a higher risk. Hypertension, COPD, and anxiety were associated with lower hazard in the adjusted model.

XSL•FO
RenderX

Table . Risk for entering dialysis in the patients with ADPKD[a].

	Cox model	
	aHR[b] (95% CI)	*P* value
Sex		
Ref=Female	1	—[c]
Male	1.55 (1.13 - 2.11)	.006
Age group		
Ref=<55	1	—
41 - 55	3.01 (2.12 - 4.30)	<.001
56 - 65	4.13 (2.63 - 6.50)	<.001
≥66	4.63 (2.71 - 7.92)	<.001
CCI[d] group		
Ref=12	1	—
345	1.42 (1.01 - 1.98)	.04
6+	1.28 (0.74 - 2.21)	.39
Underlying disease		
Cerebrovascular diseases		
Intracranial aneurysms	2.12 (0.71 - 6.31)	.18
Ischemic stroke	1.22 (0.59 - 2.49)	.59
Hemorrhagic stroke	1.03 (0.57 - 1.86)	.91
Cardiovascular diseases		
Hypertension	0.70 (0.53 - 0.94)	.02
AF[e]	0.66 (0.32 - 1.32)	.24
CHF[f]	1.81 (1.29 - 2.54)	<.001
PVD[g]	1.25 (0.79 - 2.00)	.34
Arrhythmia	0.98 (0.69 - 1.40)	.92
Gastrointestinal diseases		
Ischemic heart disease	1.09 (0.80 - 1.48)	.58
Liver cirrhosis	0.91 (0.48 - 1.72)	.77
Peptic ulcer bleeding	0.67 (0.42 - 1.09)	.11
Diverticulosis	0.82 (0.38 - 1.79)	.62
Cholangitis	1.10 (0.61 - 1.96)	.76
Acute pancreatitis	1.18 (0.69 - 2.04)	.55
Renal and urological diseases		
UTI[h]	0.84 (0.62 - 1.14)	.26
AKI[i]	1.69 (1.19 - 2.41)	.003
COPD[j]	0.61 (0.44 - 0.86)	.005
Pneumonia	0.86 (0.65 - 1.12)	.26
Asthma	0.74 (0.47 - 1.15)	.18
Metabolic diseases		
Dyslipidemia	0.78 (0.51 - 1.01)	.12
DM[k]	1.38 (0.97 - 1.96)	.07

XSL•FO
RenderX

	Cox model	
	aHR[b] (95% CI)	*P* value
Gout	1.49 (1.08 - 2.05)	.01
Mental diseases		
Dementia	0.71 (0.37 - 1.34)	.29
Anxiety	0.53 (0.34 - 0.82)	.004
Depression	1.33 (0.82 - 2.16)	.26
Others		
Glaucoma	1.02 (0.60 - 1.74)	.93
Anemia	4.33 (3.25 - 5.78)	<.001

[a]ADPKD: autosomal dominant polycystic kidney disease.
[b]aHR: adjusted hazard ratio.
[c]Not applicable.
[d]CCI: Charlson Comorbidity Index.
[e]AF: atrial fibrillation.
[f]CHF: congestive heart failure.
[g]PVD: peripheral vascular disease.
[h]UTI: urinary tract infection.
[i]AKI: acute kidney injury.
[j]COPD: chronic obstructive pulmonary disease.
[k]DM: diabetes mellitus.

Feature Importance Across ML Models

Figure 2 presents the relative importance of variables in predicting dialysis risk, as determined by the 6 ML models. A consensus emerged across the different algorithms, highlighting several key predictors. The top-ranked predictors spanned comorbidity burden and medication patterns, including anticoagulant use, hypertension, CCB use, higher CCI category, loop diuretics, sodium bicarbonate, anemia, and ACEI or ARB use, together with age and sex. Additional contributors included cardiorespiratory comorbidities, UTI, gout, and neuropsychiatric conditions and related medications.

XSL•FO
RenderX

Figure 2. Overall consensus feature ranking aggregated across 6 machine learning algorithms. Predictors were ranked within each algorithm using the model's native importance criterion; standardized ranks were then aggregated to generate an overall consensus ranking. ACEI: angiotensin-converting enzyme inhibitor; AF: atrial fibrillation; ARB: angiotensin II receptor blocker; CCB: calcium channel blocker; CCI: Charlson Comorbidity Index; COPD: chronic obstructive pulmonary disease; DM: diabetes mellitus; NSAID: nonsteroidal anti-inflammatory drug; PVD: peripheral vascular disease; UTI: urinary tract infection.

XSL•FO
RenderX

Performance Comparison of ML Models

Table 3 summarizes performance for 6 ML algorithms under 2 feature selection strategies: model-specific feature selection and a consensus feature set derived from rank aggregation. Using model-specific feature selection, XGBoost achieved the highest accuracy (0.9748) with specificity of 1, with an AUC of 0.9436 and a Brier score of 0.0254. Using the consensus feature set, performance improved across models. XGBoost achieved the highest accuracy (0.9832) and F_1-score (0.8), whereas MARS achieved the highest AUC (0.9762) and CatBoost achieved the lowest Brier score (0.0217), indicating strong discrimination with favorable calibration across top-performing models.

Table . The model predictions with six different machine learning algorithms.

Methods	Accuracy	Kappa	Sensitivity	Specificity	AUC[a]	F_1-score	Brier score
Variables selected by its own algorithm							
LGR[b]	0.9034	0.2552	0.4167	0.9292	0.7806	0.303	0.0771
RF[c]	0.9580	0.4790	0.4167	0.9867	0.9338	0.500	0.0352
MARS[d]	0.9664	0.6189	0.5833	0.9867	0.9558	0.6364	0.0260
CART[e]	0.9202	0.3037	0.4167	0.9469	0.9023	0.3448	0.0588
XGBoost[f]	0.9748	0.6551	0.5000	1	0.9436	0.6667	0.0254
CatBoost	0.9622	0.5521	0.5000	0.9867	0.9672	0.5714	0.0245
Variables selected based on the average results from six ML algorithms.							
LGR	0.9496	0.4735	0.5000	0.9735	0.9288	0.5000	0.0557
RF	0.9706	0.6174	0.5000	0.9956	0.9611	0.6316	0.0302
MARS	0.9790	0.7512	0.6667	0.9956	0.9762	0.7619	0.0232
CART	0.9454	0.4899	0.5833	0.9646	0.9399	0.5185	0.060
XGBoost	0.9832	0.7916	0.6667	1	0.9546	0.8000	0.0219
CatBoost	0.9790	0.7512	0.6567	0.9956	0.9676	0.7619	0.0217

[a]AUC: area under the curve.

[b]LGR: logistic regression.

[c]RF: random forest.

[d]MARS: multivariate adaptive regression splines.

[e]CART: classification and regression trees.

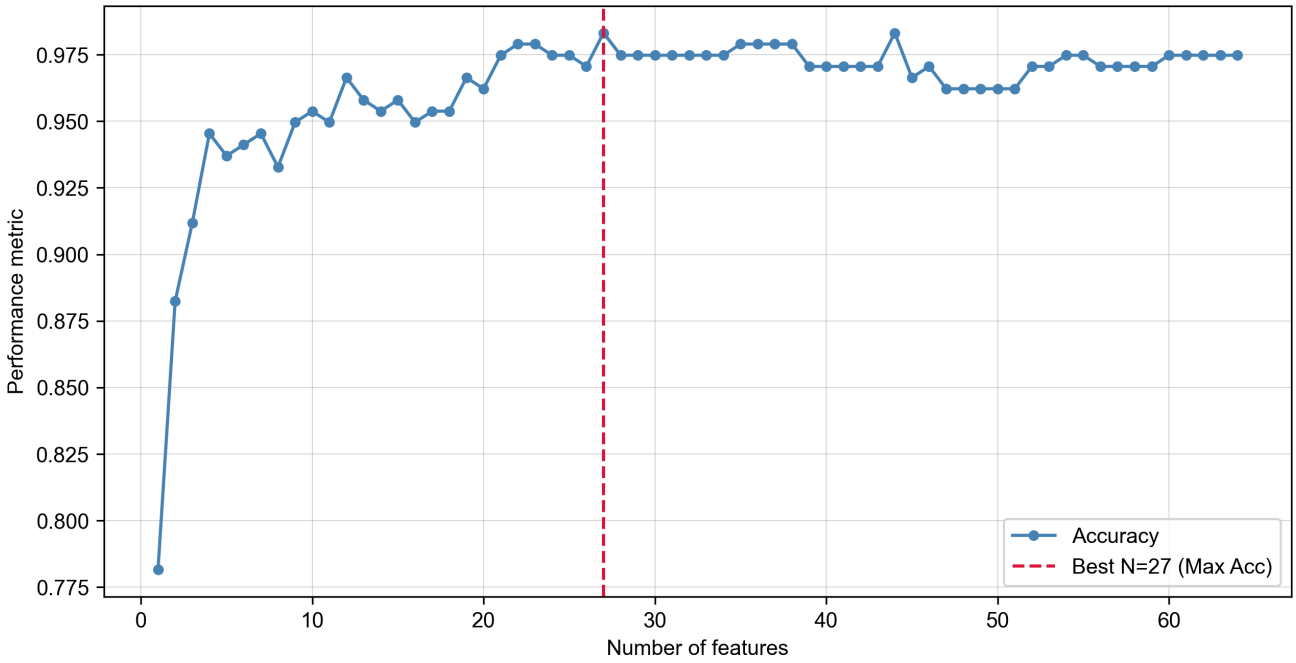
[f]XGBoost: extreme gradient boosting.

Optimization of the XGBoost Model via Feature Selection

Figure 3 shows the change in test set accuracy as top-ranked predictors from the consensus ranking were sequentially added to the XGBoost model. Accuracy increased rapidly as the highest-ranked predictors were included and reached its maximum with 27 predictors. This feature count was selected as the optimal balance between parsimony and predictive performance, and the 27-feature set was used for the final XGBoost model.

XSL•FO
RenderX

Figure 3. Optimization of extreme gradient boosting model performance by sequential feature addition. The plot shows the change in model accuracy (y-axis) as the number of top-ranked predictor variables (x-axis). The dashed vertical line indicates the point at which maximum accuracy was achieved with 27 variables.



Key Predictors in the Final XGBoost Model

The final XGBoost model used 27 baseline predictors selected during feature optimization (Table 4). This set included demographic and comorbidity burden measures (age, sex, and CCI category), major clinical conditions linked to renal risk or frailty (anemia, hypertension, ischemic heart disease, gout, pneumonia, COPD, UTI, anxiety, and depression), and multiple medication exposures. Key medication predictors included antihypertensive classes (ACEI or ARB, CCB, beta-blockers, and alpha-blockers), agents reflecting advanced renal management or metabolic status (loop diuretics, sodium bicarbonate, potassium-sparing diuretics, and rapid-acting insulin), and medications that may relate to underlying diseases (anticoagulants, antiplatelet agents, NSAIDs, sedative hypnotics, tranexamic acid, febuxostat, and statins). It is crucial to note that these medication-related predictors should be interpreted as proxies for underlying disease severity and complexity (ie, confounding by indication), rather than as direct causal risk factors.

XSL•FO
RenderX

Table . The variables selected for the best accuracy result in XGBoost. The rank is the average ranking from the average results of LGR[a], RF[b], MARS[c], CART[d], XGBoost[e], and CatBoost.

Rank	Variable
1	Age
2	CCI[f] scores
3	Anticoagulants
4	Loops diuretics
5	Alpha-blockers
6	NSAID[g]
7	Sex
8	CCB[h]
9	Sedative hypnotics
10	Anemia
11	ACEI[i]/ARB[j]
12	Hypertension
13	Gout
14	Tranexamic acid
15	Febuxostat
16	Ischemic heart disease
17	Statin
18	Antiplatelet agents
19	Beta-blockers
20	Pneumonia
21	Sodium bicarbonate
22	Potassium-sparing diuretics
23	Rapid-acting insulins
24	Anxiety
25	COPD[k]
26	Depression
27	UTI[l]

[a]LGR: logistic regression.
[b]RF: random forest.
[c]MARS: multivariate adaptive regression splines.
[d]CART: classification and regression trees.
[e]XGBoost: extreme gradient boosting.
[f]CCI: Charlson Comorbidity Index.
[g]NSAID: nonsteroidal anti-inflammatory drug.
[h]CCB: calcium channel blocker.
[i]ACEI: angiotensin-converting enzyme inhibitors.
[j]ARB: angiotensin receptor blocker.
[k]COPD: chronic obstructive pulmonary disease.
[l]UTI: urinary tract infection.

XSL•FO
RenderX

Discussion

Principal Findings

This retrospective cohort study used data from the NHIRD to predict ESRD progression in ADPKD using 6 ML algorithms, with ESRD operationalized as initiation of maintenance dialysis. On the held-out temporal test set, XGBoost demonstrated the best overall performance after feature optimization (accuracy 0.9832; F_1 score 0.8). The final optimized XGBoost model was derived by sequentially adding top-ranked predictors and achieved peak accuracy with 27 variables. The most informative predictors reflected underlying disease severity and care intensity, including age, CCI, anemia, hypertension, ischemic heart disease, gout, pneumonia, COPD, UTI, anxiety, and depression, as well as medication exposures such as anticoagulants, loop diuretics, ACEI/ARB, CCB, and febuxostat.

Interpretation of Major Clinical Risk Factors

Hypertension is a common early manifestation of ADPKD, often developing before renal function declines. It affects up to 86.6% of patients with ADPKD, frequently appearing around 30 years of age, with up to 20% of them diagnosed before the age of 20 [24,25]. The underlying mechanisms involve intrarenal ischemia, renin-angiotensin-aldosterone system (RAAS) activation, and endothelial dysfunction. Progressive cyst expansion leads to renal hypoperfusion and sustained RAAS stimulation, further exacerbating hypertension and accelerating kidney damage [26,27]. This mechanistic framework is consistent with our results, where hypertension and cardiovascular comorbidity markers, including ischemic heart disease and related medication patterns, contributed meaningfully to dialysis risk prediction.

Age and overall comorbidity burden were among the highest-ranked predictors in both the consensus feature ranking and the final XGBoost feature set. This pattern suggests that administrative claims-based models capture the cumulative burden of systemic illness, frailty, and health care utilization that often accompanies advanced kidney disease trajectories [28,29]. In our cohort, ischemic heart disease and cardiovascular medication proxies (eg, anticoagulants, antiplatelet agents, beta-blockers, and loop diuretics) were repeatedly selected, indicating that cardiovascular disease severity and its treatment intensity provide informative signals of near-term dialysis risk [30,31].

In our cohort, anemia showed the strongest association with dialysis initiation in the adjusted Cox model and was consistently selected by the ML algorithms. A study conducted in Japan found that, despite higher hemoglobin levels in ADPKD than in other CKD types, anemia remains a key predictor of renal disease progression. In nondialysis patients with ADPKD, lower hemoglobin levels were associated with faster kidney function decline, with men with hemoglobin levels less than 12 g/dL and women with hemoglobin levels less than 11 g/dL at higher risk [32]. Together, these findings support anemia as a clinically accessible marker of advanced disease and progression risk in ADPKD.

AKI and infection-related diagnoses such as pneumonia and UTI also contributed to the prediction. These events may accelerate loss of residual kidney function through hemodynamic instability, inflammatory injury, or nephrotoxic exposures, and they may also serve as markers of vulnerability and more frequent clinical encounters. In parallel, medication predictors such as sodium bicarbonate and NSAIDs plausibly reflect clinically recognized metabolic acidosis and nephrotoxic exposure that often accompany advanced CKD and precede dialysis initiation [33].

Medication Predictors as Markers of Disease Severity

As highlighted earlier, a critical aspect of interpreting our findings is the potential for confounding by indication, particularly for medication-related predictors. In a retrospective cohort study, the use of a specific medication is often a marker of underlying disease severity rather than a cause of the outcome itself. Therefore, while these medications emerged as strong predictors in our model, they should be viewed primarily as indicators of a higher underlying risk profile, and our findings do not imply a causal relationship or recommend changes in prescribing practices. Consistent with prior work using administrative claims for kidney failure prediction, these medication patterns should be interpreted primarily as proxies for clinical complexity rather than as causal determinants of dialysis initiation [31].

Renin-angiotensin system inhibitors have been established as the first-line treatment due to their benefits in hypertension control [34]. In our model, the selection of additional antihypertensive classes (eg, alpha-blockers and CCB) and loop diuretics likely reflects difficult-to-control blood pressure, volume management needs, or concomitant cardiovascular disease rather than direct causal effects on renal decline. Similarly, anticoagulant and antiplatelet use may act as proxies for atrial fibrillation, vascular disease, or prior cardiovascular events, conditions that commonly cluster with advanced CKD and a higher risk of dialysis initiation.

Tranexamic acid was also selected among the final XGBoost predictors. In ADPKD, clinically significant gross hematuria is a common and clinically relevant event, and tranexamic acid has been used in small series for severe hematuria, although evidence remains limited [2,35]. Therefore, in claims-based modeling, tranexamic acid likely functions as a proxy for clinically significant hematuria or bleeding-related care that clusters with more advanced disease trajectories, rather than as a causal determinant of dialysis initiation.

Hyperuricemia is increasingly implicated in CKD progression, and our model identified gout and febuxostat as informative predictors of dialysis. This finding does not suggest a causal link but rather reflects prescribing patterns that align with both international guidelines and local reimbursement policies. Major clinical practice guidelines for ADPKD recommend allopurinol as the first-line urate-lowering therapy [2]. This standard is strictly reflected in Taiwan's National Health Insurance system, where reimbursement for the second-line agent, febuxostat, was restricted during our study period. These restrictions likely limited its use primarily to patients with substantial preexisting conditions, most notably advanced chronic kidney disease

XSL•FO
RenderX

(defined as eGFR <45 mL/min/1.73 m^2) or other complications such as urate nephrolithiasis or severe tophaceous gout. Consequently, the predictive signal from febuxostat in our model arises because it serves as a powerful proxy, identifying a patient subgroup with higher baseline renal risk and greater disease complexity. Further prospective studies are needed to determine if urate-lowering therapy itself modifies the disease course in ADPKD.

Statins, primarily known for their cholesterol-lowering effects, also possess anti-inflammatory and antioxidant properties that could be beneficial in PKD. Chronic inflammation and oxidative stress play roles in the pathophysiology of PKD, contributing to cyst growth and renal function decline. Evidence regarding the efficacy of statins in PKD is equivocal. Some cohort studies and animal research suggest that statins may slow the decline in renal function in patients with PKD by reducing inflammation and oxidative stress [9]. However, findings from definitive clinical trials in PKD populations, which are scarce, remain inconclusive.

Comparison With Existing Risk Stratification Tools

Our ML approach should be compared with existing clinical risk tools for ADPKD, such as the Mayo Imaging Classification and the PROPKD score. The primary advantage of our model lies in its broad accessibility. By leveraging nationwide administrative data, it bypasses the need for resource-intensive assessments, such as MRI for TKV measurement or genetic sequencing, making it a scalable and cost-effective tool for initial risk stratification across a large population. The main disadvantage, however, is the lack of direct pathophysiological data. Without TKV or genetic data, our model relies on surrogate markers of disease severity and cannot offer the same level of mechanistic insight or precision as the gold-standard tools. Therefore, our model should not be viewed as a replacement for these established methods but rather as a complementary instrument. Its optimal clinical role would be to serve as a first-line screening tool to efficiently identify high-risk patients who could then be prioritized for more definitive prognostic evaluations like imaging or genetic testing.

Strengths and Limitations

To our knowledge, this study represents the first nationwide ML-based analysis of dialysis risk in ADPKD using administrative data. It provides a scalable framework for population-level risk stratification. The use of the NHIRD, a large population-based dataset with long-term follow-up, ensures high generalizability to the nationwide population of Taiwan.

Several limitations of this study should be acknowledged. First, its retrospective design precludes the establishment of causality. Second, and most importantly, our study is constrained by the nature of the NHIRD, which lacks the gold-standard prognostic markers for ADPKD including genetic data (PKD1/PKD2 status) and imaging biomarkers like TKV, and key laboratory values such as eGFR or proteinuria. Without direct access to TKV or genetic data, our model inherently relies on surrogate markers such as comorbidity burden and medication patterns to infer underlying risk. Despite the use of a washout period and validation via the Catastrophic Illness Patient Registry, truly incident ADPKD cases cannot be fully guaranteed in claims-based data. We therefore wish to emphasize that our model is proposed as a complementary tool for broad-based risk stratification, not as a replacement for comprehensive clinical evaluation. Third, because observation windows vary by index year, the binary labeling framework may under-ascertain events among patients with shorter follow-up and can bias absolute risk interpretation. In addition, calibration curves were not presented, limiting assessment of absolute risk calibration. Fourth, because deaths occurring before ESRD were labeled as nonevents in the binary ML framework, patients with high competing mortality may be misclassified as having low renal risk. Fifth, the measurement of medication exposure, quantified by prescription days, is a proxy that does not capture actual patient adherence, dosage, or treatment adjustments. This limits our ability to make definitive conclusions about the role of specific pharmacological agents. Finally, our model was developed using a Taiwanese population, and its performance and the relative importance of predictors may vary across different ethnic and geographic groups, which may limit generalizability.

Future Work

Future studies should prioritize prospective validation in diverse external cohorts. Methodologically, fixed-horizon prediction and time-to-event approaches, including survival ML and competing-risk or multistate frameworks, should be evaluated to better align modeling with clinical trajectories. External validation datasets that include laboratory, genetic, and imaging biomarkers (eg, eGFR, proteinuria, PKD1/PKD2 status, and TKV) will enable integrated, multitiered risk stratification and more complete assessment of calibration and clinical utility. Moreover, instance-level explainability methods (eg, SHAP) can be explored to support clinical interpretability.

Conclusions

This study developed and evaluated 6 ML models to classify dialysis initiation risk in a nationwide ADPKD cohort using administrative claims data and a prespecified temporal validation design. On the held-out temporal test set, XGBoost showed the best overall performance. The most informative predictors primarily reflected disease severity and care intensity captured in claims data, including age, comorbidity burden, anemia, and hypertension, whereas medication variables should be interpreted as noncausal proxies of clinical complexity.

Acknowledgments

No generative artificial intelligence tools were used to generate the manuscript text, analyses, figures, tables, or multimedia appendix. All study design decisions, analyses, interpretations, and final wording were produced and verified by the authors.

XSL•FO
RenderX

Funding

This study was supported by the Shin-Kong Wu Ho-Su Memorial Hospital Research Foundation (2022SKHADR035) and The National Science and Technology Council, Taiwan (NSTC 112-2221-E-030 -009 -MY3). The funding source had no role in the study design, data collection, analysis, or interpretation of the results.

Data Availability

The dataset used in this study belongs to the Taiwan Ministry of Health and Welfare (MOHW). Due to data protection regulations, the dataset cannot be shared by the corresponding author upon request. However, researchers interested in accessing this dataset may apply for access through the website of the National Health Informatics Project of the MOHW [36].

Authors' Contributions

Conceptualization: MC, MHT, YWF
Data curation: MHT, HHL
Formal analysis: CHC, MC, YCH, CL
Funding acquisition: MC, YWF
Investigation: MHT, YWF
Methodology: CHC, MC, YCH, HHL, BCS, CL
Resources: MC, YCH
Software: MC, YCH
Supervision: MC, YWF
Validation: MHT, YCH, BCS
Writing – original draft: CHC, MC
Writing – review and editing: MHT, BCS, CL, YWF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods and tables, including model implementation details for the six machine learning algorithms, ICD-9-CM/ICD-10-CM codes for comorbidities (Table S1), ATC codes and drug lists (Table S2), and tuned hyperparameters with software packages and versions (Table S3).
[DOCX File, 38 KB - medinform_v14i1e80343_app1.docx]

References

1. Harris PC, Torres VE. Polycystic kidney disease. Annu Rev Med 2009;60:321-337. [doi: 10.1146/annurev.med.60.101707.125712] [Medline: 18947299]
2. Kidney Disease: Improving Global Outcomes (KDIGO) ADPKD Work Group. KDIGO 2025 clinical practice guideline for the evaluation, management, and treatment of autosomal dominant polycystic kidney disease (ADPKD). Kidney Int 2025 Feb;107(2S):S1-S239. [doi: 10.1016/j.kint.2024.07.009] [Medline: 39848759]
3. Jafar TH, Stark PC, Schmid CH, et al. Progression of chronic kidney disease: the role of blood pressure control, proteinuria, and angiotensin-converting enzyme inhibition: a patient-level meta-analysis. Ann Intern Med 2003 Aug 19;139(4):244-252. [doi: 10.7326/0003-4819-139-4-200308190-00006] [Medline: 12965979]
4. Benck U, Krüger B, Schmitt WH. Blood pressure in early autosomal dominant polycystic kidney disease. N Engl J Med 2015 Mar 5;372(10):975-977. [doi: 10.1056/NEJMc1500332] [Medline: 25738679]
5. Kramers BJ, Koorevaar IW, Drenth JPH, et al. Salt, but not protein intake, is associated with accelerated disease progression in autosomal dominant polycystic kidney disease. Kidney Int 2020 Oct;98(4):989-998. [doi: 10.1016/j.kint.2020.04.053] [Medline: 32534051]
6. Nowak KL, Steele C, Gitomer B, Wang W, Ouyang J, Chonchol MB. Overweight and obesity and progression of ADPKD. Clin J Am Soc Nephrol 2021 Jun;16(6):908-915. [doi: 10.2215/CJN.16871020] [Medline: 34117082]
7. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. N Engl J Med 2012 Dec 20;367(25):2407-2418. [doi: 10.1056/NEJMoa1205511] [Medline: 23121377]
8. Takiar V, Nishio S, Seo-Mayer P, et al. Activating AMP-activated protein kinase (AMPK) slows renal cystogenesis. Proc Natl Acad Sci U S A 2011 Feb 8;108(6):2462-2467. [doi: 10.1073/pnas.1011498108] [Medline: 21262823]
9. Fassett RG, Robertson IK, Ball MJ, Geraghty DP, Coombes JS. Effect of atorvastatin on kidney function in chronic kidney disease: a randomised double-blind placebo-controlled trial. Atherosclerosis 2010 Nov;213(1):218-224. [doi: 10.1016/j.atherosclerosis.2010.07.053] [Medline: 20810109]

XSL•FO
RenderX

10. Irazabal MV, Rangel LJ, Bergstralh EJ, et al. Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials. J Am Soc Nephrol 2015 Jan;26(1):160-172. [doi: 10.1681/ASN.2013101138] [Medline: 24904092]
11. Cornec-Le Gall E, Audrézet MP, Rousseau A, et al. The PROPKD score: a new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. J Am Soc Nephrol 2016 Mar;27(3):942-951. [doi: 10.1681/ASN.2015010016] [Medline: 26150605]
12. Niel O, Boussard C, Bastard P. Artificial intelligence can predict GFR decline during the course of ADPKD. Am J Kidney Dis 2018 Jun;71(6):911-912. [doi: 10.1053/j.ajkd.2018.01.051] [Medline: 29609979]
13. Kim Y, Bu S, Tao C, Bae KT, HALT Polycystic Kidney Disease Study Group. Deep learning-based automated imaging classification of ADPKD. Kidney Int Rep 2024 Jun;9(6):1802-1809. [doi: 10.1016/j.ekir.2024.04.002] [Medline: 38899202]
14. Sharma K, Rupprecht C, Caroli A, et al. Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. Sci Rep 2017 May 17;7(1):2049. [doi: 10.1038/s41598-017-01779-0] [Medline: 28515418]
15. Taylor J, Thomas R, Metherall P, et al. An artificial intelligence generated automated algorithm to measure total kidney volume in ADPKD. Kidney Int Rep 2024 Feb;9(2):249-256. [doi: 10.1016/j.ekir.2023.10.029] [Medline: 38344736]
16. Lin LY, Warren-Gash C, Smeeth L, Chen PC. Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiol Health 2018;40:e2018062. [doi: 10.4178/epih.e2018062] [Medline: 30727703]
17. Kalatharan V, McArthur E, Nash DM, et al. Diagnostic accuracy of administrative codes for autosomal dominant polycystic kidney disease in clinic patients with cystic kidney disease. Clin Kidney J 2021 Feb;14(2):612-616. [doi: 10.1093/ckj/sfz184] [Medline: 33623686]
18. Nan-Ping Y, Yi-Hui L, Chi-Yu C, et al. Comparisons of medical utilizations and categorical diagnoses of emergency visits between the elderly with catastrophic illness certificates and those without. BMC Health Serv Res 2013 Apr 26;13:152. [doi: 10.1186/1472-6963-13-152] [Medline: 23622501]
19. Wang SW, Li LC, Su CH, Yang YH, Hsu TW, Hsu CN. Association of statin and its lipophilicity with cardiovascular events in patients receiving chronic dialysis. Clin Pharmacol Ther 2020 Jun;107(6):1312-1324. [doi: 10.1002/cpt.1722] [Medline: 31715017]
20. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Springer; 2009. [doi: 10.1007/978-0-387-84858-7]
21. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Presented at: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); Dec 3-8, 2018 URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf [accessed 2026-02-24]
22. Huang YC, Ho CW, Chou WR, Chen M. A framework to predict second primary lung cancer patients by using ensemble models. Ann Oper Res 2025 May;348(1):373-397. [doi: 10.1007/s10479-023-05691-x]
23. Chen KY, Huang YC, Liu CK, Li SJ, Chen M. Machine learning-driven prediction of medical expenses in triple-vessel PCI patients using feature selection. BMC Health Serv Res 2025 Jan 20;25(1):105. [doi: 10.1186/s12913-025-12218-6] [Medline: 39833782]
24. Rahbari-Oskoui F, Williams O, Chapman A. Mechanisms and management of hypertension in autosomal dominant polycystic kidney disease. Nephrol Dial Transplant 2014 Dec;29(12):2194-2201. [doi: 10.1093/ndt/gft513] [Medline: 24463189]
25. Helal I, Reed B, Mettler P, et al. Prevalence of cardiovascular events in patients with autosomal dominant polycystic kidney disease. Am J Nephrol 2012;36(4):362-370. [doi: 10.1159/000343281] [Medline: 23038404]
26. Chapman AB, Stepniakowski K, Rahbari-Oskoui F. Hypertension in autosomal dominant polycystic kidney disease. Adv Chronic Kidney Dis 2010 Mar;17(2):153-163. [doi: 10.1053/j.ackd.2010.01.001] [Medline: 20219618]
27. Schrier RW, Abebe KZ, Perrone RD, et al. Blood pressure in early autosomal dominant polycystic kidney disease. N Engl J Med 2014 Dec 11;371(24):2255-2266. [doi: 10.1056/NEJMoa1402685] [Medline: 25399733]
28. Kim DH, Schneeweiss S, Glynn RJ, Lipsitz LA, Rockwood K, Avorn J. Measuring frailty in Medicare data: development and validation of a claims-based frailty index. J Gerontol A Biol Sci Med Sci 2018 Jun 14;73(7):980-987. [doi: 10.1093/gerona/glx229] [Medline: 29244057]
29. Walker H, Day S, Grant CH, et al. Representation of multimorbidity and frailty in the development and validation of kidney failure prognostic prediction models: a systematic review. BMC Med 2024 Oct 11;22(1):452. [doi: 10.1186/s12916-024-03649-9] [Medline: 39394084]
30. Chuah A, Walters G, Christiadi D, et al. Machine learning improves upon clinicians' prediction of end stage kidney disease. Front Med 2022;9. [doi: 10.3389/fmed.2022.837232]
31. Li Y, Al-Sayouri S, Padman R. Towards interpretable end-stage renal disease (ESRD) prediction: utilizing administrative claims data with explainable AI techniques. AMIA Annu Symp Proc 2024;2024:664-673. [Medline: 40417492]
32. Ushio Y, Kataoka H, Sato M, et al. Association between anemia and renal prognosis in autosomal dominant polycystic kidney disease: a retrospective study. Clin Exp Nephrol 2020 Jun;24(6):500-508. [doi: 10.1007/s10157-020-01856-1] [Medline: 32036465]
33. Goraya N, Wesson DE. Clinical evidence that treatment of metabolic acidosis slows the progression of chronic kidney disease. Curr Opin Nephrol Hypertens 2019 May;28(3):267-277. [doi: 10.1097/MNH.0000000000000491] [Medline: 30681417]

XSL•FO
RenderX

34. Torres VE, Abebe KZ, Chapman AB, et al. Angiotensin blockade in late autosomal dominant polycystic kidney disease. N Engl J Med 2014 Dec 11;371(24):2267-2276. [doi: 10.1056/NEJMoa1402686] [Medline: 25399731]
35. Yao Q, Wu M, Zhou J, et al. Treatment of persistent gross hematuria with tranexamic acid in autosomal dominant polycystic kidney disease. Kidney Blood Press Res 2017;42(1):156-164. [doi: 10.1159/000474961] [Medline: 28395294]
36. National health informatics project of the MOHW. Ministry of Health and Welfare. URL: https://dep.mohw.gov.tw/DOS/mp-113.html [accessed 2026-02-27]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ADPKD: autosomal dominant polycystic kidney disease
AI: artificial intelligence
AKI: acute kidney injury
ARB: angiotensin II receptor blocker
AUC: area under the curve
CCB: calcium channel blocker
CCI: Charlson Comorbidity Index
COPD: chronic obstructive pulmonary disease
ESRD: end-stage renal disease
HR: hazard ratio
ML: machine learning
MRI: magnetic resonance imaging
NHIRD: National Health Insurance Research Database
NSAID: nonsteroidal anti-inflammatory drug
PKD: polycystic kidney disease
TKV: total kidney volume
UTI: urinary tract infection
XGBoost: extreme gradient boosting

Edited by A Coristine, A Benis; submitted 09.Jul.2025; peer-reviewed by SB Duan, SS Chai; revised version received 14.Feb.2026; accepted 17.Feb.2026; published 16.Mar.2026.

Please cite as:
Chang CH, Chen M, Tsai MH, Huang YC, Liou HH, Shia BC, Liang C, Fang YW
Machine Learning Prediction of Progression to Dialysis in Patients With Polycystic Kidney Disease: Population-Based Retrospective Cohort Study
JMIR Med Inform 2026;14:e80343
URL: https://medinform.jmir.org/2026/1/e80343
doi:10.2196/80343

© Cheng-Hao Chang, Mingchih Chen, Ming-Hsien Tsai, Yen-Chun Huang, Hung-Hsiang Liou, Ben-Chang Shia, Chingying Liang, Yu-Wei Fang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mild Cognitive Impairment Detection System Based on Unstructured Spontaneous Speech: Longitudinal Dual-Modal Framework

Yu-Shan Liao[1], MSc; Thiri Wai[2], MSc; Ting-Yun Liao[1], MSc; Ho-Ling Chang[1], MSc; Yu-Ling Chang[3], PhD; Li-Chen Fu[2,4], MSc, PhD

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

[2]Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec 4, Roosevelt Rd, Taipei, Taiwan

[3]Department of Psychology, National Taiwan University, Taipei, Taiwan

[4]Center of Artificial intelligence Research, University of Tsukuba, Tsukuba, Ibaraki, Japan

Corresponding Author:
Li-Chen Fu, MSc, PhD
Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec 4, Roosevelt Rd, Taipei, Taiwan

Abstract

Background: In recent years, the incidence of cognitive diseases has also risen with the significant increase in population aging. Among these diseases, Alzheimer disease constitutes a substantial proportion, placing a high-cost burden on health care systems. To give early treatment and slow the progression of patient deterioration, it is crucial to diagnose mild cognitive impairment (MCI), a transitional stage.

Objective: In this study, we use autobiographical memory (AM) test speech data to establish a dual-modal longitudinal cognitive detection system for MCI. The AM test is a psychological assessment method that evaluates the cognitive status of subjects as they freely narrate important life experiences.

Methods: Identifying hidden disease-related information in unstructured, spontaneous speech is more difficult than in structured speech. To improve this process, we use both speech and text data, which provide more clues about a person's cognitive state. In addition, to track how cognition changes over time in spontaneous speech, we introduce an aging trajectory module. This module uses local and global alignment loss functions to better learn time-related features by aligning cognitive changes across different time points.

Results: In our experiments on the Chinese dataset, the longitudinal model incorporating the aging trajectory module achieved area under the receiver operating characteristic curve of 0.85 and 0.89 on 2 datasets, respectively, showing significant improvement over cross-sectional, single time point models. We also conducted ablation studies to verify the necessity of the proposed aging trajectory module. To confirm that the model not only applies to AM test data, we used part of the model to evaluate the performance on the ADReSSo dataset, a single time point semistructured data for validation, with results showing an accuracy exceeding 0.88.

Conclusions: This study presents a noninvasive and scalable approach for early MCI detection by leveraging AM speech data across multiple time points. Through dual-modal analysis and the introduction of an aging trajectory module, our system effectively captures cognitive decline trends over time. Experimental results demonstrate the method's robustness and generalizability, highlighting its potential for real-world, long-term cognitive monitoring.

(JMIR Med Inform 2026;14:e80883) doi:10.2196/80883

KEYWORDS

mild cognitive impairment; longitudinal speech analysis; multimodal fusion; Alzheimer disease detection; deep learning; autobiographical memory test

Introduction

Alzheimer disease (AD) is a neurodegenerative disorder often associated with memory decline, language impairment, and a diminishing capacity to perform daily activities independently. However, AD remains incurable. Current research indicates that approximately 47 million people were diagnosed with dementia in 2015 [1], and it is expected that the number will triple by 2050 [2]. Such forecasts pose a significant burden on both patients' families and health care systems.

XSL•FO
RenderX

Within the spectrum of cognitive health, an intermediary stage exists between a state of normalcy and the onset of AD, denoted as mild cognitive impairment (MCI). Patients with MCI do not exhibit significant cognitive symptoms compared with those with AD. Those patients are still capable of performing daily activities independently. According to the study by Petersen et al [3], an estimated 16% of individuals with MCI progress to dementia within a year. Therefore, early identification of MCI holds significant value in instigating interventions aimed at impeding the advancement of dementia. However, the detection of MCI presents a notable challenge due to the subtle manifestation of its symptoms.

While biomarkers have demonstrated sensitivity in identifying MCI, such as positron emission tomography and magnetic resonance imaging [4,5], many of them necessitate invasive procedures. The associated costs of data collection for such methods are relatively high. Conversely, leveraging voice data for cognitive assessment offers a simpler and more suitable approach for long-term monitoring of cognitive status. Our approach relies on using autobiographical memory (AM) test data as the primary input [6], wherein interviewers engage participants in a series of questions and answers regarding their life experiences. This method, resembling a casual conversation, is more approachable than other psychological test methods because it reduces the testing stress experienced by participants.

Longitudinal analysis for MCI detection based on speech data remains relatively unexplored compared with other data types, such as magnetic resonance imaging, resulting in a scarcity of research in this area. The research proposed by Al-Hameed et al [7] focuses on extracting solely acoustic features from speech data collected across multiple visits from DementiaBank, which primarily pertains to speech associated with picture description tasks. They aim to predict participants' Mini-Mental State Examination scores and cognitive labels. Following feature selection, they used support vector machine or stochastic gradient descent for prediction. Integrating data from 3 visits achieved a 90% accuracy in distinguishing between healthy control (HC) and MCI in a binary classification task. However, due to label imbalance in the dataset, accuracy as a metric may not be entirely equitable. This study demonstrates that incorporating features from previous visits enhances prediction accuracy for subsequent visits.

Using the same dataset, another survey by Ammar et al [8] concentrates on extracting a wide range of linguistic features related to fluency and richness for Mini-Mental State Examination score prediction. However, unlike the former study, they do not leverage the correlation between previous visits. In the research proposed by Laguarta and Subirana [9], 16 physiological biomarkers were identified from speech data, represented in a personalized subject feature map. This feature map is capable of illustrating and monitoring cognitive changes over time. Achieving 93.6% accuracy in the ADReSS datasets, this study is the only one among them that visually depicts cognitive decline.

Yamada et al [10] used phone calls for long-term participant tracking, manually transcribing the conversations to extract linguistic features. They focused on extracting topic and word repetition features across different calls. Their model achieved an area under the receiver operating characteristic curve (AUROC) of 0.91 in classifying between AD and HC. In our previous work proposed by Chang et al [11], the author first extracted traditional acoustic feature sets, extended Geneva Minimalistic Acoustic Parameter Set [12], from AM audio files collected across multiple visits. The author then used TabNet [13] to learn the cognitive information embedded within these feature sets. To further enhance the model, an aging alignment loss was applied, encouraging cognitive representations in the latent space to be closer to each other. The approach achieved an accuracy of 0.66 and an F_1-score of 0.64.

In this work, we use AM data as the primary input for our model. We analyze the AM data from both linguistic and acoustic perspectives and further establish an aging trajectory module (ATM). By introducing the ATM, which captures cognitive decline across different time points through direction embeddings, our framework provides more representative modeling of longitudinal changes. The key contributions of this work are summarized as follows:

1. Unlike most studies focusing on single time point analysis, this work introduces a longitudinal framework using AM speech data to track cognitive changes over time.
2. The system jointly analyzes linguistic and acoustic features from unstructured AM speech, offering a richer representation of cognitive status than unimodal approaches.
3. Proposed ATM, which models cognitive decline by aligning temporal features across visits through a specially designed alignment loss applied to direction embeddings.

Methods

Problem Setting

There are 2 problem settings in this work: longitudinal type and cross-sectional type. The longitudinal problem setting is our main focus. The cross-sectional problem setting evaluates whether our model is general enough to tackle different tasks of detecting cognitive impairment disease detection tasks.

Longitudinal Problem Setting

In our problem setting, we have the dataset D={D1,D2,…,DV}with the corresponding label set L={L1,L2,…,LV}, where V means the total visit times, and D_k means the dataset at the kth visit. Each D_k is expressed as Dk={x1,k, x2,k, …, xN,k}, where N represents the number of the subjects involved in the test, andxi,k={xi,k1, xi,k2, …,xi,kM} represents the M stories provided by subject i at the kth visit; note that the number of M may vary across different visits. The corresponding labels set Lk={l1,k, l2,k, …, lN,k} shows the cognitive status of each subject, with each $l_{i,k}$ indicating whether the subject is an HC or has MCI.

During the data preprocessing, we extract the feature set {xi,km,a, xi,km,t, xi,km,l} from the raw audio xi,km, where xi,km,a, xi,km,t, xi,km,l represents the acoustic feature set, textual data, and linguistic feature set, respectively. Subsequently, the model input is the sequential visit data defined

XSL•FO
RenderX

in equation 1, where xi,1m1∈ xi,1, xi,2m2∈ xi,2, and xi,VmV∈ xi,V.

(0)m2, . . xmV={xm1m1,xm2m2, . . ,xmVmV}

It is worth noting that stories from each visit are randomly paired across visits within each subject. This design ensures that the model learns subject-level temporal progression rather than relying on similarities driven by specific story topics or linguistic content. Random pairing is not an arbitrary source of noise but a deliberate mechanism aligned with our longitudinal modeling objective. Because story content varies substantially across visits and participants, enforcing semantic matching or averaging all embeddings within a visit would bias the model toward topic-dependent features, flatten meaningful variability in linguistic complexity, and increase the risk of overfitting to specific story topics. In contrast, random pairing operates strictly within each subject, preserving the temporal order while exposing the model to multiple sampled transitions across training epochs. This provides a richer distribution of temporal transitions, enabling the framework to learn visit-level cognitive progression directions that remain robust to story content and reflect each subject's distribution of linguistic changes over time.

Formally, our model can be viewed as a mapping function fW, where *W* represents the set of learnable parameters. The function fWprojects the input sequence {xi,1m1,xi,2m2,…,xi,vmv} into the label of the last visit l(i,V). In this study, the maximum number of visits per subject, *V*, is 3, following the available structure of the dataset.

Cross-Sectional Problem Setting

In the cross-sectional problem setting, the primary difference from the longitudinal problem setting is that data collected from the same subject at different interview visits are treated as data from independent subjects. Thus, we have the dataset D={x1,x2, …, xCN} with the corresponding label set L={l1,l2, …, lCN}, where CN is the sum of all subjects from all visits, and x_i represents the raw audio for the subject *i* with the cognitive label l_i. The cognitive label l_i indicates that the subject is either an HC or under the targeted condition (MCI or AD). During the data preprocessing, we extract the feature set {xia,xit,xil} from the raw audio x_i. Our cross-sectional model aims to map the model input {xia,xit,xil} into the cognitive label l_i.

Overview

The system overview is shown in Figure 1. The goal of the system is to detect the cognitive condition of subjects, specifically whether they have MCI, using speech data. Before training, the speech data first undergo data preprocessing, which involves converting the audio files to text using automatic speech recognition (ASR), extracting additional linguistic features, and deep learning–based acoustic feature extraction. These steps generate additional text data and acoustic data, which are then intertwined to form dual-modal data pairs serving as the input for the model training. In addition, to effectively capture the features of aging and enhance the detection of potential MCI, the model input should have speech data for this time as well as points of follow-up data. This philosophy leads to our proposed longitudinal analysis for MCI detection.

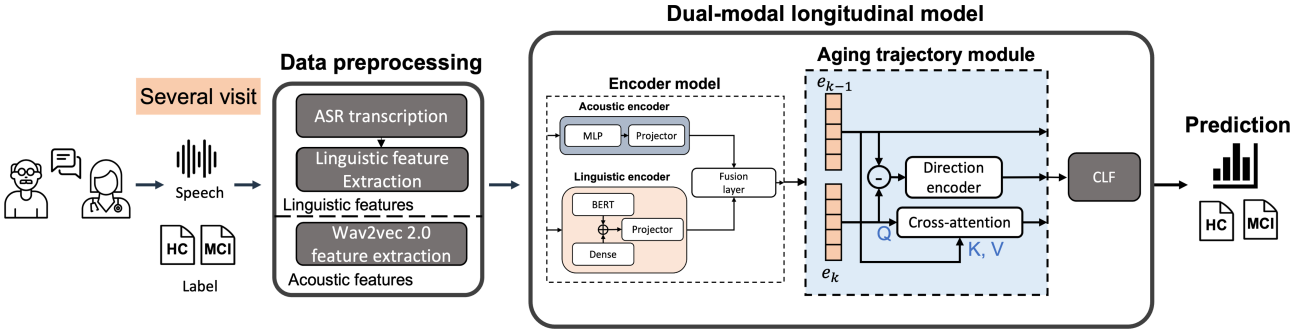


Figure 1. Overall framework of the proposed longitudinal multimodal model. The model consists of two main modules: (1) encoder model, combining a CNN-based acoustic encoder and a BERT-based linguistic encoder to generate fused multimodal embeddings; and (2) aging trajectory module, which models longitudinal changes between consecutive visits using a direction encoder and cross-attention before classification. ASR: automatic speech recognition; BERT: bidirectional encoder representation for transformer; CLF: classifier; CNN: convolutional neural network; HC: healthy control; MCI: mild cognitive impairment.

Data Preprocessing

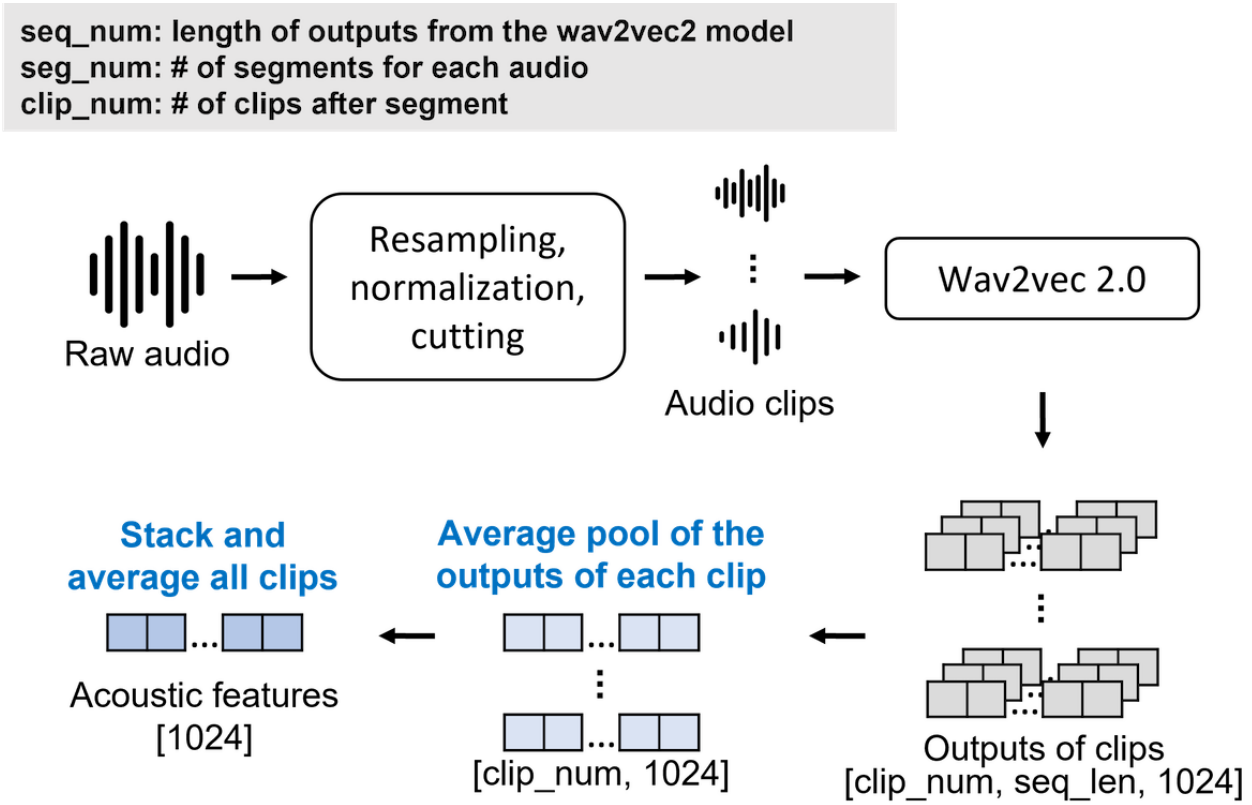
Before the data are input into the model, they need to be preprocessed to ensure data quality and to extract useful features. After data collection is completed, given that the data are based on the entire interview process of each subject during a single visit, we first segment each story recording from the interview. To prepare the audio file of each subject, we segment the audio files based on the story topics provided by the subject, cutting the corresponding recall and probing audio files.

Acoustic Feature Extraction

The preprocessing of audio data is shown in Figure 2. First, the notation xi,km denotes a raw audio waveform, which means the *m*th audio file of the subject *i* collected at the *k*th interview visit undergoes basic processing, including normalization and resampling. Normalization is done using zero-mean unit-variance normalization. Next, the audio is segmented into *t*-second clips with a 5-second overlap. Each audio file can be denoted as xi,km={xi,km,1,xi,km,2,…,xi,km,seg_num}, where seg_num represents the number of clips to which the raw audio file can be divided, and the total number of segments varies for each audio file.

XSL•FO
RenderX

Figure 2. Overview of acoustic preprocessing. Each raw audio recording is divided into multiple fixed-length clips and processed by the wav2vec2.0 feature extractor to obtain frame-level representations. For each segment, the outputs from all clips are stacked and averaged to produce a single clip-level feature. The averaged clip features are then pooled across segments to generate the final visit-level wav2vec2.0 embedding, which is used as input to the acoustic encoder.



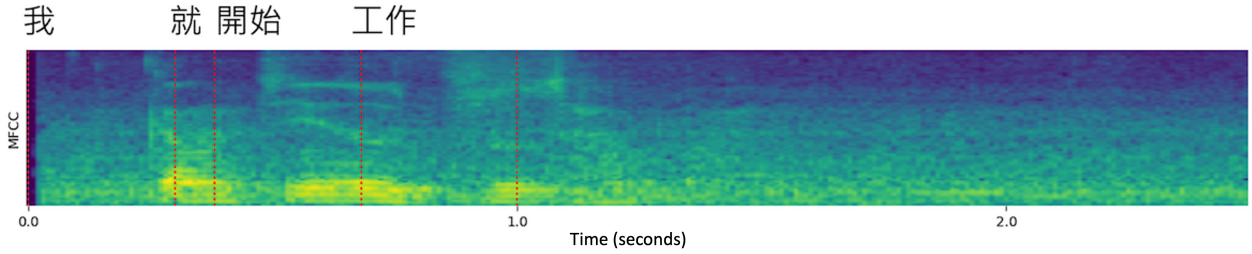
In our research, wav2vec 2.0 model [14] is used as an acoustic feature extractor, and we leverage the pretrained model available on Hugging Face. Using the wav2vec 2.0 model in advance to extract deep acoustic representation features offers the advantage of saving memory space and making the subsequent model structure lighter, thereby increasing training efficiency. Each clip is fed into the wav2vec 2.0 model, producing an output with dimensions [seq_len, 1024], where seq_len is determined by the convolution layer in wav2vec 2.0 and the clip length, which is computed by dividing the clip length and the multiplication of the stride length in the convolution layers. After obtaining the output of each clip, the first average pooling is performed, resulting in a 1024-dimensional acoustic feature for each clip. Then, all clip outputs are stacked together, and a final average pooling is conducted to obtain the final acoustic features xi,km,a of the entire raw audio file; hence, the dimension of xi,km,a is 1024.

Linguistic Feature Extraction

The preprocessing of linguistic data can be divided into 2 stages: generating the transcript and extracting features from the generated text. Since not all audio files in the dataset have manual transcripts, we use ASR to generate the transcripts for our research. We use the open-source tool kit "whisper-timestamped," which provides both timestamps and the corresponding text, as illustrated in Figure 3. Previous research [15-17] has confirmed that pauses are an indicator in AD-related diseases, where patients with MCI or AD tend to exhibit more frequent and longer pauses. In our study, we classify pauses based on the interval duration between sentences into 3 types: short pause (duration <0.5 seconds), medium pause (duration between 0.5 and 2 seconds), and long pause (duration >2 seconds) [18]. In addition to pauses, we calculate the mean length of utterance, which is defined as the average number of words spoken in a sentence. In this step, we will generate transcripts with pause information and 4 numerical linguistic features. Besides pause information, we provide additional syntax information to enrich the linguistic features. Previous research [19] has shown that the sentence structures used by individuals with AD and HC are different. Therefore, we use part-of-speech (POS) tagging in this study. After generating the text, we use the open-source tool kit "ckipnlp" for word parsing and POS tagging. The text segments are then categorized into 5 POS types: noun,

XSL•FO
RenderX

Figure 3. Example of how the result of whisper-timestamped, where the timestamp of the Mandarin speech is aligned with the transcript, is used for the temporal segmentation to extract linguistic features. MFCC: mel frequency cepstral coefficients.



verb, adjective, pronoun, and other. In this step, we count the total number of these POS taggers.

In addition, due to the nature of the AM test, subjects can share stories on unrestricted topics. To facilitate the model to capture more meaningful information from the transcripts, we apply the topic insertion technique proposed by our previous work [11,18]. This technique is to add a sentence with the mentioned topic at the beginning of the transcript, such as “I’m going to share a story of...” The tester specifies the topic during data collection. This step helps ensure that the subsequent transcript is thematic. After the whole preprocessing of linguistic data, from the raw audio xi,km, we finally obtain the transcripts with topics and pauses xi,km,t, along with 9 additional linguistic features xi,km,l, which include mean length of utterance; the number of short, medium, and long pauses; and 5 types of POS tags.

Encoder Module

Acoustic Encoder

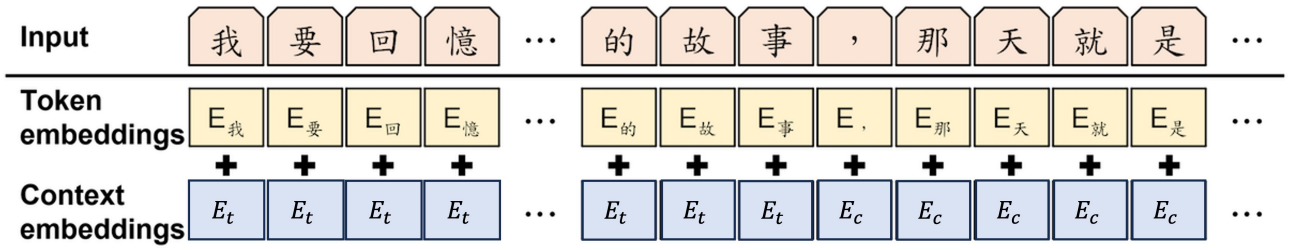
During the data preprocessing stage, 1024-dimensional deep learning–based acoustic features, denoted as xi,km,a, were extracted using the wav2vec 2.0 model. These high-level representations were then passed to a lightweight convolutional neural network–based acoustic encoder designed to capture hierarchical temporal-spectral patterns. The encoder consists of three 1D convolutional layers with kernel size 3 and stride 1, each followed by batch normalization, rectified linear unit activation, and max-pooling with a kernel size and stride of 2, progressively reducing the feature resolution by a factor of 8. The convolutional layers use 16, 32, and 64 filters, respectively, to model increasingly abstract local dependencies in the acoustic embedding. The resulting feature maps are flattened and projected through a fully connected layer that maps the representation to a 128-dimensional latent space. This projection enhances computational efficiency, mitigates overfitting, and provides a compact and informative acoustic embedding ai,km for multimodal fusion.

Linguistic Encoder

In the linguistic encoder, the ASR transcript xi,km,t and 9-dimension features xi,km,l generated during data preprocessing are sent to the BERT [20] model and a linear layer, respectively, for initial encoding. During data preprocessing, to ensure that the subjects’ speech has a thematic structure, we add a topic at the beginning of the transcript. In BERT, we borrow the concept of segment embeddings from the original paper [20]. Here, we refer to it as “context embedding” to enhance the model’s understanding of the topic and content. We divide the transcript into 2 parts: topic and context. The topic part is labeled as 0, and the context part as 1. These values are then encoded into context embeddings E_t and E_c through a learning embedding layer. This design helps the model determine whether there is a sufficient correlation between the topic and content. Research in AM tests indicates that episodic memory significantly declines in patients with cognitive impairments [21,22]. Therefore, we can assess the subject’s cognitive state by examining whether the subject’s story contains details related to the topic.

As illustrated in Figure 4, context embeddings are element-wise added to the original BERT token embeddings for further processing. A linear layer is used as a numerical encoder for linguistic features xi,km,l. Finally, the output from BERT and the learned numerical embeddings are concatenated and processed through a projector to obtain the final linguistic embedding ti,km.

Figure 4. Overview of the BERT embedding. Each input token is first mapped to a token embedding, which is combined with additional context embeddings to capture both semantic meaning and conversational context. BERT: bidirectional encoder representation for transformer.

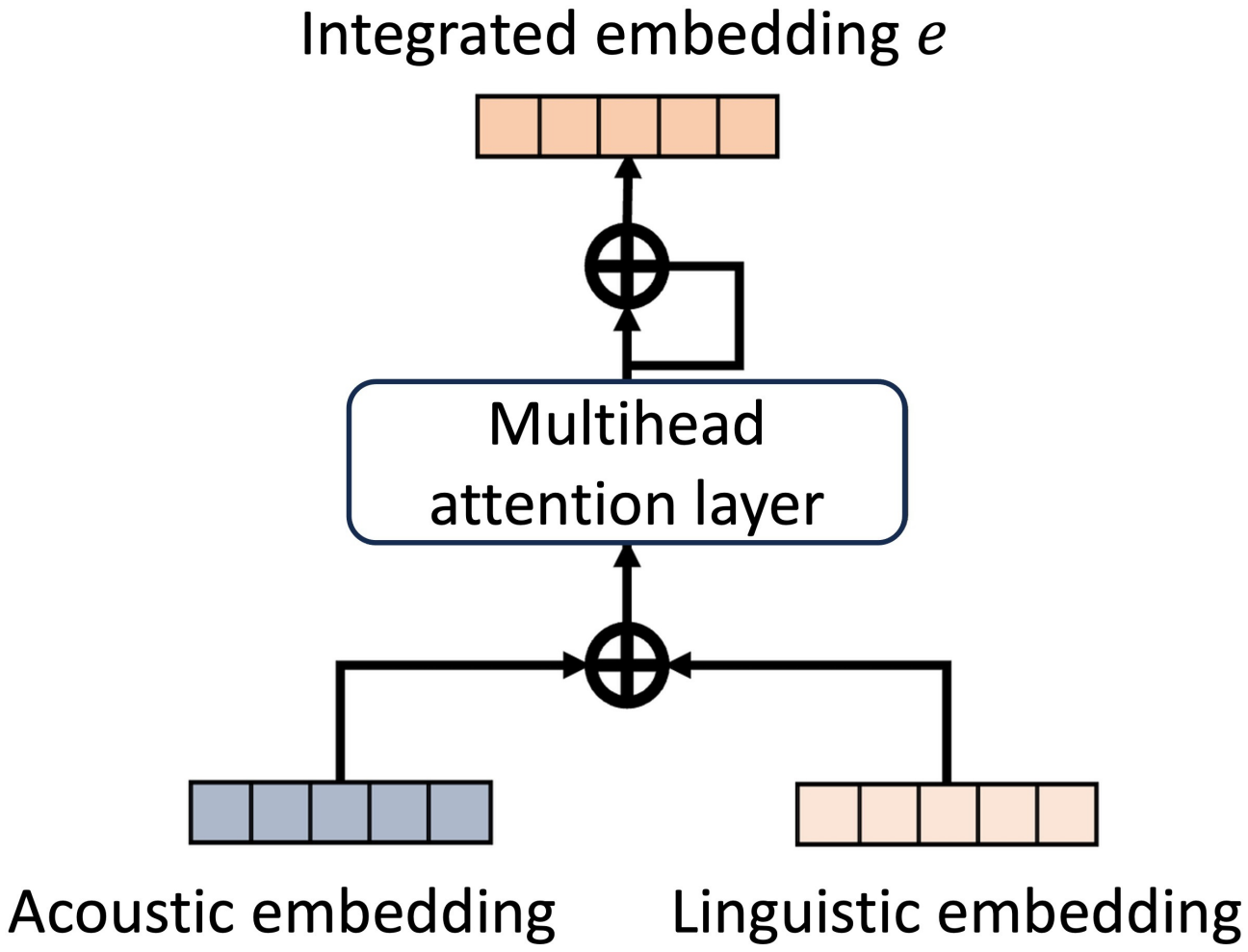


XSL•FO
RenderX

Fusion Layer

In this study, we adopt a self-attention fusion method to enable the model to capture information between different modalities more effectively. The concept is illustrated in Figure 5. After obtaining the acoustic embedding ai,km and the linguistic embedding ti,km, these 2 embeddings are added together to form a combined representation, which is then fed into a multihead self-attention layer. Through the calculations of the self-attention layer, we obtain the final integrated embedding ei,km. This design helps the model capture deeper interrelated information between different features.

Figure 5. Overview of the fusion layer. Acoustic and linguistic embeddings are integrated through a multihead attention mechanism, producing a joint embedding that captures cross-modal interactions and complementary cognitive cues.

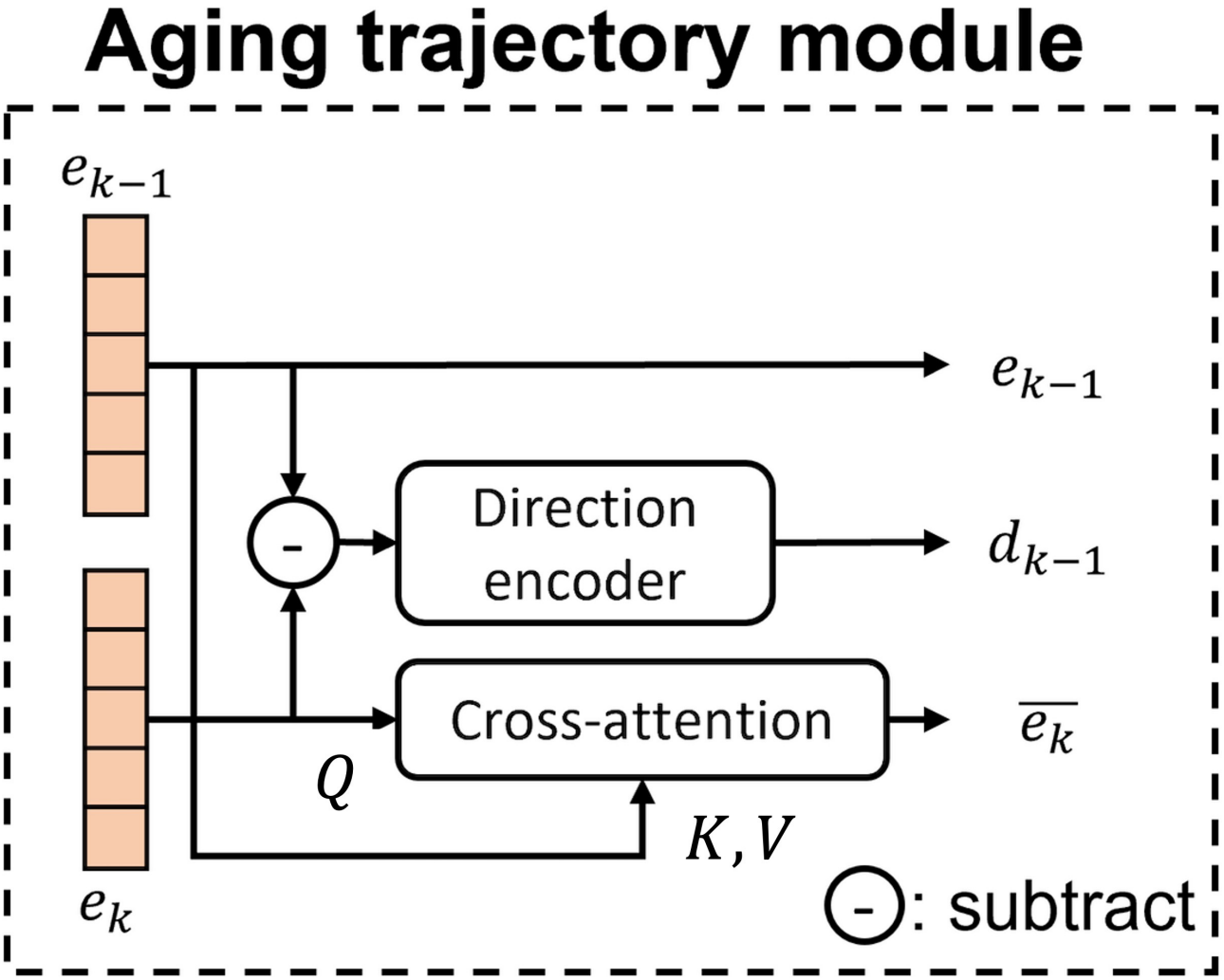


Aging Trajectory Module

The ATM is a key component of our longitudinal model, designed to capture cognitive changes across multiple visits. The structure is shown in Figure 6. In our study, each subject contributes up to 3 visits. Therefore, the model's initial input is in the form of a triplet, {xi,1m1, xi,2m2,xi,3m3}. After passing them through the encoder model process, this becomes the embedding triplet of the same form, {ei,1m1, ei,2m2,ei,3m3}. In this module stage, the embeddings are fed in pairs, such as {ei,1m1,ei,2m2}or {ei,2m2,ei,3m3}, so that the model focuses on learning the relationship between visits *k* and *k*-1. To model these relationships, we use a cross-attention mechanism, where the embedding from the *k*th visit serves as the query (*Q*), and the embedding from the (*k*-1)th visit serves as the key (*K*) and value (*V*). This mechanism allows us to capture the differences in feature values between 2 visits. The reason for using the *k*th visit embedding as the query and the (*k*-1)th visit embedding as the key and value is that, in practice, we can retrieve information only from the past data and not future data. This approach ensures that the model can access information only from the previous visits, reflecting a realistic scenario where future data are unavailable. By doing so, the model learns to focus on the most relevant past information that can be referenced by the current visit, enabling the tracking of changes and peeking of patterns of cognitive decline over time. This method effectively highlights the differences and similarities between visits, providing a more nuanced understanding of the subject's cognitive progression.

XSL•FO
RenderX

Figure 6. Overview of the aging trajectory module. Consecutive visit embeddings are compared through cross-attention and a direction encoder to capture cognitive differences between visits.



The concept of contrasts is quite straightforward. We aim to explore a deeper understanding of the relationship between the (*k*-1)th and *k*th visits. To achieve this, we use a cross-attention mechanism. In this setup, the embedding from the *k*th visit is used as the query (*Q*), whereas the embedding from the (*k*-1)th visit serves as the key (*K*) and value (*V*). This mechanism allows us to capture the differences in feature values between 2 visits. The reason for using the *k*th visit embedding as the query and the (*k*-1)th visit embedding as the key and value is that, in practice, we can retrieve information only from the past data and not future data. This approach ensures that the model can access information only from the previous visits, reflecting a realistic scenario where future data are unavailable. By doing so, the model learns to focus on the most relevant past information that can be referenced by the current visit, enabling the tracking of changes and peeking of patterns of cognitive decline over time. This method effectively highlights the differences and similarities between visits, providing a more nuanced understanding of the subject's cognitive progression.

To quantify the differences between visits, we introduce a direction encoder f_D, composed of several linear layers. The difference between embeddings is projected into a latent space to represent an aging-oriented change vector *d*, formulated as equation 2, where di,k−1mk−1,mk means the direction from mk-1th story in the (*k*-1)th visit to the mkth story in the *k*th visit. mk=1,2,…,Mi,k, where Mi,k is the total number of stories for subject *i* in the *k*th visit. Similarly, mk-1=1,2,…,Mi,k-1, where Mi,k-1 is the total number of stories for subject *i* in the (*k*-1)th visit.

(2)di,k−1mk−1,mk= fD(ei,kmk−ei,k−1mk−1)

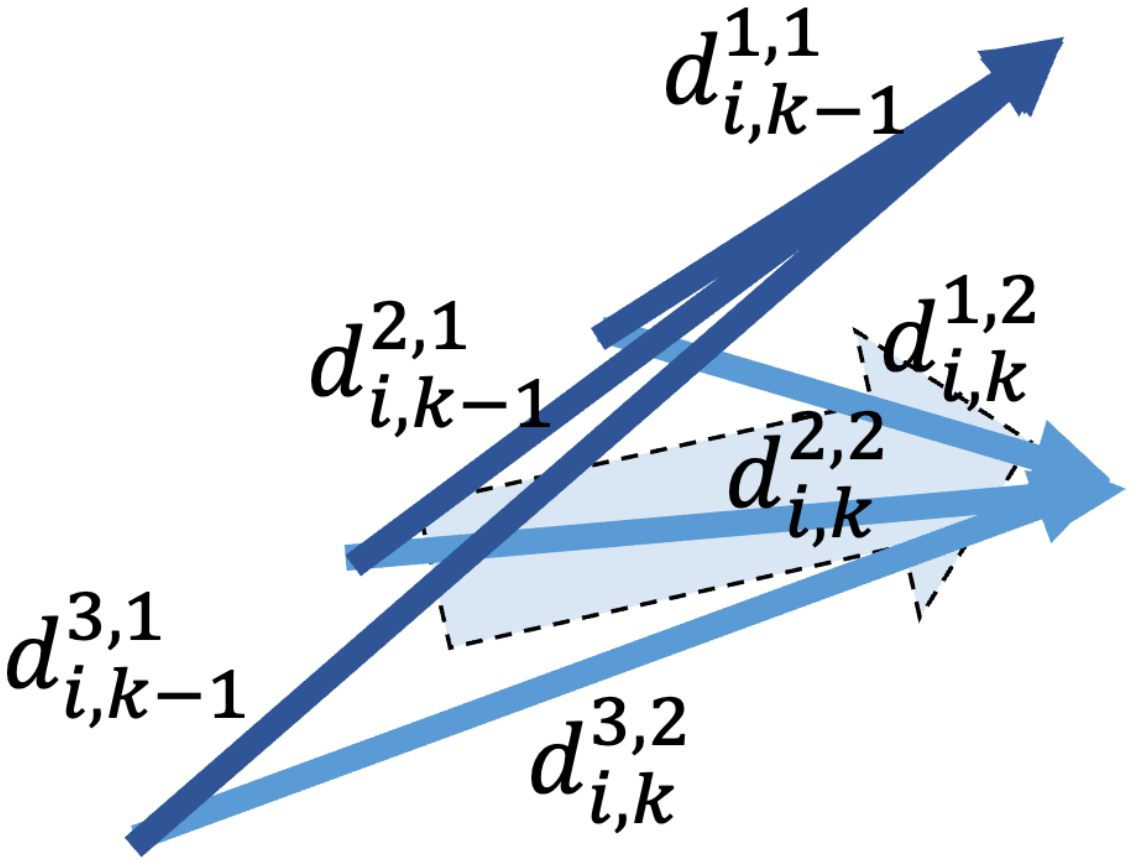
In addition, during the training process, motivated by the research [11,23,24], we introduce 2 types of losses: subject alignment loss and group alignment loss. These alignment losses are designed to align both individual and group targets, helping the direction encoder learn more accurate aging-oriented changes. We expect that by applying these losses, the resulting direction *d* will carry more representative meaning, aiding the longitudinal model in accurately assessing the subject's cognitive condition.

Subject Alignment Loss

In the subject alignment loss L_S, the concept is to ensure that the same subject's internal directions are similar. In our problem setting, the stories of the input triplet are randomly paired in different visits. Therefore, it is expected that the directions from the same subject may vary in the latent space. However, the cognitive decline of a subject should be consistent. Figure 7 shows an illustration of the concept.

XSL•FO
RenderX

Figure 7. Illustration of the concept for subject alignment loss. It visualizes how the subject alignment loss ensures that the overall trajectory directions of the subject remain consistent, although the internal trajectory of story-level embeddings for the same subject may show slight variations across different visits.



Therefore, subject alignment loss aims to reduce the differences in direction within the same subject. The loss used here is cosine similarity loss. The process of calculating the subject alignment loss is to first calculate the average subject direction for each visit as shown in equation 2. We then compute the subject alignment loss by summing the alignment loss of all subjects as defined in equation 3. The generalized subject alignment loss is defined as equation 3, where di,k−1− represents the average direction of every audio file in the (*k*-1)th visit to *k*th visit from the subject *i* in the latent space, and *N* represents the total number of subjects. $L_{S,i}$ represents the subject alignment loss for subject *i*, and cos θA,B=A.BAB is the cosine similarity function. In our case, since the dataset contains 3 visits, our *V* is 3, and the loss is applied to the pairs (first visit and second visit) and (second visit and third visit).

(3)di,k−1−=1Mi,k−1×Mi,k∑mk=1Mi,k∑mk−1=1Mi,k−1di,k−1mk−1,mk

LS=∑i=1NLS,i

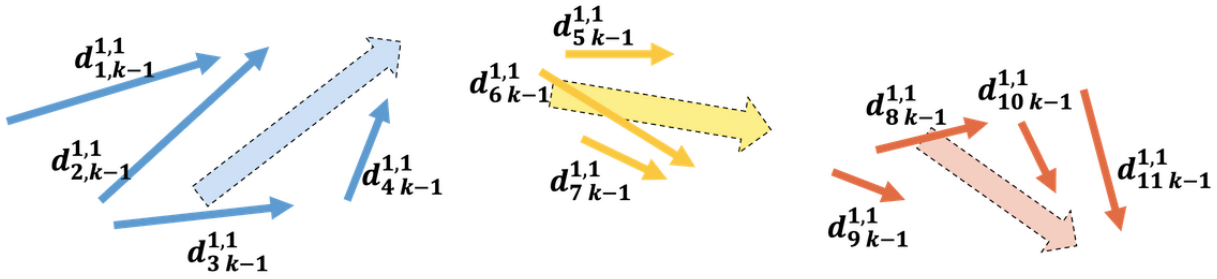
(4) =∑i=1N∑k=2V∑mk=1Mi,k∑mk−1=1Mi,k−11−cosθ(di,k−1−,di,k−1mk−1,mk)

Group Alignment Loss

In the group alignment loss L_G, the concept is to ensure that subjects with the same cognitive label have similar cognitive progression changes as illustrated in Figure 8. Unlike the subject alignment loss, which focuses on the internal alignment of directions within a subject, the group alignment loss emphasizes the external alignment of directions between different subjects. Specifically, we compare the cognitive change directions of subjects with the same label transitions. Here, we use the Supervised Contrastive Loss (SupConLoss) [25] to encourage the directions of subjects with the same label changes to be closer to one another. The label transitions considered are HC to HC, HC to MCI, and MCI to MCI. The process of calculating the group alignment loss is using equations 5 and 6, where *N* represents the total number of subjects, and $L_{G,i}$ represents each group alignment loss for subject *i*. SupConLoss(·, ·) is the SupConLoss function. lk-1,ki is the label transition type of subject *i* from (k − 1)th visit to *k*th visit.

XSL•FO
RenderX

Figure 8. Illustration of the concept for group alignment loss. It visualizes how the group alignment loss enforces consistency in progression directions among subjects sharing the same cognitive label. Blue represents the group from healthy control (HC) to HC, yellow represents the group from HC to mild cognitive impairment (MCI), whereas orange represents the group from MCI to MCI.



$$L_G=\sum_{i=1}^{N} L_{G,i}=\sum_{i=1}^{N}\sum_{k=2}^{V}\sum_{m_k=1}^{M_k}\sum_{l_{m_k}=1}^{M_{k-1}} SupConLoss(e_{i,k-1,m_k},e_{i,k,l_{m_k}})$$

$$\text{(6)}\ L_{G,i}=\{\text{0 if } H_{i,k-1}\text{ and } H_{i,k}\text{ or } M_{i,k-1}\text{ and } M_{i,k}\text{; otherwise } \sum_{m_k=1}^{M_k}\sum_{l_{m_k}=1}^{M_{k-1}} SupConLoss(e_{i,k-1,m_k},e_{i,k,l_{m_k}})\}$$

To summarize, group alignment loss L_G aims to minimize the distance between the cognitive progression directions of different subjects with the same cognitive label transitions types. This helps ensure that subjects exhibiting similar cognitive changes have consistent representations, thereby improving the model's ability to generalize across different subjects. By leveraging SupConLoss, we can effectively capture and align these progression patterns, leading to more accurate and robust cognitive state predictions.

Classifier and Overall Loss

After the ATM, we obtain the final outputs {e−i,1m1,e−i,2m2,e−i,3m3,di,1m1,m2, di,2m2,m3}, where e−i,kmk represents the enhanced latent representation after the cross-attention layer. These representations are then concatenated and fed into a classifier. The classifier comprises a batch normalization layer and a fully connected layer to produce the final prediction. Finally, the overall loss function can be formulated by equation 7, where Lclass is the binary classification loss, and α, δ, and γ are the hyperparameters of loss weight for L_E, L_G, and L_S, respectively. The hyperparameters are fine-tuned, and the optimal values are 0.01, 0.001, and 0.001.

$$\text{(7)}\ L= L_{class}+\alpha L_E+\delta L_G+\gamma L_S$$

The encoder loss L_E is a contrastive loss designed to enhance the representativeness of the embeddings generated by the encoder and is defined by equation 8, where N is the total number of the subjects, V is the total number of visits, and M_k is the total number of stories in the kth visit. This loss ensures that embeddings with the same cognitive status are closer to one another while pushing apart embeddings with different labels.

$$\text{(8)}\ L_E=\sum_{i=1}^{N} L_{E,i}=\sum_{i=1}^{N}\sum_{k=2}^{V}\sum_{m_k=1}^{M_k} SupConLoss(e_{i,k,m_k},l_{i,k})$$

Dataset

Dataset Collection

Our experiments primarily focus on comparisons using our self-collected National Taiwan University Autobiographical Memory Test (NTU-AM) dataset. Previous studies [6,11,26,27] have confirmed that the AM test effectively tracks changes in cognitive function by evaluating subjects' memory functions. It can be used as a tool for MCI testing. During the test, subjects are asked to recall important life stories from different periods in their past. Compared with formal memory testing methods, this process resembles a general conversation, which helps reduce the stress experienced by subjects.

For the collected data used in this study, professional psychologists conducted one-on-one interviews. During these interviews, a microphone is used to record the conversation between the subject and the interviewer. Furthermore, subjects are asked first to talk about the titles of the important events they want to share (eg, "I want to share the story about when I was admitted to college at the age of 18 and my test result were announced"). The process of describing the events can be divided into two stages, namely, recall and probing.

1. Recall: The subject is first asked to roughly describe the events leisurely and then is encouraged to share as many details of the event as possible.
2. Probing: The interviewer then asks even more detailed questions based on what the subject mentioned in the recall stage. These questions cover various categories, such as "what," "when," "where," emotion or thought, and time integration.

During a single interview, the same subject can share multiple stories as a matter of fact. To sum up, the overall process of the AM interview test can be described as follows: (1) the subject decides the number and content of stories to be shared, (2) shares the recall content of each story, and then (3) the interviewer probes each story with a series of detailed questions.

Moreover, to track the cognitive changes of subjects over a long term, they are invited to take the AM test again after at least 6 months. In our current research, our longitudinal data include up to 3 rounds of AM interview test data, that is, each subject has 1 test data and 2 follow-up data. All data were collected by professional psychologists with institutional review board approval, ensuring high-quality data.

The National Taiwan University Autobiographical Memory Test Cross-Sectional (NTU-AM-CS) dataset was collected under a cross-sectional scenario, with the demographic distribution of participants shown in Table 1. This dataset includes 111 subjects, consisting of 56 HC and 55 MCI participants. Each participant could provide multiple important life stories during a single visit.

XSL•FO
RenderX

Table . The demographic distribution of participants in the National Taiwan University Autobiographical Memory Test Cross-Sectional dataset.

Group	N	Male	Female	Age (years), mean (SD)	Years of education, mean (SD)	MMSE[a], mean (SD)
HC[b] (n=56)	56	18	38	71.3 (5.4)	16.2 (9.8)	28.7 (1.6)
MCI[c] (n=55)	55	17	38	70.7 (6.9)	13.1 (7.1)	27.1 (2.0)
Total (N=111)	111	35	76	71.0 (6.2)	14.6 (8.7)	27.9 (2.0)

[a]MMSE: Mini-Mental State Examination.

[b]HC: healthy control.

[c]MCI: mild cognitive impairment.

The longitudinal scenario dataset, National Taiwan University Autobiography Memory Test Longitudinal (NTU-AM-LG), involves 54 of the original 111 subjects who participated in follow-up studies, including 28 HC and 26 MCI participants. The types of data collected are the same as in the cross-sectional scenario, encompassing both recall and probing audio files. The key difference is that NTU-AM-LG includes story audio files from 3 visits. The demographic distribution for this dataset is also shown in Table 2.

Table . The demographic information of the National Taiwan University Autobiography Memory Test Longitudinal dataset.

Group (N=54)	Male	Female	Age (years), mean (SD)	Years of education, mean (SD)	MMSE[a], mean (SD)
HC[b] (n=28)	12	16	74.0 (5.9)	14.3 (2.7)	28.4 (1.5)
MCI[c] (n=26)	7	19	72.5 (6.4)	14.0 (9.6)	27.5 (1.9)
Total	19	35	73.3 (6.1)	14.1 (7.0)	28.0 (1.8)

[a]MMSE: Mini-Mental State Examination.

[b]HC: healthy control.

[c]MCI: mild cognitive impairment.

Ethical Considerations

The collection of experimental data from human participants was reviewed and approved by the National Taiwan University Hospital Research Ethics Committee B (protocol no. 202105013RINB, approved on June 18, 2021). All participants provided written informed consent prior to data collection. The original consent and institutional review board approval explicitly permitted the secondary analysis of deidentified recordings and transcripts for research purposes, without requiring additional consent. All data were fully deidentified before analysis, and no personal or identifiable information was retained in the research dataset. Participant privacy was protected by storing all files on secure, access-restricted servers, and only aggregated results are reported. Participants received NTD 160 (equivalent to US $5.09) per hour as compensation for their time and contribution. No identifiable participant images are included in this manuscript.

Experiment Dataset

The summary of the collected dataset can be provided. The total number of datasets is shown in Table 3, where the unit of measurement is the number of stories. Since each story includes both recall and probing phases, the quantities are identical for recall and probing data.

Table . Primary dataset statistics.

Dataset	Visit 1		Visit 2		Visit 3		Total	
	HC[a]	MCI[b]	HC	MCI	HC	MCI	HC	MCI
NTU-AM-CS[c]	134	122	N/A[d]	N/A	N/A	N/A	134	122
NTU-AM-LG[e]	44	30	50	40	76	54	170	124

[a]HC: healthy control.

[b]MCI: mild cognitive impairment.

[c]NTU-AM-CS: National Taiwan University Autobiography Memory Test Cross-Sectional.

[d]N/A: not applicable.

[e]NTU-AM-LG: National Taiwan University Autobiography Memory Test Longitudinal.

XSL•FO
RenderX

NTU-AM-CS is a cross-sectional dataset that includes a total of 111 subjects (56 HCs and 55 with MCI). Each subject participated in a single AM interview, during which they could share multiple stories. Therefore, NTU-AM-CS does not include repeated follow-up visits.

NTU-AM-LG is the longitudinal subset derived from NTU-AM-CS. A portion of participants agreed to return for additional follow-up interviews, resulting in up to 3 visits per subject. As shown in Table 3, the number of stories per visit varies because the volume of data depends on how many stories participants chose to share during each session. In some cases, more stories were collected in visit 3 than in earlier visits, which explains the higher data volume observed.

To further clarify disease progression patterns, the number of subjects transitioning across cognitive status categories between visits, including stable (HC→HC, MCI→MCI) and progressive or recovered (HC→MCI, MCI→HC) cases, is shown in Table 4.

Table . Cognitive status transitions between consecutive visits in the National Taiwan University Autobiography Memory Test Longitudinal dataset across visits 1 and 2 and visits 2 and 3.

Transition	Number of cases (subject based), n =108
HC[a] to HC	50
HC to MCI[b]	10
MCI to HC	8
MCI to MCI	40

[a]HC: healthy control.

[b]MCI: mild cognitive impairment.

For additional experiments, we primarily used a semistructured speech dataset. The National Taiwan University Hospital Picture Description Test (NTUH-PIC) and National Taiwan University Hospital Memory Test (NTUH-MEM) datasets, which we collected ourselves, include speech data from picture description and logical memory tests, respectively. Both datasets feature 2 cognitive labels: HC and MCI. In the picture description task, participants are required to describe the scenario depicted in a specific picture as accurately and detailed as possible. In the NTUH-PIC dataset, the well-known "cookie theft" picture [28] is used. For the logical memory test, participants first listen to a story and then must repeat it from memory. The third dataset, ADReSSo [29], was introduced as part of the cognitive challenge at INTERSPEECH 2020. This dataset also involves a picture description test, which also used "cookie theft" [28] but differs from the other datasets in that it includes only 2 cognitive labels: HC and AD. Due to data access limitations, we used only the training set for evaluation in this case. By incorporating these additional datasets, we aim to comprehensively evaluate the robustness and generalizability of our model across various types of speech data and cognitive assessments. The statistics of the additional dataset are shown in Table 5.

Table . Additional dataset statistics.

Dataset	Total	HC[a]	MCI[b]	Language	Test
NTUH-PIC[c]	80	40	40	Chinese	Picture Description Test
NTUH-MEM[d]	104	52	52	Chinese	Logical Memory Test
ADReSSo (Train)	166	79	87	English	Picture Description Test

[a]HC: healthy control.

[b]MCI: mild cognitive impairment.

[c]NTUH-PIC: National Taiwan University Hospital Picture Description Test.

[d]NTUH-MEM: National Taiwan University Hospital Memory Test.

Experiments

We conducted a series of experiments to evaluate our proposed longitudinal multimodal model under different data conditions and to verify the contribution of each model component.

Experiments on NTU-AM-LG

We evaluated the model on the longitudinal NTU-AM-LG dataset using 2 speaking tasks: recall and probing. For each task, we compared four variants of our model:

1. Ours (audio): acoustic branch only, wav2vec 2.0 to acoustic encoder to classifier.
2. Ours (text): linguistic branch only, BERT to linguistic encoder to classifier.
3. Ours (CS): cross-sectional problem setting
4. Ours (LG): full proposed longitudinal multimodal model with direction encoding and alignment losses.

XSL•FO
RenderX

These models were compared against standard baselines (BERT, wav2vec 2.0, and the prior AM-based work [11]), and the corresponding results are reported in the "Results" section.

An ablation study on NTU-AM-LG is made to examine the contribution of the ATM, direction embedding, and the 2 alignment losses, and we performed ablation experiments by removing (1) the direction, (2) the subject alignment loss, (3) the group alignment loss, and (4) both alignment losses.

Experiments on NTU-AM-CS

To demonstrate that the proposed encoders and fusion strategy also work without longitudinal information, we trained the same acoustic and linguistic encoders in the cross-sectional problem setting (Our [CS]) and compared them with BERT, wav2vec 2.0, and the prior AM method [11]. An ablation study is also conducted to highlight the importance of linguistic features and context embeddings.

Experiments on Additional Datasets

We further tested whether our proposed model transfers to other cognitive-speech datasets (NTUH-PIC and NTUH-MEM). We compared our method with that of Lin et al [18].

Finally, to enable comparison with existing AD or MCI speech work on a public dataset, we evaluated our model on the ADReSSo corpus.

Experimental Setup

In the subsequent experiments, due to the limitations of the dataset size, we will present the averaged performance of 10-fold cross-validation. To avoid data leakage, we will split the training set and validation set by subject for the NTU-AM-CS and NTU-AM-LG datasets. It ensures that audio files from different visits of the same subject will not appear in the training set and validation set simultaneously.

For the NTUH-PIC, NTUH-MEM, and ADReSSo datasets, since each subject provides only 1 audio file, there is no need to split the data by subject. We will perform down-sampling on the majority class in the training set during the experimental process to stabilize the model's learning capability. All experiments are conducted on an NVIDIA RTX 4090 GPU.

Evaluation Metrics

In the experimental section, we will use 6 different evaluation metrics to measure the performance of the model: accuracy, F_1-score, precision, sensitivity (recall), specificity, and AUROC. These are common indicators used to assess the predictive ability of a model. Each metric provides a different perspective on the effectiveness of the model.

In the NTU-AM dataset, since the dataset is split by subject, each evaluation set from every fold may have a similar number of subjects, but the number of audio files provided by each subject can vary. This can result in label imbalance in the final soft-voting results. Therefore, the most critical metrics in the subsequent experiments will be F_1-score and AUROC. These 2 metrics provide a fairer assessment in situations with label imbalance. As for other experiments, the primary metrics of interest will be accuracy and F_1-score.

Baseline

In the subsequent experiments in the NTU-AM dataset, we will use a fine-tuned BERT model and the wav2vec 2.0 model as baselines. Since these 2 models are the backbones of our proposed model, in the NTU-AM-LG dataset, the baseline model will be composed of these 2 main models as encoders. The embeddings generated from the 3 visits will be concatenated and passed through the same classifier. For the NTU-AM-CS dataset, the backbone model will serve as the encoder, followed by a classifier. However, these models will not have additional loss functions to assist training during the training process, relying only on the basic binary cross-entropy loss.

Implementation Details

The probing data are interview data, which include both the interviewer's and the subject's voices. Since our study aims to determine whether the subject has a tendency for MCI based on their audio data, the interviewer's speech is considered noise during training. To remove the interviewer's portion, we use the tool kit "pyannote" to split the audio into different speaker segments. This process results in audio files without the interviewer's voice. These cleaned files are then used to generate the linguistic features during the data preprocessing stage.

We used the "openai/whisper-large-v2" version from Hugging Face for the ASR tool. For data preprocessing, the pretrained Wav2vec 2.0 model used for the English dataset is "facebook/wav2vec2-base-960h," while for the Chinese dataset, we used "jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn," which are also from HuggingFace. We used a 15-second clip segmentation with a 5-second overlap. Considering the limitations of the Wav2vec model, we resampled the audio files to 16,000 Hz and standardized them to mono-channel.

For the linguistic model, the BERT model used for the English dataset is "google-bert/bert-base-uncased," and for the Chinese dataset, it is "hfl/chinese-bert-wwm-ext," which is also the version used for the BERT baseline. In the Wav2vec baseline model, we used "jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn." The hyperparameters used for implementing the proposed model are shown in Table 6.

XSL•FO
RenderX

Table . Hyperparameters used for implementation.

Hyperparameters	Value
BERT[a] attention dropout rate	0.3
BERT linear layer dropout rate	0.3
Maximum sentence (token) length	512
Dimension of audio embedding	128
Dimension of text embedding	128
Dimension of direction embedding	32
Learning rate	1e-5
Number of layers for attention-fusion	1
Number of layers for cross-attention	1
Batch size for training	8
Batch size for evaluation	8
Maximum epoch number	150
Early stop	10
α for encoder loss	0.01
δ for group alignment loss	0.001
γ for subject alignment loss	0.001

[a]BERT: bidirectional encoder representation for transformer.

Results

Results of NTU-AM-LG

We experiment with the model's effectiveness on both recall and probing data. Given the limited research on longitudinal speech, we primarily compare our model against baselines such as BERT, Wav2vec 2.0, and the unimodal model within our longitudinal framework. In addition, we present an ablation study to demonstrate the necessity of each component in our model. We discuss the results for recall data and probing data separately. It is important to note that both the cross-sectional (CS) and longitudinal (LG) models are trained on the same NTU-AM-LG dataset, sharing identical subjects and feature representations. The difference lies only in the problem formulation: the CS model performs single-visit classification, while the LG model performs a prognostic task that leverages prior visit embeddings to predict the final cognitive state. Thus, the performance difference reflects the use of incorporating longitudinal context rather than architectural or data disparities.

Table 7 shows the performance of our proposed model and the baselines for the recall data. First, when the input consists only of audio data, our proposed model "Ours (audio)" outperforms the previous work [11] on most metrics. Notably, there are improvements of 5% and 6% in F_1-score and AUROC, respectively.

XSL•FO
RenderX

Table . Results of recall data on National Taiwan University Autobiography Memory Test Longitudinal dataset.

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Audio only						
Chang et al [11]	0.66	0.64	0.74	0.67	0.65	0.63
Wav2Vec2	0.59	0.58	0.53	0.78[b]	0.43	0.60
Ours (audio)	0.67	0.69	0.73	0.73	0.66	0.69
Text only						
BERT[c]	0.73	0.65	0.77	0.64	0.82	0.73
Ours (text)	0.75	0.72	0.83	0.73	0.72	0.73
Text + Audio						
Ours (CS[d])	0.72	0.72	0.68[b]	0.76[b]	0.74	0.66
Ours (LG[e])	0.82[b]	0.78[b]	0.95[b]	0.72[b]	0.97[b]	0.85[b]

[a]AUROC: area under the receiver operating characteristic curve.

[b]Highest value for each metric.

[c]BERT: bidirectional encoder representation for transformer.

[d]CS: cross-sectional.

[e]LG: longitudinal.

As for the model using only linguistic data as input, our proposed linguistic longitudinal model, denoted as "Ours (text)," outperforms BERT, demonstrating that including extra-linguistic features and context embedding helps improve the model's performance with 7% increase in the F_1-score.

Finally, our model, which is denoted as "Ours (LG)" in Table 7, significantly outperforms other models across most metrics for the dual-modal longitudinal model that combines both data types. Specifically, there is up to a 16% improvement in accuracy, a 14% improvement in F_1-score, and a 22% improvement in AUROC. Since the previous work used only unimodal data, examining AM data from different perspectives benefits MCI detection. We also compare the cross-sectional setting results of NTU-AM-LG, denoted as "Ours (CS)" in Table 7, with the longitudinal settings, confirming that long-term tracking of subjects' cognitive trajectories in NTU-AM-LG helps improve the model's prediction accuracy. Specifically, there are improvements of 10% in accuracy, 6% in F_1-score, and 19% in AUROC.

Table 8 shows the results of the probing data. For the audio data, our proposed model outperforms wav2vec 2.0 in most metrics, with improvements of 9%, 18%, and 10% in accuracy, F_1-score, and AUROC, respectively. This indicates that instead of directly using wav2vec 2.0's output representation for classification, extracting features using wav2vec 2.0 followed by a convolutional neural network–based acoustic encoder structure can better capture the hidden cognitive patterns in audio data.

XSL•FO
RenderX

Table . Result of probing data on National Taiwan University Autobiography Memory Test Longitudinal dataset.

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Audio only						
Wav2vec2	0.76	0.53	0.60	0.53	0.87	0.70
Ours (audio)	0.85	0.71	0.67	0.78	0.83	0.80
Text only						
BERT[b]	0.85	0.72	0.83	0.71	0.94	0.82
Ours (text)	0.84	0.80	0.86	0.81	0.82	0.82
Text + Audio						
Ours (CS[c])	0.71	0.67	0.61	0.75	0.63	0.68
Ours (LG[d])	0.89[e]	0.83[e]	0.90[e]	0.82[e]	0.95[e]	0.89[e]

[a]AUROC: area under the receiver operating characteristic curve.

[b]BERT: bidirectional encoder representation for transformer.

[c]CS: cross-sectional.

[d]LG: longitudinal.

[e]Values in boldface are the highest values for each metric.

For the linguistic data results, our model performs comparably with BERT, with improvements of 8%, 3%, and 10% in F_1-score, precision, and sensitivity, respectively. This suggests that the inclusion of linguistic features and the context embedding layer can also aid in MCI detection on probing data.

Finally, in the dual-modal data tests, the proposed longitudinal model achieved the best performance, excelling in every metric. Specifically, it showed improvements of up to 13%, 30%, and 19% in accuracy, F_1-score, and AUROC, respectively. Comparing the longitudinal model with the cross-sectional model, it is evident that the longitudinal model continues to perform better. This indicates that incorporating input from multiple time points and considering the concept of aging direction can enhance the model's predictions, resulting in performance increases of 12%, 16%, and 21% in accuracy, F_1-score, and AUROC, respectively.

Ablation Study

In the longitudinal setting, the most crucial idea is the alignment loss we proposed. Therefore, this section examines the contribution of alignment loss by removing each loss individually. The results are shown in Table 9. In the table, "Without direction" refers to not including direction embedding as the classifier input; namely, the input would be ei,1m1, ei,2m2, ei,3m3 treating each time point's data as independent.

XSL•FO
RenderX

Table . Result of the ablation study for our proposed model on National Taiwan University Autobiography Memory Test Longitudinal dataset.

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Recall data						
Without direction	0.79	0.74	0.71	0.83	0.67	0.75
Ours (LG[b])	0.82[c]	0.78[c]	0.95[c]	0.72	0.97[c]	0.85[c]
Subject alignment loss	0.71	0.67	0.68	0.77	0.63	0.70
Group alignment loss	0.68	0.69	0.68	0.89[c]	0.47	0.68
Both alignment loss	0.65	0.59	0.64	0.62	0.63	0.63
Probing data						
Without direction	0.79	0.70	0.68	0.76	0.78	0.77
Ours (LG)	0.89[c]	0.83[c]	0.9[c]	0.82[c]	0.95[c]	0.89[c]
Subject alignment loss	0.78	0.59	0.61	0.59	0.92	0.75
Group alignment loss	0.69	0.54	0.63	0.55	0.8	0.68
Both alignment loss	0.65	0.53	0.60	0.53	0.87	0.70

[a]AUROC: area under the receiver operating characteristic curve.

[b]LG: longitudinal.

[c]Highest value for each metric in recall data and probing data, respectively.

The results show that incorporating alignment losses with direction embedding improves performance across both recall and probing data. For instance, on recall data, our full model improves AUROC by 10% (0.85 vs 0.75) and F_1-score by 4% compared with "Without direction." On probing data, the gains are even larger, with AUROC improved by 12% (0.89 vs 0.77) and F_1-score improved by 13% (0.83 vs 0.70). Importantly, removing the group alignment loss results in the sharpest performance drop, with AUROC decreasing by 17% (0.68 vs 0.85) and 21% (0.68 vs 0.89) for recall and probing data, respectively. Similarly, removing both alignment losses further reduces AUROC to 0.63 (recall) and 0.70 (probing), confirming their necessity.

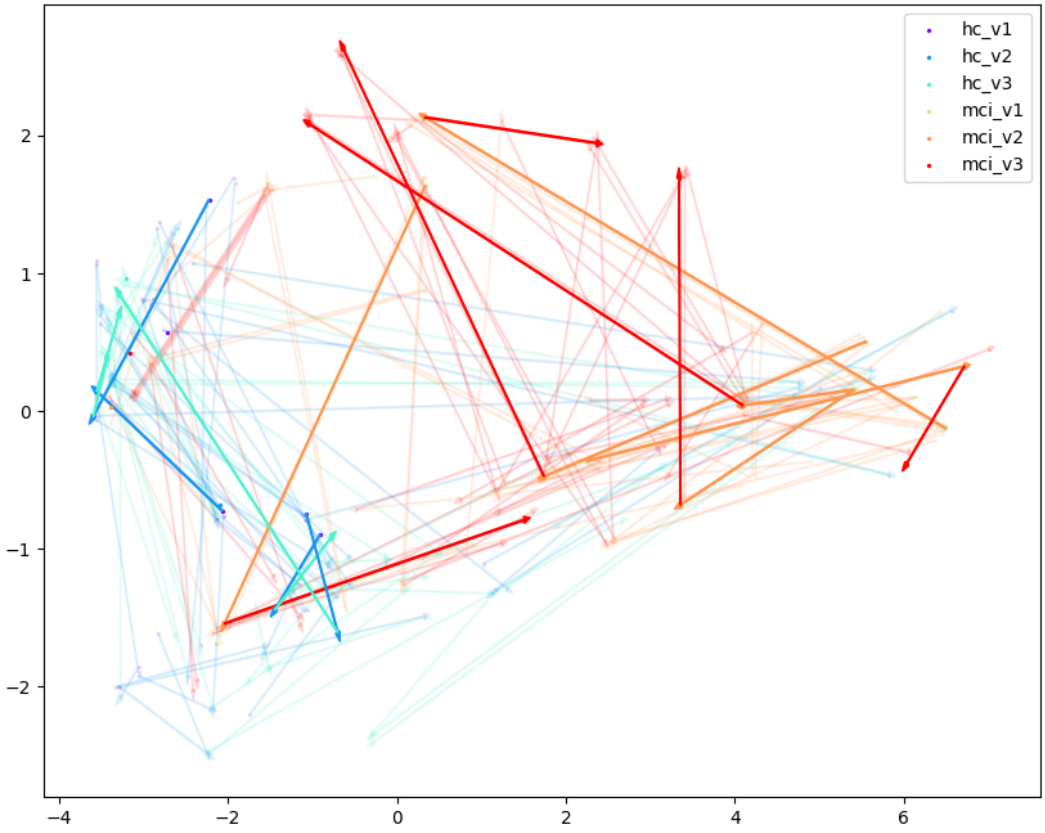
Overall, these results highlight the complementary importance of both alignment losses that rely on the strong supervisory signal of ground-truth label transitions to capture temporal progression patterns in longitudinal settings. Removing either the subject alignment loss or the group alignment loss results in significant performance degradation, and removing both leads to the sharpest decline. This indicates that while the subject alignment loss stabilizes intrasubject trajectory consistency, the group alignment loss aligns intersubject cognitive progression trends across similar diagnostic groups.

Visualization

We visualize the latent space distribution of subjects across different visits to better interpret the role of the alignment loss. After passing through all modules, we extracted the final embeddings, applied principal component analysis for dimensionality reduction, and projected the results into a 2D space. The arrows in Figures 9 and 10 represent the trajectory of each subject's embeddings across consecutive visits, thereby illustrating how cognitive states evolve over time.

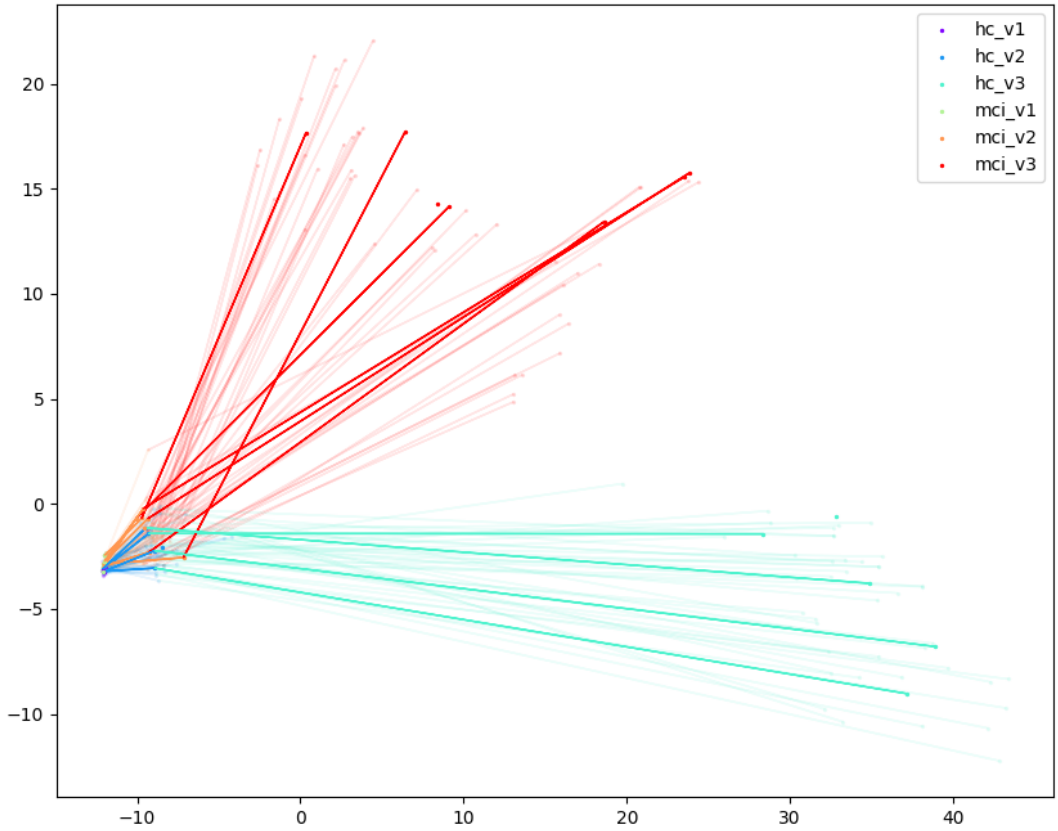
XSL•FO
RenderX

Figure 9. Principal component analysis projection of subject embeddings without alignment loss. The x- and y-axes represent the first 2 principal components. Colors denote diagnostic group and visit (blue tones = healthy control visits 1 - 3; red tones = mild cognitive impairment visits 1 - 3). Arrows illustrate the trajectory of each subject's embeddings. The scattered trajectories reflect high variability and weak temporal consistency without alignment loss.



XSL•FO
RenderX

Figure 10. Principal component analysis projection of subject embeddings with alignment loss. The x- and y-axes show the first 2 principal components. Colors indicate diagnostic group and visit (blue tones = healthy control visits 1 - 3; red tones = mild cognitive impairment visits 1 - 3). Arrows illustrate the trajectory of each subject's embeddings. With alignment loss, trajectories become far more coherent and directionally consistent, indicating improved temporal stability and clearer separation between healthy control and mild cognitive impairment progression patterns.



As shown in Figure 9 (without alignment loss), the trajectories are scattered and inconsistent, with large variability both within and across groups. This makes it difficult to distinguish between HC and MCI or to capture stable temporal patterns. In contrast, Figure 10 (with alignment loss) reveals more coherent trajectories: subjects within the same group (HC or MCI) show consistent progression directions across visits, while the separation between HC and MCI becomes clearer. Importantly, this indicates that the alignment loss encourages the model to capture cognitive progression trends rather than random variability in story content, reflecting interpretable patterns of decline (for MCI) and stability (for HC).

These results demonstrate that the alignment loss not only improves metrics but also enhances interpretability by aligning temporal trajectories in a way that reflects clinically meaningful differences in cognitive change. Feature-level analyses could provide even more intuitive examples of how the ATM aligns specific cognitive markers in speech. Future feature-level analyses could further enhance this by providing more intuitive illustrations of how the ATM captures and aligns specific cognitive markers in speech.

Results of NTU-AM-CS

In addition to the main longitudinal experiment, we conducted further evaluations in a cross-sectional setting to assess the performance of our proposed model. Here, the primary comparison is with our previous work [11], which also used audio and text data. However, a key difference is that the prior work relied on manual transcripts, which are less scalable and require special preprocessing for the transcripts. The results are shown in Tables 10 and 11.

XSL•FO
RenderX

Table . Results of recall data on National Taiwan University Autobiographical Memory Test Cross-Sectional dataset[c].

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Audio only						
Wav2vec2	0.65	0.42	0.41	0.48	0.47	0.51
Ours (audio)	0.57	0.60	0.52	0.72	0.45	0.58
Text only						
BERT[b]	0.67	0.56	0.62	0.61	0.67	0.64
Ours (text)	0.69	0.66	0.65	0.68	0.69	0.69[c]
Text + Audio						
Chang et al [11]	0.75[c]	0.61	0.76[c]	0.55	0.85[c]	0.68
Ours (CS[d])	0.74	0.72[c]	0.74	0.69[c]	0.83	0.68

[a]AUROC: area under the receiver operating characteristic curve.

[b]BERT: bidirectional encoder representation for transformer.

[c]Highest value of each evaluation metric.

[d]CS: cross-sectional.

Table . Results of probing data on National Taiwan University Autobiographical Memory Test Cross-Sectional dataset[c].

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Audio only						
Wav2vec2	0.55	0.49	0.65	0.43	0.66	0.61
Ours (audio)	0.61	0.65	0.56	0.84	0.43	0.64
Text only						
BERT[b]	0.69	0.64	0.66	0.72	0.67	0.69
Ours (text)	0.69	0.66	0.67	0.64	0.75[c]	0.65
Text + Audio						
Chang et al [11]	0.77[c]	0.64	0.62	0.73	0.68[c]	0.65
Ours (CS[d])	0.73	0.71[c]	0.67[c]	0.77[c]	0.68	0.72[c]

[a]AUROC: area under the receiver operating characteristic curve.

[b]BERT: bidirectional encoder representation for transformer.

[c]Highest value of each evaluation metric.

[d]CS: cross-sectional.

Our proposed method performs better or comparably on critical metrics such as F_1-score and AUROC in the results of recall data. Specifically, there is an 11% improvement in F_1-score. However, the precision is not as good as in the study by Chang et al [11]; our sensitivity metric performs better. For probing data, our method also yields better results in F_1-score and AUROC. However, the overall performance of the NTU-AM-CS dataset is not as promising as that of the NTU-AM-LG dataset. Detecting MCI features in AM data is challenging and reinforces the idea that long-term data relationships aid in model prediction.

Ablation Study

We investigated the contributions of linguistic features and context embedding in the encoder module for the prediction ability. The results are shown in Table 12. It demonstrates the impact of removing individual components on the model's predictions. For example, when linguistic features are removed, there is a noticeable decrease in specificity and a 6% drop in F_1-score. The most significant impact is seen when context embedding is removed, resulting in a 9% decrease in F_1-score and substantial declines in most other metrics. Only the sensitivity metric performs better, but overall, the original configuration still performs best in terms of F_1-score, which balances sensitivity and precision.

XSL•FO
RenderX

Table . Results of the ablation study for recall data on the National Taiwan University Autobiographical Memory Test Cross-Sectional dataset.

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
Ours (CS[b])	0.74[c]	0.72[c]	0.74[c]	0.69[c]	0.83[c]	0.68[c]
Linguistic features	0.62	0.66	0.57	0.84[c]	0.45	0.62
Context embedding	0.59	0.63	0.53	0.80	0.45	0.60

[a]AUROC: area under the receiver operating characteristic curve.

[b]CS: cross-sectional.

[c]Highest values of each evaluation metric.

These results indicated that linguistic features provide auxiliary information that aids the model's decision-making, particularly in identifying negative samples. Context embedding supports the model's understanding of the story by distinguishing between the story topic and content through embeddings.

Result of Additional Dataset

In the experiment with an additional dataset, our main goal was to confirm whether the method our model uses to encode data is also applicable to other datasets related to cognitive tests. This means whether the way we encoded each data contributes to MCI detection. In Table 13, we present several results related to the detection of MCI. Our primary comparison targets in the NTUH-PIC and NTUH-MEM datasets are as per the study by Lin et al [18], which use textual information, including pauses, for MCI prediction. From the NTUH-PIC experimental results, there was an improvement of 11% in accuracy and 18% in F_1-score.

Table . Results of the National Taiwan University Hospital dataset.

Dataset	Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity	AUROC[a]
NTUH-PIC[b]	Lin et al [18]	0.68	0.63	0.81[c]	0.58	0.78[c]	N/A[e]
	Ours (CS[d])	0.79[c]	0.81	0.80	0.88[c]	0.7	0.79
NTUH-MEM[f]	Lin et al [18]	0.68	0.66	0.75	0.7	0.65	N/A
	Ours (CS)	0.78[c]	0.76[c]	0.82[c]	0.73[c]	0.83[c]	0.78

[a]AUROC: area under the receiver operating characteristic curve.

[b]NTUH-PIC: National Taiwan University Hospital Picture Description Test.

[c]Highest values for each evaluation metric in its respective dataset.

[d]CS: cross-sectional.

[e]N/A: not applicable.

[f]NTUH-MEM: National Taiwan University Hospital Memory Test.

Similarly, in the NTUH-MEM dataset, a 10% improvement was observed in both accuracy and F_1-score compared with the study by Lin et al [18], indicating overall better performance. This experiment demonstrates that our data-encoding method effectively extracts latent representations suitable for MCI detection. Viewing data from multiple modalities provides consistent benefits for MCI detection. In addition, the performance differences observed in the NTU-AM-CS experiment indicate that using AM data for MCI detection is relatively challenging.

Due to the nature of the self-collected dataset, making comparisons with other works is challenging. Therefore, we present results on an open dataset in the final experiment. However, the ADReSSo dataset provides only HC and AD data. Table 14 shows our results alongside those of other state-of-the-art methods. Overall, our approach yields the best results, with a 4% improvement in both accuracy and F_1-score. In addition, our precision achieves the best performance. Once again, it confirms that our data-encoding method contributes to predictive outcomes in AD and MCI detection.

XSL•FO
RenderX

Table . Result of the ADReSSo dataset[a].

Method	Accuracy	F_1-score	Precision	Sensitivity	Specificity
Syed et al [30]	0.84	0.84	N/A[b]	0.79	0.90
Rohanian et al [31]	0.84	N/A	N/A	N/A	N/A
Zhu et al [32]	0.83	0.83	0.83	0.84[a]	N/A
Ying et al [33]	0.84	0.84	0.85	0.84[a]	N/A
Ours (CS[c])	0.88[a]	0.88[a]	0.93[a]	0.84[a]	0.92[a]

[a]Highest value in each evaluation metric.

[b]N/A: not applicable.

[c]CS: cross-sectional.

Discussion

Principal Results

In this study, we proposed a longitudinal framework for MCI detection using AM test speech data. The key innovation is the ATM, which leverages direction embeddings with subject and group alignment losses to model cognitive changes across multiple visits.

Our experiments demonstrate that the longitudinal framework substantially outperforms cross-sectional approaches. Compared with single-visit models, the longitudinal model achieved notable improvements, with AUROC gains of up to 21%, F_1-score increases of 16%, and accuracy improvements of 12%. These results confirm that incorporating multi–time point information provides a clearer picture of subtle cognitive trajectories that cannot be fully captured using only cross-sectional data.

The ablation studies further highlight the effectiveness of the alignment mechanisms. Removing the alignment losses led to considerable performance degradation, particularly a 17% - 21% drop in AUROC when the group alignment loss was excluded. This underscores the critical role of alignment in ensuring temporal consistency and improving the robustness of the framework.

In addition to evaluations on our self-collected dataset, we validated the framework on the publicly available ADReSSo dataset, which uses English picture description speech. Despite differences in language and task, our model achieved more than 88% accuracy, suggesting that the approach generalizes well across languages and cognitive assessment protocols.

In our analysis, the distribution of false positives and false negatives provides additional context for interpreting the model's sensitivity and specificity. False positives, which reduce specificity, were often linked to healthy subjects who displayed atypical speech characteristics such as disfluencies, pauses, or reduced lexical richness that mimicked early cognitive decline. On the other hand, false negatives, which lower sensitivity, were frequently observed in MCI subjects who retained strong narrative skills or enriched their responses with contextual details, thereby masking subtle impairments. These observations illustrate how individual variability in speech patterns can differentially impact sensitivity and specificity, underscoring both the strengths and the boundary conditions of the framework. While the alignment-based approach improves overall robustness, further refinement of feature representations and the inclusion of personalized modeling strategies may help balance sensitivity and specificity more effectively.

Taken together, these results highlight both the methodological and clinical significance of the proposed approach. Improvements in AUROC of 10% - 15% can translate into earlier identification of at-risk individuals and a reduction in false negatives in real-world screening. This demonstrates the potential of longitudinal, dual-modal speech analysis as a noninvasive, scalable tool for long-term cognitive monitoring, with the flexibility to be integrated into routine memory clinic assessments or used as a low-cost prescreening solution to complement standardized neuropsychological testing.

Limitations

Although the proposed framework shows promising performance for early MCI detection, several limitations remain. One major limitation of this study lies in the relatively small sample size of the NTU-AM-LG dataset, which includes only 54 subjects. Although each subject contributes multiple audio recordings across different visits, the limited number of unique participants poses inherent challenges for deep learning models, including a higher risk of overfitting, model instability, and low statistical power. As a result, while the proposed model demonstrates promising performance trends, its robustness and generalizability should be interpreted with caution. To mitigate these risks, we adopted strict subject-independent data splits and cross-validation to avoid data leakage and improve reliability. Nevertheless, future studies with larger and more diverse longitudinal datasets are essential to confirm the model's generalization capability and to further validate its clinical applicability. Furthermore, we also validate our approach on the publicly available ADReSSo dataset, although it is limited to a cross-sectional study. Cross-lingual robustness also presents a challenge, as our dataset is in Chinese while ADReSSo is in English. Linguistic and cultural factors such as narrative structure, emotional expression, and lexical choices may influence AM speech patterns and, in turn, affect model generalizability. Nevertheless, our results demonstrate encouraging cross-lingual generalizability. We highlight this as an open challenge and a direction for future work, where

XSL•FO
RenderX

cross-lingual adaptation and multilingual modeling will be crucial to improving robustness across diverse populations.

Moreover, the current longitudinal dataset includes 3 visits per participant, which restricts the ability to capture long-term trajectories of cognitive change. Collecting data across additional time points would allow the development of models better suited to characterizing the progression of cognitive decline.

This study also acknowledges potential confounding factors arising from the use of automatic speech processing tools. The *whisper-timestamped* model and *pyannote* diarization tool kit were used for transcription and speaker segmentation, respectively; however, the resulting ASR and diarization errors were not formally measured. Such errors, especially the higher ASR error rates likely to occur in participants with MCI due to disfluency or reduced articulation, and diarization inaccuracies that may introduce noise, could impact the extracted linguistic and acoustic features. These unaddressed confounders represent an inherent limitation of this study.

Further Studies

Based on the current findings, several directions can be explored in future work. First, extending the framework to multilingual settings will be essential for improving cross-lingual generalizability, as language and cultural factors strongly influence AM narratives. Second, adapting the model for low-resource scenarios will enhance its applicability in real-world clinical practice, where large-scale longitudinal datasets are often unavailable. Techniques such as transfer learning, data augmentation, and self-supervised pretraining could make the system more robust under such constraints.

Third, integrating AM speech analysis with other noninvasive modalities, such as wearable sensor data, offers the potential to capture complementary behavioral and physiological signals (eg, activity levels, sleep patterns, and heart rate variability). Such multimodal fusion could provide a more holistic view of cognitive health and improve early detection of decline. Finally, future studies should focus on personalized modeling of cognitive trajectories, acknowledging that the pace and pattern of decline vary across individuals. Personalized approaches may increase the clinical relevance of the framework and support tailored interventions.

Conclusions

We developed a dual-modal longitudinal analysis system to enhance the accuracy of diagnosing MCI. The system consists of an encoder and an ATM. We used a pretrained Wav2vec 2.0 model for speech data and BERT for text data to generate latent representations. Our proposed ATM generates significantly improved performance for MCI prediction, with F_1-scores of 0.78 and 0.83 on the recall and probing datasets, respectively. Combining data from both modalities improved understanding of the AM test data, with up to 19% and 12% improvements in F_1-score compared with unimodal models. In addition, ablation experiments verified the necessity of context embedding and linguistic features in the linguistic encoder. Our model's effectiveness was consistently superior across different datasets.

Acknowledgments

This research would not have been possible without the support of the

National Science and Technology Council of Taiwan and the Center of Artificial Intelligence and Advanced Robotics, National Taiwan University. We acknowledge that generative artificial intelligence was used solely to assist with improving grammar, clarity, and English language expression in the preparation of this manuscript. The authors reviewed and approved all edits to ensure accuracy and integrity of the scientific content.

Funding

This research was funded by the National Taiwan University (grant number NTU-ICRP-115L7505), National Science and Technology Council of Taiwan and the Center for Artificial Intelligence and Advanced Robotics, National Taiwan University (grant numbers NSTC 114-2634-F-002 -005, NSTC 114-2223-E-002 -002, and NSTC 112-2410-H-002-201-MY3).

Data Availability

The autobiographical memory speech data used in this study are not publicly available due to ethical and privacy restrictions involving patient privacy and informed consent. The ADReSSo dataset is publicly accessible for DementiaBank members.

Conflicts of Interest

None declared.

References

1. Prince M, Guerchet M, Prina M. The Epidemiology and Impact of Dementia-Current State and Future Trends: WHO Thematic Briefing; 2015. URL: https://hal.science/hal-03517019/document [accessed 2023-12-30]
2. World Alzheimer Report 2019: Attitudes to Dementia: Alzheimer's Disease International; 2019. URL: https://www.alzint.org/u/WorldAlzheimerReport2019.pdf [accessed 2024-01-19]

XSL•FO
RenderX

3. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology (ECronicon) 2010 Jan 19;74(3):201-209. [doi: 10.1212/WNL.0b013e3181cb3e25] [Medline: 20042704]
4. Chandra A, Dervenoulas G, Politis M, Initiative ADN. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. J Neurol 2019 Jun;266(6):1293-1302. [doi: 10.1007/s00415-018-9016-3] [Medline: 30120563]
5. Geda YE. Mild cognitive impairment in older adults. Curr Psychiatry Rep 2012 Aug;14(4):320-327. [doi: 10.1007/s11920-012-0291-x] [Medline: 22773365]
6. Levine B, Svoboda E, Hay JF, Winocur G, Moscovitch M. Aging and autobiographical memory: dissociating episodic from semantic retrieval. Psychol Aging 2002 Dec;17(4):677-689. [Medline: 12507363]
7. Al-Hameed S, Benaissa M, Christensen H. Detecting and predicting Alzheimer's disease severity in longitudinal acoustic data. Presented at: Proceedings of the 4th International Conference on Bioinformatics Research and Applications; Dec 8-10, 2017; Barcelona Spain p. 57-61. [doi: 10.1145/3175587.3175589]
8. Ben Ammar R, Ben Ayed Y. A language-based approach for predicting Alzheimer disease severity. In: International Conference on Research Challenges in Information Science: Springer; 2021:283-294. [doi: 10.1007/978-3-030-75018-3_19]
9. Laguarta J, Subirana B. Longitudinal speech biomarkers for automated Alzheimer's detection. Front Comput Sci 2021 Apr 8;3. [doi: 10.3389/fcomp.2021.624694]
10. Yamada Y, Shinkawa K, Shimmei K. Atypical repetition in daily conversation on different days for detecting Alzheimer disease: evaluation of phone-call data from regular monitoring service. JMIR Ment Health 2020 Jan 12;7(1):e16790. [doi: 10.2196/16790] [Medline: 31934870]
11. Chang HL, Wai T, Liao YS, Lin SY, Chang YL, Fu LC. A dual-modal fusion framework for detection of mild cognitive impairment based on autobiographical memory. IEEE J Biomed Health Inform 2025 Jun;29(6):4474-4485. [doi: 10.1109/JBHI.2025.3540207] [Medline: 40031689]
12. Eyben F, Scherer KR, Schuller BW, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 2016 Apr 1;7(2). [doi: 10.1109/taffc.2015.2457417]
13. Arik S, Pfister T. TabNet: attentive interpretable tabular learning. AAAI 2021;35(8):6679-6687. [doi: 10.1609/aaai.v35i8.16826]
14. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems 2020:12449-12460 [FREE Full text]
15. Yuan J, Cai X, Bian Y, Ye Z, Church K. Pauses for detection of Alzheimer's disease. Front Comput Sci 2021;2:624488. [doi: 10.3389/fcomp.2020.624488]
16. Pistono A, Pariente J, Bézy C, Lemesle B, Le Men J, Jucla M. What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. Neuropsychologia 2019 Feb 18;124:133-143. [doi: 10.1016/j.neuropsychologia.2018.12.018] [Medline: 30593773]
17. Pastoriza-Domínguez P, Torre IG, Diéguez-Vide F, et al. Speech pause distribution as an early marker for Alzheimer's disease. Speech Commun 2022 Jan;136:107-117. [doi: 10.1016/j.specom.2021.11.009]
18. Lin SY, Chang HL, Wai T, Fu LC, Chang YL. Contrast-enhanced automatic cognitive impairment detection system with pause-encoder. Presented at: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC); Oct 9-12, 2022; Prague, Czech Republic p. 2796-2801. [doi: 10.1109/SMC53654.2022.9945085]
19. Sangchocanonta S, Vongsurakrai S, Sroykhumpa K, et al. Development of Thai picture description task for Alzheimer's screening using part-of-speech tagging. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov:2104-2109. [doi: 10.1109/EMBC46164.2021.9629861]
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) 2019:4171-4186 [FREE Full text] [doi: 10.18653/v1/N19-1423]
21. Berna F, Schönknecht P, Seidl U, Toro P, Schröder J. Episodic autobiographical memory in normal aging and mild cognitive impairment: a population-based study. Psychiatry Res 2012 Dec 30;200(2-3):807-812. [doi: 10.1016/j.psychres.2012.03.022] [Medline: 22520853]
22. Irish M, Lawlor BA, O'Mara SM, Coen RF. Impaired capacity for autonoetic reliving during autobiographical event recall in mild Alzheimer's disease. Cortex 2011 Feb;47(2):236-249. [doi: 10.1016/j.cortex.2010.01.002] [Medline: 20153463]
23. Ouyang J, Zhao Q, Adeli E, Sullivan EV, Pfefferbaum A, Zaharchuk G, et al. Self-supervised longitudinal neighbourhood embedding. In: Medical Image Computing and Computer Assisted Intervention– MICCAI 2021: 24th International Conference 2021:80-89. [doi: 10.1007/978-3-030-87196-3_8]
24. Zhao Q, Liu Z, Adeli E, Pohl KM. Longitudinal self-supervised learning. Med Image Anal 2021 Jul;71:102051. [doi: 10.1016/j.media.2021.102051] [Medline: 33882336]
25. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. arXiv. Preprint posted online on Apr 23, 2020. [doi: 10.48550/arXiv.2004.11362]

XSL•FO
RenderX

26. Leyhe T, Müller S, Milian M, Eschweiler GW, Saur R. Impairment of episodic and semantic autobiographical memory in patients with mild cognitive impairment and early Alzheimer's disease. Neuropsychologia 2009 Oct;47(12):2464-2469. [doi: 10.1016/j.neuropsychologia.2009.04.018] [Medline: 19409401]
27. Urbanowitsch N, Gorenc L, Herold CJ, Schröder J. Autobiographical memory: a clinical perspective. Front Behav Neurosci 2013 Dec 10;7:194. [doi: 10.3389/fnbeh.2013.00194] [Medline: 24339804]
28. Goodglass H, Kaplan E, Weintraub B. Boston Diagnostic Aphasia Examination: Lippincott Williams & Wilkins; 2001.
29. Luz S, Haider F, Fuente SDL, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: The ADReSSo Challenge. MedRxiv. Preprint posted online on 2021. [doi: 10.1101/2021.03.24.21254263]
30. Syed ZS, Syed MSS, Lech M, Pirogova E. Tackling the ADRESSO challenge 2021: the MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech. Presented at: Interspeech 2021; Aug 30 to Sep 3, 2021; Brno, Czechia p. 3815-3819. [doi: 10.21437/Interspeech.2021-1572]
31. Rohanian M, Hough J, Purver M. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. Presented at: Interspeech 2021; Aug 30 to Sep 3, 2021 URL: https://www.isca-archive.org/interspeech_2021/ [doi: 10.21437/Interspeech.2021-1633]
32. Zhu Y, Obyat A, Liang X, Batsis JA, Roth RM. WavBERT: exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection. Interspeech 2021;2021:3790-3794. [doi: 10.21437/interspeech.2021-332] [Medline: 37063977]
33. Ying Y, Yang T, Zhou H. Multimodal fusion for Alzheimer's disease recognition. Appl Intell 2023 Jun;53(12):16029-16040. [doi: 10.1007/s10489-022-04255-z]

Abbreviations

AD: Alzheimer disease
AM: autobiographical memory
ASR: automatic speech recognition
ATM: aging trajectory module
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representation for transformer
CS: cross-sectional
LG: longitudinal
MCI: mild cognitive impairment
NTU-AM: National Taiwan University Autobiographical Memory Test
NTU-AM-CS: National Taiwan University Autobiographical Memory Test Cross-Sectional
NTUH-MEM: National Taiwan University Hospital Memory Test
NTUH-PIC: National Taiwan University Hospital Picture Description Test
POS: part-of-speech

Edited by A Coristine; submitted 26.Jul.2025; peer-reviewed by B Jing, J Shi; revised version received 21.Nov.2025; accepted 25.Nov.2025; published 15.Jan.2026.

Please cite as:
Liao YS, Wai T, Liao TY, Chang HL, Chang YL, Fu LC
Mild Cognitive Impairment Detection System Based on Unstructured Spontaneous Speech: Longitudinal Dual-Modal Framework
JMIR Med Inform 2026;14:e80883
URL: https://medinform.jmir.org/2026/1/e80883
doi:10.2196/80883

© Yu-Shan Liao, Thiri Wai, Ting-Yun Liao, Ho-Ling Chang, Yu-Ling Chang, Li-Chen Fu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Anesthetic Depth Assessment via Unsupervised Machine Learning in Processed Electroencephalography Analysis: Novel Methodological Study

Po-Yu Huang[1], MD; Wei-Lun Hong[2], MS; Hui-Zen Hee[3], MD; Wen-Kuei Chang[1], MD; Ching-Hung Lee[2,4], PhD; Chien-Kun Ting[1,5], MD, PhD

[1]Department of Anesthesiology, Taipei Veterans General Hospital, No. 201, Sec 2, Shipai Rd, Beitou District, Taipei City, Taiwan

[2]Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan

[3]Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[4]Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan

[5]Institute of Emergency and Critical Care Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan

Corresponding Author:
Chien-Kun Ting, MD, PhD
Department of Anesthesiology, Taipei Veterans General Hospital, No. 201, Sec 2, Shipai Rd, Beitou District, Taipei City, Taiwan

Abstract

Background: General anesthesia comprises 3 essential components—hypnosis, analgesia, and immobility. Among these, maintaining an appropriate hypnotic state, or anesthetic depth, is crucial for patient safety. Excessively deep anesthesia may lead to hemodynamic instability and postoperative cognitive dysfunction, whereas inadequate anesthesia increases the risk of intraoperative awareness. Electroencephalography (EEG)-based monitoring has therefore become a cornerstone for evaluating anesthetic depth. However, processed electroencephalography (pEEG) indices remain vulnerable to various sources of interference, including electromyographic activity, interindividual variability, and anesthetic drug effects, which can yield inaccurate numerical outputs.

Objective: With recent advances in machine learning, particularly unsupervised learning, data-driven methods that classify signals according to inherent patterns offer new possibilities for anesthetic depth analysis. This study aimed to establish a methodology for automatically identifying anesthesia depth using an unsupervised, machine learning–based clustering approach applied to pEEG data.

Methods: Standard frontal EEG data from participants undergoing elective lumbar spine surgery were retrospectively analyzed, yielding more than 16,000 data points. The signals were filtered with a fourth-order Butterworth bandpass filter and transformed using the fast Fourier transform to estimate power spectral density. Normalized band power ratios for delta, high-theta, alpha, and beta frequencies were extracted as input features. Fuzzy C-Means (FCM) clustering (c=3, m=2) was applied to categorize anesthetic depth into slight, proper, and deep clusters.

Results: FCM clustering successfully identified 3 physiologically interpretable clusters consistent with EEG dynamics during progressive anesthesia. As anesthesia deepened, frontal alpha oscillations became more prominent within a delta-dominant background, while beta activity decreased with loss of consciousness. The fuzzy membership values quantified transitional states and captured the continuum of anesthetic depth. Visualization confirmed strong correspondence among cluster transitions, Patient State Index trends, and spectral density patterns.

Conclusions: This study demonstrates the feasibility of using unsupervised machine learning to enhance anesthetic depth assessment. By applying FCM clustering to pEEG data, this approach improves the understanding of anesthesia depth and integrates effectively with existing monitoring modalities. The proposed FCM-based method complements current EEG indices and may assist anesthesia practitioners and even nonanesthesia professionals in assessing anesthetic depth to enhance patient safety.

(JMIR Med Inform 2026;14:e77830) doi:10.2196/77830

KEYWORDS

artificial intelligence; cluster analysis; electroencephalography; general anesthesia; intraoperative monitoring; machine learning

XSL•FO
RenderX

Introduction

General anesthesia induces temporary loss of consciousness characterized by unconsciousness, immobility, and autonomic regulation of pain [1,2]. Maintaining an appropriate depth of anesthesia is essential for patient safety. Insufficient anesthesia may lead to intraoperative awareness, autonomic hyperactivity, and significant psychological distress [3]. Conversely, excessive anesthesia has been associated with neuroinflammation [4], neuronal injury, and postoperative cognitive dysfunction or delirium [5,6]. Therefore, continuous monitoring of anesthetic depth during surgery is of critical importance. Electroencephalography (EEG)-based monitoring serves as a valuable tool for this purpose. Real-time EEG analysis provides insight into anesthetic states, as oscillatory patterns vary with the anesthetic dose, type, and patient characteristics [7,8]. The Perioperative Quality Initiative 6 Workgroup, led by Chan, details the available EEG monitors [9]. Understanding perioperative neural activity enables anesthesiologists to optimize anesthesia delivery and enhance patient outcomes.

Artificial intelligence is revolutionizing anesthesiology, with techniques such as anesthesia depth monitoring [10]. Unsupervised machine learning operates without predefined hypotheses and identifies patterns and structures within datasets to classify patients, medications, and groups [11]. Unlike supervised machine learning, which correlates the input data with known outcomes, unsupervised machine learning categorizes instances based on inherent patterns. In EEG analyses, unsupervised machine learning methods such as fuzzy clustering have been used to analyze sleep stages [12], detect epileptic spikes [13], and identify significant elements in multichannel recordings [14].

Anesthesia management fundamentally involves 3 key components: hypnosis, analgesia, and muscle relaxation. During induction and maintenance, a reproducible spectral shift occurs in the EEG, wherein frontal alpha oscillations become prominent within a delta-dominant background, delta power progressively increases with deepening anesthesia, and beta activity decreases as consciousness is lost [8,15-17]. Power spectral metrics—including relative alpha, beta, and delta power, power ratios, and spectral entropy—have been shown to reliably distinguish between conscious and anesthetized states [18-20]. Inhalational agents such as sevoflurane and desflurane consistently produce dose-dependent increases in frontal alpha and delta power [17,21-23], providing a physiologic basis for categorizing anesthesia into light, proper, and deep stages. However, current perioperative EEG-based indices, such as the Bispectral Index (BIS) [24,25] and Patient State Index (PSi) [26,27], primarily reflect the hypnotic component but can be influenced by various factors, potentially leading to misleading processed numerical outputs. Therefore, this study aimed to enhance anesthetic depth assessment by focusing on the hypnosis dimension and applying unsupervised machine learning methods, specifically Fuzzy C-Means (FCM) clustering, to processed electroencephalography (pEEG) data.

The primary objective of this study was to establish a methodology for automatically identifying anesthesia depth through a visualized presentation, leveraging machine learning–based clustering techniques. By applying unsupervised learning to pEEG data, this approach aims to enhance the classification and understanding of anesthesia depth and facilitate patient-specific anesthesia management.

Methods

Recruitment

Data were collected at the Veterans General Hospital, Taipei, with the approval of the Institutional Review Board (number 2021-09-008BC). More than 16,000 data points were collected from 10 participants (Participants 1 - 10) who underwent elective lumbar spinal surgery in the prone position. The ages of the participants ranged from 44 to 78 years, with the American Society of Anesthesiologists' physical status classifications ranging from class I to class III. Participants with conditions such as pregnancy, neurological disorders, ongoing infections, a history of central nervous system medication use, respiratory failure, heart failure, or renal failure were excluded.

Masimo SedLine [28] brain function monitoring was performed with the participant in the supine position. Induction included fentanyl at 1 - 3 mcg/kg as an adjunct for intubation, followed by propofol at 1 - 3 mg/kg as an induction bolus. Rocuronium was administered at an intubation dose of 0.6 - 1 mg/kg for muscle relaxation. The participant was then intubated.

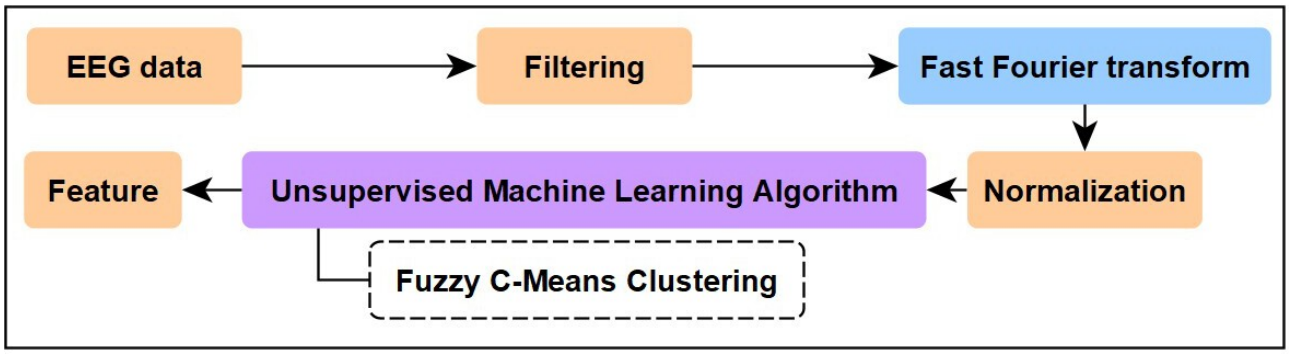
During maintenance, the inhalational agent sevoflurane was administered, with dosage adjustments titrated based on the PSi levels and pEEG density spectral array (DSA) patterns. The participant was then placed in a prone position for lumbar spinal surgery. After the completion of the surgery, the participant was returned to the supine position, and the inhalation agent was gradually titrated off. Rocuronium administration was reversed using sugammadex. Subsequently, the participants were extubated and transferred to the postanesthetic care unit for further monitoring and recovery.

Data Collection and Processing

Figure 1 outlines the preprocessing steps for the EEG data in our study.

XSL•FO
RenderX

Figure 1. Signal processing flow diagram. Flow diagram illustrating preprocessing and feature-extraction steps applied to processed electroencephalography data obtained from participants who underwent elective lumbar spine surgery at Taipei Veterans General Hospital. Raw frontal electroencephalography signals were filtered with a fourth-order Butterworth bandpass filter, and power spectral density was estimated using the Welch method with a 30-second window and 2-second overlap. Normalized delta, high-theta, alpha, and beta band power ratios were extracted as input features for unsupervised clustering using the Fuzzy C-Means (FCM) algorithm (*c=3*, *m=2*). EEG: electroencephalography.



Signal Processing and Feature Extraction

EEG tracings from the 4 Masimo SedLine channels were averaged into 1-dimensional signals and synchronized for consistency. Each participant's EEG data were recorded at a sampling frequency of 178 Hz, with a data collection period ranging from 3 to 5 hours, depending on the duration of anesthesia. A fourth-order Butterworth bandpass filter (0.5 - 30 Hz) was applied to minimize the ripple and achieve a maximally flat response. Power spectral density was estimated using the Welch method implemented in Python (SciPy library), using a 30-second window with a 2-second sliding overlap to ensure both temporal continuity and frequency resolution. Additional preprocessing specifications are available in Multimedia Appendix 1. The Welch method and the fast Fourier transform were used to estimate the power spectral density, facilitating the extraction of alpha (8 - 12 Hz), beta (13 - 30 Hz), delta (0.5 - 4 Hz), and high-theta (6 - 8 Hz) band powers. Normalization was performed by calculating the ratio of each band's power to the sum of the power of these 4 specific bands (delta+high-theta+alpha+beta), ensuring that the sum of the feature vector elements equals 1. The 4 - 6 Hz and 12 - 13 Hz bands were excluded based on their limited variation across states. Removing these transitional zones established a "spectral buffer" to reduce noise, allowing the algorithm to prioritize the more discriminative delta and alpha and high-theta features. Manual artifact rejection was conducted by board-certified anesthesiologists and engineers to identify and exclude segments exhibiting abnormal, noisy epochs indicative of artifacts.

FCM Clustering Algorithm

FCM clustering categorizes data points into clusters with membership values between 0 and 1, accommodating the uncertainty in the cluster assignment. It was performed with 3 clusters (c=3), a fuzziness parameter (m) of 2, random initialization of cluster centers, and a convergence threshold of 0.005. The model iteratively updates the centroids and membership values until convergence, guided by a fuzziness parameter (m) that regulates cluster ambiguity. A global FCM model was established to generate fixed global cluster centroids. These fixed centroids were then applied to all participants to calculate membership values. Further implementation details are provided in Multimedia Appendix 1. This method extracts delta, alpha, high-theta, and beta band power ratios from pEEG signals to partition the data into 3 statistically distinct clusters. These clusters were subsequently labeled as "slight," "proper," and "deep" based on their physiological spectral characteristics. The capability of FCM to manage data ambiguity makes it suitable for nuanced depth assessments under general anesthesia.

Ethical Considerations

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and was approved by the Institutional Review Board of Taipei Veterans General Hospital, Taiwan (number: 2021-09-008BC). Written informed consent was obtained from all participants prior to enrollment. The data used in this study were deidentified before analysis to ensure participant confidentiality and privacy. No identifiable personal information or images were included in the manuscript or supplementary materials. Participants did not receive any financial compensation for their participation in the study.

Results

Overview

In this study, an unsupervised learning method, FCM, was used to analyze the pEEG data from participants with complete datasets. Table 1 summarizes the demographic information and baseline characteristics of the participants. The analysis focused on a feature set comprising the delta-band power ratio and the sum of the alpha and high-theta band power ratios. The results revealed 3 distinct clusters. Based on the centroids' spectral features, we assigned post hoc clinical labels to these clusters representing different anesthesia depths: slight (green), proper (blue), and deep (red). This clustering approach was validated using visual representation to confirm the effectiveness of the feature set in differentiating anesthesia depths. Subsequently, the same method was applied to individual participant data, categorizing each participant's pEEG into 1 of the 3 clusters, thus providing a detailed estimation of the anesthesia depth for each individual.

XSL•FO
RenderX

Table . Demographic information and baseline characteristics. Demographic and intraoperative characteristics of 10 participants included in this retrospective observational study of processed electroencephalography–based anesthetic-depth assessment. All participants underwent elective lumbar spine surgery at Taipei Veterans General Hospital, Taiwan. Variables include age, sex, height, weight, BMI, American Society of Anesthesiologists classification, comorbidities, intraoperative blood loss, operation time, and postoperative hospital stay.

Participant	Age (years)	Sex	Height (cm)	Body weight (kg)	BMI (kg/m^2)	ASA[a] classification	Comorbidities	Blood loss (mL)	Operation time (minutes)	Postoperative stay (days)
1	63	F[b]	152.2	63.7	27.5	I	None	300	250	7
2	78	F	150.5	53.5	23.6	II	Hepatitis	200	280	6
3	65	M[c]	168	89.8	31.8	II	None	300	360	4
4	75	M	170	70	24.2	II	Hypertension, dementia	660	405	7
5	67	M	160	64	25.0	II	Hypertension	200	285	6
6	45	M	171	69.9	23.9	I	None	[d]—	170	2
7	61	M	160.3	77.7	30.2	II	Hypertension, diabetes mellitus	1300	195	6
8	68	F	157.7	71.1	28.6	II	Hypertension, diabetes mellitus	450	210	5
9	44	M	173.6	67.6	22.4	I	None	50	375	29
10	71	M	167	83	29.8	III	Hypertension, coronary artery disease	530	320	6

[a]ASA: American Society of Anesthesiologists.

[b]F: female.

[c]M: male.

[d]Not available.

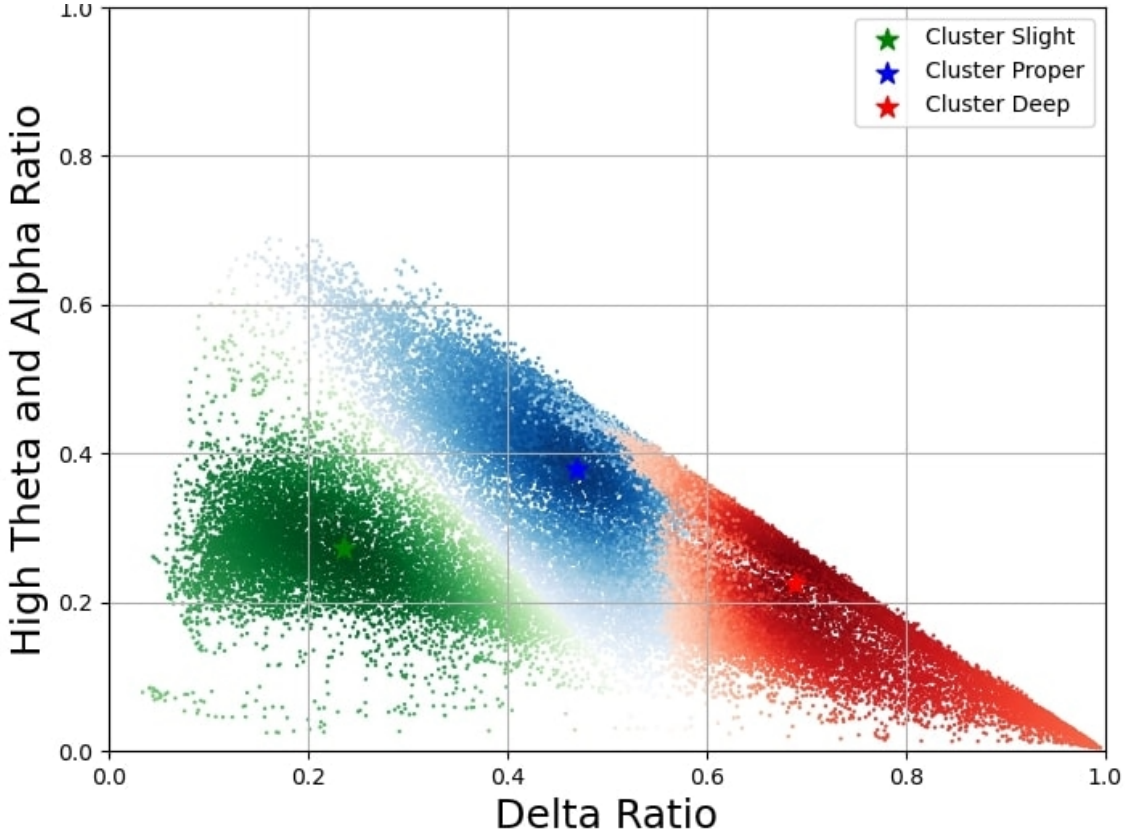
FCM Clustering Analysis of pEEG Features

FCM clustering was applied to all participants, using alpha, beta, delta, and high-theta band power as the selected features for clustering. Figure 2 illustrates the FCM clustering results for all participants, with the delta band power ratio plotted on the *x*-axis and the sum of the alpha and high-theta band power ratios on the *y*-axis. Each data point represents a single time point of a participant's pEEG signal and is clustered into 3 distinct groups: slight anesthesia depth (green), proper anesthesia depth (blue), and deep anesthesia depth (red). The contribution of the beta band power ratio can be inferred as the sum of the delta, alpha, high-theta, and beta band power ratios equals 1, although not explicitly displayed. This visual depiction offers insights into the distribution of a participant's single time points within each anesthesia depth cluster.

After establishing the efficacy of the selected features and parameters through a collective analysis of the participants, our investigation shifted focus to the individual application of this methodology. The fixed global cluster centroids were then applied to the data of Participants 1 - 10 to calculate membership values and simplified estimation results. Following the methodology outlined, where all time indices share a common time scale, the pEEG data were combined with PSi data retrieved from the Masimo SedLine brain function monitoring system, and the time-frequency images were compared within the same figure, with each participant outputting a distinct figure. Within these figures, Participants 4 and 6 are described in the article. By incorporating this comprehensive approach, our study verified the applicability of the selected features and enhanced confidence in the adaptability of our methodology to diverse clinical scenarios and datasets, strengthening the reliability of our conclusions and paving the way for further exploration and validation within the anesthesia depth assessment domain.

XSL•FO
RenderX

Figure 2. Fuzzy C-Means clustering results for all participants. Two-dimensional clustering distribution of processed electroencephalography features from participants undergoing elective lumbar spine surgery. The *x*-axis represents the delta-power ratio and the *y*-axis the combined alpha and high-theta power ratios. Fuzzy C-Means clustering (*c=3*, *m=2*) identified 3 physiologically interpretable clusters corresponding to anesthesia-depth levels: slight (green), proper (blue), and deep (red). Each point denotes a 30-second windowed segment of electroencephalography data, demonstrating the continuum of anesthetic states observed across the study population.



Participant 4 and Participant 6: FCM Clustering Analysis

Participant 4 underwent the same data retrieval and processing procedures as the other participants, with raw EEG data obtained from the Masimo SedLine brain function monitoring system. These data were subjected to filtering, fast Fourier transformation, normalization, and subsequent FCM clustering. The culmination of these processes is depicted in Figure 3, which comprises multiple panels, sharing a synchronized timescale on the *x*-axis.

Figure 3A shows the PSi of Participant 4 obtained from the Masimo SedLine brain function monitoring system. The *x*-axis represents the time scale, whereas the *y*-axis displays the PSi values, reflecting the PSi over time.

Figure 3B shows the results of FCM clustering for Participant 4. The *y*-axis illustrates the estimated membership values and denotes the degree of belongingness of each cluster. This belongingness quantifies the "fuzziness" of the FCM clustering, indicating the confidence level of each assignment of the data point to a particular cluster. It effectively quantifies the "color shading" shown in Figure 2, where darker shades represent higher confidence levels in cluster assignments. For example, a membership value of 1 signifies full belongingness to the cluster, and a value of 0.8 indicates 80% confidence in the assignment. This "soft" labeling approach allows for a more nuanced understanding of the clustering results, providing insights into the uncertainty associated with each classification of the data point.

Figure 3C represents a modified rendition of Figure 3B, specifically tailored to illustrate the anesthesia depth classification for Participant 4. Along the *y*-axis, this plot delineates the categorization into slight, proper, or deep anesthesia groups at individual time points. Figure 3B provides a detailed depiction of the continuous membership values, reflecting the degree of belonging to each cluster: slight, proper, and deep. However, this comprehensive display introduces the challenge of intuitively discerning the predominant cluster at a single point. While this approach offers a more complete dataset, it complicates the task of determining the cluster affiliations for each data point. Hence, the modified version in Figure 3C simplifies this interpretation by highlighting the cluster with the "higher value." This adaptation facilitates a more straightforward assessment of the anesthesia depth classification by presenting a clearer distinction. This clarity enhances the utility of visualization in aiding clinician decision-making during patient monitoring and management.

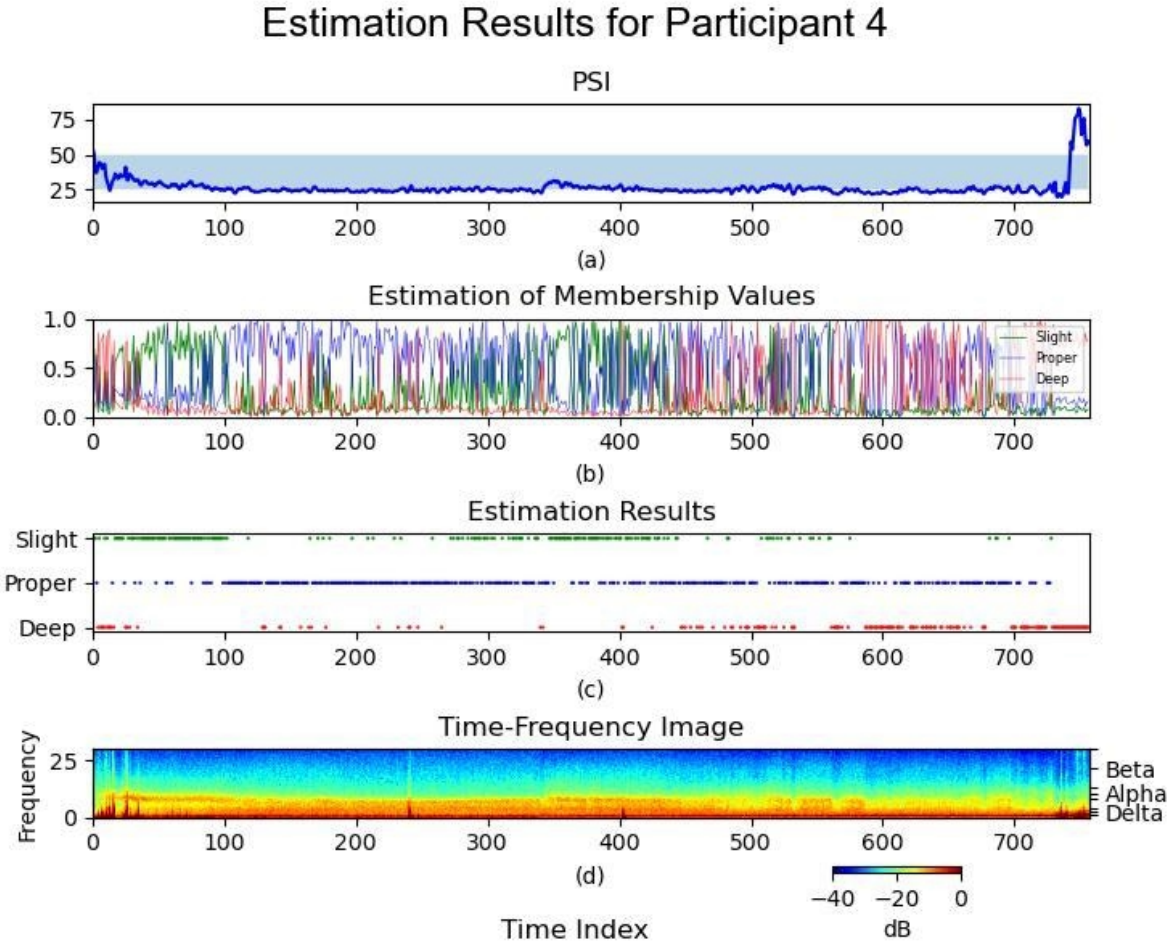
Figure 3D presents a time-frequency image derived from the raw EEG signal without processing, providing insights into the frequency components of the EEG signal over time. Overall, Figure 3 provides a comprehensive analysis of the EEG signals and anesthesia depth classification for Participant 4, contributing

XSL•FO
RenderX

to the understanding of brain function monitoring and anesthesia management.

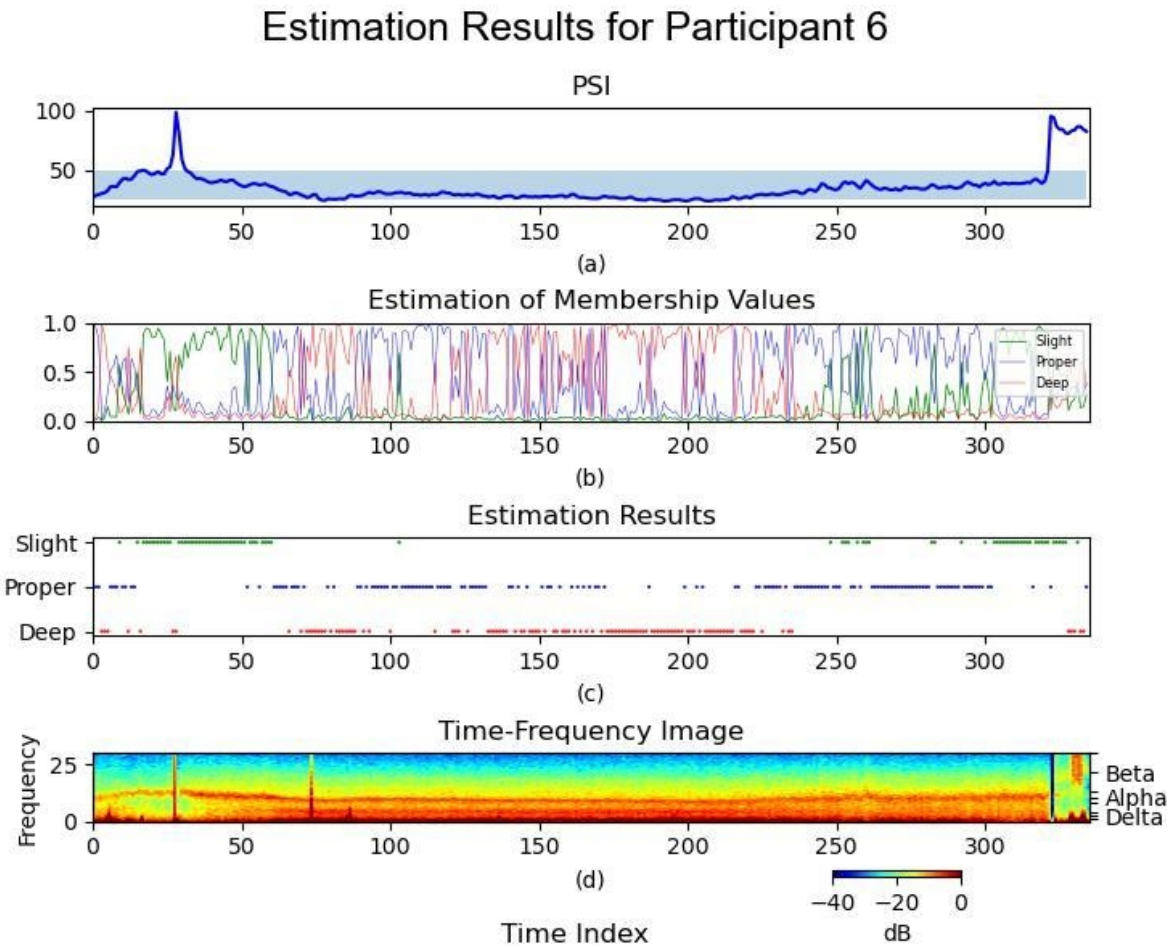
Participant 6 then underwent identical signal processing and analysis procedures, resulting in Figure 4. This comprehensive Figures 4A-D underwent the same rigorous process as shown in Figures 3A-D.

Figure 3. Fuzzy C-Means clustering analysis of Participant 4. Representative analysis of Participant 4 (male, 75 years, American Society of Anesthesiologists class II) who underwent lumbar spine surgery at Taipei Veterans General Hospital. (**A**) shows the Patient State Index trend over time; (**B**) displays Fuzzy C-Means membership values representing the degree of belongingness to each cluster; (**C**) illustrates the simplified classification into slight, proper, and deep anesthesia levels; (**D**) presents the corresponding electroencephalography time-frequency spectrogram. These plots depict temporal transitions in anesthetic depth.



XSL•FO
RenderX

Figure 4. Fuzzy C-Means clustering analysis of Participant 6. Representative analysis of Participant 6 (male, 45 years, American Society of Anesthesiologists class I) who underwent elective lumbar spine surgery at Taipei Veterans General Hospital. (**A**) shows the Patient State Index trend over time. (**B**) displays Fuzzy C-Means membership values representing the degree of belongingness to each cluster. (**C**) illustrates the simplified classification into slight, proper, and deep anesthesia levels. (**D**) presents the corresponding electroencephalography time-frequency spectrogram. These plots depict temporal transitions in anesthetic depth. PSi: Patient State Index.



Discussion

Principal Findings

This study explored the feasibility of improving anesthetic depth assessment by applying unsupervised machine learning to pEEG data. EEG patterns vary with the depth of anesthesia [10]. PEEG-derived indices, such as the BIS and PSi, are susceptible to various influences, presenting limitations in their interpretation [24]. Purdon et al [8] characterized the EEG signatures of intravenous and inhaled anesthetics, providing additional insights into the depth of anesthesia. In Figure 2, the delta ratio quantifies cortical suppression, as delta oscillations reflect deep anesthesia and thalamocortical inhibition. The theta and alpha ratios were selected for their modulation by anesthetics in the frontal cortex. Alpha oscillations indicate stable unconsciousness under propofol and sevoflurane, correlating with GABAergic inhibition, while theta oscillations become prominent at higher anesthetic concentrations, particularly with sevoflurane, signaling transitions between moderate and deep anesthesia. Figures 3 and 4 depict Participants 4 and 6, respectively, demonstrating the combination of PSi values, anesthesia depth clustering values, simplified clustering results, and processed time-frequency images. These visual presentations illustrate patient-specific anesthesia depth, facilitating the automatic identification of anesthesia depth.

Postoperative delirium and neurocognitive complications after spinal surgery remain significant concerns [29], and perioperative EEG monitoring may contribute to their management [30]. In this study, one participant (Participant 4) had a history of cognitive dysfunction due to dementia. All participants underwent lumbar spinal surgery in the prone position with sevoflurane as the maintenance agent. The operation time ranged from 170 to 405 minutes. Most participants were discharged within 1 week, and no other cognitive impairments were documented.

Limitations

This study was constrained by a small sample size. While the extensive dataset of over 16,000 EEG epochs ensures high within-subject data density and supports the internal stability of the model for the analyzed participants, we acknowledge that the limited number of independent participants restricts the external generalizability of the findings. Previous studies have indicated that a sampling frequency of 100 Hz is sufficient for EEG-based analysis [31]. While a larger sample size could

XSL•FO
RenderX

further refine the FCM clustering results presented in Figure 1, the primary determinants of clustering effectiveness in single-subject EEG analysis are the duration and density of the data. With 3 to 5 hours of continuous EEG recordings per participant, our study provided sufficient granularity for meaningful intrasubject clustering, as demonstrated in Figures 3 and 4. Our primary objective was to explore the feasibility and establish the methodological framework for applying unsupervised machine learning to processed EEG data.

García et al [32] explored EEG activity and delineated an analgesia-related axis alongside hypnosis during general anesthesia, revealing distinct changes in EEG patterns. We acknowledge the limitation of a single-dimensional index in capturing the full spectrum of neural activity during anesthesia. However, our primary goal is to demonstrate the feasibility of unsupervised clustering within a clinical framework, focusing on the hypnosis axis—one of anesthesia's 3 key components. This study used frontal EEG signals, as used in processed systems such as BIS and SedLine. Frontal EEG monitoring inherently limits spatial coverage but remains the clinical standard due to its accessibility and reliability. The frontal cortex exhibits predictable EEG dynamics—α-spindle emergence and δ-wave dominance—with increasing anesthetic depth, correlating with loss of consciousness and hypnotic adequacy [33-35].

While this study does not include external validation or direct quantitative comparisons, this reflects the broader challenge in anesthetic depth classification, which lacks a universal standard. Although no intraoperative recall was observed, determining the optimal depth of anesthesia remains complex. Conventional monitors, such as BIS and SedLine, along with DSA visualization, aid clinical judgment but do not serve as definitive references. The study introduces a complementary tool to enhance anesthetic depth assessment. This is reflected in Figures 3 and 4, where multiple monitoring modalities are integrated to provide additional insights.

Finally, the frame-by-frame analysis maximizes temporal sensitivity but may result in rapid state transitions, or "flickering," as observed in Figure 3C. For clinical application, future iterations would likely require temporal smoothing algorithms to stabilize the output and minimize alarm fatigue. Despite strong visual correspondence with PSi trends, standardized quantitative metrics were not calculated due to methodological constraints. Specifically, the significant time delay inherent in the PSi algorithm [36] limits direct linear correlation with the frame-by-frame FCM method, and imposing rigid thresholds for concordance rates contradicts the continuous, fuzzy nature of anesthetic depth.

Comparison With Prior Work

Unsupervised learning involves algorithms that identify patterns or structures within a dataset, facilitating the discovery of novel classification methods [10]. Systematic reviews have highlighted the effectiveness of unsupervised clustering methods in various medical applications. Clustering analysis has revealed pivotal features underlying the progression from early stage to advanced Alzheimer disease [37]. Unsupervised learning models applied in cardiac resynchronization therapy can accurately discern patient clusters and are valuable for phenotyping and predicting treatment responses [38]. In amyotrophic lateral sclerosis research, the k-means score, a hierarchical clustering method, has emerged as the most widely used method for stratification and progression prediction [39]. We predefined 3 practical categories—slight, proper, and deep anesthesia—to emulate the structured reasoning process that governs intraoperative anesthetic management. Clinicians continuously evaluate anesthetic depth and determine whether to increase, maintain, or decrease anesthetic delivery. Each titration is typically incremental and followed by a brief observation period to allow pharmacokinetic redistribution and physiological stabilization before reevaluation. Although the equilibration interval varies among individuals, such judgments integrate both the instantaneous anesthetic state and its temporal evolution [36]. FCM clustering offers distinct advantages for modeling this continuum. It enables classification into clinically interpretable states while concurrently generating fuzzy membership values that quantify the degree to which a given epoch inclines toward a "lighter" or "deeper" level of anesthesia. This property captures the continuous and overlapping nature of anesthetic depth and mirrors the gradual, feedback-driven adjustments that characterize intraoperative anesthesia management.

In our continuous monitoring data, transitions between anesthetic states occurred gradually, consistent with the physiological continuum of anesthesia depth. Abrupt shifts from "light" directly to "deep" anesthesia were uncommon and, when observed, were generally attributable to identifiable physiological perturbations—such as bolus drug administration, intense noxious stimulation, or sudden hemodynamic changes—rather than modeling artifacts. Processed EEG indices may also be transiently affected by electromyographic [40], electrocardiographic [41], or mechanical vibration artifacts, particularly during spine surgery. These interferences are usually recognizable in the raw EEG tracing and can be excluded from analysis.

Near-equal memberships across clusters (≈33% each) are uncommon and typically occur when a feature vector lies between centroids or during transient noise, reflecting a transitional or indeterminate phase rather than a distinct anesthetic state. Interpretation of anesthetic depth is inherently pattern-based, as clinicians assess temporal trends and spectral features in the DSA—for instance, alpha dominance suggests adequate hypnosis, whereas delta and theta predominance or burst suppression indicate a deeper anesthetic level [42]. Therefore, a single epoch with near-equal memberships would not influence clinical judgment, which relies on the overall evolution of EEG patterns rather than isolated points.

Prior research has focused on developing and enhancing automated system controls, including closed-loop anesthesia. Components such as measurement devices, controllers, and actuators are central to these advancements [43]. pEEG-derived indices, such as the BIS [43,44] and PSi [45], frequently serve as key targets. This study introduced unsupervised learning clustering as a potential new approach for assessing anesthesia depth, offering additional avenues for improving automated anesthesia control systems.

XSL•FO
RenderX

The novel clustering approach provides an alternative method for assessing anesthesia depth, particularly focusing on the "hypnosis" axis, as depicted in Figure 3B and C. These figures illustrate the EEG density spectral array data from Participant 4, suggesting segmentation into slight, proper, and excessively deep anesthesia depths. Unlike traditional PSi values, which provide singular metrics, the representation of Figure 3B with fuzziness on the *y*-axis shows nuanced membership toward anesthesia depth clusters. The graphical depiction in Figure 3C highlights the peak values, indicating the period of deep anesthesia. This approach offers a novel tool for assisting clinical judgment in anesthesia depth assessment.

Among conventional unsupervised clustering paradigms, FCM, Gaussian Mixture Models (GMMs), and density-based algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) exhibit fundamentally distinct analytical mechanisms. FCM minimizes a distance-weighted objective function while permitting graded memberships across clusters—an essential property for anesthesia-related data in which EEG and physiological variables vary continuously and overlap across anesthetic states. In contrast, GMM represents data as a superposition of Gaussian components and assumes that each cluster conforms to an elliptical, symmetric distribution. However, anesthetic EEG features often display non-Gaussian characteristics, including skewness and heavy tails arising from nonlinear pharmacodynamic effects and interindividual variability, rendering GMM less robust and prone to boundary instability. Density-based algorithms such as DBSCAN, although effective in identifying discrete or spatially separated clusters, depend on fixed density thresholds and therefore perform poorly with continuous-depth data, where transitions between anesthetic states are gradual and density gradients evolve smoothly. Considering these properties, FCM provides a mathematically rigorous and physiologically consistent framework for delineating the smooth, overlapping transitions that define anesthetic depth.

Building upon these theoretical considerations, we initially explored k-means clustering but found it less flexible in classifying anesthetic depth. When visualizing frequency band proportions, k-means struggled to capture transition zones effectively, reinforcing our decision to use FCM for a more adaptable classification. The FCM algorithm stands out in unsupervised machine learning because of its adaptability and effectiveness in increasing data volumes. It iteratively adjusts centroids and clusters, creating personalized representations tailored to individuals, analogous to bespoke garments [46]. This adaptability ensures an accurate assessment of anesthesia depth dynamics, which is crucial for personalized healthcare.

Conclusions

In conclusion, this study established a machine learning–based methodology for anesthesia depth assessment, demonstrating its feasibility and providing preliminary insights into classification, visualization, and patient-specific management. By applying FCM clustering to processed EEG data, this novel approach enhances the understanding of anesthesia depth patterns and integrates with existing monitoring modalities, including PSi values and time-frequency image visualization for a more comprehensive assessment. Although preliminary, these findings lay the foundation for future research and refinement.

Acknowledgments

The authors declare the use of generative artificial intelligence in the writing process. According to the GAIDeT taxonomy (2025), the task of "Proofreading and editing" was delegated to ChatGPT (OpenAI) under full human supervision, specifically for grammatical error correction and phrasing improvements. The authors reviewed and edited the output as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the National Science and Technology Council (NSTC), Taiwan [Grant number: NSTC 112-2314-B-A49 -068 -MY3].

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PYH, WLH, CHL, CKT
Methodology: PYH, WLH, HZH, CHL, CKT
Investigation: PYH, HZH, WLH
Formal analysis: PYH, WLH
Data curation: HZH
Writing – original draft: PYH
Writing – review and editing: CKT, CHL
Supervision: WKC, CKT, CHL
Project administration: WKC, CKT, CHL

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed data preprocessing workflow and Fuzzy C-Means (FCM) parameter settings. Electromyography signals were processed via the Welch method (*30-s windows, 2-s overlap*). Normalized delta, high-theta, alpha, and beta band-power ratios were extracted as input features. The FCM algorithm ($c = 3$, $m = 2$) was applied with random initialization of cluster centers to identify three anesthesia depth clusters. Measures of dispersion for the delta band within each cluster are summarized in Table S1. Figures S1 and S2 show additional two-dimensional clustering distributions, demonstrating consistent three-cluster separation across different frequency-band pairings.

[PDF File, 304 KB - medinform_v14i1e77830_app1.pdf]

References

1. Brown EN, Lydic R, Schiff ND. General anesthesia, sleep, and coma. N Engl J Med 2010 Dec 30;363(27):2638-2650. [doi: 10.1056/NEJMra0808281] [Medline: 21190458]
2. Egan TD, Svensen CH. Multimodal general anesthesia: a principled approach to producing the drug-induced, reversible coma of anesthesia. Anesth Analg 2018 Nov;127(5):1104-1106. [doi: 10.1213/ANE.0000000000003743] [Medline: 30335656]
3. Ghoneim MM, Block RI, Haffarnan M, Mathews MJ. Awareness during anesthesia: risk factors, causes and sequelae: a review of reported cases in the literature. Anesth Analg 2009 Feb;108(2):527-535. [doi: 10.1213/ane.0b013e318193c634] [Medline: 19151283]
4. Zhou X, Dong S, Xu Y. Molecular mechanisms of propofol-induced cognitive impairment: suppression of critical hippocampal pathways. J Neurochem 2025 Apr;169(4):e70070. [doi: 10.1111/jnc.70070] [Medline: 40265596]
5. Aldecoa C, Bettelli G, Bilotta F, et al. Update of the European Society of Anaesthesiology and Intensive Care Medicine evidence-based and consensus-based guideline on postoperative delirium in adult patients. Eur J Anaesthesiol 2024 Feb 1;41(2):81-108. [doi: 10.1097/EJA.0000000000001876] [Medline: 37599617]
6. Da X, Li X, Dong A, Chen L, Ma Y, Li X. Effect of electroencephalogram-guided anesthesia on postoperative delirium in older adults after surgery: a systematic review and meta-analysis. Front Neurol 2025;16:1638282. [doi: 10.3389/fneur.2025.1638282] [Medline: 40948652]
7. Bennett C, Voss LJ, Barnard JPM, Sleigh JW. Practical use of the raw electroencephalogram waveform during general anesthesia: the art and science. Anesth Analg 2009 Aug;109(2):539-550. [doi: 10.1213/ane.0b013e3181a9fc38] [Medline: 19608830]
8. Purdon PL, Sampson A, Pavone KJ, Brown EN. Clinical electroencephalography for anesthesiologists: Part I: background and basic signatures. Anesthesiology 2015 Oct;123(4):937-960. [doi: 10.1097/ALN.0000000000000841] [Medline: 26275092]
9. Chan MTV, Hedrick TL, Egan TD, et al. American Society for Enhanced Recovery and Perioperative Quality Initiative Joint Consensus Statement on the role of neuromonitoring in perioperative outcomes: electroencephalography. Anesth Analg 2020 May;130(5):1278-1291. [doi: 10.1213/ANE.0000000000004502] [Medline: 31764163]
10. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology 2020 Feb;132(2):379-394. [doi: 10.1097/ALN.0000000000002960] [Medline: 31939856]
11. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
12. Nasibov E, Ozgören M, Ulutagay G, Oniz A, Kocaaslan S. On the analysis of BIS stage epochs via fuzzy clustering. Biomed Tech (Berl) 2010 Jun;55(3):147-153. [doi: 10.1515/BMT.2010.009] [Medline: 20156029]
13. Inan ZH, Kuntalp M. A study on fuzzy C-means clustering-based systems in automatic spike detection. Comput Biol Med 2007 Aug;37(8):1160-1166. [doi: 10.1016/j.compbiomed.2006.10.010] [Medline: 17145054]
14. Krajca V, Petránek S, Patáková I, Värri A. Automatic identification of significant graphoelements in multichannel EEG recordings by adaptive segmentation and fuzzy clustering. Int J Biomed Comput 1991;28(1-2):71-89. [doi: 10.1016/0020-7101(91)90028-d] [Medline: 1889908]
15. Fehrlin ES, Hight D, Kaiser HA, et al. A pilot investigation evaluating relative changes in fronto-occipital alpha and beta spectral power as measurement of anesthesia hypnotic depth. Anesth Analg 2023 Sep 1;137(3):656-664. [doi: 10.1213/ANE.0000000000006398] [Medline: 36961823]
16. Gutiérrez R, Maldonado F, Egaña JI, Penna A. Electroencephalographic alpha and delta oscillation dynamics in response to increasing doses of propofol. J Neurosurg Anesthesiol 2022 Jan 1;34(1):79-83. [doi: 10.1097/ANA.0000000000000733] [Medline: 33060553]
17. Koch S, Stegherr AM, Mörgeli R, et al. Electroencephalogram dynamics in children during different levels of anaesthetic depth. Clin Neurophysiol 2017 Oct;128(10):2014-2021. [doi: 10.1016/j.clinph.2017.07.417] [Medline: 28837907]

XSL•FO
RenderX

18. Liu Q, Chen YF, Fan SZ, Abbod MF, Shieh JS. Improved spectrum analysis in EEG for measure of depth of anesthesia based on phase-rectified signal averaging. Physiol Meas 2017 Feb;38(2):116-138. [doi: 10.1088/1361-6579/38/2/116] [Medline: 28033111]
19. Gu Y, Liang Z, Hagihira S. Use of multiple EEG features and artificial neural network to monitor the depth of anesthesia. Sensors (Basel) 2019 May 31;19(11):2499. [doi: 10.3390/s19112499] [Medline: 31159263]
20. Chen YF, Fan SZ, Abbod MF, Shieh JS, Zhang M. Electroencephalogram variability analysis for monitoring depth of anesthesia. J Neural Eng 2021 Nov 17;18(6):066030. [doi: 10.1088/1741-2552/ac3316] [Medline: 34695812]
21. Bruhn J, Röpcke H, Hoeft A. Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia. Anesthesiology 2000 Mar;92(3):715-726. [doi: 10.1097/00000542-200003000-00016] [Medline: 10719951]
22. Bruhn J, Lehmann LE, Röpcke H, Bouillon TW, Hoeft A. Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. Anesthesiology 2001 Jul;95(1):30-35. [doi: 10.1097/00000542-200107000-00010] [Medline: 11465580]
23. Röpcke H, Rehberg B, Koenen-Bergmann M, Bouillon T, Bruhn J, Hoeft A. Surgical stimulation shifts EEG concentration-response relationship of desflurane. Anesthesiology 2001 Mar;94(3):390-399. [doi: 10.1097/00000542-200103000-00006] [Medline: 11374596]
24. Hajat Z, Ahmad N, Andrzejowski J. The role and limitations of EEG-based depth of anaesthesia monitoring in theatres and intensive care. Anaesthesia 2017 Jan;72 Suppl 1(Suppl 1):38-47. [doi: 10.1111/anae.13739] [Medline: 28044337]
25. Russell IF. The ability of bispectral index to detect intra-operative wakefulness during total intravenous anaesthesia compared with the isolated forearm technique. Anaesthesia 2013 May;68(5):502-511. [doi: 10.1111/anae.12177] [Medline: 23521699]
26. Prichep LS, Gugino LD, John ER, et al. The Patient State Index as an indicator of the level of hypnosis under general anaesthesia. Br J Anaesth 2004 Mar;92(3):393-399. [doi: 10.1093/bja/aeh082] [Medline: 14742326]
27. Obara S, Oishi R, Nakano Y, et al. Update on the SedLine® algorithm for calculating the Patient State Index of older individuals during general anesthesia: a randomized controlled trial. Minerva Anestesiol 2021 Jul;87(7):774-785. [doi: 10.23736/S0375-9393.21.14929-6] [Medline: 33938673]
28. Operator's manual: sedline sedation monitor. Masimo Corporation. 2018. URL: https://techdocs.masimo.com/globalassets/techdocs/pdf/lab-7373f_master.pdf [accessed 2026-01-16]
29. Müller J, Nowak S, Weidemeier M, et al. Duration of surgery and intraoperative blood pressure management are modifiable risk factors for postoperative neurocognitive disorders after spine surgery: results of the Prospective CONFESS Study. Spine (Phila Pa 1976) 2023 Aug 15;48(16):1127-1137. [doi: 10.1097/BRS.0000000000004722] [Medline: 37195031]
30. Seo JS, Park SW, Lee YS, Chung C, Kim YB. Risk factors for delirium after spine surgery in elderly patients. J Korean Neurosurg Soc 2014 Jul;56(1):28-33. [doi: 10.3340/jkns.2014.56.1.28] [Medline: 25289122]
31. Eagleman SL, Drover CM, Li X, MacIver MB, Drover DR. Offline comparison of processed electroencephalogram monitors for anaesthetic-induced electroencephalogram changes in older adults. Br J Anaesth 2021 May;126(5):975-984. [doi: 10.1016/j.bja.2020.12.042] [Medline: 33640118]
32. García PS, Kreuzer M, Hight D, Sleigh JW. Effects of noxious stimulation on the electroencephalogram during general anaesthesia: a narrative review and approach to analgesic titration. Br J Anaesth 2021 Feb;126(2):445-457. [doi: 10.1016/j.bja.2020.10.036] [Medline: 33461725]
33. Hagihira S. Changes in the electroencephalogram during anaesthesia and their physiological basis. Br J Anaesth 2015 Jul;115 Suppl 1(Suppl 1):i27-i31. [doi: 10.1093/bja/aev212] [Medline: 26174297]
34. Hight DF, Kaiser HA, Sleigh JW, Avidan MS. Continuing professional development module: an updated introduction to electroencephalogram-based brain monitoring during intended general anesthesia. Can J Anaesth 2020 Dec;67(12):1858-1878. [doi: 10.1007/s12630-020-01820-3] [Medline: 33190217]
35. Zhang Y, Wang Y, Cheng H, et al. EEG spectral slope: a reliable indicator for continuous evaluation of consciousness levels during propofol anesthesia. Neuroimage 2023 Dec;283:120426. [doi: 10.1016/j.neuroimage.2023.120426]
36. Obert DP, Schneider F, Schneider G, et al. Performance of the SEDLine monitor: age dependency and time delay. Anesth Analg 2023 Oct 1;137(4):887-895. [doi: 10.1213/ANE.0000000000006369] [Medline: 36727845]
37. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci 2019;13:31. [doi: 10.3389/fncom.2019.00031] [Medline: 31178711]
38. Nazar W, Szymanowicz S, Nazar K, et al. Artificial intelligence models in prediction of response to cardiac resynchronization therapy: a systematic review. Heart Fail Rev 2024 Jan;29(1):133-150. [doi: 10.1007/s10741-023-10357-8] [Medline: 37861853]
39. Tavazzi E, Longato E, Vettoretti M, et al. Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: a systematic review. Artif Intell Med 2023 Aug;142:102588. [doi: 10.1016/j.artmed.2023.102588] [Medline: 37316101]
40. Lichtenfeld F, Kratzer S, Hinzmann D, García PS, Schneider G, Kreuzer M. The influence of electromyographic on electroencephalogram-based monitoring: putting the forearm on the forehead. Anesth Analg 2024 Jun 1;138(6):1285-1294. [doi: 10.1213/ANE.0000000000006652] [Medline: 37756246]

XSL•FO
RenderX

41. Tatum WO, Dworetzky BA, Freeman WD, Schomer DL. Artifact: recording EEG in special care units. J Clin Neurophysiol 2011 Jun;28(3):264-277. [doi: 10.1097/WNP] [Medline: 21633252]
42. Weyer C, Prötzl E, Kinateder T, et al. The strength of alpha oscillations in the electroencephalogram differently affects algorithms used for anesthesia monitoring. Anesth Analg 2021 Dec 1;133(6):1577-1587. [doi: 10.1213/ANE.0000000000005704] [Medline: 34543237]
43. Brogi E, Cyr S, Kazan R, Giunta F, Hemmerling TM. Clinical performance and safety of closed-loop systems: a systematic review and meta-analysis of randomized controlled trials. Anesth Analg 2017 Feb;124(2):446-455. [doi: 10.1213/ANE.0000000000001372] [Medline: 27482773]
44. Wang D, Song Z, Zhang C, Chen P. Bispectral index monitoring of the clinical effects of propofol closed-loop target-controlled infusion: systematic review and meta-analysis of randomized controlled trials. Medicine (Baltimore) 2021 Jan 29;100(4):e23930. [doi: 10.1097/MD.0000000000023930] [Medline: 33530193]
45. Castellanos Peñaranda C, Casas Arroyave FD, Gómez FJ, et al. Technical and clinical evaluation of a closed loop TIVA system with SEDLineTM spectral density monitoring: multicentric prospective cohort study. Perioper Med (Lond) 2020;9(1):1. [doi: 10.1186/s13741-019-0130-2] [Medline: 31921411]
46. Nayak J, Naik B, Behera HS. Fuzzy c-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In: Computational Intelligence in Data Mining: New Delhi: Springer; 2015, Vol. 2:978-981. [doi: 10.1007/978-81-322-2208-8_14]

Abbreviations

BIS: Bispectral Index
DBSCAN: Density-Based Spatial Clustering of Applications with Noise
DSA: density spectral array
EEG: electroencephalography
EMG: electromyography
FCM: Fuzzy C-Means
GMM: Gaussian Mixture Model
pEEG: processed electroencephalography
PSi: Patient State Index

Edited by A Coristine; submitted 20.May.2025; peer-reviewed by CG Altintop, TP Kao; accepted 24.Dec.2025; published 06.Feb.2026.

Please cite as:
Huang PY, Hong WL, Hee HZ, Chang WK, Lee CH, Ting CK
Enhancing Anesthetic Depth Assessment via Unsupervised Machine Learning in Processed Electroencephalography Analysis: Novel Methodological Study
JMIR Med Inform 2026;14:e77830
URL: https://medinform.jmir.org/2026/1/e77830
doi:10.2196/77830

© Po-Yu Huang, Wei-Lun Hong, Hui-Zen Hee, Wen-Kuei Chang, Ching-Hung Lee, Chien-Kun Ting. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Concordance of Death Information Between Two Health Systems Serving the Same Region and the Social Security Administration Death Master File: Retrospective Observational Study

Zhan Wang[1], PhD; Mahanaz Syed[1], PhD; Shweta Bansal[2], MD; Meredith Nahm Zozus[1], PhD; Eric Eisenstein[3,4], DBA; Muayad Hamidi[1], MBChB

[1]Clinical Research Informatics Division, Department of Population Health Sciences, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX, United States

[2]Division of Nephrology, Department of Medicine, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX, United States

[3]Department of Medicine, Duke University, Durham, NC, United States

[4]School of Health Information Science, University of Victoria, Victoria BC, CN, Canada

Corresponding Author:
Zhan Wang, PhD
Clinical Research Informatics Division, Department of Population Health Sciences, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX, United States

Abstract

Background: There are multiple sources that document the date of a person's death in the United States. Unfortunately, this seemingly simple information is either incomplete or costly to obtain. Nevertheless, information on such events is critical for both health systems and clinical studies to assess the outcomes of operational and therapeutic interventions.

Objective: As part of a larger assessment of the quality of multisource death information, we compared the death data from two health systems serving the same region and the Social Security Administration Death Master File (SSADMF).

Methods: This study linked death records for patients seen in either health system with the SSADMF from 2007 to 2020 to identify concordant and discordant death data among sources. Analyses included cross-system matching, classification of death records by overlap, and calculation of agreement using the Fleiss κ.

Results: Among 904,581 matched patients, only 209 (0.02%) deaths were confirmed by all 3 sources. Large proportions of deaths were uniquely recorded by a single source: 54.32% (10,697/19,691) by health system A, 5.16% (1017/19,691) by health system B, and 20.17% (3972/19,691) by the SSADMF. The Fleiss κ was negative (−0.312), reflecting less agreement than expected by chance.

Conclusions: While this study is not generalizable, it showed that, without processes in place to obtain external data regarding patient deaths, health care facility death information should not be relied upon as a complete list of those who have died. The discordances observed highlight the potential for significant gaps in death reporting within health care systems, which could impact the accuracy of mortality-based analyses and quality assessments.

(JMIR Med Inform 2026;14:e85136) doi:10.2196/85136

KEYWORDS

death record concordance; health systems; mortality data quality; electronic health record data quality; EHR data quality; Social Security Administration Death Master File

Introduction

Mortality is the ultimate outcome of health care and wellness interventions. As such, mortality data are important to many clinical studies. In the era of learning health systems and real-world evidence, institutions increasingly rely on routinely collected electronic health record (EHR) data to evaluate care delivery, monitor outcomes, and iteratively improve practice [1,2]. However, the accuracy and completeness of mortality information within these data remain uncertain, particularly for deaths that occur outside individual health systems [3,4].

Our clinical data warehouse, serving a large academic health center in the southwestern United States, is used for approximately 100 clinical studies per year. Many of these studies focus on our local population and address whether they improve or worsen particular outcomes and include tracking or intervening in health disparities in our region, including observational comparative effectiveness studies, local quality

XSL•FO
RenderX

improvement initiatives, and program evaluations targeting population health and health disparities in the surrounding region. These data are used primarily for retrospective and hybrid retrospective-prospective designs, where mortality end points are often ascertained from EHRs and linked administrative sources rather than from dedicated, trial-specific follow-up. Because findings from such studies can inform guideline development, institutional policies, and clinical decision support within learning health systems, inaccurate or incomplete mortality data can directly affect patient care.

In clinical research evaluating treatments for diseases with low mortality rates (eg, due to short follow-up), errors in the few available death data points can introduce significant bias, potentially compromising the validity of the study [5]. Clinical studies are increasingly using EHR real-world data in the United States, where recent regulatory guidance requires assessment of accuracy and completeness [1]. Unfortunately, many institutions lack comprehensive information about death for patients who die outside of their facilities, making the use of EHR real-world data less reliable in these studies. The increasing trend of in-home end-of-life care further limits mortality data in EHRs [6].

Researchers and clinicians in the United States have used public data sources such as the National Death Index (NDI) or the Social Security Administration Death Master File (SSADMF) to obtain mortality data [7-10]. The reports of NDI sensitivity range from 87% to 97%, with SSADMF sensitivity ranging from 71% to 83%, although performance varies by cohort, identifiers available, and period [3,11-15]. Importantly, a Social Security Administration policy change in November 2011 removed protected state death records from the publicly disseminated Death Master File, eliminating approximately 4.2 million existing records and sharply reducing capture of more recent deaths, with continuing implications for the completeness of SSADMF-based mortality ascertainment [16]. These national resources; near ubiquitous EHR adoption; and significant strides in EHR interoperability, such as health information exchanges supporting the long-used Health Level Seven Admission, Discharge, and Transfer message that can indicate a death, imply good availability of reliable death information. However, obtaining high-quality data for mortality end points in clinical studies remains a significant challenge [17,18].

These national sources are also costly and logistically constrained. The costs of obtaining external death information range from 15 cents per patient per year (NDI) to an approximately US $3000 initial fee and US $7000 annual subscription fee per user for the Limited Access SSADMF [4]. The NDI data are only available for research, their use is time limited, and reuse is constrained through a data use agreement. These limitations lead some health systems to purchase and routinely integrate such data and others to rely solely on local or regional sources [3].

Even when national mortality data are purchased and linked, there is no single contemporary gold standard for deaths occurring outside a given health system, and both national and local sources show imperfect sensitivity and date accuracy [19]. Increasing trends toward in-home and hospice end-of-life care further reduce the proportion of deaths directly captured within hospital EHR systems [20]. Consequently, one can reasonably argue that all health systems lack fully comprehensive information about out-of-hospital deaths, and that practical strategies rely on combining multiple imperfect data sources to approximate a gold standard for mortality ascertainment.

Academic health centers need a sustainable way to maintain high-quality mortality data linked to health care data. For example, our cancer registry requires death information, as do many clinical trials. In this context, we posed two major questions. (1) What is the accuracy of death data, such as occurrence and date? (2) What is the quality and relative contribution from available death data sources? To begin, we compared patient deaths reported in our health system, a neighboring health system serving the same region, and the SSADMF.

Methods

Data Sources and Linkage

This study evaluated the concordance of death information between 2 health systems serving the same geographic region and the SSADMF over a 13-year period (January 1, 2007, to July 10, 2020), as determined by the availability of data from both health systems. Patient records from the 2 health systems were linked using three approaches: (1) an existing identifier crosswalk; (2) matching first name, last name, gender, and birth date; and (3) matching medical record number. To link SSADMF data with the health system records, five distinct hash-based tokens were used: (1) last name, first initial of first name, gender, and birth date; (2) Soundex last name, Soundex first name, gender, and birth date; (3) last name, first name, gender, and birth date; (4) Social Security number, gender, and birth date; and (5) Social Security number and first name. For linkage, a match on both token 1 and token 2 was required, whereas tokens 3, 4, and 5 were each sufficient for linkage when matched individually.

Death Record Categorization

Death records were classified into mutually exclusive groups to capture concordance and discordance across the 3 sources. For the matched patient cohort (patients identified in both health systems), deaths were categorized as follows: (1) concordant deaths were deaths recorded in all 3 sources (health system A, health system B, and the SSADMF), and (2) discordant deaths were deaths recorded in only 1 or 2 sources among the matched patients, further subdivided into those recorded only in health system A, recorded only in health system B, recorded only in the SSADMF, and recorded in combinations of 2 sources (eg, health systems A and B, health system A and the SSADMF, or health system B and the SSADMF). For patients present in a single health system, deaths recorded exclusively in that system or in the SSADMF were also identified. This approach allowed us to distinguish between deaths that were consistently captured across all sources and those that were discordant, reflecting differences in reporting or population coverage.

To further investigate discordant death records, we analyzed “lost contact” patients, defined as individuals who had at least

XSL•FO
RenderX

1 encounter at 1 facility during the observation period (January 1, 2007, to July 10, 2020) but no subsequent encounters at that facility for at least 12 consecutive months after their last visit there while maintaining contact with the other facility after those 12 months. This analysis helped identify potential gaps in follow-up or differences in patient retention between systems.

In addition, we performed 2 exploratory analyses: date discrepancies of more than 1 month for deaths recorded between sources and postdeath encounters occurring more than 1 year after the recorded death date. The first analysis indicated inconsistencies in death reporting between sources, which can affect the accuracy of mortality statistics and clinical research. For postdeath encounters, we compared the death dates from the 3 sources to the encounter dates from both health systems. The analysis revealed potential issues in death recording or patient management, suggesting inaccuracies in encounter or death records or possible misuse of deceased individuals' health care credentials.

κ Analysis for Intersource Agreement

To measure the consistency of death reporting across the 3 data sources, we calculated observed agreement; expected agreement; and the Fleiss κ, a statistic used to evaluate agreement among multiple raters with categorical data. Each patient in the matched cohort was classified as either "dead" or "not dead" by each source. We then grouped patients into 8 mutually exclusive categories based on the combination of sources that reported their death. These metrics allowed us to assess the level of concordance in death classification across health system A, health system B, and the SSADMF.

Visualization

A 3-set Venn diagram was constructed to illustrate the overlap and unique contributions of each data source in identifying deaths within the matched patient cohort. Each region of the diagram was annotated with the corresponding patient counts, providing a comprehensive view of concordance and discordance in death ascertainment across health system A, health system B, and the SSADMF. This visualization facilitated the identification of patterns and gaps in death reporting among the 3 sources.

Ethical Considerations

This study was reviewed by the Institutional Review Board of the University of Texas Health Science Center at San Antonio and determined to meet the criteria for exemption from institutional review board review in accordance with Title 45 of the Code of Federal Regulations §46.104 [21] on February 21, 2024 (STUDY00000154). As this research involved secondary analysis of existing death data and no direct interaction with human participants, informed consent was not required. All data were handled in accordance with institutional and national ethical standards to ensure the privacy and confidentiality of the individuals' information.

Results

Health system A reported 1,708,399 unique patients, including 21,623 (1.27%) reported deaths, while health system B reported 1,535,936 unique patients, including 8740 (0.57%) reported deaths (Figure 1). A total of 11,052 death records from the SSADMF were linked to the combined cohort, encompassing both matched and unmatched patient records from the 2 health systems.

Among the 904,581 matched patients, health system A reported 14,659 (1.62%) deaths, health system B reported 3951 (0.44%) deaths, and the SSADMF identified 5295 (0.59%) deaths (Figure 2). Collectively, the total number of unique death records across these sources was 19,691. Figure 2 shows that 1.06% (209/19,691) of deaths were recorded by all 3 sources, whereas 54.32% (10,697/19,691) of deaths were identified exclusively by health system A, 5.16% (1017/19,691) were identified exclusively by health system B, and 20.17% (3972/19,691) were identified exclusively by the SSADMF. The proportion of death overlap between health system A and the SSADMF, health system B and the SSADMF, and the 2 health systems was 5.44% (1071/19,691), 0.22% (43/19,691), and 13.62% (2682/19,691), respectively. Health system A contributed 74.45% (14,659/19,691) of the unique deaths observed in at least one source, health system B contributed 20.07% (3951/19,691), and the SSADMF contributed 26.89% (5295/19,691).

XSL•FO
RenderX

Figure 1. Patients and deaths reported among health systems A and B and the Social Security Administration Death Master File (SSADMF; encompassing matched and unmatched patient records).

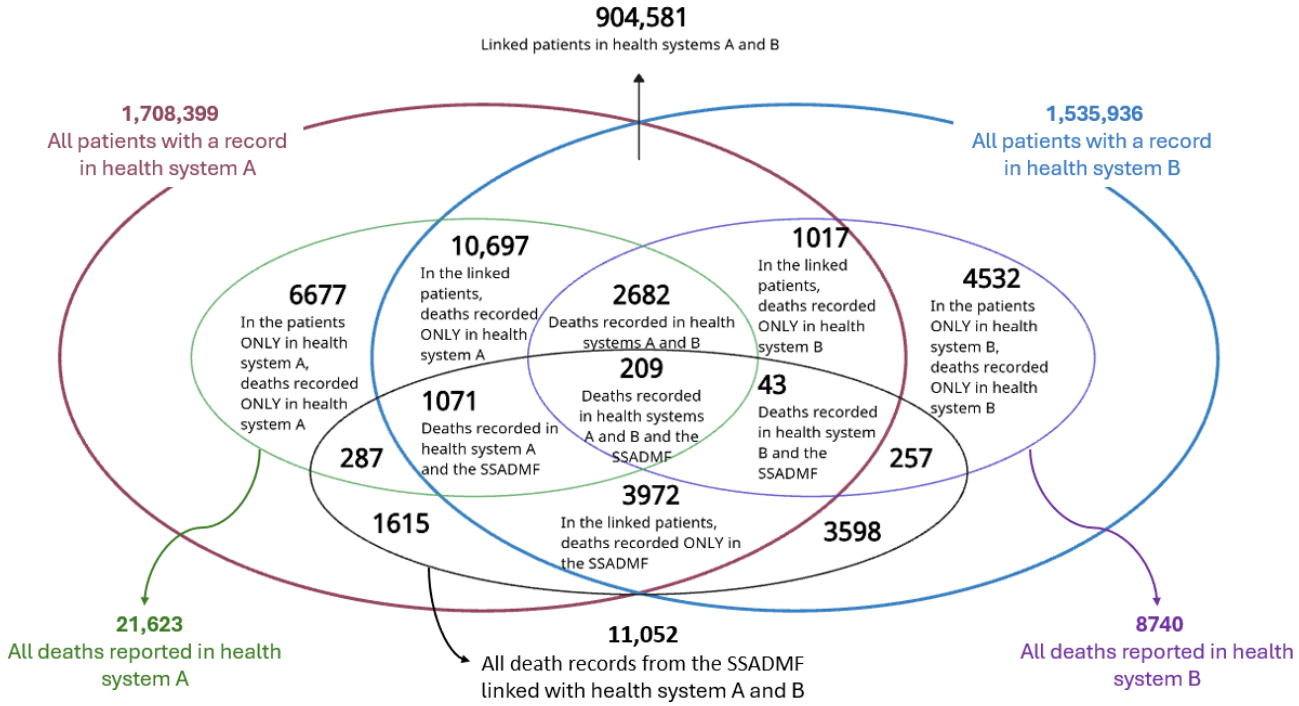
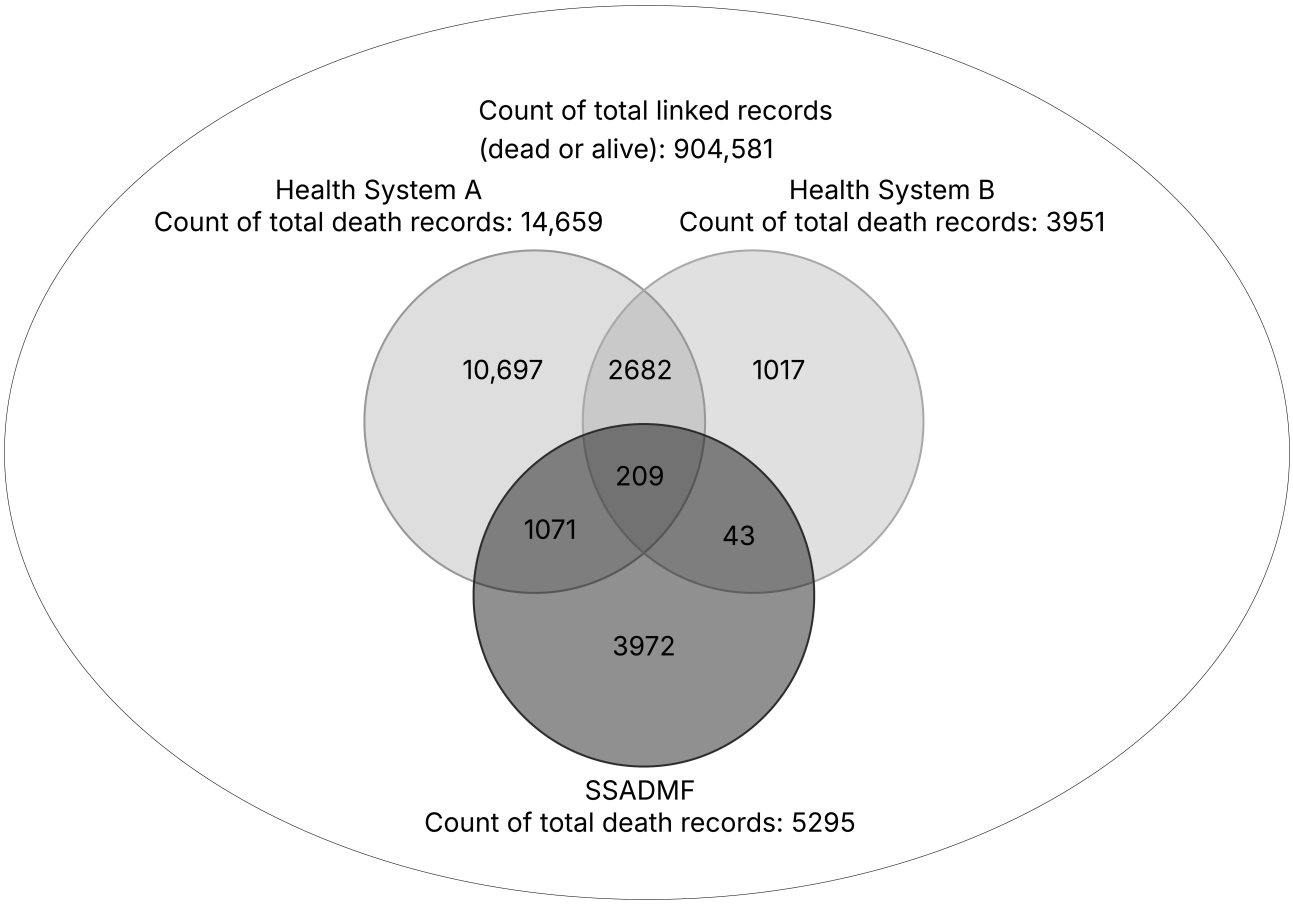


Figure 2. Three-set Venn diagram for the matched patient records. SSADMF: Social Security Administration Death Master File.



To quantify overall agreement in death classification among the 3 sources, the Fleiss κ was calculated using the distribution of death status assignments across all matched patients (Table 1). The observed agreement, defined as the proportion of patients for whom all 3 sources concurred on death status (either all "dead" or all "not dead"), was 0.9771. The expected agreement by chance based on the marginal probabilities of each source assigning "dead" or "not dead" was 0.9826. The Fleiss κ, summarizing agreement beyond chance, was −0.312, with a 95% CI from −0.330 to −0.294.

XSL•FO
RenderX

Table . Distribution of death status among the 3 sources in the matched cohort (N=904,581).

Overlap category	Patients, n (%)
Deaths in all 3 sources	209 (0.02)
Deaths in health systems A and B only	2682 (0.30)
Deaths in health system A and the SSADMF[a] only	1071 (0.12)
Deaths in health system B and the SSADMF only	43 (0.00)
Deaths in health system A only	10,697 (1.18)
Deaths in health system B only	1017 (0.11)
Deaths in the SSADMF only	3972 (0.44)
Not dead in any source	884,890 (97.82)

[a]SSADMF: Social Security Administration Death Master File.

Further investigation of discrepancies in matched patients revealed that, of the 11,768 patients recorded as deceased in health system A but not in health system B, 4796 (40.75%) lost-contact patients in health system B switched to health system A. Similarly, of the 1060 patients recorded as deceased in health system B but not in health system A, 44 (4.15%) lost-contact patients in health system A switched to health system B. These lost-contact patterns suggest that patient migration between facilities may contribute to the discordance in death reporting.

To explore the potential benefit of data quality checking rules, we identified 0.20% (39/19,691) of cases in which the date of death recorded in health system A differed by 1 month or more (median 2, range 1 to 20 months) from that recorded in health system B. Similarly, 0.23% (45/19,691) of the cases showed a discrepancy of at least 1 month (median 29; range 1-122 months) between the date of death recorded in health system A and the SSADMF, whereas 0.02% (3/19,691) of the cases showed a discrepancy of at least 1 month (median 52; range 49-122 months) between the date of death recorded in health system B and the SSADMF.

An analysis of postdeath encounters revealed that 1.22% (179/14,659) of the cases from health system A had visits recorded in either of the 2 health systems more than 1 year following their documented date of death. Similarly, 0.35% (14/3951) of such cases were identified in health system B, and 8.48% (449/5295) of cases were identified within the SSADMF.

Discussion

Principal Findings

Death records from 2 major health systems serving the same region and the SSADMF provide new insights into the complexities and limitations of mortality reporting in health care research. Ideally, each health system would maintain a complete and accurate record of all deaths, including precise date of death for every deceased patient. However, our study revealed substantial variability and discordance in mortality ascertainment across these 3 sources, underscoring the challenges inherent in relying solely on health care facility records for vital status determination.

Significant discrepancies were observed in both the identification of deaths and the recording of death dates among the 2 health systems and the SSADMF. The Venn diagram analysis demonstrated that a considerable number of deaths were uniquely captured by only 1 data source, with relatively few cases identified by all 3. This pattern highlights the risk of systematic bias and under-ascertainment when mortality surveillance depends on a single institutional record system. The presence of deaths recorded exclusively by the SSADMF (deaths not captured in either health system) emphasizes the value of integrating external sources to enhance completeness.

In Texas, hospitals are required to report deaths that occur within their facilities to the local registrar, who then forwards this information to the Texas Department of State Health Services Vital Statistics Section. This process ensures that deaths occurring in hospitals are captured in the official state vital records. However, a critical regulatory change further complicates the concordance between health system records and the SSADMF. Deaths reported solely to state vital record offices after November 1, 2011, may not appear in the publicly available SSADMF even if they are accurately captured in hospital or state databases. These reporting practices and regulatory constraints may explain the observed discrepancies between the health system data and the SSADMF. Deaths that occur in hospitals and are reported to the state may be present in the state's vital records but absent from the SSADMF if they are considered protected state records. Conversely, deaths reported directly to the Social Security Administration through other means (such as funeral directors or family members) may appear in the SSADMF but not in hospital records if the patient was not under active care at the time of death or if the hospital was not notified.

Analysis of lost-contact patients suggests that the discrepancies in death reporting may be partially explained by differences in patient engagement with each health system over time. Patients who continued to receive care primarily from one system or had no recorded visits in the other system were more likely to have their death recorded by the system with which they last interacted, potentially due to more up-to-date information or discovering deaths through postcare follow-up or billing interactions. The higher proportion of health system B–to–health system A switching compared to health system A–to–health

XSL•FO
RenderX

system B switching explains part of health system A's substantially greater volume of death records overall. These findings suggest that patient migration between systems contributes to death reporting discordance, although the reasons for migration require further investigation. While these findings offer insight into the discrepancies, they do not fully account for the substantial differences observed. This substantial difference suggests that additional factors such as institutional practices, death documentation processes, or data management systems may significantly influence death reporting.

The presence of postdeath encounters in both systems is particularly concerning. These instances could represent data entry errors, identity mix-ups, fraudulent use of a deceased patient's identity and health care coverage by another individual or entity, or other root causes. These should be investigated to identify opportunities for improvement. The results reported in this paper emphasize the importance of cross-referencing death information from multiple sources to identify discrepancies and incompleteness.

The integration of the SSADMF into our analysis allowed for a more robust assessment of concordance across sources. A high observed agreement was driven primarily by the predominance of patients classified as "not dead" by all sources. In the context of this highly imbalanced outcome, the negative κ should not be interpreted as strong paradoxical disagreement but rather as a statistical artifact of extreme class imbalance where chance agreement among "not dead" classifications is very high. Therefore, the Venn diagrams and Table 1 provide a more informative description of how often each source captured deaths and how frequently those deaths were uniquely identified by one source vs shared across sources. The finding is that the 3 sources showed limited concordance in identifying deaths themselves.

Limitations

A persistent limitation of our study is the absence of a definitive gold standard for mortality ascertainment. Without such a reference, we are unable to calculate accuracy statistics such as sensitivity and specificity and cannot definitively identify false positives or false negatives. While some potential false positives may be inferred from the presence of postdeath encounters, the true extent of missed deaths (false negatives) remains unknown and likely requires external validation. Errors in patient linkage, crosswalks, or record attribution may also contribute to misclassification. In addition, differences in linkage methodology may influence the observed discordance: the SSADMF linkage required agreement on 2 composite name–demographic tokens recommended by the data provider to balance tolerance for name variation, a design that may reduce missed matches but carries a higher risk of false linkages when names are similar yet pertain to different individuals. Future work should incorporate more reliable data sources, including the NDI, state death records, obituary data, and family-reported information. These additional sources would provide a more comprehensive and accurate picture of mortality, allowing for a better assessment of false positives and false negatives in health care facility death data.

We report this evaluation as a marker and example of significant discordance in death data among 2 health systems serving the same region and the SSADMF. While not an unexpected or groundbreaking result, such assessments serve as important reminders that, although we have come a long way toward EHR adoption and interoperability of health information, we have a long way to go before even the most basic health care data can be assumed to be accurate and complete.

Conclusions

This study demonstrates that death data from health care facilities, even when supplemented with data from the SSADMF, exhibit substantial discordance and should not be presumed to be accurate or complete without corroboration from additional sources. The persistent discrepancies among the 3 sources highlight the limitations and potential unreliability of depending solely on the EHR. Despite advances in EHR adoption and interoperability, obtaining high-quality, comprehensive mortality data for research remains challenging. Researchers and clinicians should not rely solely on facility death data for mortality end points in observational studies unless death information is systematically and prospectively collected as part of the study protocol; instead, they are strongly encouraged to incorporate external sources, such as the NDI or state vital records, to improve accuracy and completeness. Further research is needed to understand the causes of discordance and develop strategies for enhancing the quality of mortality information in clinical and epidemiological studies.

Acknowledgments

The authors are grateful to Mrs Lissa A Persson for her assistance with the phase of quality control of the coding for this study. Generative artificial intelligence was not used in any portion of manuscript generation.

Funding

This research was funded by the Clinical and Translational Science Awards Program (grant ID UM1TR004538-01).

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: MNZ (lead), ZW (equal), EE (supporting), SB (supporting)

XSL•FO
RenderX

Data curation: ZW (lead), MS (supporting)
Formal analysis: ZW (lead), MS (supporting), MH (supporting)
Funding acquisition: MNZ
Investigation: ZW
Methodology: MNZ (lead), ZW (equal), EE (supporting), SB (supporting)
Project administration: MH (lead), MNZ (equal), ZW (supporting)
Resources: ZW (lead), SB (equal), MNZ (equal)
Supervision: MNZ (lead), SB (supporting)
Validation: ZW (lead), MH (equal)
Visualization: ZW (lead), MH (equal)
Writing—original draft: ZW (lead), MS (supporting), MH (supporting)
Writing—review and editing: ZW (lead), MNZ (supporting), EE (supporting), SB (supporting), MS (supporting), MH (supporting)

Conflicts of Interest

None declared.

References

1. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. U.S. Food and Drug Administration. 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory [accessed 2026-03-15]
2. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv Res 2021 Mar 5;21(1):200. [doi: 10.1186/s12913-021-06215-8] [Medline: 33663508]
3. Foster SO, Schubauer-Berigan M, Waters KM. The specificity of the National Death Index and Social Security Administration death master file when information on Social Security number is lacking. Am J Epidemiol 2000;151(11):S43 [FREE Full text]
4. Hanna DB, Pfeiffer MR, Sackoff JE, Selik RM, Begier EM, Torian LV. Comparing the National Death Index and the Social Security Administration's Death Master File to ascertain death in HIV surveillance. Public Health Rep 2009;124(6):850-860. [doi: 10.1177/003335490912400613] [Medline: 19894428]
5. Lerman MH, Holmes B, St Hilaire D, et al. Validation of a mortality composite score in the real-world setting: overcoming source-specific disparities and biases. JCO Clin Cancer Inform 2021 Apr;5:401-413. [doi: 10.1200/CCI.20.00143] [Medline: 33830788]
6. Cross SH, Warraich HJ. Changes in the place of death in the United States. N Engl J Med 2019 Dec 12;381(24):2369-2370. [doi: 10.1056/NEJMc1911892] [Medline: 31826345]
7. Fisher SG, Weber L, Goldberg J, Davis F. Mortality ascertainment in the veteran population: alternatives to the National Death Index. Am J Epidemiol 1995 Feb 1;141(3):242-250. [doi: 10.1093/oxfordjournals.aje.a117426] [Medline: 7840098]
8. Mumma MT, Cohen SS, Sirko JL, Ellis ED, Boice JD Jr. Obtaining vital status and cause of death on a million persons. Int J Radiat Biol 2022;98(4):580-586. [doi: 10.1080/09553002.2018.1539884] [Medline: 30412007]
9. Schisterman EF, Whitcomb BW. Use of the Social Security Administration Death Master File for ascertainment of mortality status. Popul Health Metr 2004 Mar 5;2(1):2. [doi: 10.1186/1478-7954-2-2] [Medline: 15003125]
10. Sohn MW, Arnold N, Maynard C, Hynes DM. Accuracy and completeness of mortality data in the Department of Veterans Affairs. Popul Health Metr 2006 Apr 10;4:2. [doi: 10.1186/1478-7954-4-2] [Medline: 16606453]
11. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. Ann Epidemiol 2002 Oct;12(7):462-468. [doi: 10.1016/s1047-2797(01)00285-x] [Medline: 12377423]
12. Curtis MD, Griffith SD, Tucker M, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res 2018 Dec;53(6):4460-4476. [doi: 10.1111/1475-6773.12872] [Medline: 29756355]
13. Levin MA, Lin HM, Prabhakar G, McCormick PJ, Egorova NN. Alive or dead: validity of the Social Security Administration Death Master File after 2011. Health Serv Res 2019 Feb;54(1):24-33. [doi: 10.1111/1475-6773.13069] [Medline: 30520023]
14. Moghanaki D, Howard LE, De Hoedt A, et al. Validity of the National Death Index to ascertain the date and cause of death in men having undergone prostatectomy for prostate cancer. Prostate Cancer Prostatic Dis 2019 Dec;22(4):633-635. [doi: 10.1038/s41391-019-0146-1] [Medline: 31036926]
15. Skopp NA, Smolenski DJ, Schwesinger DA, Johnson CJ, Metzger-Abamukong MJ, Reger MA. Evaluation of a methodology to validate National Death Index retrieval results among a cohort of U.S. service members. Ann Epidemiol 2017 Jun;27(6):397-400. [doi: 10.1016/j.annepidem.2017.05.004] [Medline: 28641759]
16. Change in public Death Master File records. U.S. Department of Commerce. 2011. URL: https://ladmf.ntis.gov/docs/import-change-dmf.pdf [accessed 2026-03-15]
17. Eisenstein EL, Curtis LH, Prather K, et al. Using death as an endpoint. In: Starman K, editor. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials: NIH Health Care Systems Research Collaboratory; 2019.

XSL•FO
RenderX

18. Eisenstein EL, Prather K, Greene SJ, et al. Death: the simple clinical trial endpoint. Stud Health Technol Inform 2019;257:86-91. [Medline: 30741178]
19. Jamal-Allial A, Sponholtz T, Vojjala SK, et al. Validation of mortality data sources compared to the National Death Index in the healthcare integrated research database. Pragmat Obs Res 2025;16:19-25. [doi: 10.2147/POR.S498221] [Medline: 39936035]
20. Eisenstein EL, Sapp S, Harding T, et al. Ascertaining death events in a pragmatic clinical trial: insights from the TRANSFORM-HF trial. J Card Fail 2022 Oct;28(10):1563-1567. [doi: 10.1016/j.cardfail.2022.01.020] [Medline: 35181553]
21. §46.104 Exempt Research. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104 [accessed 2026-03-26]

Abbreviations

EHR: electronic health record
NDI: National Death Index
SSADMF: Social Security Administration Death Master File

Edited by A Benis; submitted 01.Oct.2025; peer-reviewed by CB Whitmore, J Tenenbaum, O Akinpeloye; revised version received 29.Jan.2026; accepted 17.Feb.2026; published 07.Apr.2026.

Please cite as:
Wang Z, Syed M, Bansal S, Zozus MN, Eisenstein E, Hamidi M
Concordance of Death Information Between Two Health Systems Serving the Same Region and the Social Security Administration Death Master File: Retrospective Observational Study
JMIR Med Inform 2026;14:e85136
URL: https://medinform.jmir.org/2026/1/e85136
doi:10.2196/85136

© Zhan Wang, Mahanaz Syed, Shweta Bansal, Meredith Nahm Zozus, Eric Eisenstein, Muayad Hamidi. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Factors Associated With the Stalled Implementation of a Ground-Up Electronic Health Record System in South Africa: Qualitative Insights From the E-Tick Case Study Using the Consolidated Framework for Implementation Research (CFIR)

Campion Zharima[1], PhD; Frances Griffiths[2], PhD; Jane Goudge[1], PhD

[1]Centre for Health Policy, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2]Centre for Primary Health Care Studies, Warwick Medical School, University of Warwick, Medical School Building, CV4 7AL, Coventry, United Kingdom

Corresponding Author:
Frances Griffiths, PhD
Centre for Primary Health Care Studies, Warwick Medical School, University of Warwick, Medical School Building, CV4 7AL, Coventry, United Kingdom

Abstract

Background: Electronic health records (EHRs) have the potential to improve service delivery through record keeping and monitoring health outcomes. As countries move toward universal health coverage, digital health tools such as EHRs are essential for achieving this goal. However, EHR implementation in middle-income countries like South Africa faces obstacles.

Objective: This study explores the reasons behind a stalled implementation of the electronic tick register (E-tick) system (an electronic version of a paper primary health care register to record services provided), using the Consolidated Framework for Implementation Research.

Methods: Using a qualitative design, in-depth interviews were conducted with 38 participants to explore their perceptions and experiences, and the factors surrounding the success and stalling of E-ticks. Participants included managers, stakeholders, implementers, and end users from the 3 implementation clinics. Data was collected using semistructured interview guides. The Thematic and Consolidated Framework for Implementation Research framework analysis (innovation, inner setting, individual characteristics, implementation process, and outer setting) was applied.

Results: The E-tick system was designed to improve data quality in paper health registers, addressing inaccuracies in reporting to district and provincial health departments (Innovation domain). Implementers iteratively developed the system through user input from managers and clinicians, and stakeholder engagement of software developers, funders, health managers, and decision-makers from the provincial health department (individual characteristics). Although the system was initially well adopted by end users, it stalled primarily due to outer setting factors, which included a change of developers, funding cuts, and limited support at the provincial health department level due to capacity gaps, political appointments, and mistrust stemming from corruption and abuse of the tender system. Moreover, resistance to leveraging lessons from locally developed small-scale systems further constrained institutional support for the E-tick.

Conclusions: Although successful implementation of EHRs can be facilitated by strong user engagement and co-design, outer setting factors such as governance, funding, and policy alignment can pose significant threats to sustainability. This underscores the importance of effective synergy between top-down and bottom-up processes for successful implementation.

(JMIR Med Inform 2026;14:e73831) doi:10.2196/73831

KEYWORDS

electronic health records; implementation; barriers; facilitators; Consolidated Framework for Implementation Research; CFIR

Introduction

Overview

Electronic health records (EHRs) hold the promise of improving service delivery, monitoring and evaluation of health interventions, and record-keeping for public health [1]. As more countries work towards universal health coverage, leveraging digital health tools like EHRs becomes more important for enhancing service delivery and improving quality of care [2,3]. The World Health Organization's global digital health strategy emphasizes the need for accelerated adoption of EHRs to achieve health goals [4]. However, EHR implementation in many middle-income countries remains slow due to financial and technological challenges [5].

XSL•FO
RenderX

Implementing EHRs is an evolving process that requires adaptive and context-specific policies to meet ever-changing needs [6], and a clear vision to move from pilot to scaled implementation [2]. In some high-income countries where implementation has occurred [7], a "top-down" approach has typically been used, where systems are preconfigured, and users must adapt. Although this offers coordination and central oversight, it can limit flexibility. In contrast, bottom-up approaches emphasize co-design with end users and adaptation to local workflows, which often enhances usability and adoption [8]. Research suggests that successful EHR implementation in complex health systems requires balancing national coordination with local innovation, as this allows centralized oversight, while engaging local health communities and tailoring systems to user needs [8,9].

South Africa has made policy advances, such as the National Digital Health Strategy (NDHS) (2019 - 2024) [10], with considerable resources as a middle-income country and sufficient technical expertise, but progress toward a national EHR has been slow. This is because policy direction and technical capacity are sometimes impeded by systemic and political barriers. Within this context, most primary health care facilities continue to rely on manual paper-based systems, which are labor-intensive and prone to errors in data collection [11,12]. However, bottom-up efforts have been made to improve data quality and accuracy. These include Tier.net, an electronic record system used with patients living with HIV and tuberculosis, and the electronic tick register (E-tick), the electronic record system we studied. The E-tick is a digital version of the tick register, a book in each health care facility for recording patient attendance and care delivered. In this paper, we use the Consolidated Framework for Implementation Research (CFIR) to explore an attempted implementation of the E-tick, a system conceptually similar to an EHR, examining how its implementation stalled despite a consultative approach in its implementation.

Background

Digital Health Policy in South Africa

As a middle-income country, South Africa has made policy strides in line with the global developments, which should facilitate the adoption of EHRs, such as the National Health Act 61 of 2003 and the NDHS (2019 - 2024) [10,13]. The Health Act outlines the role of the National Department of Health in coordinating health information systems across national, provincial, and district levels [10]. Although policy direction is set at the national level, the federal system means that provinces can independently determine their own health care spending. Therefore, implementation largely occurs at provincial and district levels, with each maintaining its own health information systems within the national framework. This provincial autonomy has contributed to fragmentation, as different provinces and regions develop EHR systems through separate procurement processes [14]. This fragmentation is further compounded by a political economy of procurement, in which tenders are prone to corruption, mismanagement, and the absence of oversight structures [15]. These factors directly affect which innovations are funded, how resources are distributed, and whether systems can successfully reach scale, as funding decisions often prioritize large-scale centralized systems.

South Africa uses 2 health information systems in the public sector: the Health Patient Registration System (HPRS) and the District Health Information Software (DHIS; developed through a collaboration between the University of the Western Cape and the University of Oslo), jointly developed by the University of the Western Cape and the University of Oslo [16]. The HPRS is a manual patient registry that records patient contact details as well as legal identification numbers, while the DHIS is an electronic system used to collect aggregated data on services provided to patients in public health facilities [17]. The DHIS data is first collected manually, then captured into the DHIS software, where it is submitted to the district, provincial, and national health departments. The method used to record primary health care data is a tick register, a manual paper-based system where clinicians mark or tick off the services provided in a large logbook with columns for dates and types of services offered (antenatal care, vaccinations, or chronic disease management). However, this process can be arduous, time-consuming, and prone to many recording errors. Studies evaluating the DHIS have confirmed persistent data gaps, citing poor documentation, manual entry under heavy workloads, and errors as major contributors to discrepancies [18].

South African Health System

South Africa plans to implement a National Health Insurance fund to facilitate better purchasing decisions through more formal contracting relationships rather than the historical allocation of budgets by provinces [19]. In this case, EHRs are key in monitoring patient outcomes and assessing the value of care purchased. Although most provinces have had some type of operational EHR system in public hospitals [20], there is no national EHR at the primary care level, and different architectures still pose a challenge to their integration for nationally coordinated use [21].

Provincial health departments in South Africa are often run by political appointees, who work under directives from their affiliated political party. This has led to pervasive corruption in the country's health system, which has been reported at various levels of government [22]. For example, 150 million rand awarded to a communications company during the COVID pandemic was misappropriated, and the health minister was suspended as a result [23]. In another case, the provincial chief financial officer who exposed a 1-billion-rand tender fraud was assassinated [24]. Such cases illustrate how poor financial management systems and political appointments enable corruption, which acts as an impediment to the public trust in the health system [25]. For digital health innovations, such dynamics increase the risk that systems are chosen for political or financial interests rather than their impact or merit, which affects the health department's ability to work with other stakeholders, or source funds for large-scale projects, including projects to implement EHRs. Moreover, the absence of countermeasures such as accountability and transparency makes it difficult to implement large-scale projects like EHRs.

XSL•FO
RenderX

The E-Tick System

An electronic version of the paper-based tick register (E-tick) was developed and piloted in 3 public health care facilities in the Ekurhuleni district between 2017 and 2021. Clinicians record services provided directly into the digital system on a patient's record, which also allows facility managers to monitor daily activities such as headcount and services offered. Although this is not a full EHR, the system can be expanded to include diagnosis and lab results and to integrate with other systems that are currently used in public health facilities. Against this backdrop, the E-tick system sought to provide an alternative that could improve data accuracy at the facility level while feeding directly into the existing provincial and national reporting systems. However, such progress is dependent on not only user adoption factors but also broader contextual dynamics [26]. Thus, by situating the E-tick in relation to both micro-level realities and macro-level policy and political economy, we aim to provide a more holistic understanding of why grassroots digital health interventions may stall or fail.

Consolidated Framework for Implementation Research (CFIR)

Implementation research seeks to understand why some implementation efforts succeed and others do not, and the role of the context [27]. To explore the key factors affecting implementation in our case study, we used the CFIR, which integrates existing theories of implementation and behavior change into a unified framework [28]. The CFIR is a meta-theoretical tool with standardized constructs to examine enablers and barriers to implementation across different stages (preimplementation, during implementation, and postimplementation) [29].

The CFIR was developed through a review of published theories and reports on factors influencing implementation [30]. Hence, its strength lies in its broad applicability and potential for generalizability, which are useful for informing practice across diverse settings. The framework has 5 broad domains (Table 1), which include: innovation domain, inner setting domain, outer setting domain, individuals domain, and the implementation process domain, which together capture multilevel determinants of implementation [31]. Due to the framework's flexibility, researchers can tailor it to specific interventions and contexts [32], and it has been widely adopted in research, applied across different research designs, and cited in over 300 published articles [29]. Recent updates have been made to the framework, using responses from authors who used the framework, leading to refined construct names and definitions to make them more encompassing [31]. In this paper, we draw on the updated CFIR to guide both our analysis and reporting of findings.

Table . Consolidated Framework for Implementation Research (CFIR) domains by Damschroder et al [28].

CFIR domains	Definition
Innovation	• The "thing" being implemented, such as a new treatment, program, or service. • Key factors that affect innovation include its source, evidence base, advantages over existing practices, adaptability, trialability, complexity, design, and cost.
Individuals	• The roles and characteristics of people involved in innovation. • Key roles include high-level leaders, mid-level leaders, opinion leaders, facilitators, and implementation teams who guide and support the innovation. • It also highlights innovation deliverers and recipients.
Inner setting	• The environment where an innovation is implemented, such as a hospital or school. • Key elements include structural characteristics like physical and IT infrastructure, work organization, and staffing levels.
Outer setting	• The broader context or setting within which an organization or project operates, such as a hospital system or school district. • Factors influencing implementation in the Outer Setting include large-scale disruptions, local attitudes, socioeconomic conditions, external partnerships, and regulatory policies.
Implementation process	• The activities and strategies to introduce and sustain an innovation. • Key components include forming teams to collaborate on tasks, assessing the needs of both innovation deliverers and recipients, and evaluating the context to identify barriers and facilitators. • Planning, tailoring strategies, engaging participants, testing small-scale changes, and reflecting on outcomes are critical steps.

XSL•FO
RenderX

Methods

This study adopted a qualitative design to explore the perceptions and experiences of external stakeholders, senior health managers, implementers, and end users of the E-tick system.

Participants and Sampling

We recruited a total of 38 participants, categorized into three groups: (1) managers and stakeholders, (2) implementation team, and (3) end users. The senior health managers and external stakeholders category included decision makers and higher-level people who were not directly involved in the system's implementation but provided insights into the context and the decisions impacting it. The implementation team comprised individuals who were responsible for initiating and deploying the system and introducing it to the users, while the end user team consisted of the clinicians and managers from 3 recipient facilities. Participants from the managers and stakeholders, as well as the implementation team, were purposively selected through referrals from key individuals and recruited via email. For end users, convenience sampling was used, with the researcher recruiting available users with the aid of gatekeepers during facility visits. Our sample intentionally included stakeholders from multiple levels of the health system to minimize elite bias, ensuring that the perspectives of senior health managers were balanced with those of frontline users and mid-level implementers. Below is a summary table of the participants in this study (Table 2).

Table . Study participants (N=38).

Category and roles	Count (n)
Senior managers and stakeholders	
National, provincial and district health managers	7
Parastatal staff	2
NGO[a] managers and staff	4
Academics	2
Private sector managers	2
Implementers	
Originators	2
IT Technician	1
Data capturers	3
End users	
Facility managers	3
Administrative officer	1
Nurses and midwives	9
Admin clerks	2

[a]NGO: Non-Governmental Organization.

Data Collection and Management

Data collection took place between November 2021 and June 2022, which was after the E-tick project had stalled. Thus, participants, particularly implementers and end users, reflected retrospectively on the system's development, piloting, and discontinuation. Due to COVID-19 health restrictions at the time, some of the interviews with senior health managers and the implementation team were conducted via web using video conferencing software. Interviews with end users were conducted in person in health care facilities during work hours and were conducted primarily in English, which is the standard working language in the facilities. However, participants were free to use local language terms or expressions, which were translated during transcription to preserve meaning.

Different semistructured interview guides were used for the 3 categories of participants. For the managers, we inquired about the broader digital health policy landscape, existing health information systems, government readiness, and feasibility for EHR implementation and the necessary next steps to advance EHR adoption in South Africa. For the implementation team and end users, the interview guides covered topics on the purpose and origins of the E-tick, its piloting rollout, experiences with implementation and daily use, training and support, perceived benefits and challenges, and overall recommendations for improvement. Interviews varied in length, ranging from 10 to 64 minutes. Shorter interviews (under 20 min) were mostly with busy frontline staff, including clerks and clinicians during clinic hours, who provided concise but relevant insights into system use. Longer interviews were with senior health managers with authority over policy alignment, resource allocation, and procurement processes, and implementers who facilitated implementation in the 3 health facilities. Despite these differences, data saturation was reached across stakeholder groups, as no new themes emerged in later interviews. All interviews were audio recorded, and recordings were stored in

XSL•FO
RenderX

a password-protected computer to ensure privacy. Audio recordings were transcribed using transcription software and checked for accuracy, and transcripts were anonymized by removing participant-identifying information.

Data Analysis

The first author led the coding, followed by regular debriefing meetings with coauthors, who are also academic supervisors, to review coding and theme development, which strengthened reliability. In this case, coauthors read all the transcripts, and discrepancies were discussed until consensus was reached. The interview transcripts were coded using NVivo qualitative analysis software (QSR International). Through this process, emergent issues were identified and used to develop the themes, following a 6-step thematic analysis process [33]. An inductive approach guided the thematic analysis, with broad topics derived from the interview guide, leading to the 4 stages of implementation reported in this study (Purpose and origins, Initial development, Piloting and expansion, and Halt). The accounts from the two originators formed the foundation, providing the background story and backbone of the implementation narrative, whose detailed recollections provided a chronological anchor for mapping implementation processes. These accounts were not taken at face value, but were supplemented by other participants' accounts, which were used to triangulate and integrate the information into a cohesive story.

After coding, the CFIR was applied to the analysis, linking the framework's 5 domains to the implementation process as reported in the study's findings (Table 3) [31]. The CFIR framework guided our analysis and reporting of findings. In this case, CFIR constructs were operationalized as sensitizing concepts to guide coding, with data excerpts being coded against relevant constructs, while leaving space for inductive subthemes to emerge. Thus, the analysis combined deductive coding (using CFIR constructs) with inductive coding to capture novel themes. Multiple coders reviewed transcripts, with regular team discussions to minimize confirmation bias.

Table . Implementation stages and applied Consolidated Framework for Implementation Research (CFIR) domains.

Stages of implementation	Description	Applied CFIR domains
Origins and purpose of the system (2015 - 2017)	Conception of E-tick[a] by system originators; designed to digitize the existing tick register and reduce errors.	• Innovation • Individuals • Inner setting • Outer setting
Initial development (2017 - 2019)	Early technical build and adaptation of the system, involving stakeholder input and limited resource mobilization.	• Individuals • Inner setting • Outer setting • Implementation process
Piloting (2019 - 2020)	Deployment of E-tick system in 3 health facilities.	• Individuals • Inner setting • Implementation process
Halt (2021)	Expansion beyond pilot sites stalled due to barriers to scale up and institutionalization.	• Inner setting • Outer setting

[a]E-tick: electronic tick register.

Ethical Considerations

Ethics approval for this study was granted by the University of the Witwatersrand Medical Human Research Ethics Committee (M210213). Additional permissions were obtained from the Ekurhuleni district research committee and the Gauteng Provincial Department of Health (GP_202106_035). The study was conducted in accordance with local, institutional, and national regulations governing research involving human participants, as well as the principles of the Declaration of Helsinki. All participants provided written informed consent prior to participation. The study purpose, procedures, potential risks, benefits, and participants' rights, including the right to withdraw at any time without penalty, were explained before consent was obtained. Participant privacy and confidentiality were maintained as interviews were conducted in private settings, and no identifying information was included in transcripts. All data were anonymized, securely stored on password-protected devices, and were only accessible to the research team. Participants did not receive any compensation for their participation.

Results

Overview

This section presents the findings, which are ordered according to the domains of the CFIR: innovation, inner setting, individual characteristics, implementation process, and outer setting. This structure allows us to highlight how factors at multiple levels and stages of implementation shaped the processes, outcomes, and ultimately the stalling of the E-tick system. To illustrate the distribution of themes across the framework, Multimedia Appendix 1 provides a summary of implementation stages, CFIR domains and applicable constructs, thematic findings, and illustrative quotes.

XSL•FO
RenderX

Origins and Purpose of the E-Tick: Innovation and Individuals Domains (2015-2017)

Identifying the Problem

The development and implementation of the E-tick was initiated and led by a senior district manager and supported by a senior technician who was employed by the funder, labeled as Originators 1 and 2. The senior manager explained the rationale behind the project:

> *Annually, the National Department of Health audits our data for accuracy, and we've always had poor audit reports. The biggest problem was the quality of the data. That's why we started the E-tick, to improve the data quality.* [Originator 1]

The source of poor data quality was traced back to the tick register: "The data was very poorly filled into the tick register, the services provided not counted properly." (Originator 1). This was the stage where most errors occurred: "The data issues started in the consultation room, and then the counting and the transferring of the data into the DHIS accounted for about eighty percent of the mistakes." (Originator 1). Some clinicians echoed these remarks as one shared: "The purpose was to reduce workload and be able to say, reduce patients' waiting time by doing away with pen and paper and to be able to keep our records digitally, which makes it easier to keep them." (Nurse 12 – AR).

The second originator further describes how decision makers often overlook the importance of effective data collection processes:

> *The problem in South Africa is that the people who decide where the funding goes don't take data management records management seriously. They want this wealth of data so they put in these big paper registers, but they don't look at the process it takes to get that data.* [Originator 2]

This includes having too many registers:

> *We had twenty-seven different registers, like the antenatal register for a mother and child, and a chronic disease register for HIV. The problem was that while completing all the registers, users were creating their own registers because some information was missing.* [Originator 1]

Nurses similarly described the burden, with one stating:

> *So with the E-tick it was one less writing system… our patients were registered, all the follow up visits are registered, and… it gave a clear history of the patient.* [Nurse 13 – MM]

Finding the Solution

Against this backdrop, the idea of the E-tick began with an invitation to a staff development project to discuss potential solutions to data quality problems:

> *We received an invitation because we already recognized a problem with data quality in our tick registers. They proposed several ways to enhance the system, involving different levels of information. While some suggestions were overly complex and resembled an Electronic Health System.* [Originator 1]

Eventually, the originators decided to make use of an Excel-based E-tick system:

> *The organization that developed the District Health Information System …had an Electronic Tick Register and we thought we can use that to improve the data quality. But the system was just an excel sheet.* [Originator 1]

Thus, the originators built on the Excel-based system:

> *We decided to improve the front end, the user interface for the clients. That's how we developed the E-tick. It was basically to collect the data electronically and to be able to feed it directly into the DHIS, in a way that is easy for the nurses in the consultation rooms.*
>
> *[Originator 1]*

End users confirmed that the system was simpler and helpful. One nurse explained:

> *It was user friendly. If only we would use E-tick only… we will definitely be perfect.* [Nurse 10 – AR]

Another shared:

> *And it was also user friendly. It wasn't like I'm talking from somebody that's not very computer literate, basically… it made it very easy, accessible.* [Nurse 13 – MM]

Initial Development of E-Tick System: Inner Setting and Individual Domains (2017-2019)

Stakeholder Engagement

The first originator approached a nonprofit organization for funding to get the project off the ground: "After our experience, we approached organization X for financial assistance and shared our ideas with them, as [they were already] our long-term funding partner deeply involved in data collection" (Originator 1). After the funders agreed, this was followed by meetings with software developers for planning: "Once the funders agreed, we met with [software] developers to explore system development possibilities, discussed utilizing the HPRS as our registration system." (Originator 1). It was at this point that the second originator, who worked for the funding organization, became involved. Together, the two originators represented 2 groups of stakeholders, one with the Department of Health and all staff at recipient facilities, and the other with the funding organization and developers:

> *[Originator 1] was very good on the clinician side and I was very good on the software side. I had been working with my team [funders] for many years before we started the E-tick, and I developed many systems with them.* [Originator 2]

As an iterative process, the originators worked closely between clinicians and developers:

XSL•FO
RenderX

> *In a normal software development cycle, you scope your business requirements, hand it over to the team, and they will go develop what they think you want and then return to show it to you. We took a different approach. We were involved in the development, sitting with them every step of the way instead of waiting for weeks for changes. This worked extremely well.* [Originator 2]

Several clinicians further confirmed this engagement, with one manager explaining:

> *The first early step was that we had meetings which included facility managers, other stakeholders and our partner [funder], to explain to us what wants to be introduced, which is when we were first given the idea of the E-tick and what it's supposed to do.* [Administrative officer 19 – RK]

The originators then selected the facilities for piloting involving managers in the process: "We used random selection and cluster sampling, through a session with all operational clinic managers, who actively participated in this process." (Originator 1). A manager from one of the facilities described their involvement: "We were called to a district meeting, and this was pitched, the intention was discussed with us... So the facilities that were going to be pilots were selected." (Assistant Director 17 – RK).

System Development

After selecting the facilities for piloting, the system was taken to the recipient facilities as one clinician recalls:

> *They called us for like a mini meeting to inform us about E-tick and told us that we were one of the only clinics currently using it, and they were piloting to see if it's going to work.* [Nurse 10 – AR]

In the meetings, one of the goals was to find ways to streamline the patient consultation:

> *We focused on what made it easy for clinicians to help the patients without them waiting too long. But a bonus that came out of the system is that the system was also able to retrieve a patient's entire history.* [Originator 2]

The clinicians provided feedback:

> *The clinicians had great ideas. The first time we had the screens ready, we took it back to the clinicians and asked them if this is what they wanted, they approved it and gave suggestions for changes, which we did.* [Originator 2]

Several engagements were made with the provincial Department of Health to gain support and input on issues such as data storage:

> *When we began the pilot, we engaged the Provincial office and kept them informed about the program. A provincial manager suggested storing the data on provincial servers instead. However, they weren't prepared to proceed at that time.* [Originator 1]

To ensure this was possible later, the developers ensured the system was compliant:

> *We looked at the Departments of Health security and ICT standards, we worked with the Department of Health ICT team as well, and their data team.* [Originator 2]

However, this did not go as planned, as the data center was not ready in time:

> *The idea was to eventually transfer the system and its data to the department's data center once it was established. However, when we developed it, there was no data center for us to put it in, and we had complied with all the national standards for the developing systems.* [Originator 2]

Piloting and Expansion: Inner Setting and Implementation Process Domains (2019-2020)

Planning and Configuration

In all 3 chosen facilities, the originators conducted an infrastructural assessment before beginning the pilot:

> *We had to check that all computers were working, and that they all had network. The developers didn't want to have a local server for each clinic because it's too complicated. So they changed the system to an online system that worked with the cloud.* [Originator 1]

Some of the existing infrastructure was from the Department of Health:

> *We only had to get the computers for the Provincial clinics. So that already made life much easier. Because every clinic, every clinician had a computer, or most of them.* [Originator 1]

A technician from the health department was involved in configuring them:

> *My role was to check infrastructure, to see if it's there. But they also installed some private infrastructure because of some policies with our network which would require a lot of approvals before we can run it.* [IT Technician]

This was further supplemented by the funding to improve the infrastructure:

> *They [funders] assisted us and made sure that each room had a network cable. Sometimes there were no plugs, so they put in the plugs. They put in the routers.* [Originator 1]

The planning and configurations also informed decisions on devices that were to be used:

> *We wanted to see what the nurses and the clinicians felt more comfortable with, the tablets or the desktops. Due to a lack of space in some consultation rooms, which were too small to put a big desktop, it was easier to mount a tablet onto the wall.* [Originator 2]

Training and Learning

A comprehensive user training was provided in recipient facilities, which was tailored for specific roles:

XSL•FO
RenderX

> *There were different training categories, because the roles were different like, clerks, clinicians and so on. It was [also] not just one training, it was several; this was because the facility was working on shifts, so they had to accommodate everyone. During rollout, there was another in-service training just to remind the staff.* [Nurse 17 – RK]

Moreover, the training was supplemented by regular visits for support and monitoring:

> *Then we had regular visits from [the originator], often enough for us to deal with whatever issues we were having on the system.* [Nurse 2 – MM]

The training was simple and accommodated varying levels of skills:

> *Some people could not use computers, but it was made as basic as possible to include those that are not computer literate.* [Nurse 17 – RK]

However, some users' competency and reactions to the training were influenced by age:

> *For the young ones, it was easy for them. For the old ones, it was a problem because they wanted to understand what they were ticking as they were pressing.* [Data capturer - 1]

In response, the originators sought assistance from data capturers, acting as an immediate source of help:

> *We had person in each clinic who would be there all day long. So for the first day or two they would sit with them [older users] when they completed it, until they became comfortable with it.* [Originator 1]

Support and Feedback Channels

To supplement the training, the originators had different people involved to provide support to end users, including IT technicians:

> *We provided desktop support, we were the first line of support. My role firstly was on all applications and IT projects running within the district.* [IT Technician]

While data capturers were already involved, especially in the early stages, admin people also assisted with various tasks:

> *We also had three admin people, one for each clinic that assisted in the clinic to help people with the training, setting up the system and checking that the computers are all working.* [Originator 1]

The originators also established communication channels to receive feedback from users:

> *We had a little book for users to report issues and to add anything, and an E-tick admin person was there. I also got information from data capturers and I would also go and check on everybody and they would give me feedback. We would report back to them monthly, usually.* [Originator 1]

Perceived Benefits of System Use

End users reported several benefits of the system, with some sharing how it improved continuity of care:

> *At times you find that the client may have been seen previously but the file could not be found. But if the client's details were captured on the previous visit, you could see it in the E-tick. So it helped with continuation of care.* [Nurse 11 – AR]

For other users, such as clerks, this reduced patient waiting times:

> *It also reduced the waiting period for the patient, because the first thing that we do is to register the patient and create a folder on the E-tick system. It's easier to use than retrieving the manual files, searching the files in the cabinets. When you're using an E-tick, you just search on the system.* [Admin Clerk 8 – AR]

Facility managers also found operational benefits, such as quicker service delivery:

> *The benefits were that the patient consultation times were shortened. It was a little bit faster. That was the best advantage that I've seen from the E-tick.* [Operations Manager 18 – RK]

Some users shared that the system improved data quality and access to it:

> *It improved the reporting of data. We had very reliable, consistent, and accurate data, at all times of the day.* [Administrative officer 19 – RK]

For some, the system enabled the monitoring of the clinic's performance:

> *We are able to evaluate ourselves on the number of patients that we are seeing. When our stats are sent to the DHIS, they can see that we are really performing.* [Nurse 4 – MM]

Perceived Challenges of System Use

Users reported several challenges with using the system, including the simultaneous use of the system with the manual:

> *To be brutal, they [users] were against it at first. They did not understand why we're implementing extra work for them.* [Administrative officer 19 – RK]

For some participants, the early days before mastering the system were difficult:

> *If you're not used to it yet, it takes time to get used to anything, especially if you're overloaded with a lot of patients.* [Nurse 14 – MM]

In some cases, short staffing led nurses to do administrative work, as one nurse shared:

> *I think it was mostly shortage of staff, sometimes I had to enter a new patient into the E-tick as a nurse, which would waste time and to delay the queue. You'd end up having to turn some patients back. So it needs us to be well staffed.* [Nurse 12 – AR]

XSL•FO
RenderX

In response, the originators designated roles for receptionists, clerks, and data capturers as the first point of contact for patient registrations.

There were concerns about the hardware and infrastructure that already existed in the facilities: "The [DOH's] ICT equipment was very outdated. Some of the computers are so old and they haven't been updated." (Originator 1 – R). Some users had challenges with the network infrastructure:

> *Our network system was failing us a lot. You'd be prepared to start the day and notice that you have no network.* [Nurse 12 – AR]

In response to this challenge, a back capture function was introduced, which allowed information to be captured into the system at a later stage. In one site, hardware was frequently stolen:

> *First disadvantage was theft. I'd say more than more than 70% of the tablets disappeared. All the rooms that had extension cords, but in less than two weeks, all the rooms had no extension cords. They also stole the chargers.* [Administrative officer 19 – RK]

In one clinic, the piloting coincided with ongoing protests by nurses:

> *When we were trying to start the system, the nurses and others would be outside complaining about the conditions in the clinic [a protest organized by unions]. Most of the nurses were not working. There were other issues in the clinic. I think they saw the E-tick as another thing that was worth protesting about, because it increased their workload.* [Originator 1 – R]

Halt and Barriers to Scale Up and Institutionalization: Outer Setting Domain (2021)

Changing Developers

Despite its relative success in implementation, several complications affected the continued use and development of the system. One of the complications came when there was a change in the software development company:

> *Originally [funders] were working with [a software developing company], who were helping with developing district programs. They were taken over by [another company] so the people that worked on our system were taken off the project, and a new team took over the development.* [Originator 1 – R]

This affected further developments:

> *We needed to update the system to make sure it's fine. Then it got stuck in that development phase because of the new developers that we have.* [Originator 1 – R]

Some of the previous work was undone by the new developers:

> *There were a lot of upgrades to the system, such as integration with the DHIS. But they removed some of the functionalities that were working well in the process. We were never happy with the final products that we got from them.* [Originator 1 – R]

Funding Cuts

The change of developers brought higher costs:

> *They [funders] agreed to pay for the additional development, but the new developers came with a lot of red tape and were quite expensive. The funding got cut and they could only give us a small amount to continue with the development.* [Originator 1 – R]

Moreover, the funders were eventually withdrawn from operating in the health district:

> *They [funders] moved out of the district, because their funding got canceled completely. So what's happening with their funding will affect our funding.* [Originator 1 – R]

This had an impact on keeping the system operational:

> *So because it's not open source, it was built on the Microsoft platform and the cost to keep it running is high. It's also web based so no offline capability, and we need constantly running servers.* [Originator 2 – T]

Thus, the system needed new funders:

> *We're looking for partners to fund the service, license costs and the implementation in facilities. To date, we haven't found anyone, so it's been put to sleep.* [Originator 2 – T]

Provincial and District Support

The originators had buy-in from district officials, who were in support of the system:

> *I had meetings with my chief director, and my director, and they are still very keen. The clinics are still very keen… So the buy-in is there, even after two to three years, but the funding is not there.* [Originator 1 – R]

However, the provincial health office's priority was to develop a new system, as the department technician explained:

> *Currently the department is rolling out an HIS platform where all other applications will be layered under and integrated into.* [IT Technician]

He further explained how locally developed systems were not as prioritized:

> *There were some contrasting and opposing views. The system was started from scratch but the department in most cases doesn't want that. They want to go on the market and buy something on the shelf. I don't think approvals were going to be granted.* [IT Technician]

The participant further mentions how the IT staff are often not seen as essential to the role of delivering care:

> *Sometimes different services compete for resources. They will say it's essential when it's a clinic, a doctor*

XSL•FO
RenderX

> *or medication, and then we [IT personnel] come last.* [IT Technician]

The originators also expressed their frustration with navigating such terrain:

> *So this is where we are stuck now. I think what is important to me is how difficult it is to roll out a system within the Department of Health, in the politics, the corruption and the agendas of people. It makes the story very complicated because we developed it the bottom up, seeing the need, with the buy-in from the district, and from all the role players.* [Originator 1 – R]

Broader Context (Outer Setting Domain)

Senior managers and external stakeholders shared insights into the Department of Health's challenges in implementing EHRs in South Africa. One manager notes how implementation requires a joint and collaborative effort of various stakeholders:

> *I think it's a complex sociotechnical project to do which requires many different role players and expertise and costs, which needs high level support from the National Department of Health which has been lacking.* [NPO manager]

The provincial health departments hold most of the decision-making power in digital health spending:

> *The national office can come up with guidelines and frameworks and ask for reports but the actual spending on health information systems is done at provincial.* [NPO manager]

In this case, the provincial health department had decided to prioritize an off-the-shelf system:

> *There is a provincial initiative and anything new that comes in has to talk to this one, there should be interoperability…I saw it [E-tick], it works. I tried to support it but we didn't get provincial support from the CIO.* [Provincial manager]

However, NGOs have been supporting and implementing local efforts to build systems (such as E-tick) in health care facilities:

> *NGOs are doing better than the districts and the provincial health in terms of implementing efficient systems. Most efficient systems are systems that have been introduced by NGO's instead of our provincial department.* [Provincial IT technician]

Some participants raised a concern over political appointments in key decision-making positions:

> *I don't believe that we have the right people in government with the right industry experience. They are just given the title and suddenly, they now have to make decisions. I think it's very wrong and it sets them up for failure.* [NGO Technician]

A health department technician added:

> *Most of our activities are driven by politicians while our head of department has his own strategies. We get our mandate from politicians.* [DoH Technician]

As a result, some decisions are made with bias:

> *There are bureaucratic officials in Department of Health national, provincial and local. If they do not like you and your project, they will say you have to go through SETA [a parastatal] which is extraordinarily slow, not particularly technically competent, and staggeringly expensive.* [NPO manager]

Moreover, one academic shared their experience with the Department of Health's reluctance:

> *I think there's not sufficient trust on the part of Government certainly in my experience. They are not coming to partner with us. In fact, they're pushing us away. We only get close when we work with NGO's and so on. The Government definitely keeps us at arms-length.* [Health Academic]

Corruption and inefficiencies have also created barriers, with the tender system being one of the issues:

> *It [corruption] is part of the tender system in South Africa, unfortunately the tender system is open to abuse.* [Provincial health manager - IT]

Moreover, past corruption scandals have tainted the reputation of the Department of Health:

> *The perception is that you can't trust the National Department any longer to do anything. No one is going to give the National Department a big pot of money to run a tender until it can show that it's got systems and people in place who can do it without squandering the money.* [National health manager]

This mistrust has also led to a reluctance to work within the department:

> *Now people are reluctant to work for the department for several reasons. They fear being embroiled in controversy and putting their personal integrity at stake.* [National health manager]

Discussions

Principal Findings

The E-tick system, an innovation, was developed to address the issue of poor data quality stemming from errors in data collection, which were common with the use of paper-based registers. The implementation process involved an iterative co-design process with input from all involved individuals, including end users and stakeholders, to ensure users' needs were met and to streamline the data entry process while supporting scalability and integration with other systems. Despite the system's successful piloting and user adoption in the inner setting, the outer setting posed challenges in funding, change of developers, and lack of support from the provincial health government, leading to the implementation stalling. Using the CFIR, we explored the interaction of different elements, actors, and stakeholders and how they facilitated or impeded implementation. As such, we organize this discussion using CFIR domains, highlighting both facilitators and barriers.

XSL•FO
RenderX

Innovation

Poor data quality is a problem that commonly affects routine monitoring of health outcomes [34]. The E-tick system had a clear relative advantage over paper-based registers to improve data quality and reporting, as errors that frequently compromised data quality and routine reporting were minimized, and users perceived the system as more streamlined than paper-based manual registers. At a national scale, this aligns with South Africa's policies that promote the digitization of record keeping and the establishment of a national EHR [10,35], a crucial component for strategic purchasing and health monitoring. The E-tick's co-design process enhanced usability and fostered buy-in, which is critical in adoption [36], yet innovation alone was insufficient to ensure sustainability.

Individuals

At the level of individuals, the originators showed strong commitment by providing vision, technical expertise, and persistence in driving the implementation of the E-tick system. Leadership practices are known to impact staff motivation, especially when they take a participatory approach [37,38]. Their leadership was instrumental in mobilizing resources, persuading managers and funders to support the system's adoption, acting as champions along with data capturers. Research shows how champions are crucial in promoting technology adoption through building enthusiasm and providing peer support [39]. End users also improved confidence in using the system as a result, especially after undergoing training, which sustained motivation during the pilot. However, while the successful adoption of the E-tick was due to the contribution of many actors, including external stakeholders, individuals in the inner setting were more affected by decisions in the outer setting. This reflects a broader challenge in digital health implementation in South Africa, that for locally developed systems to succeed, appropriate actors must be in place to provide leadership, vision, and sustained support, within an enabling environment that fosters continuity and long-term success [37,40].

Inner Setting

Within the inner setting, several facilitators supported early implementation, such as the co-design process and extensive user training. The originators successfully engaged a wide range of stakeholders, including senior managers and funders, which was pivotal in overcoming initial financial and political barriers. Involving end users was also crucial in enhancing the system's usability and promoting buy-in, an important aspect when ensuring effective user adoption [41,42]. Moreover, users' recognition of the system's value in improving data quality reinforced their motivation to adopt the system. Studies in other low- to middle-income countries show that organizational commitment is a critical success factor to sustain user motivation during piloting phases [43]. However, the sustained use of the E-tick required stable infrastructure, which often posed challenges for many users in the form of interrupted internet connectivity and power cuts. This mirrors challenges reported in other LMICs, where infrastructural fragilities, such as poor internet connectivity and unstable power supply, are common challenges to EHR adoption [44]. Moreover, without dedicated budgets to overcome these challenges, such as the use of mobile internet devices and backup power solutions, this creates uncertainty about sustained use.

Implementation Process

The implementation process revealed a mix of strengths and gaps, which were closely linked to the role of individuals and the inner setting. A key strength was the iterative co-design approach, which ensured that the system was not only technically functional but also responsive to user needs and compatible with existing workflows. Participatory design and user engagement are widely recognized in implementation science as best practice, as they enhance user ownership, adaptability, and long-term adoption, especially as most implementation efforts fail due to user resistance [45,46]. Thus, involving both clinicians and managers at early stages of the implementation process helped to align the E-tick system with user realities, reducing resistance to change. However, this process was undermined by a lack of a long-term financing mechanism and failed provincial support, despite the demonstrated proof of concept at the pilot phase. Therefore, while the implementation process was well executed, the system lacked the institutional anchoring and support for long-term use, a critical success factor of digital health efforts [47].

Outer Setting

We found that the most significant challenges were encountered in the outer setting, particularly the withdrawal of financial support, the change of developers, and the lack of support from the provincial health government due to its lack of capacity. Studies show that one of the main challenges to the initiation of digital health initiatives, including EHRs in South Africa, has been the poor governance and lack of political will [48,49]. For instance, the appointment of people without the required skills and competence has led to a chain of events that affect institutions such as the Gauteng provincial health department [50]. As a result, corruption has eroded the public's trust in the government. These issues are not isolated but rather reflective of the entrenched political economy dynamics that routinely affect digital health innovations in South Africa. Political appointments, who are put in place because they are easily manipulated, often make decisions driven by political interests rather than professional judgment, leading to the political control of government institutions and their spending [51]. As such, patronage networks often produce politically aligned rather than skilled appointments. This leaves room for corruption, irregular spending, fraud, and poor oversight to mitigate them, with procurement processes being vulnerable to perverse incentives, with contracts awarded to less qualified but politically connected vendors. Although efforts have been made to uncover large-scale corruption in South Africa, the necessary mechanisms have not yet been put in place to deal with the challenges [52,53]. Perverse incentives undermine sustainability as officials face pressure to launch new initiatives instead of sustaining existing ones to ensure full expenditure of budgets.

The administrative fragmentation can be seen as a product of South Africa's federal system, which means provinces can develop their own health information systems (bottom-up) or acquire them through separate procurement processes

XSL•FO
RenderX

(top-down) [8]. In this case, despite the provincial government's preference for a top-down approach by opting for an off-the-shelf system over the E-tick, they were unable to implement the former. High-income countries such as the United Kingdom, the United States, and Australia have taken top-down, bottom-up, and middle-out approaches, respectively, with varying outcomes [54]. Bottom-up approaches often build on existing systems, such as the E-tick, using interoperability standards to integrate them across facilities. In contrast, top-down approaches replace local systems with a centralized EHR, but this can be too complex, limiting flexibility and local adaptability, as in the case of the United Kingdom [55]. Therefore, our analysis shows that a single centralized model is not necessarily more effective, as lessons from these experiences, including Australia, suggest that a middle-out approach is more effective, as this blends both by offering central guidance while allowing for local systems to gradually align with national plans [56].

Implementing EHRs requires significant investment, vision, robust management systems, and strong oversight to sustain them, making effective leadership essential for success. Strong leadership plays an important role in guiding decision making, mobilizing the right expertise, and establishing mechanisms to prevent corruption, which increases the likelihood of successful implementation, particularly for large-scale EHRs [57]. Middle-income countries like South Africa and China have the capacity to finance and support EHR adoption, but such efforts can be undermined by corruption and poor governance [49,58]. In some cases, poor leadership and governance stand as significant barriers to achieving universal health coverage and strengthening health systems [59].

Implications for Policy and Practice

Our findings highlight several implications for policy and practice. One of them is the strengthening of accountability in health systems through transparent procurement systems and independent oversight to neutralize the effects of corruption and political patronage. In such cases, procurement frameworks could include pilot-to-scale pathways, where bottom-up innovations such as the E-tick could be formally assessed against national priorities and considered for integration with provincial and national systems. These steps could create a more enabling environment for similar innovations to scale sustainably. Moreover, while the E-tick system was locally hosted, it also incorporated cloud functionality, aligning with the cloud-based EHR architectures promoted in both the NDHS and the national data and cloud policy of South Africa [10,60]. Such policies can be leveraged to mitigate barriers such as reliance on facility servers or hardware failures, challenges already observed with the E-tick, which hold important lessons for similar middle-income country contexts.

Our findings also carry implications for digital health equity, a concept that examines how structural inequities shape the successes and challenges of digital health interventions [61]. In South Africa's dual health system, where private-sector facilities often have better infrastructure and resources than the public sector, grassroots innovations like E-tick are left to face compounded challenges, especially when implemented in low-resource contexts [62]. The stalled scale-up reflects not only outer system or structural barriers by the government but also the broader inequities in digital infrastructure, institutional support, and resourcing across the country's health system [63].

Limitations

Our study incorporates perspectives from a diverse range of stakeholders, including higher-level managers, decision makers, academics, and end users from recipient facilities. This diversity helped in capturing nuanced perspectives from different positions within the system and in closing information gaps through triangulation. A limitation of our study is that data collection from the implementation team, especially clinicians, occurred over a year after the system had halted. As a result, some participants struggled to accurately recall past events, while others had moved to different jobs by the time of data collection. As a result, formal member checking with participants was not possible due to staff turnover and the discontinuation of the E-tick project. However, we validated emerging findings through iterative review with coauthors who included experts in South African health systems and originators who led the implementation from the beginning and had contextual expertise.

Conclusions

In conclusion, the E-tick system is a valuable innovation that was developed from the ground up with user input. While the system showed promise, its long-term sustainability was undermined by external factors such as funding cuts and poor support, political appointments, and corruption, which led to the unwillingness of provincial officials to support local initiatives such as the E-tick. Such bureaucratic inefficiencies also affected the capacity to move forward with another system. This study demonstrates that without alignment to broader factors such as governance, corruption, and skilled leadership, innovative systems such as the E-tick will struggle to be implemented.

Acknowledgments

We would like to thank the study participants for generously sharing their time and experiences. We also acknowledge the support of the implementation team and the provincial and district health authorities for facilitating access to study sites.

XSL•FO
RenderX

Funding

This work is part of a PhD study funded by the South African National Research Foundation (NRF) through the SARChI programme for health systems and policy research at the Centre for Health Policy, School of Public Health, University of the Witwatersrand. Scholarship number: PMDS2205046516. The funders were not involved in writing this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of findings by implementation stage, alignment with Consolidated Framework for Implementation Research (CFIR) domains, applicable constructs, thematic findings, and illustrative quotes.
[DOCX File, 22 KB - medinform_v14i1e73831_app1.docx]

References

1. Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature. J Med Internet Res 2021 Dec 23;23(12):e26323. [doi: 10.2196/26323] [Medline: 34941544]
2. Kumar M, Mostafa J. Electronic health records for better health in the lower- and middle-income countries. LHT 2020 Mar 10;38(4):751-767. [doi: 10.1108/LHT-09-2019-0179]
3. Mehl G, Labrique A. Prioritizing integrated mHealth strategies for universal health coverage. Science 2014 Sep 12;345(6202):1284-1287. [doi: 10.1126/science.1258926] [Medline: 25214614]
4. Global Strategy on Digital Health 2020-2025, 1st edition: World Health Organization; 2021.
5. Kumar M, Mostafa J. Research evidence on strategies enabling integration of electronic health records in the health care systems of low- and middle-income countries: a literature review. Int J Health Plann Manage 2019 Apr;34(2):e1016-e1025. [doi: 10.1002/hpm.2754] [Medline: 30762907]
6. Wilson D, Sheikh A, Görgens M, Ward K, Bank W. Technology and universal health coverage: examining the role of digital health. J Glob Health 2021;11:16006. [doi: 10.7189/jogh.11.16006] [Medline: 34912559]
7. Fragidis LL, Chatzoglou PD. Implementation of a nationwide electronic health record (EHR). Int J Health Care Qual Assur 2018 Mar 12;31(2):116-130. [doi: 10.1108/IJHCQA-09-2016-0136] [Medline: 29504871]
8. Eason K, Dent M, Waterson P, Tutt D, Thornett A. Bottom-up and middle-out approaches to electronic patient information systems: a focus on healthcare pathways. J Innov Health Inform 2013 Dec 20;20(1):51-56. [doi: 10.14236/jhi.v20i1.47]
9. Robertson A, Cresswell K, Takian A, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation. BMJ 2010 Sep 1;341(sep01 3):c4564. [doi: 10.1136/bmj.c4564] [Medline: 20813822]
10. National Department of Health South Africa. National digital health strategy for South Africa 2019–2024. : Pretoria South Africa; 2019.
11. Horwood C, Luthuli S, Mapumulo S, et al. Challenges of using e-health technologies to support clinical care in rural Africa: a longitudinal mixed methods study exploring primary health care nurses' experiences of using an electronic clinical decision support system (CDSS) in South Africa. BMC Health Serv Res 2023 Jan 13;23(1):30. [doi: 10.1186/s12913-022-09001-2] [Medline: 36639801]
12. Masana N, Muriithi GM. Adoption of an integrated cloud-based electronic medical record system at public healthcare facilities in free-state, South Africa. Presented at: 2019 Conference on Information Communications Technology and Society (ICTAS); Mar 6-8, 2019; Durban, South Africa p. 1-6. [doi: 10.1109/ICTAS.2019.8703606]
13. Hassim A, Heywood M, Honermann B. The National Health Act 61 of 2003: a guide. : Siber Ink; 2008.
14. Chuma K, Sibiya P. Digital health ecosystem framework to address fragmentation of the health system in South Africa. AJNM 2022 Nov 10. [doi: 10.25159/2520-5293/11547]
15. Fourie D, Malan C. Public procurement in the South African economy: addressing the systemic issues. Sustainability 2020 Oct 20;12(20):8692. [doi: 10.3390/su12208692]
16. Garrib A, Stoops N, McKenzie A, et al. An evaluation of the district health information system in rural South Africa. S Afr Med J 2008 Jul;98(7):549-552. [Medline: 18785397]
17. English R, Masilela T, Barron P, Schönfeldt A. Health information systems in South Africa. In: Padarath A, English R, editors. South African Health Review 2011: Health Systems Trust; 2011:81-90.
18. Oyebola OG, Debra J, Thubelihle M. PMTCT data management and reporting during the transition phase of implementing the rationalised registers in Amathole District, Eastern Cape Province, South Africa. Int J Environ Res Public Health 2022 Nov 28;19(23):15855. [doi: 10.3390/ijerph192315855] [Medline: 36497931]
19. National department of health. national health insurance for south africa: towards universal health coverage. Gov Gaz. 2015 URL: https://www.gov.za/sites/default/files/gcis_document/201512/39506gon1230.pdf [accessed 2023-08-12]

XSL•FO
RenderX

20. Wright G, O'Mahony D, Cilliers L. Electronic health information systems for public health care in South Africa: a review of current operational systems. J Health Inform Afr 2017 Nov 5;4(1). [doi: 10.12856/JHIA-2017-v4-i1-164]
21. Olu O, Muneene D, Bataringaya JE, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? a perspective. Front Public Health 2019;7:341. [doi: 10.3389/fpubh.2019.00341] [Medline: 31803706]
22. Rispel LC, de Jager P, Fonn S. Exploring corruption in the South African health sector. Health Policy Plan 2016 Mar;31(2):239-249. [doi: 10.1093/heapol/czv047] [Medline: 26104821]
23. Zali M. Health department suspends officials over Digital Vibes contract. Mail Guard. 2021. URL: https://mg.co.za/news/2021-09-30-health-department-suspends-officials-over-digital-vibes-contract/ [accessed 2024-11-11]
24. Ndaba B. Babita Deokaran killed for exposing R1bn corruption at Tembisa Hospital, says Panyaza Lesufi. 2022. URL: https://www.iol.co.za/pretoria-news/news/babita-deokaran-killed-for-exposing-r1bn-corruption-at-tembisa-hospital-says-panyaza-lesufi-fbdb09a1-6e0e-4006-bbb3-a90f17b22279 [accessed 2024-11-11]
25. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254. [doi: 10.2196/11254] [Medline: 30545807]
26. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022 Jan 27;5(1):13. [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
27. McAlearney AS, Walker DM, Livaudais-Toman J, Parides M, Bickell NA. Challenges of implementation and implementation research: learning from an intervention study designed to improve tumor registry reporting. SAGE Open Med 2016;4:2050312116666215. [doi: 10.1177/2050312116666215] [Medline: 27635248]
28. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
29. Kirk MA, Kelley C, Yankey N, Birken SA, Abadie B, Damschroder L. A systematic review of the use of the consolidated framework for implementation research. Implementation Sci 2015 Dec;11(1):72. [doi: 10.1186/s13012-016-0437-z]
30. Keith RE, Crosson JC, O'Malley AS, Cromp D, Taylor EF. Using the consolidated framework for implementation research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. Implement Sci 2017 Feb 10;12(1):15. [doi: 10.1186/s13012-017-0550-7] [Medline: 28187747]
31. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
32. Damschroder LJ, Lowery JC. Evaluation of a large-scale weight management program using the consolidated framework for implementation research (CFIR). Implement Sci 2013 May 10;8(1):51. [doi: 10.1186/1748-5908-8-51] [Medline: 23663819]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. Radebe L, Haeri Mazanderani A, Sherman GG. Evaluating patient data quality in South Africa's national health laboratory service data warehouse, 2017-2020: implications for monitoring child health programmes. BMC Public Health 2022 Jun 29;22(1):1266. [doi: 10.1186/s12889-022-13508-y] [Medline: 35768861]
35. Kante M, Ndayizigamiye P. Internet of medical things, policies and geriatrics: an analysis of the national digital health strategy for South Africa 2019–2024 from the policy triangle framework perspective. Scientific African 2021 Jul;12:e00759. [doi: 10.1016/j.sciaf.2021.e00759]
36. Duffy A, Boroumandzad N, Sherman AL, Christie G, Riadi I, Moreno S. Examining challenges to co-design digital health interventions with end users: systematic review. J Med Internet Res 2025 Mar 14;27:e50178. [doi: 10.2196/50178] [Medline: 40085834]
37. Gilson L, Agyepong IA. Strengthening health system leadership for better governance: what does it take? Health Policy Plan 2018 Jul 1;33(suppl_2):ii1-ii4. [doi: 10.1093/heapol/czy052] [Medline: 30053034]
38. Open Mindsets: Participatory Leadership for Health: World Health Organization; 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/251458/9789241511360-eng.pdf
39. Pettersen S, Eide H, Berg A. The role of champions in the implementation of technology in healthcare services: a systematic mixed studies review. BMC Health Serv Res 2024 Apr 11;24(1):456. [doi: 10.1186/s12913-024-10867-7] [Medline: 38605304]
40. Katurura MC, Cilliers L. Electronic health record system in the public health care sector of South Africa: a systematic literature review. Afr J Prim Health Care Fam Med 2018 Nov 20;10(1):e1-e8. [doi: 10.4102/phcfm.v10i1.1746] [Medline: 30456963]
41. Cresswell K, Morrison Z, Crowe S, Robertson A, Sheikh A. Anything but engaged: user involvement in the context of a national electronic health record implementation. Inform Prim Care 2011;19(4):191-206. [doi: 10.14236/jhi.v19i4.814] [Medline: 22828574]

XSL•FO
RenderX

42. Hartzler A, McCarty CA, Rasmussen LV, et al. Stakeholder engagement: a key component of integrating genomic information into electronic health records. Genet Med 2013 Oct;15(10):792-801. [doi: 10.1038/gim.2013.127] [Medline: 24030437]
43. Xiong S, Lu H, Peoples N, et al. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023 Feb 1;6(1):12. [doi: 10.1038/s41746-023-00764-4] [Medline: 36725977]
44. Hui CY, Abdulla A, Ahmed Z, et al. Mapping national information and communication technology (ICT) infrastructure to the requirements of potential digital health interventions in low- and middle-income countries. J Glob Health 2022 Dec 29;12:04094. [doi: 10.7189/jogh.12.04094] [Medline: 36579436]
45. Zharima C, Mhlanga S, Abdulla S, Goudge J, Griffiths F. What engagement strategies are useful in facilitating the implementation of electronic health records in health care settings? a rapid review of qualitative evidence synthesis using the normalization process theory. Digit Health 2024;10:20552076241291286. [doi: 10.1177/20552076241291286] [Medline: 39497787]
46. Voorheis P, Petch J, Pham Q, Kuluski K. Maximizing the value of patient and public involvement in the digital health co-design process: a qualitative descriptive study with design leaders and patient-public partners. In: König LM, editor. PLOS Digit Health 2023 Oct;2(10):e0000213. [doi: 10.1371/journal.pdig.0000213] [Medline: 37878566]
47. Bostan S, Johnson OA, Jaspersen LJ, Randell R. Contextual barriers to implementing open-source electronic health record systems for low- and lower-middle-income countries: scoping review. J Med Internet Res 2024 Aug 1;26:e45242. [doi: 10.2196/45242] [Medline: 39088815]
48. Rheeder R, Department of Moral Development of Society, Faculty of Theology, North-West University, Potchefstroom, South Africa. Corruption in the public health sector in South Africa: a global bioethical perspective. SAJBL 2021 Dec 31;14(3):79-83. [doi: 10.7196/SAJBL.2021.v14i3.693]
49. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. Front Digit Health 2023;5:1207602. [doi: 10.3389/fdgth.2023.1207602] [Medline: 37600481]
50. Fana TE, Goudge J. Managing under austerity: a qualitative study of management-union relations during attempts to cut labour costs in three South African public hospitals. J Health Organ Manag 2024 Mar 5;38(9):89-105. [doi: 10.1108/JHOM-11-2022-0324] [Medline: 38448233]
51. Swanepoel PCF. The slippery slope to state capture: cadre deployment as an enabler of corruption and a contributor to blurred party-state lines. Law Democr Dev 2022;25(1):1-23. [doi: 10.17159/2077-4907/2021/ldd.v25.15]
52. Mcebisi J. South Africa needs these five fundamentals to change course and start rebuilding. Daily Maverick. 2022. URL: https://www.dailymaverick.co.za/article/2022-10-21-south-africa-needs-these-five-fundamentals-to-change-course-and-start-rebuilding/ [accessed 2024-06-18]
53. Kulati P. South Africans are sick to death of persistent corruption by politicians. Daily Maverick. 2025. URL: https://www.dailymaverick.co.za/opinionista/2025-07-21-south-africans-sick-to-death-of-persistent-corruption-by-politicians/ [accessed 2025-07-28]
54. Coiera E. Building a national health IT system from the middle out. J Am Med Inform Assoc 2009;16(3):271-273. [doi: 10.1197/jamia.M3183] [Medline: 19407078]
55. Cresswell KM, Worth A, Sheikh A. Comparative case study investigating sociotechnical processes of change in the context of a national electronic health record implementation. Health Informatics J 2012 Dec;18(4):251-270. [doi: 10.1177/1460458212445399] [Medline: 23257056]
56. Boudreau KJ. Let a Thousand Flowers Bloom? An Early Look at Large Numbers of Software App Developers and Patterns of Innovation. Organization Science 2012 Oct;23(5):1409-1427. [doi: 10.1287/orsc.1110.0678]
57. Alzghaibi HA, Hutchings HA. Exploring facilitators of the implementation of electronic health records in Saudi Arabia. BMC Med Inform Decis Mak 2022 Dec 7;22(1):321. [doi: 10.1186/s12911-022-02072-5] [Medline: 36476224]
58. Fu H, Lai Y, Li Y, Zhu Y, Yip W. Understanding medical corruption in China: a mixed-methods study. Health Policy Plan 2023 Apr 11;38(4):496-508. [doi: 10.1093/heapol/czad015] [Medline: 36798965]
59. Kodali PB. Achieving universal health coverage in low- and middle-income countries: challenges for policy post-pandemic and beyond. Risk Manag Healthc Policy 2023;16:607-621. [doi: 10.2147/RMHP.S366759] [Medline: 37050920]
60. National policy on data and cloud (notice 2533 of 2024, government gazette no50741). Government of South Africa.: Republic of South Africa: Department of Communications and Digital Technologies; 2024. URL: https://www.gov.za/sites/default/files/gcis_document/202406/50741gen2533.pdf [accessed 2024-09-21]
61. Gordon T, Booysen F, Mbonigaba J. Socio-economic inequalities in the multiple dimensions of access to healthcare: the case of South Africa. BMC Public Health 2020 Mar 4;20(1):289. [doi: 10.1186/s12889-020-8368-7] [Medline: 32131787]
62. Pillay Y, Motsoaledi PA. Digital health in South Africa: innovating to improve health. BMJ Glob Health 2018;3(Suppl 2):e000722. [doi: 10.1136/bmjgh-2018-000722] [Medline: 29713513]
63. Kim KK, Backonja U. Digital health equity frameworks and key concepts: a scoping review. J Am Med Inform Assoc 2025 May 1;32(5):932-944. [doi: 10.1093/jamia/ocaf017] [Medline: 39936843]

XSL•FO
RenderX

Abbreviations

CFIR: Consolidated Framework for Implementation Research
DHIS: District Health Information Software
E-tick: electronic tick register
EHR: electronic health record
HPRS: Health Patient Registration System
NDHS: National Digital Health Strategy

Edited by A Benis; submitted 13.Mar.2025; peer-reviewed by K Adegoke, Y Chu, Z Zhang; revised version received 29.Oct.2025; accepted 16.Nov.2025; published 12.Jan.2026.

Please cite as:
Zharima C, Griffiths F, Goudge J
Exploring Factors Associated With the Stalled Implementation of a Ground-Up Electronic Health Record System in South Africa: Qualitative Insights From the E-Tick Case Study Using the Consolidated Framework for Implementation Research (CFIR)
JMIR Med Inform 2026;14:e73831
URL: https://medinform.jmir.org/2026/1/e73831
doi:10.2196/73831

© Campion Zharima, Frances Griffiths, Jane Goudge. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of Quality Indicators for the Correct Use of Electronic Medical Records in Primary Care: Modified Delphi Study

Rico Paridaens[1], MD; Steve Van den Bulck[1,2], MD, PhD; Michel De Jonghe[3], MD; Benjamin Fauquert[4], MD; Liesbeth Meel[5], MSc; Willem Raat[1], MD, PhD; Bert Vaes[1], MD, PhD

[1]Academisch Centrum voor Huisartsgeneeskunde, Katholieke Universiteit Leuven (KU Leuven), Leuven, Flanders, Belgium

[2]Research Group Healthcare and Ethics, Hasselt University, Hasselt, Belgium

[3]Centre Académique de Médecine Générale, Université catholique de Louvain, Leuven, Belgium

[4]Unité de Recherche en Soins Primaires, Vrije Universiteit Brussel, Brussels, Belgium

[5]Belgian Centre for Evidence-Based Medicine (CEBAM), Leuven, Belgium

Corresponding Author:
Rico Paridaens, MD
Academisch Centrum voor Huisartsgeneeskunde
Katholieke Universiteit Leuven (KU Leuven)
Kapucijnenvoer 7
Block G, A and N/H, box 7001
Leuven, Flanders, 3000
Belgium
Phone: 32 16324010
Email: rico.paridaens@kuleuven.be

Abstract

Background: When used correctly, electronic medical records (EMRs) can support clinical decision-making, provide information for research, facilitate coordination of care, reduce medical errors, and generate patient health summaries. Studies have reported large differences in the quality of EMR data.

Objective: Our study aimed to develop an evidence-based set of electronically extractable quality indicators (QIs) approved by expert consensus to assess the good use of EMRs by general practitioners (GPs) from a medical perspective.

Methods: The RAND-modified Delphi method was used in this study. The TRIP and MEDLINE databases were searched, and a selection of recommendations was filtered using the specific, measurable, assignable, realistic, and time-bound principles. The panel comprised 12 GPs and 6 EMR developers. The selected recommendations were transformed into QIs as percentages.

Results: A combined list of 20 indicators and 30 recommendations was created from 9 guidelines and 4 review articles. After the consensus round, 20 (100%) indicators and 20 (67%) recommendations were approved by the panel. All 20 recommendations were transformed into QIs. Most (16, 40%) QIs evaluated the completeness and adequacy of the problem list.

Conclusions: This study provided a set of 40 EMR-extractable QIs for the correct use of EMRs in primary care. These QIs can be used to map the completeness of EMRs by setting up an audit and feedback system, and to develop specific (computer-based) training for GPs.

(JMIR Med Inform 2026;14:e80057) doi:10.2196/80057

KEYWORDS

electronic medical records; primary care; quality assessment; Delphi method; electronic audit and feedback

Introduction

With the advancement of digital technologies, electronic medical records (EMRs) have become the preferred method for recording, storing, and retrieving medical information [1]. The National Alliance for Health Information Technology defined the EMR as “an electronic record of health-related information on an individual that can be created, gathered, managed and consulted by authorised clinicians and staff within one health care organisation” [2,3]. Several studies have shown that the correct use of EMRs may improve the quality and efficiency of care and reduce mortality [4-6]. When used correctly, EMRs can support clinical decision-making (eg, monitoring medication safety) [7-9], facilitate care coordination, reduce medical errors,

XSL•FO
RenderX

and provide information for research [1,10]. EMR data can also be used to generate patient health summaries (PHSs). The PHS is a minimum set of clinically relevant data that can be transferred to other health care workers to support the continuity of care and deliver safe and high-quality care to patients [11,12].

Unfortunately, an observational analysis conducted by the Vlaams Agentschap Zorg en Gezondheid, based on approximately 2000 general practitioners' (GPs) consultations during weekends and holidays in Belgium between 2019 and 2020, revealed that only 32% of patients had a PHS. Several studies have shown that the amount of encoded data registered in EMRs is low (eg, the completeness of height and weight is <76%) [13-16]. Hamade et al [17] suggested that targeting EMR data quality or the use of EMR functions may significantly improve EMR use.

However, there remains a gap in knowledge regarding how to improve the correct use of EMRs [17]. Recently, Ngugi et al [18] developed and validated a set of indicators for the use of EMRs, focusing on the implementation of the EMR, in low- and middle-income countries. There are some guidelines on the good use of EMRs, but none are up to date [19,20]. Appropriate EMR training programs for health care workers and students can improve the use of EMRs; however, an exhaustive list of learning points is lacking [2,21]. Electronic audit and feedback (eA&F) can also be used to improve the correct use of the EMR and, more particularly, the quality of the data stored in the EMR [22]. Ivers et al [23] defined eA&F as "an electronic summary of the clinical performance of health care over a specified period of time aimed at providing information to health professionals to allow them to assess and adjust their performance." To this end, a validated set of quality indicators (QIs) is used. QIs are measurable items that refer to the structures, processes, and outcomes of care [24]. An evidence-based list of QIs, approved by expert consensus, could help develop appropriate training programs and eA&F interventions to improve EMR use [17,25,26]. This study aimed to develop an evidence-based set of electronically extractable QIs approved by expert consensus to assess the good use of EMRs by GPs from a medical perspective.

Methods

Study Design

The RAND-modified Delphi method was used to develop QIs for the use of EMRs from January 2024 to December 2024 based on the method used by Van den Bulck et al [27-29]. The process included the following: (1) development of a questionnaire based on recommendations and existing QIs from evidence-based guidelines and (2) rating procedure by an expert panel in 3 rounds, namely individual written questionnaire, online consensus meeting, and final appraisal.

Selection of Recommendations and Existing QIs

The TRIP database and MEDLINE were searched for guidelines and indicator sets using the following search terms: "Electronic medical records" AND ("Primary Health Care" OR "Primary Care" OR "General Practice") AND ("Standard of Care" OR "guideline" OR "recommendation" OR "quality indicator" OR "quality of health care" OR "quality of care"). The search was also extended to Google Scholar (accessed on April 2, 2024). The sources were selected by the primary researcher using the criteria specified in Textbox 1. Sources were not filtered based on the quality of the methodology because of the limited methodological information in any of the sources found in the first search. An exhaustive list of all indicators and recommendations relevant to EMR quality was compiled. For inclusion, indicators and recommendations were filtered using the specific, measurable, acceptable, realistic, and time-bound (SMART) principle [27-29]. A QI is defined by Grypdonck et al as "a measurable tool to assess the quality of care, including outcome-, process- and patient-oriented indicators" [30]. For recommendations, the following definition provided by the World Health Organization was used: "Recommendations are statements designed to help end-users make informed decisions on whether, when and how to undertake specific actions such as clinical interventions, diagnostic tests or public health measures, to achieve the best possible individual or collective health outcomes" [31]. Screening was performed by 2 independent researchers: a GP (primary researcher) and a methodologist. If there was a difference in the results for an indicator or recommendation between the 2 researchers, they discussed it until a consensus was reached. All remaining items (indicators and recommendations) were then categorized into topics according to the organization of the PHS in Belgium as explained by Domus Medica [19], and comparable with categories defined by Jabaaij et al [13] in the EMR scan for general practitioners: completeness and adequacy of the problem list, encoded registration in the EMR, completeness and actuality of medication list, risk factors and drug monitoring, patient identification and contact information, vaccination status, and patients' choices.

XSL•FO
RenderX

Textbox 1. Selection criteria used to filter articles in the literature.

Inclusion criteria

- Language: English, Dutch, or French
- Evaluation level: microlevel (evaluation at the patient level)
- Source age: indicator sets—any age; recommendations—maximum 10 years
- Target population: primary care providers

Exclusion criteria

- Language: other than English, Dutch, or French
- Evaluation level: mesolevel or macrolevel (regional, national, or other)
- Source age: recommendations published more than 10 years ago
- Target population: hospital physicians

Panel Selection

GPs and EMR developers were invited to participate in the panel via email. Using a purposive sampling technique, priority was given to representing all involved parties at the regional (Flanders, Wallonia, and Brussels) and occupational levels (solo GPs, GPs in group practices, programmers, and academic personnel). The final panel consisted of 18 participants: 11 (61%) Flemish-speaking participants and 7 (39%) French-speaking participants. There were 12 GPs (recently retired: n=1, 6%; worked in a community health center: n=1, 6%; had an individual practice: n=2, 11%; and worked in a group practice: n=8, 44%) and 6 (33%) EMR developers from different EMRs. They were provided with information about the aims and methodology of this study.

Written Questionnaire Round

Completion

An online LimeSurvey questionnaire was presented to a multidisciplinary expert panel (from August 2024 to November 2024) [32]. Each participant was asked to score each item for their capability to measure the quality of EMR use in primary care on a 9-point Likert scale (1=lowest score; 9=highest score). They were asked to base their assessment on the extractability of the indicator or recommendation from the EMR and its possible positive impact on the quality of care. Panel members could also label an item as “not assessable.” At the end of each topic, they were asked to give the top 5 (prioritization) of the items. If a topic did not contain enough items to create the top 5, participants were asked to order half of the indicators in the top X. Finally, all participants could write remarks and add new indicators.

Analysis

The results of the online survey were analyzed using a Node.js (OpenJS Foundation) script to subdivide each item into 3 categories: high, uncertain, or low potential as a QI [33]. The script generated a report for each participant with their responses, a general overview of the calculated parameters (Table 1), and a category based on these parameters (Table 2).

Table 1. Defined parameters to score each quality indicator and recommendation [28].

Parameter	Calculation
Total number of responses	N responses
Likert score	
For each score, the total number of responses	N responses for score
Median Likert score	Median Likert score of all responses
Highest ratings (%)	N responses with a score ≥7 divided by the total number of responses
Prioritization (top %)	Sum of the individual top X scores (eg, first place=score 3, second place=score 2, third place=score 1, and items not ranked in the top 3=score 0) divided by the maximum possible score (N responses × highest possible top score)

XSL•FO
RenderX

Table 2. Selection criteria for indicators and recommendations [28].

Criterion	Selection[a]	Discussion[a]	No selection[a]
Likert score: median	≥7	Any	Any
Likert score: distribution	≥70% with score ≥7	≥30% with score <3	Other
Prioritization (top %)	≥20%	≥1%	<1%

[a]Indicators and recommendations must fulfill all criteria, not only some, to be categorized as either "selection," "discussion," or "no selection."

Consensus Meeting Round

Each panel member received an individual feedback report 2 weeks before the consensus meeting. It also included new recommendations and comments. The report contained all items with a color code representing their potential to reach a consensus (Multimedia Appendix 1).

An online consensus meeting was held using Microsoft Teams [34]. To improve the efficiency of the consensus meeting, focus was given to the items categorized as "uncertain," together with the newly added indicators proposed by the panelists. Items with high potential were automatically selected for the final selection unless panel members requested a decision-making discussion. Items with low potential were excluded from the final selection unless the panel members requested deliberation.

All accepted indicators and recommendations were adjusted to the conclusions of the consensus meeting and conformed to the SMART criteria [27-29]. The recommendations were transformed into QIs as percentages. The final report was sent to the panel members for final acceptance in Dutch and French. The final indicator set was translated into English by the researchers.

Ethical Considerations

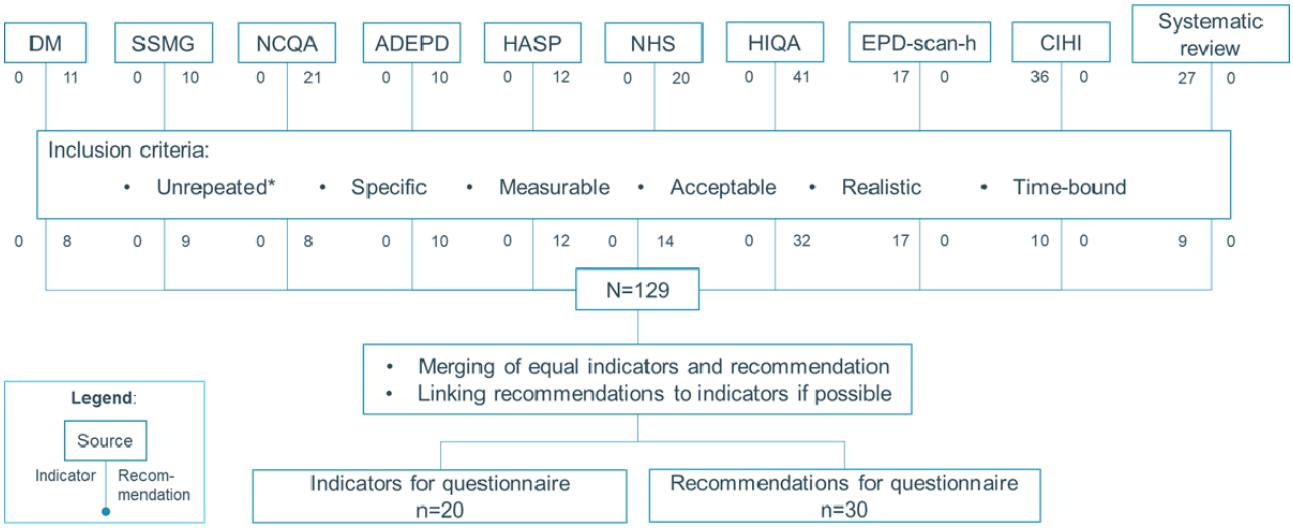
This study was approved on June 25, 2024, by the Social and Societal Ethics Committee of KU Leuven (G-2024-8020-R2(AMD)). The project was also vetted and approved by KU Leuven's Privacy and Ethics platform in view of the principles and obligations laid down in the General Data Protection Regulation of the European Parliament and the Council of 27 April 2016 [35]. Participants were asked to digitally sign an informed consent form before participation. All data were pseudonymized after the written questionnaire round using a unique identifier. Personal information was separated from other research data in a second file with password protection. The personal data was used to provide the participants the feedback report, to invite them for the consensus meeting round, and to approve the final report. All personal information was deleted after the study was finished. No compensation was provided to the participants.

Results

Extraction of QI and Recommendations

A total of 4 review articles listed QIs for the use of EMRs, containing 27 QIs [13,15,36-38]. Nine guidelines contained 53 QIs and 125 recommendations [11,19,20,39-42]. Where possible, duplicate items were removed, and indicators were linked to similar recommendations in the guidelines. This resulted in a combined list of 20 QIs and 30 recommendations (Figure 1 [11,13,15,19,20,36-42]; Multimedia Appendices 2-4).

Figure 1. Overview of the recommendation and indicator extraction. *Unique: repetitive recommendations were removed. ADEPD: NHG-Richtlijn Adequate dossiervorming met het elektronisch patiëntdossier; CIHI: Canadian Institute for Health Information; DM: Domus Medica; HASP: NHG Richtlijn Informatie-uitwisseling tussen huisarts en medisch specialist; HIQA: Health Information and Quality Authority; NCQA: National Committee for Quality Assurance; NHS: National Health Service; SSMG: Société Scientifique de Médecine Générale [11,14,16,20,21,38-44].



XSL•FO
RenderX

Questionnaire Round

All panel members (18/18, 100%) completed the questionnaire. On the basis of these categories, out of 50 items, 21 (42%) were selected, 27 (54%) required further discussion, and 2 (4%) were excluded.

Consensus Round

Of the 18 panel members, 13 (72%) were present for the consensus meeting—9 (50%) GPs and 4 (22%) EMR developers—8 (44%) of whom were Flemish speaking and 5 (28%) French speaking. In the panel round, none of the preselected indicators (n=7) or recommendations (n=14) were discarded. Of the 10 indicators labeled as "discussion," 7 (70%) were selected, and 1 (10%) indicator was further divided into 7 specific indicators (Multimedia Appendix 1; indicator 3). Of the 17 recommendations labeled as "discussion," 6 (35%) were selected. This resulted in a list of 20 indicators and 20 recommendations (Multimedia Appendix 1). During the consensus meeting, some items that were not yet extractable were included because the panel members felt that their clinical importance should prompt EMR providers to adapt their software.

Final Evaluation

The 20 recommendations were transformed into QIs and added to the 20 other QIs, resulting in 16 (40%) QIs regarding the completeness and adequacy of the problem list, 4 (10%) QIs regarding encoded registration in the EMR, 3 (8%) QIs regarding the completeness and actuality of the medication list, 6 (15%) QIs regarding risk factors and drug monitoring, 4 (10%) QIs regarding patient identification and contact information, 3 (8%) QIs regarding vaccination status, and 4 (10%) QIs regarding patient choices (Textbox 2; Multimedia Appendix 5). All panel members present in the consensus meeting (13/13, 100%) approved the final report.

XSL•FO
RenderX

Textbox 2. Quality indicators for the good use of electronic medical records (EMRs) in primary care.

Topic 1: completeness and adequacy of the problem list

1.1 The mean of active problem items listed on a patient's problem list

1.2 Percentage of items on the problem list that are linked to an encoded diagnosis

1.3 Percentage of patients with medication for thyroid disease in their medication list with an encoded diagnosis of thyroid disease on the problem list

1.4 Percentage of patients with medication for epilepsy in their medication list with an encoded diagnosis of epilepsy on the problem list

1.5 Percentage of patients with medication for Parkinson disease in their medication list with an encoded diagnosis of Parkinson disease on the problem list

1.6 Percentage of patients with medication for depression in their medication list with an encoded diagnosis of depression on the problem list

1.7 Percentage of patients with medication for cardiovascular disease in their medication list with an encoded diagnosis of cardiovascular disease on the problem list

1.8 Percentage of patients with medication for asthma or chronic obstructive pulmonary disease in their medication list with an encoded diagnosis of asthma or chronic obstructive pulmonary disease on the problem list

1.9 Percentage of patients with medication for diabetes in their medication list with an encoded diagnosis of diabetes on the problem list

1.10 Percentage of encoded items in the problem list with an additional description or comment

1.11 Percentage of encoded items in the problem list for which the estimated or actual date of onset has been registered

1.12 Percentage of inactive items on the problem list for which the estimated or actual date the condition was resolved has been registered

1.13 Percentage of surgeries and other procedures that have been registered as an encoded item

1.14 Percentage of surgeries and other procedures in the patient history that have been linked to the related episode of care

1.15 Percentage of surgeries and other procedures in the patient's history with an additional description or comment

1.16 Percentage of surgeries and other procedures in the patient summary for which the date and/or time on which the procedure was or is intended to be performed is registered

Topic 2: encoded registration in the EMR

2.1 The mean number of subcontacts created per contact during the observation period

2.2 Percentage of subcontacts during the observation period connected with an episode of care

2.3 Percentage of subcontacts (registered as consultation, home visit, or teleconsultation) with a subjective, objective, evaluation, and planning item

2.4 The number of patients with a global medical record in the practice with changes in the status of tasks (procedures or actions) over the total number of patients with a global medical record in the practice

Topic 3: completeness and actuality of the medication list

3.1 Percentage of medications on the medication list labeled as *active* that are no longer active

3.2 Percentage of active medications on the medication list that are linked to a health condition on the problem list

3.3 Percentage of active medications on the medication list for which the dosage and treatment regimen are complete

Topic 4: risk factors and drug monitoring

4.1 Percentage of patients with at least 1 registration for a drug allergy or intolerance

4.2 Percentage of patients with a registration for a risk factor; risk factors include clinical information that is imperative to know so that the life or health of the patient is not threatened, for example, prophylaxis for adrenal crisis (Addison disease), endocarditis, bleeding disorders, endoprosthesis, patients who are immunocompromised, (functional) asplenia, thrombosis, or particularly resistant microorganisms

4.3 Percentage of patients with at least 1 encoded registration for a physical examination (weight, height, blood pressure, and pulse)

4.4 Percentage of patients with at least 1 registration of an item for social history

4.5 Percentage of patients aged 12 years and older for whom there is at least 1 notation concerning the use of cigarettes, alcohol, and substances

4.6 Percentage of patients with at least 1 registration of an item for family medical history

Topic 5: patient identification and contact information

5.1 Percentage of patients with a registration for hospital preference

5.2 Percentage of patients with at least 1 registration of contact information for a contact person, caregiver, or person designated as representative

5.3 Percentage of patients with at least 1 registration of a health care professional in their care team

XSL•FO
RenderX

5.4 Percentage of patients with at least 1 registration for address, email, employer, home and work phone, marital status, registration status on patient platform (eg, Helena [43]), or cohousing people and animals

Topic 6: vaccination status

6.1 Percentage of patients aged 7 years and older who have a registration for all the recommended primary childhood vaccines

6.2 Percentage of patients aged 65 years and older who have a registration for a yearly influenza vaccine

6.3 Percentage of vaccinations for which the date the vaccination was administered is registered

Topic 7: patient choices

7.1 Percentage of patients with a registration of treatment preferences: euthanasia request or refusal for intubation, resuscitation, organ donation, vaccination, or blood transfusion

7.2 Percentage of patients with at least 1 contact with a subjective, objective, evaluation, or planning item linked to an encoded diagnosis regarding treatment preferences or end-of-life care

7.3 Percentage of patients with at least 1 registration of a goal for personalized care (goal-oriented care)

7.4 Percentage of patients with at least 1 registration regarding preferences for preventive health care

Discussion

Principal Findings and Comparison With Prior Work

This study used the RAND-modified Delphi method to validate a set of 40 QIs for the correct use of EMRs by GPs based on the indicators and recommendations found in the literature. These indicators are (potentially) EMR extractable and evaluate key aspects of EMR use by GPs: completeness and adequacy of the problem list, encoded registration, completeness and actuality of the medication list, risk factors and drug monitoring, patient identification and contact information, vaccination status, and patient choices. Using the SMART principle, we ensured that the QIs could be evaluated in clinical practice [27-29,44-46].

Previously, Jabaaij et al [38] developed the EMR scan for general practitioners in the Netherlands to evaluate EMR use. However, the method they used to define QIs was unclear. Ngugi et al [18] developed and validated a set of indicators for the use of EMRs in low- and middle-income countries. Their QIs mostly evaluated the general use of the EMRs (eg, indicator 7, "% of required data elements contained in EHRs"). Our study focuses on how data are registered and encoded in EMRs, focusing on which data are registered [18]. Other indicator sets were either outdated (aged >20 years) [36], focused on clinical performance instead of use of the EMR [15], or evaluated EMR use on the mesolevel or macrolevel instead of the microlevel [37]. Hamade et al [17] noted that while there are many studies on the implementation of EMR, there is a scarcity of studies on its use.

Most of the selected indicators evaluated the completeness and adequacy of the problem list. The problem list provides important information for decision-making [47]. Our indicators are primarily process indicators because they focus on the process of EMR registration (ie, use). However, these QIs cannot be directly linked to patient outcomes. Process indicators are useful for quality assessment. They are more sensitive to differences in quality than outcome measures and are easier to interpret [48]. Recent literature shows an interest in eA&F [22,29], dashboard interface features to support reflection on practice ("barometers") [27-29], and computer-based decision support in primary care [49]. The quality of data in EMRs impacts the effectiveness of these systems. Several studies have shown that improving data quality may improve the quality and efficiency of care and reduce mortality [4-6]. Unfortunately, data show that physicians spend approximately two-thirds of their work time interacting with EMRs in their offices. This administrative burden negatively impacts clinicians' wellness, leading to an increased chance of burnout. Recently, the interest in the use of artificial intelligence for analyzing data, identifying hidden information, identifying risks, and providing suggestions for diagnosis has increased rapidly. This can greatly facilitate the process for GPs to easily improve data quality without spending more time entering data in the EMR and less time interacting with patients [50]. GPs should not need to adapt their working habits to the capabilities of their medical software; rather, EMRs should be adapted to the needs of GPs to facilitate the correct registration of data [51]. An important lever to encourage EMR developers to align their software more closely with clinicians' data entry and retrieval needs lies in regulatory frameworks, such as homologation. This is a certification procedure to evaluate the mandatory standards and certification requirements that EMRs must fulfill to be approved for use within the health care system [52]. By embedding QIs derived through expert consensus, such as those developed in this study, into these frameworks, regulators can create explicit performance benchmarks that EMRs must meet. This alignment incentivizes developers to design functionalities that facilitate accurate, complete, and efficient data capture and reporting by clinicians [50].

Strengths and Limitations

Our panel included members working in all parts of Belgium and in all types of practices, giving it broad support. Because our panel included health professionals and EMR developers, we were able to define clinically relevant indicators that considered the technical requirements for extractability. During the consensus meeting, the EMR developers emphasized that these indicators could possibly stimulate GPs to use the features they have already developed more extensively. We are convinced that their opinion was a valuable addition to the panel. EMR developers were the most capable of assessing whether a QI could be automatically extracted from the EHR data. However, we cannot exclude a potential bias caused by their

XSL•FO
RenderX

inclusion (eg, the ease or cost of addressing a recommended QI outweighing the importance of making that QI available from a public health perspective). The scarcity of literature on the initial set of QIs and recommendations is a double-edged sword for our study. On the one hand, it limits the strengths of the initial set of QIs and recommendations because none of our sources used a systematic approach to evaluate the validity of their QIs and recommendations, had a well-explained methodology [13,15,36-38], or could be evaluated using the Appraisal of Guidelines for Research and Evaluation II method [11,19,20,39-42]. On the other hand, this allowed us to conduct a broad and general search on the topic, which could increase its universal applicability. There are many differences between countries in terms of how and which data are registered [11,19,20,39-42]. We attempted to generalize the QIs and recommendations as much as possible. However, adaptation to the local context is necessary. Another important limitation is that some indicators are not currently extractable. However, as mentioned earlier, these indicators reflect aspects of clinical practice that are important enough to warrant adaptation of EMRs to include this information.

Future Research Directions

This set of indicators is important for future research on EMR use. Developing a list of QIs is the first step in implementing a strategy and realizing a quality loop. The next essential step is to conduct practice testing for operational validity [29]. These results can be used to create an eA&F system [22,23]. This set of QIs can also be used to develop specific (computer based) training for GPs based on a set of learning goals [2]. It should be noted that our QIs can be used for the correct use of EMRs but cannot be used to quantify the quality of care provided by GPs. Further research is necessary to evaluate whether the correct use of the EMR, as proposed in this set of QIs, also affects the quality of care provided to patients and its possible effect on the data in the PHS.

Conclusions

This study provided a set of 40 EMR-extractable QIs for the correct use of EMRs in primary care based on international guidelines and approved by GPs and EMR developers. These QIs can be used as a framework to measure and improve the quality and completeness of EMRs in primary care.

Acknowledgments

The authors would like to thank all the participating panel members for the many hours they spent in this study.

Funding

This study was conducted by UZ Leuven and KU Leuven at the request of Rijksinstituut voor ziekte- en invaliditeitsverzekering–Institut national d’assurance maladie-invalidité and funded by Rijksinstituut voor ziekte- en invaliditeitsverzekering–Institut national d’assurance maladie-invalidité as part of a national project to improve medical care in Belgium.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RP conducted the literature search and organized each step of the Delphi procedure as part of his PhD project. LM assisted during each step of the Delphi procedure. SVdB oversaw the whole process and helped to draft the manuscript. BV and WR supervised the entire process from the beginning to the end and helped draft the manuscript. MDJ and BF were part of the panel during the Delphi procedure and helped draft the final manuscript. All the authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the feedback report for panel members.
[PDF File (Adobe PDF File), 693 KB - medinform_v14i1e80057_app1.pdf]

Multimedia Appendix 2
Online questionnaire - Dutch version.
[PDF File (Adobe PDF File), 1776 KB - medinform_v14i1e80057_app2.pdf]

Multimedia Appendix 3
Online questionnaire - French version.
[PDF File (Adobe PDF File), 1722 KB - medinform_v14i1e80057_app3.pdf]

XSL•FO
RenderX

Multimedia Appendix 4
Online questionnaire - English version (automated generated translation by DeepL Translator; https://www.deepl.com/nl/translator).
[PDF File (Adobe PDF File), 2142 KB - medinform_v14i1e80057_app4.pdf]

Multimedia Appendix 5
Final indicator set.
[PDF File (Adobe PDF File), 182 KB - medinform_v14i1e80057_app5.pdf]

References

1. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood) 2005;24(5):1103-1117. [doi: 10.1377/hlthaff.24.5.1103] [Medline: 16162551]
2. Samadbeik M, Fatehi F, Braunstein M, Barry B, Saremian M, Kalhor F, et al. Education and training on electronic medical records (EMRs) for health care professionals and students: a scoping review. Int J Med Inform 2020 Oct;142:104238. [doi: 10.1016/j.ijmedinf.2020.104238] [Medline: 32828034]
3. Defining key health information technology terms. Alliance. 2008. URL: https://tigerstandards.pbworks.com/f/HITTermsFinalReport.pdf [accessed 2025-01-25]
4. Zhang Y, Yu P, Shen J. The benefits of introducing electronic health records in residential aged care facilities: a multiple case study. Int J Med Inform 2012 Oct;81(10):690-704. [doi: 10.1016/j.ijmedinf.2012.05.013] [Medline: 22749424]
5. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752 [FREE Full text] [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
6. Bogale TN, Derseh L, Abraham L, Willems H, Metzger J, Abere B, et al. Effect of electronic records on mortality among patients in hospital and primary healthcare settings: a systematic review and meta-analyses. Front Digit Health 2024 Jun 26;6:1377826 [FREE Full text] [doi: 10.3389/fdgth.2024.1377826] [Medline: 38988733]
7. Kortteisto T, Raitanen J, Komulainen J, Kunnamo I, Mäkelä M, Rissanen P, et al. Patient-specific computer-based decision support in primary healthcare--a randomized trial. Implement Sci 2014 Jan 20;9:15 [FREE Full text] [doi: 10.1186/1748-5908-9-15] [Medline: 24444113]
8. Upshur RE, Tracy S. Chronicity and complexity: is what's good for the diseases always good for the patients? Can Fam Physician 2008 Dec;54(12):1655-1658 [FREE Full text] [Medline: 19074692]
9. Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. AMIA Annu Symp Proc 2006;2006:324-328 [FREE Full text] [Medline: 17238356]
10. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. Acad Pediatr 2011;11(4):280-287 [FREE Full text] [doi: 10.1016/j.acap.2011.02.007] [Medline: 21622040]
11. National standard on information requirements for a national electronic patient summary. Health Information and Quality Authority. 2018 Dec. URL: https://www.hiqa.ie/sites/default/files/2019-01/PatientSummary_InfoReqs.pdf [accessed 2024-03-11]
12. Mathioudakis A, Rousalova I, Gagnat AA, Saad N, Hardavella G. How to keep good clinical records. Breathe (Sheff) 2016 Dec;12(4):369-373 [FREE Full text] [doi: 10.1183/20734735.018016] [Medline: 28210323]
13. Jabaaij L, Verheij R, Njoo K, van den Hoogen H, Tiersma W, Levelink H. Het meten van de kwaliteit van de registratie in elektronische patiënten dossiers van huisartsen met de EPD-scan-h. NIVEL. 2008. URL: https://www.nivel.nl/nl/publicatie/het-meten-van-de-kwaliteit-van-de-registratie-elektronische-patienten-dossiers-van [accessed 2025-04-23]
14. Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak 2019 Feb 12;19(1):30 [FREE Full text] [doi: 10.1186/s12911-019-0740-0] [Medline: 30755205]
15. Nemeth LS, Ornstein SM, Jenkins RG, Wessell AM, Nietert PJ. Implementing and evaluating electronic standing orders in primary care practice: a PPRNet study. J Am Board Fam Med 2012;25(5):594-604 [FREE Full text] [doi: 10.3122/jabfm.2012.05.110214] [Medline: 22956695]
16. De Clercq E, Van Casteren V, Jonckheer P, Burggraeve P, Lafontaine MF. Electronic patient record data as proxy of GPs' thoughts. Stud Health Technol Inform 2008;141:103-110. [Medline: 18953130]
17. Hamade N, Terry A, Malvankar-Mehta M. Interventions to improve the use of EMRs in primary health care: a systematic review and meta-analysis. BMJ Health Care Inform 2019 May;26(1) [FREE Full text] [doi: 10.1136/bmjhci-2019-000023] [Medline: 31142493]
18. Ngugi P, Babic A, Kariuki J, Santas X, Naanyu V, Were MC. Development of standard indicators to assess use of electronic health record systems implemented in low-and medium-income countries. PLoS One 2021 Jan 11;16(1):e0244917 [FREE Full text] [doi: 10.1371/journal.pone.0244917] [Medline: 33428656]

XSL•FO
RenderX

19. Verdonck P, Strobbe J, Steenackers J, Van Royen J, De Naeyer P, Govaerts F, et al. Het Elektronisch Medisch Dossier: aanbeveling voor goede medische praktijkvoerin. Huisarts Nu 2004;33(2):58-70 [FREE Full text]
20. Organisation de la pratique. Société Scientifique de Médecine Générale (SSMG). URL: https://www.ssmg.be/organisation-pratique/ [accessed 2024-02-01]
21. Rajaram A, Hickey Z, Patel N, Newbigging J, Wolfrom B. Training medical students and residents in the use of electronic health records: a systematic review of the literature. J Am Med Inform Assoc 2020 Jan 01;27(1):175-180 [FREE Full text] [doi: 10.1093/jamia/ocz178] [Medline: 31592531]
22. Van Den Bulck S, Spitaels D, Vaes B, Goderis G, Hermens R, Vankrunkelsven P. The effect of electronic audits and feedback in primary care and factors that contribute to their effectiveness: a systematic review. Int J Qual Health Care 2020 Dec 15;32(10):708-720. [doi: 10.1093/intqhc/mzaa128] [Medline: 33057648]
23. Ivers NM, Sales A, Colquhoun H, Michie S, Foy R, Francis JJ, et al. No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention. Implement Sci 2014 Jan 17;9:14 [FREE Full text] [doi: 10.1186/1748-5908-9-14] [Medline: 24438584]
24. Kötter T, Blozik E, Scherer M. Methods for the guideline-based development of quality indicators--a systematic review. Implement Sci 2012 Mar 21;7(1):21 [FREE Full text] [doi: 10.1186/1748-5908-7-21] [Medline: 22436067]
25. Mottaghy FM, Hertel F, Beheshti M. Will we successfully avoid the garbage in garbage out problem in imaging data mining? An overview on current concepts and future directions in molecular imaging. Methods 2021 Apr;188:1-3. [doi: 10.1016/j.ymeth.2021.02.011] [Medline: 33592236]
26. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]
27. Dewaele S, Van den Bulck S, Gerne L, Vaes B. Development of primary care quality indicators for chronic obstructive pulmonary disease using a Delphi-derived method. NPJ Prim Care Respir Med 2022 Mar 18;32(1):12 [FREE Full text] [doi: 10.1038/s41533-022-00276-w] [Medline: 35304476]
28. Van den Bulck SA, Vankrunkelsven P, Goderis G, Van Pottelbergh G, Swerts J, Panis K, et al. Developing quality indicators for chronic kidney disease in primary care, extractable from the electronic medical record. A rand-modified Delphi method. BMC Nephrol 2020 May 05;21(1):161 [FREE Full text] [doi: 10.1186/s12882-020-01788-8] [Medline: 32370742]
29. Van den Bulck SA, Vankrunkelsven P, Goderis G, Broekx L, Dreesen K, Ruijten L, et al. Development of quality indicators for type 2 diabetes, extractable from the electronic health record of the general physician. A rand-modified Delphi method. Prim Care Diabetes 2020 Feb;14(1):75-84. [doi: 10.1016/j.pcd.2019.05.002] [Medline: 31204263]
30. Grypdonck L, Aertgeerts B, Luyten F, Wollersheim H, Bellemans J, Peers K, et al. Development of quality indicators for an integrated approach of knee osteoarthritis. J Rheumatol 2014 Jun 15;41(6):1155-1162. [doi: 10.3899/jrheum.130680] [Medline: 24737907]
31. WHO guidelines. World Health Organization. URL: https://www.who.int/publications/who-guidelines [accessed 2025-06-26]
32. LimeSurvey - free online survey tool. LimeSurvey. URL: https://www.limesurvey.org/ [accessed 2025-04-23]
33. Node.js - run JavaScript everywhere. OpenJS Foundation. URL: https://nodejs.org/en [accessed 2024-04-23]
34. Microsoft 365 Apps for enterprise. Microsoft. URL: https://www.microsoft.com/en-in/microsoft-365/enterprise/microsoft-365-apps-for-enterprise-product [accessed 2025-04-23]
35. General data protection regulation (GDPR). European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:310401_2 [accessed 2025-04-23]
36. De Lusignan S, Stephens PN, Adal N, Majeed A. Does feedback improve the quality of computerized medical records in primary care? J Am Med Inform Assoc 2002;9(4):395-401 [FREE Full text] [doi: 10.1197/jamia.m1023] [Medline: 12087120]
37. Pan-Canadian primary health care indicator - update report. Canadian Institute for Health Information. 2012. URL: https://publications.gc.ca/site/eng/444237/publication.html [accessed 2024-10-29]
38. Jabaaij L, Njoo K, Visscher S, van den Hoogen H, Levelink H, Verheij R. Verbeter uw verslaglegging, gebruik de EPD-scan-h. HUWE 2009;52:240-246. [doi: 10.1007/bf03085615]
39. Guidelines for medical record documentation. National Committee for Quality Assurance. 2018. URL: https://wpcdn.ncqa.org/www-prod/wp-content/uploads/2018/07/20180110_Guidelines_Medical_Record_Documentation.pdf [accessed 2024-02-24]
40. Summary care record. National Health Service England. URL: https://digital.nhs.uk/services/summary-care-records-scr [accessed 2024-02-01]
41. NHG-Richtlijn Adequate dossiervorming met het elektronisch patiëntendossier (ADEPD). Nederlands Huisartsen Genootschap. URL: https://www.nhg.org/praktijkvoering/informatisering/richtlijn-adequate-dossiervorming-epd/ [accessed 2024-10-29]
42. Richtlijn Informatie-uitwisseling tussen huisarts en medisch specialist (Richtlijn HASP). Nederlands Huisartsen Genootschap. 2017. URL: https://www.nhg.org/wp-content/uploads/2022/10/Richtlijn-Gegevensuitwisseling-Huisarts-en-Specialist-HASP.pdf [accessed 2024-10-29]
43. Helena voor patiënten. Helena. URL: https://nl-info.helena.care/helena-voor-patienten [accessed 2025-04-23]

XSL•FO
RenderX

44. Tichelaar J, Uil den SH, Antonini NF, van Agtmael MA, de Vries TP, Richir MC. A 'SMART' way to determine treatment goals in pharmacotherapy education. Br J Clin Pharmacol 2016 Jul 07;82(1):280-284 [FREE Full text] [doi: 10.1111/bcp.12919] [Medline: 26914983]
45. Smets M, Smeets M, Van den Bulck S, Janssens S, Aertgeerts B, Vaes B. Defining quality indicators for heart failure in general practice. Acta Cardiol 2019 Aug 27;74(4):291-298. [doi: 10.1080/00015385.2018.1492659] [Medline: 30261830]
46. Doran GT. There's a SMART way to write management's goals and objectives. J Manag Rev 1981;70:35-36 [FREE Full text]
47. Summerton N. The medical history as a diagnostic technology. Br J Gen Pract 2008;58(549):273-276. [doi: 10.3399/bjgp08x279779]
48. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
49. Donnelly C, Janssen A, Vinod S, Stone E, Harnett P, Shaw T. A systematic review of electronic medical record driven quality measurement and feedback systems. Int J Environ Res Public Health 2022 Dec 23;20(1):200 [FREE Full text] [doi: 10.3390/ijerph20010200] [Medline: 36612522]
50. Bednorz A, Mak JK, Jylhävä J, Religa D. Use of electronic medical records (EMR) in gerontology: benefits, considerations and a promising future. Clin Interv Aging 2023;18:2171-2183 [FREE Full text] [doi: 10.2147/CIA.S400887] [Medline: 38152074]
51. Wirth D, Senn O, Burgstaller JM, Djalali S, Muheim L, Rohrbasser A, et al. Acceptability of components for a mandatory quality improvement framework: a survey among Swiss general practitioners. Health Serv Insights 2025 Jun 21;18:11786329251346828 [FREE Full text] [doi: 10.1177/11786329251346828] [Medline: 40546342]
52. Elektronisch Medisch Dossier of EMD. Domus Medica. URL: https://www.domusmedica.be/expertisedomein/ict/het-elektronisch-medisch-dossier-emd [accessed 2025-11-06]

Abbreviations

eA&F: electronic audit and feedback
EMR: electronic medical record
GP: general practitioner
PHS: patient health summary
QI: quality indicator
SMART: specific, measurable, assignable, realistic, and time-bound

Edited by A Benis; submitted 03.Jul.2025; peer-reviewed by L Agrawal, M Jones; comments to author 23.Oct.2025; revised version received 12.Dec.2025; accepted 17.Dec.2025; published 19.Jan.2026.

Please cite as:
Paridaens R, Van den Bulck S, De Jonghe M, Fauquert B, Meel L, Raat W, Vaes B
Development of Quality Indicators for the Correct Use of Electronic Medical Records in Primary Care: Modified Delphi Study
JMIR Med Inform 2026;14:e80057
URL: https://medinform.jmir.org/2026/1/e80057
doi:10.2196/80057
PMID:

©Rico Paridaens, Steve Van den Bulck, Michel De Jonghe, Benjamin Fauquert, Liesbeth Meel, Willem Raat, Bert Vaes. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multi-Institutional Drug Use Patterns in Hospitalized Older Patients: Retrospective Cross-Sectional Study

Chung Chun Lee[1*], MS; Grace Juyun Kim[2*], PharmD, PhD; Suhyun Kim[3,4], MS; Jee Young Hong[5], MD, PhD; Won Min Hwang[6], MD, PhD; Jong-Yeup Kim[1,7], MD, PhD; Kye Hwa Lee[8,9], MD, PhD; Kwangsoo Kim[3,10], PhD; Mingyu Kang[11], MD, PhD; Ju Han Kim[12], MD, PhD; Suehyun Lee[13], PhD

[1]Department of Biomedical Informatics, College of Medicine, Konyang University, Daejeon, Republic of Korea

[2]Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[3]Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, Seoul, Republic of Korea

[4]Department of Clinical Medical Sciences, College of Medicine, Seoul National University, Seoul, Republic of Korea

[5]Department of Preventive Medicine, College of Medicine, Jeonbuk National University, Jeonju, Republic of Korea

[6]Division of Nephrology, Department of Internal Medicine, College of Medicine, Konyang University, Daejeon, Republic of Korea

[7]Healthcare Data Science Center, Konyang University Hospital, Daejeon, Republic of Korea

[8]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea

[9]Department of Biomedical Informatics, College of Medicine, University of Ulsan, Ulsan, Republic of Korea

[10]Department of Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea

[11]Department of Internal Medicine, College of Medicine, Chungbuk National University, Cheongju, Republic of Korea

[12]Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics, College of Medicine, Seoul National University, Seoul, Republic of Korea

[13]College of IT Convergence, Gachon University, Seongnam-si, Gyeonggi-do, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Suehyun Lee, PhD
College of IT Convergence
Gachon University
1342 Seongnamdaero
Sujeong-gu
Seongnam-si, Gyeonggi-do, 13120
Republic of Korea
Phone: 82 31 750 5333
Email: leesh@gachon.ac.kr

Abstract

Background: A rapidly aging population led to an increase in the number of patients with chronic diseases and polypharmacy. Although investigations on the appropriate number of drugs for older patients have been conducted, there is a shortage of studies on polypharmacy criteria in older inpatients from multiple institutions.

Objective: The aim of this study was to examine the patterns of polypharmacy and determine the criteria for the number of drugs defining polypharmacy in the geriatric inpatient population.

Methods: Electronic health records of 4 medical institutions for patients aged 65 years and older hospitalized between January 1, 2012, and December 31, 2020, were analyzed for the study. The maximum number of drugs prescribed was obtained for each patient and, along with a literature review, was used to determine the appropriate polypharmacy level for our population.

Results: We suggest a 4-level polypharmacy category system consisting of nonpolypharmacy, polypharmacy, major polypharmacy, and excessive polypharmacy based on a review of international guidelines and polypharmacy literature. Application of this system to our study population showed that the major polypharmacy category (use of 10-19 concurrent drugs) was an appropriate threshold for polypharmacy in hospitalized patients versus the traditional threshold of 5 or more concurrent drugs. The tendency of our study population to have a higher disease and drug count supports this threshold. Frequently prescribed therapeutic subgroups in this category were antibacterials for systemic use, anesthetics, and cardiac therapy.

XSL•FO
RenderX

Conclusions: This study proposes a polypharmacy categorization system for older inpatients, which differs from the common definition of the concomitant prescription of 5 or more drugs. The older population tends to have severe conditions including those requiring major surgeries; therefore, a drug count corresponding to the definition of major polypharmacy is appropriate.

(JMIR Med Inform 2026;14:e78353) doi:10.2196/78353

KEYWORDS

polypharmacy; older patients; medication; electronic health record; Common Data Model

Introduction

South Korea is reported to have the highest rate of aging among the Organisation for Economic Cooperation and Development countries, where 8.2 of 51.8 (15.7%) million people in 2020 were aged 65 years or older [1]. The prevalence of chronic diseases in this population was found to be 8481 of 10,097 (84%) individuals aged 65 years or older, making it highly likely that the cost of health care is increasing [2]. This requires close review and consideration of effective measures to prevent and treat chronic diseases and adverse events from polypharmacy.

With the increase in lifespan and aging population in South Korea, patients commonly have multiple concomitant diseases. Consequently, patients visit 1 or more health care institutions and are prescribed medications that expose them to polypharmacy [3]. In 2006, Intercontinental Marketing Statistics Korea found that an average of 4.16 medications were prescribed per health care appointment, which was the highest among the 10 countries studied [2]. A study using the South Korean health claims data of 2016-2019 reported that in 2016, a total of 1548 of 47,631 (3.3%) individuals aged 45 years or older experienced polypharmacy, defined as the use of 10 or more medications for at least 60 days, and this rate increased gradually up to 2019, where 2012 of 48,249 (4.2%) individuals aged 45 years or older were on polypharmacy [4]. To ensure that older adult patients receive appropriate drug therapy, the Korean Health Insurance Review & Assessment Service launched the Drug Utilization Review program. This provides real-time assessment of inappropriate patient drug therapy, including the use of duplicative drug ingredients or drugs of different therapeutic classes, inappropriate drugs for older adults, or drugs not to be used in combination [5].

Considering that older adults have decreased physiological functions and often have several chronic diseases, they are at a higher risk of polypharmacy and hospitalization. The use of at least 5-6 drugs is a definition widely used for polypharmacy in the ambulatory population, and several studies report 25% or more older adults on polypharmacy [2,6-9]. However, this definition may not be practical for the hospitalized older adult population, as this demographic tends to receive more treatments for multiple diseases. Previously reported definitions of polypharmacy in the geriatric hospitalized population were the prescription of 5-9 medications and that of hyper-polypharmacy were 10 or more medications [10,11]. There were also cases of using 5 or more drugs as the definition of polypharmacy for hospitalized patients [12,13]. Although definitions of polypharmacy applicable to older hospitalized patients exist, most studies use single-institution prescription data, making the findings difficult to generalize. Studies from multiple institutions are needed to determine and apply definitions of polypharmacy in the geriatric inpatient population.

Previous studies have used various data sources and methods to define polypharmacy. Data sources included health insurance claims, electronic medical records, spontaneous adverse drug event (ADE) reports, and surveys. The methods used to define polypharmacy include simple numeric definitions, numeric definitions with terms of duration in a therapy or health care setting, and nonnumeric descriptive definitions [14]. Definitions of polypharmacy are diverse due to differing patient age, underlying disease, and disease severity. Accordingly, it is necessary to review the existing definitions of polypharmacy to understand the categories used in various clinical settings and to apply the most appropriate method when assessing the status of polypharmacy in the study population.

Kim et al [2] reported that approximately 6.30 million patients aged 65 years or older in South Korea were prescribed drugs for at least 90 days during 2018, and among this population, 2.64 (41.9%) million were prescribed 5 or more drugs, and 0.77 (13.9%) million were prescribed 10 or more drugs. Cresswell et al [15] reported an increased risk of ADEs with a higher number of prescribed drugs. The older adults were especially at risk of serious long-term outcomes, including death and disability, from these ADEs [15]. Ensuring safer therapy for this population is important. Hanlon et al [16] found that in 167 adults aged 65 years and older, ADEs occurred in 58 (35%), and the most frequent ADEs were gastrointestinal disorders. This was followed by ADEs of the nervous, respiratory, endocrine, skin, and subcutaneous systems [16]. Among the 58 patients who had an ADE, 49 (84%) patients required medical care: 37 (63%) were examined by a physician, 7 (11%) were hospitalized, and 5 (10%) visited the emergency room [16]. Mabuchi et al [17] compared the number of drugs used in patients with a spontaneous ADE report to the overall patient population in Japan using real-world data. In both populations, the number of drugs used increased with age [17].

This study used anonymized electronic health record (EHR) data converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format, a standardized data structure that accommodates retrospective observational studies [18]. The OMOP CDM data content is generated by extracting EHR data recorded during patient care, removing personal identifiable information, transforming the data into a standardized structure, and mapping hospital-specific terms to standardized, controlled vocabularies. Each CDM in the 4 institutions included patient demographic, condition, measurement, procedure, and drug prescription information [19]. Overall, we examined the characteristics and patterns of polypharmacy in older hospitalized adults across 4 health care

XSL•FO
RenderX

institutions (presented as medical institutions A-D) and created a polypharmacy categorization system for this population. Specifically, for each polypharmacy category, we characterized the drugs and drug classes prescribed to the population of interest and examined their drug use patterns.

Methods

Study Participants and Observation Period

EHR data of 4 medical institutions converted to the OMOP CDM format were used to analyze polypharmacy patterns. The study participants were patients aged 20 years or older hospitalized at one of these 4 institutions between January 1, 2012, and December 31, 2020. The number of patients aged 20 years or older who visited the medical institutions in the study period ranged from 494,875 to 3,349,795. The number of hospitalized patients, hospitalization events, hospitalized patients aged 60 years and older, and the number of drug classes prescribed to the overall hospitalized patients are presented in Table 1.

To extract study participants, we consulted clinicians from medical departments who frequently encountered polypharmacy. The patient visit types were restricted to inpatients. The earliest date of patient admission was the patient study entry date; if the discharge date of this admission and the admission date of the next admission were within 30 days, the 2 admissions were treated as a single continuous admission period. To investigate the status of discharge drug use, drug exposure data from the outpatient setting were followed up for 7 days after patient discharge. For example, if the patient had 2 admission episodes, the total admission period would be the sum of the duration of admission 1 in days, 7 days after discharge from admission 1, the duration of admission 2 in days, and 7 days after discharge from admission 2 (Figure 1).

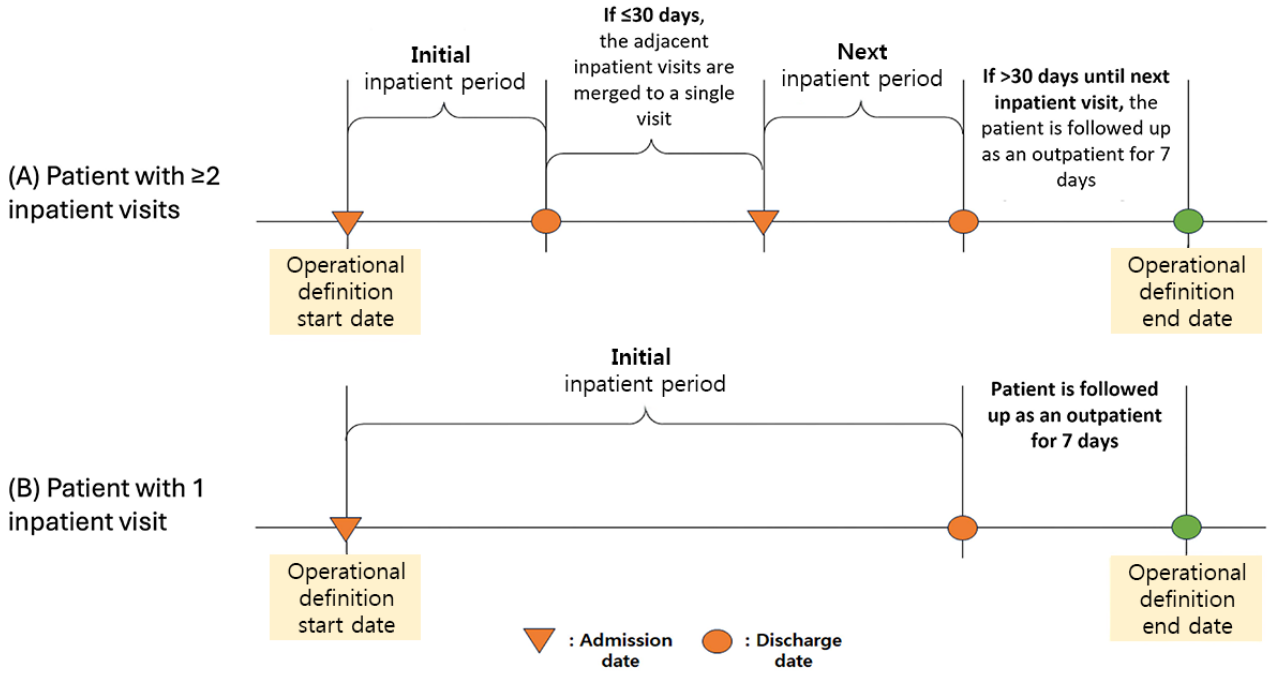
Table 1. Patient characteristics by medical institution.

	Medical institution A	Medical institution B	Medical institution C	Medical institution D
Patients, n (%)	602,483 (100)	1,944,140 (100)	3,349,795 (100)	494,875 (100)
Hospitalized patients, n (%)	169,931 (20.2)	405,956 (20.9)	419,460 (12.5)	128,700 (26)
Hospitalization events, n (%)	315,624 (100)	1,092,936 (100)	871,764 (100)	212,489 (100)
Older hospitalized adults aged ≥60 years, n (%)	48,512 (8.1)	166,071 (8.5)	113,573 (3.4)	33,997 (6.9)
Drug therapeutic subgroups (WHO[a] ATC[b] second-level codes [20]) prescribed, distinct n (%)	79 (84)	86 (91.5)	72 (76.6)	81 (81.2)

[a]WHO: World Health Organization.

[b]ATC: Anatomical Therapeutic Chemical.

Figure 1. Definition of patient observation period.



XSL•FO
RenderX

Polypharmacy Level

This study examined the distribution of patients by polypharmacy level for each medical institution. The polypharmacy levels increase by the number of concomitant drugs prescribed to the patient. The cutoff values for concomitant drug counts in each polypharmacy level were determined based on internationally recommended thresholds and existing literature on polypharmacy in older adults. The World Health Organization (WHO) and several countries defined the use of 5 or more drugs as polypharmacy [1,14]. WHO differentiates between appropriate and inappropriate polypharmacy by defining appropriate polypharmacy as evidence-based, personalized medication prescribing and inappropriate polypharmacy as irrational, excessive prescribing [21]. It emphasizes the importance of reducing inappropriate polypharmacy by continuously ensuring the appropriate, rational use of medications [21]. The European Geriatric Medicine Society states that polypharmacy patients should be closely managed through medication reviews to prevent inappropriate polypharmacy [22]. The National Institute for Health and Care Excellence recommends monitoring the benefits and harms of each drug taken by adults using 15 or more regular medications [23].

Masnoon et al [14] used concurrent drug counts, admission duration, and treatment duration to define polypharmacy, and Young et al [24] determined the category of polypharmacy by the number of concomitant drugs and the presence of high-risk drugs in the patient's drug regimen. Studies on hospitalized, older patients used polypharmacy levels other than the widespread definition of 5 or more drugs [25-27]. An investigation of the physical performance of hospitalized older patients defined polypharmacy as the use of 10 or more drugs during hospital stay [26]. A study analyzing risk factors of increased hospital stay in older adults defined polypharmacy as the concomitant use of 5 to 9 drugs and excessive polypharmacy as 10 or more drugs [25]. The polypharmacy definition for nursing home patients as well as older populations in general was also 5-9 concurrent drugs, and that of excessive polypharmacy was 10 or more concurrent drugs [27-29]. Kim et al [30] divided polypharmacy into 3 levels: concomitant use of 6 or more drugs as polypharmacy, 11 or more drugs as major polypharmacy, and 21 or more drugs as excessive polypharmacy [30].

Based on WHO recommendations and the literature search mentioned earlier, we suggested a polypharmacy categorization method to detect regimens, including an excessive number of drugs, with an aim to assist in detecting inappropriate polypharmacy (Table 2). Polypharmacy was defined as concomitant prescription of 5 or more drugs, which is consistent with the widespread definition [14] and based on previous studies on polypharmacy in older adults. The next severe polypharmacy level, which was major polypharmacy, was defined as the prescription of 10-19 drugs. This threshold was used, as it was commonly used in the older patient population [25-29]. The most severe polypharmacy level, excessive polypharmacy, was defined as the prescription of 20 or more drugs based on a previous study of polypharmacy use among older adults [30].

Table 2. Definition of polypharmacy based on the number of concomitant drugs.

Concomitant drugs, n	Polypharmacy category
<5	Nonpolypharmacy
5-9	Polypharmacy
10-19	Major polypharmacy
≥20	Excessive polypharmacy

Data Preprocessing and Data Analysis Environment

Preprocessing of the medical institution data was performed by converting the EHR to the OMOP CDM (version 5.3.1) format. We used the following inpatient data extracted from the CDM tables: person, drug_exposure, and visit_occurrence. The data for each CDM table were extracted for the years 2012-2020. Using drug_exposure OMOP CDM data, the severity of polypharmacy and drug use patterns in the patients were analyzed. Based on clinical professional consultation, the study patients were divided into those with 1 admission and those with 2 or more admissions. The operational definition of continuous admission was applied to the data of patients with multiple admissions within 30 days to obtain the final admission counts. Drug exposure data in the outpatient setting for 7 days after follow-up were extracted for each patient. We extracted the drug exposure data for all patients during their observation period and for each day of drug exposure and counted the number of drug ingredients exposed to the patient. The date of the maximum number of dispensed drug ingredients was determined. If the patient was exposed to this maximum number of drug ingredients for multiple days, the date closest to admission was used as the date of the maximum drug ingredient count. The drug ingredient count on that date was used to determine the patient's polypharmacy level. To perform the drug ingredient mapping process, we excluded patients without at least 7 days of drug exposure data after discharge.

Data processing and analyses were performed using R (version 4.1.3; R Foundation for Statistical Computing). Specifically, CDM data access, extraction, and conversion into variables were performed using R packages *RpostgreSql* (version 0.7-8) [31] and *sqldf* (version 0.4-11) [32]. R packages used for data preprocessing were *dplyr* (version 0.8.3) [33] and *data.table* (version 1.12.2) [34], and that used for descriptive analyses was *moonBook* (version 0.3.1) [35]. The *ggplot2* (version 3.2.1) [36] R package was used for data visualization. Medical institute A used 16 GB RAM in the Windows operating system (OS), medical institutes B and C used 64 GB RAM in the Windows

XSL•FO
RenderX

OS, and medical institute D used 128 GB RAM in the Windows OS.

Polypharmacy Pattern Analysis

For each patient, we mapped the WHO Anatomical Therapeutic Chemical (ATC) code to each drug exposure event in the OMOP CDM. Medical institutions A, C, and D assigned WHO ATC codes to drugs in the OMOP CDM using electronic data interchange (EDI) codes. This is possible because the EDI codes are loaded into the institutional CDM. Mapping the WHO ATC codes to EDI codes required a file with code information for drugs distributed in South Korea provided by the Korea Pharmaceutical Information Service [37]. Korea Pharmaceutical Information Service, operated by the Health Insurance Review & Assessment Service, assigns WHO ATC codes to drugs approved in South Korea. The code information file includes EDI, WHO ATC, and drug ingredient codes for each drug approved; therefore, assigning WHO ATC and drug ingredients to each drug in the OMOP CDM was possible. Traditional Korean medications and drugs without WHO-ATC, EDI, and drug ingredient codes were excluded from the mapping process.

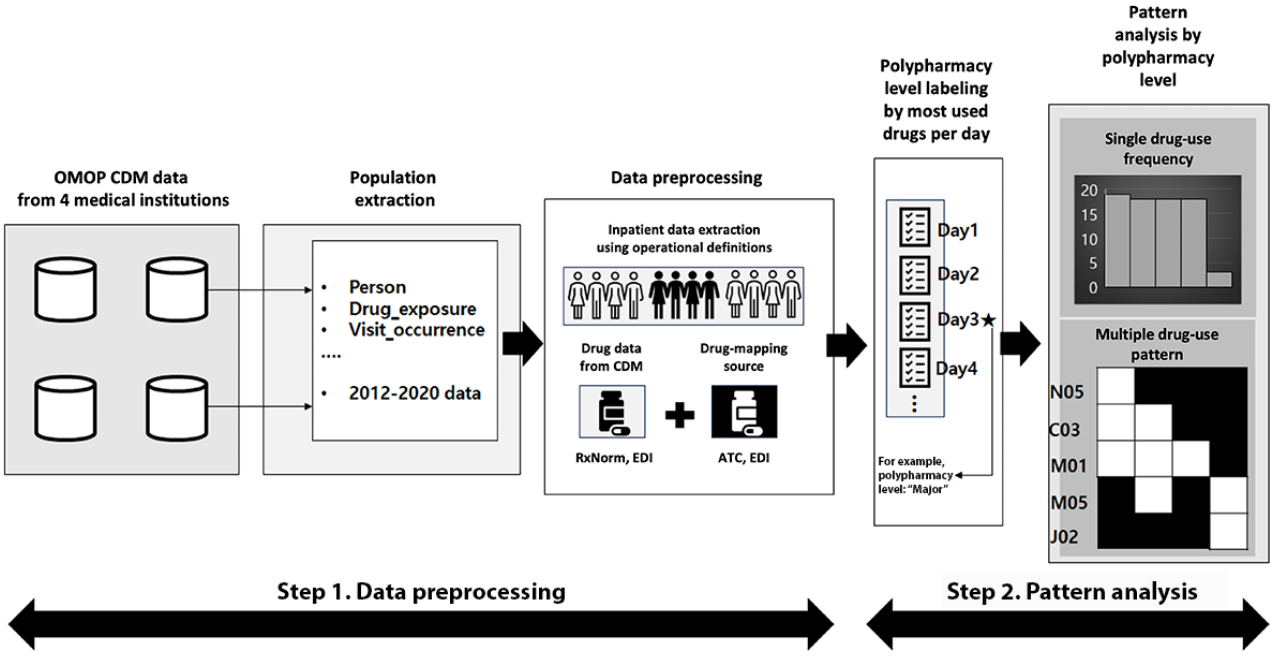
Medical institution B did not have EDI codes available in the OMOP CDM; therefore, a different approach was used to assign WHO ATC codes to the drug concept IDs. This institution maps the WHO ATC code to each drug in the CDM drug_exposure table using the concept and concept_ancestor tables. If the drug_concept_id in the drug_exposure table was a descendant_concept_id of an ancestor_concept_id indicating a fifth-level WHO ATC code, then the concept_id and WHO ATC codes were mapped to the drug_concept_id. The concept table was used to determine if the ancestor_concept_id was a WHO ATC code, and the concept_ancestor table was used to determine if drug_concept_id was a descendant_concept_id of a WHO ATC code.

The maximum daily drug ingredient count was used to determine the polypharmacy level. A maximum drug ingredient count of less than 5 drugs was considered nonpolypharmacy, 5-9 drug ingredients as polypharmacy, 10-19 drug ingredients as major polypharmacy, and 20 drug ingredients or more as excessive polypharmacy.

Demographic characteristics were analyzed using the processed CDM data. The characteristics examined included the total number of inpatients and their admissions, patient age, precautionary drug use, proportion of older adults, admission duration, and drug ingredient exposure distribution for the years 2012-2020.

We performed single and combination drug class pattern analyses using the second ATC level as the unit for the drug class. This level was selected, as it represents clinically meaningful therapeutic groups, allowing the comparison of prescribing trends across multiple health care institutions. We excluded drugs in the ATC second-level B05 (blood substitutes and perfusion solutions) for the analyses, as these drugs have a low risk of interacting with other drugs, have little clinical significance, and are used in the majority of inpatients. [38]. The drug use patterns were visualized using the *ggplot2* R package. Single-drug class use patterns were analyzed by determining the number of patients using each drug class at least once. Combination drug class patterns were analyzed by first determining the date of the highest number of drugs prescribed for each patient and examining the drug classes used on that date. The workflow of the study is depicted in Figure 2.

Figure 2. Workflow for polypharmacy pattern determination. ATC: Anatomical Therapeutic Chemical; C03: diuretics; CDM: Common Data Model; EDI: electronic data interchange; J02: antimycotics for systemic use; M01: anti-inflammatories and antirheumatics; M05: drugs for treatment of bone diseases; N05: psycholeptics; OMOP: Observational Medical Outcomes Partnership.



XSL•FO
RenderX

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and approved by the institutional review board of Konyang University Hospital (protocol code: KYUH2022-07-015-001, approved on September 21, 2022), Asan Medical Center (protocol code: 2022-0342, approved on October 21, 2022), Seoul National University Hospital (protocol code: E-2209-069-1358, approved on September 20, 2022), and Chungbuk University Hospital (protocol code: CBNUH2022-09-030, approved on October 18, 2022). Patient consent was waived, as the data were collected during the routine care of physicians and removed of personally identifiable information, therefore eliminating the risk of compromising patient privacy. There was no direct contact with any patients during the course of this study, and no compensation was provided to the participants.

Results

Polypharmacy Levels and Their Drug Use Patterns

The mean maximum daily drug counts in older adults aged 60-69, 70-79, and ≥80 years were highest in patients admitted to medical institution B followed by medical institution A. Duration of admission increased with higher age range in medical institution C. In all institutions except for institution A, the average hospitalization period of the population aged 70 years or older was longer than the population aged 60-69 years (Table 3).

Table 3. Multi-institutional maximum daily number of drugs prescribed and duration of hospitalization in older adults.

Age group (years)	Medical institution A		Medical institution B		Medical institution C		Medical institution D	
	Drugs, mean (SD)	Hospitalization period (days), mean (SD)	Drugs, mean (SD)	Hospitalization period (days), mean (SD)	Drugs, mean (SD)	Hospitalization period (days), mean (SD)	Drugs, mean (SD)	Hospitalization period (days), mean (SD)
60-69	21.3 (10.3)	24.0 (35.5)	24.6 (14.5)	19.3 (25.7)	15.9 (7.0)	17.3 (15.8)	20.0 (10.3)	15.2 (26.3)
70-79	21.5 (10.3)	25.3 (32.4)	25.2 (14.7)	19.7 (24.4)	16.1 (7.1)	18.5 (18.1)	20.3 (10.8)	18.0 (32.7)
≥80	20.9 (9.6)	22.8 (23.5)	24.5 (14.7)	19.3 (28.0)	15.5 (6.9)	19.1 (19.4)	19.8 (10.6)	18.1 (25.7)

Older Adult Polypharmacy Levels Based on Multi-Institutional CDM Data

The level of polypharmacy with the most patients was major polypharmacy in medical institutions A, C, and D, while for medical institution B, this was excessive polypharmacy (Table 4).

XSL•FO
RenderX

Table 4. Distribution of patients by polypharmacy level.

Medical institution and age group (years)	Nonpolypharmacy patients, n (%)	Polypharmacy patients, n (%)	Major polypharmacy patients, n (%)	Excessive polypharmacy patients, n (%)
Medical institution A				
60-69	730 (31.9)	2121 (24.9)	5002 (22.2)	3753 (24.7)
70-79	1123 (49.1)	4085 (48)	10,557 (46.9)	7251 (47.6)
≥80	436 (19)	2297 (27)	6936 (30.8)	4221 (27.7)
Total	2289 (100)	8503 (100)	22,495 (100)	15,225 (100)
Medical institution B				
60-69	4164 (61.4)	7975 (55.6)	22,847 (53.1)	55,016 (54)
70-79	2092 (30.8)	4887 (34.1)	15,422 (35.8)	37,130 (36.4)
≥80	528 (7.8)	1486 (10.4)	4756 (11.1)	9768 (9.6)
Total	6784 (100)	14,348 (100)	43,025 (100)	101,914 (100)
Medical institution C				
60-69	3657 (31.9)	8746 (30.9)	18,573 (31.6)	4777 (31.8)
70-79	5951 (51.9)	14,634 (51.7)	30,563 (52)	7977 (53.1)
≥80	1859 (16.2)	4927 (17.4)	9640 (16.4)	2269 (15.1)
Total	11,467 (100)	28,307 (100)	58,776 (100)	15,023 (100)
Medical institution D				
60-69	963 (26.6)	1954 (24.1)	3018 (22.8)	2173 (24.1)
70-79	1909 (52.7)	3991 (49.2)	6448 (48.7)	4609 (51.1)
≥80	752 (20.8)	2167 (26.7)	3775 (28.5)	2238 (24.8)
Total	3624 (100)	8112 (100)	13,241 (100)	9020 (100)

Multi-Institute CDM-Based Polypharmacy Pattern

The pattern of polypharmacy in the older adults was analyzed based on the use of single or combinations of drug therapeutic subgroups (WHO ATC second-level codes). The older population was defined as adults aged 65 years and older, based on the population grouping method of the National Statistical Office of Korea [1].

Single Drug Use Patterns

We examined frequently prescribed drug therapeutic subgroups using second-level WHO ATC codes as the subgroup unit. The 10 most frequently prescribed subgroups for each medical institution are presented in Figures 3-6. The horizontal axis represents the number of prescriptions for drugs included in the therapeutic subgroup specified on the vertical axis. According to the 10 most frequently prescribed therapeutic subgroups, the number of prescriptions tended to increase with higher levels of polypharmacy in all medical institutions, excluding institution C. The highest prescription frequency in this institution was observed at the major polypharmacy level.

For medical institution A, the antithrombotics (B01) therapeutic subgroup was frequently prescribed in patients in the nonpolypharmacy and polypharmacy levels compared to other levels and antibacterials for systemic use (J01) for the major and excessive polypharmacy levels. Notably, drugs in the anesthetic (N01) subgroup were most frequently prescribed to patients with excessive polypharmacy (Figure 3).

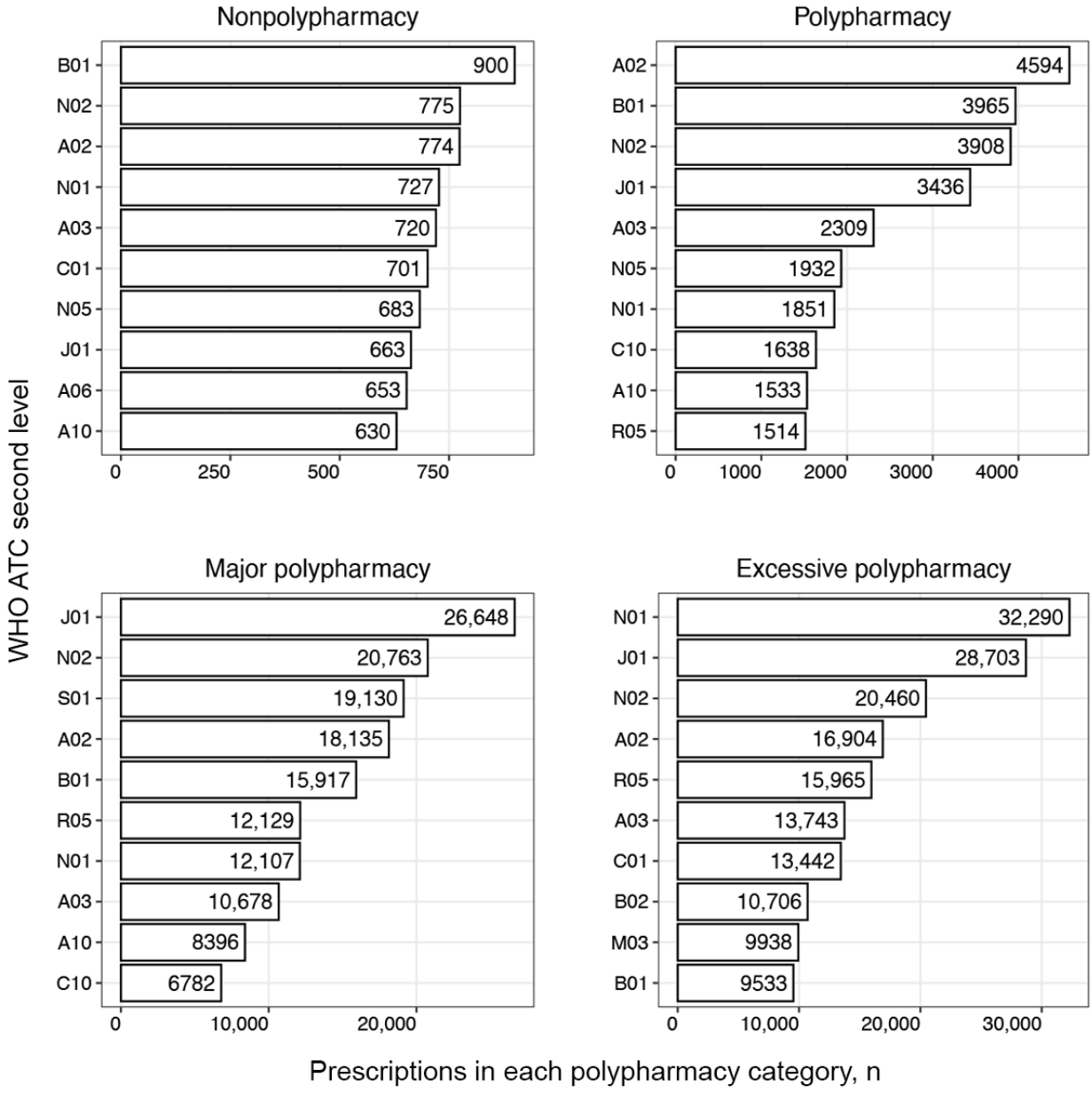
In medical institution B, patients in all polypharmacy levels except the excessive level were most frequently prescribed analgesics (N02). The most frequently prescribed therapeutic subgroups for patients with excessive polypharmacy included mineral supplements (A12) and antianemic preparations (B03). Patients in all polypharmacy levels were frequently prescribed antibacterial agents for systemic use (J01) and drugs for acid-related disorders (A02). Patients in the major polypharmacy level were frequently prescribed cardiac therapy (C01) and drugs for constipation (A06; Figure 4).

In medical institution C, the analgesic (N02) therapeutic subgroup was prescribed most frequently to patients in all polypharmacy levels. Cough and cold preparations (R05) were relatively frequently prescribed to patients in the nonpolypharmacy level, and lipid-modifying drugs (C10) and antithrombotics (B01) to patients in the polypharmacy level. Anesthetics (N01) were prescribed more frequently to patients in major and excessive polypharmacy levels than to those in other polypharmacy levels (Figure 5).

In medical institution D, antibacterials for systemic use (J01) and analgesics (N02) were prescribed frequently to patients across all polypharmacy levels. Drugs for acid-related disorders (A02) were frequently prescribed to patients in all polypharmacy levels, excluding the excessive level, whereas antihemorrhagics (B02) and anesthetics (N01) were most frequently prescribed to patients in the excessive level (Figure 6).

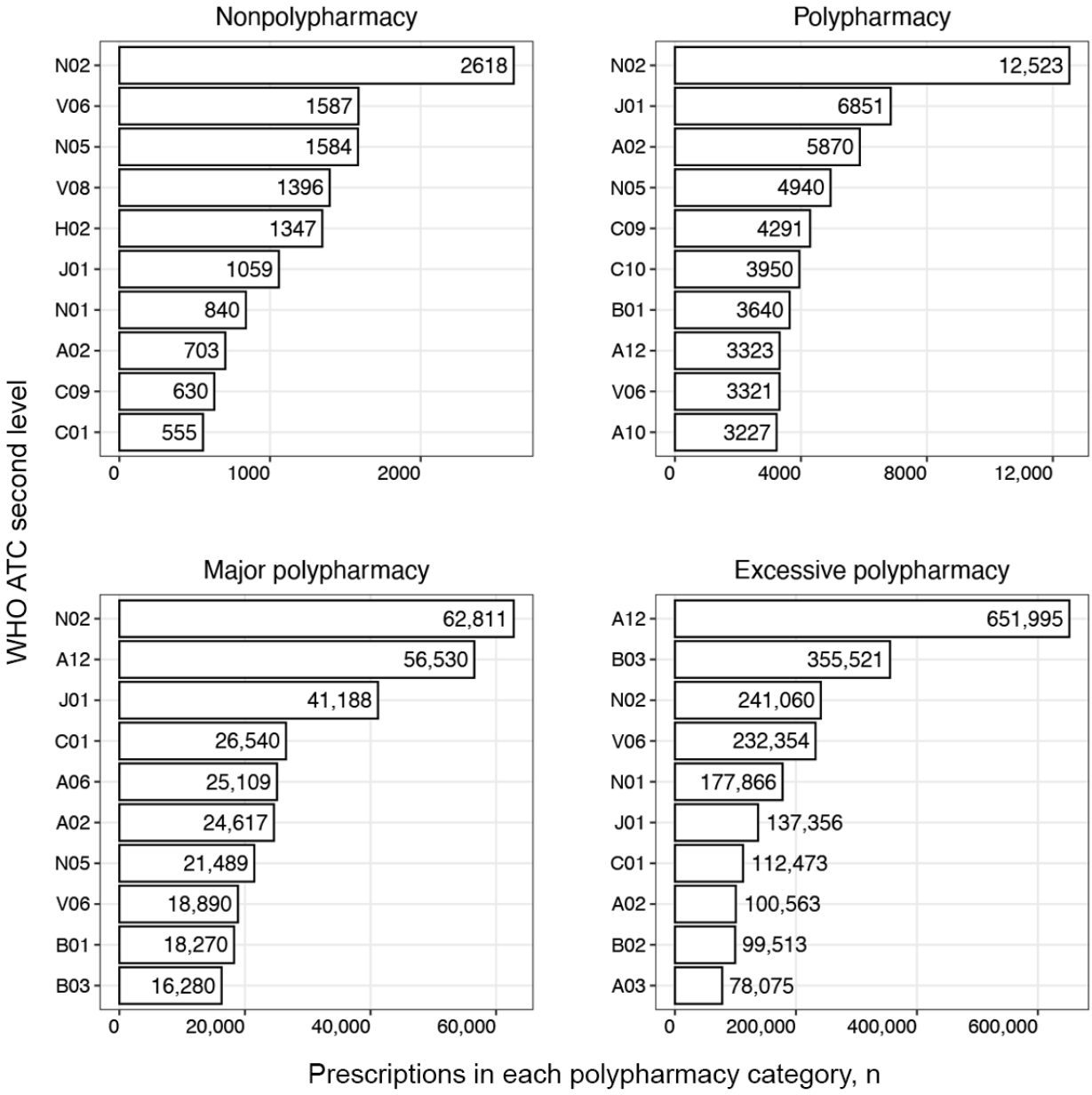
XSL•FO
RenderX

Figure 3. Single therapeutic subgroup (WHO ATC second-level) prescribing pattern in medical institution A. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A06: constipation drug; A10: antidiabetics; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; C01: cardiac therapy; C10: lipid-modifying agent; J01: systemic antibacterials; M03: muscle relaxant; N01: anesthetics; N02: analgesics; N05: psycholeptics; R05: cough and cold preparation; S01: ophthalmologicals; WHO: World Health Organization.



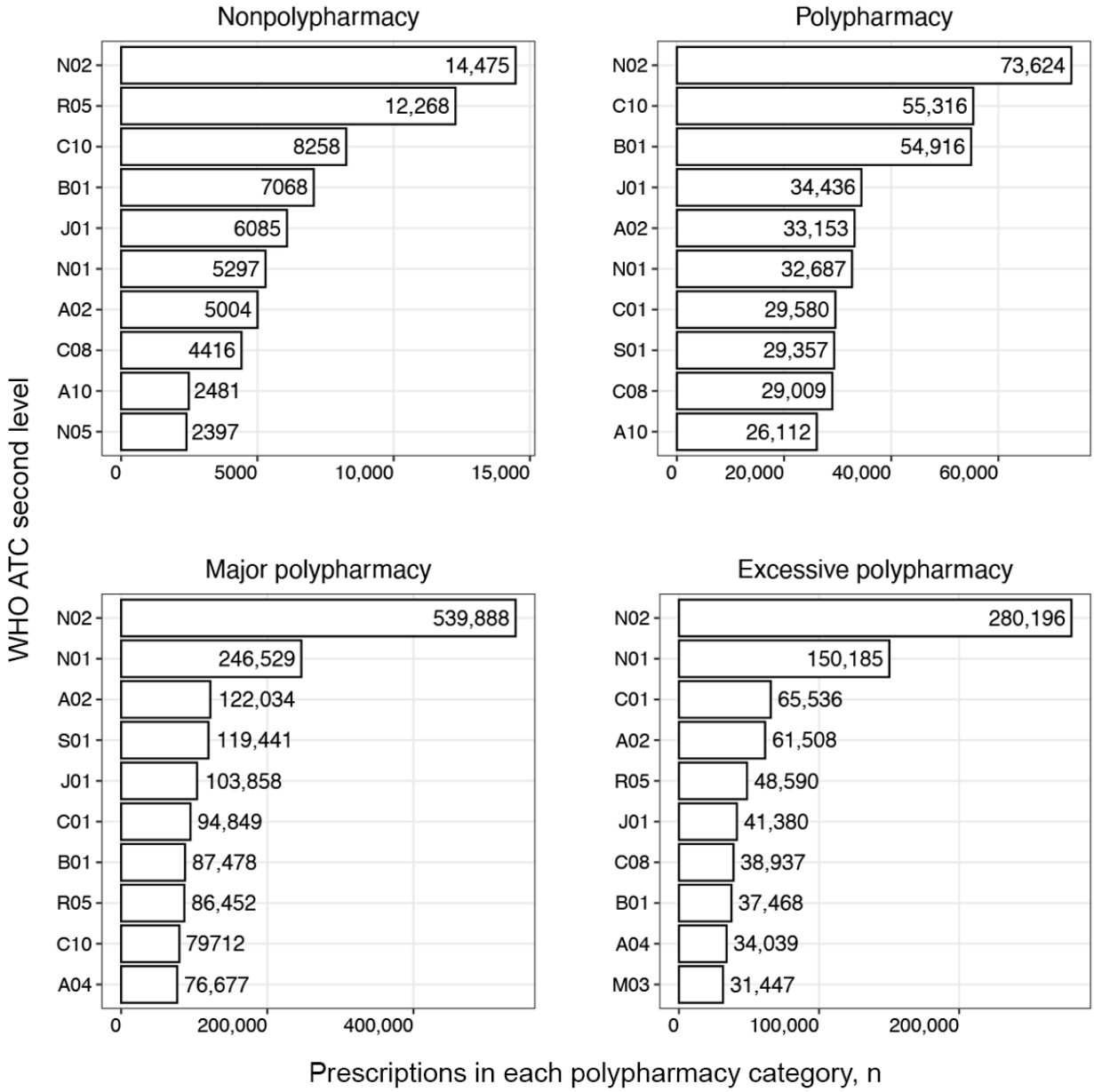
XSL•FO
RenderX

Figure 4. Single therapeutic subgroup (WHO ATC second-level) prescribing pattern in medical institution B. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A06: constipation drug; A10: antidiabetics; A12: mineral supplement; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; B03: antianemics; C01: cardiac therapy; C09: renin-angiotensin system agent; C10: lipid-modifying agent; H02: systemic corticosteroid; J01: systemic antibacterials; N01: anesthetics; N02: analgesics; N05: psycholeptics; V06: general nutrient; V08: contrast media; WHO: World Health Organization.



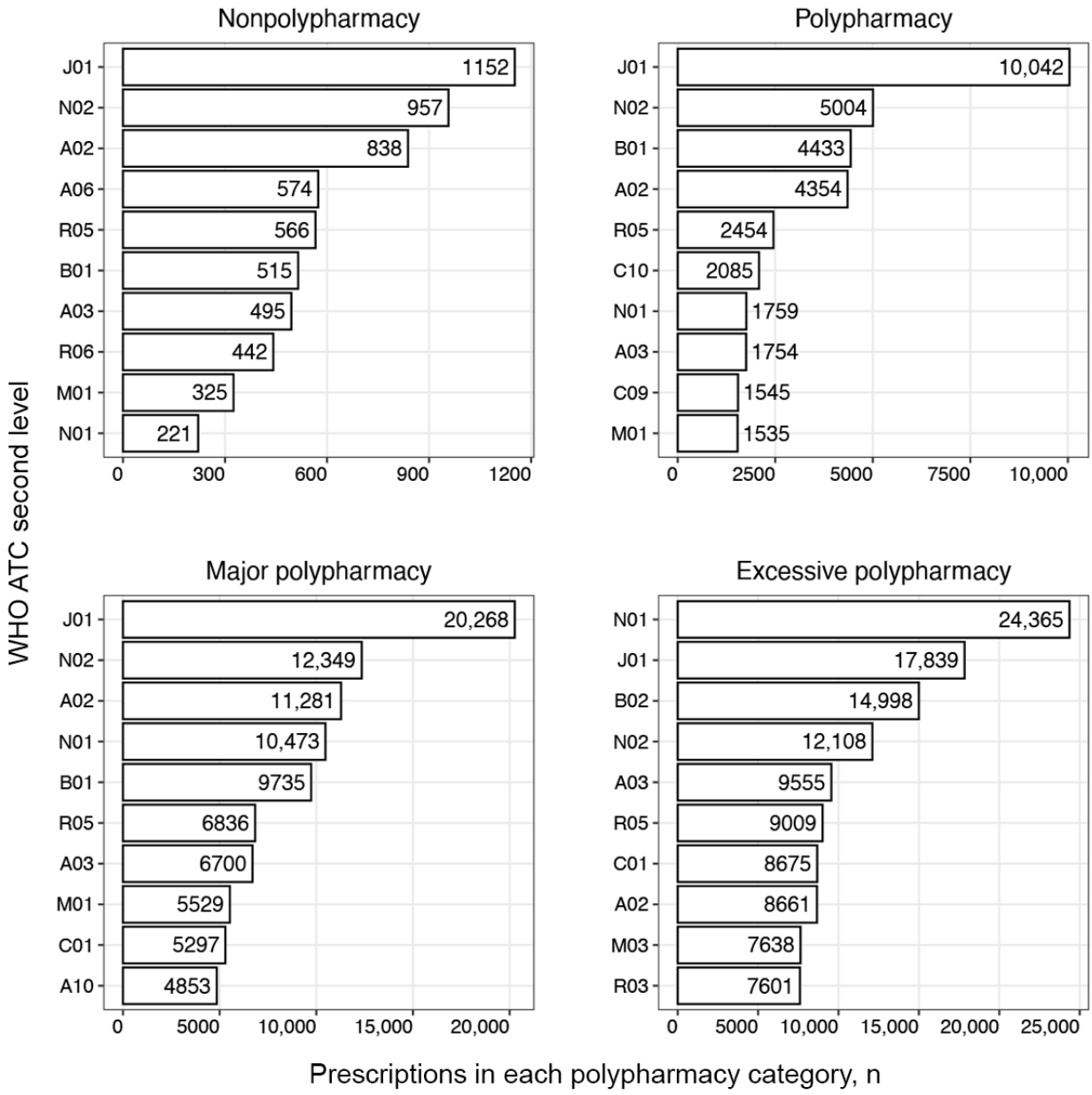
XSL•FO
RenderX

Figure 5. Single therapeutic subgroup (WHO ATC second-level) prescribing pattern in medical institution C. A02: acid-related disorder drug; A04: antiemetics and antinauseants; A10: antidiabetics; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; C01: cardiac therapy; C08: calcium channel blocker; C10: lipid-modifying agent; J01: systemic antibacterials; M03: muscle relaxant; N01: anesthetics; N02: analgesics; N05: psycholeptics; R05: cough and cold preparation; S01: ophthalmologicals; WHO: World Health Organization.



XSL•FO
RenderX

Figure 6. Single therapeutic subgroup (WHO ATC second-level) prescribing pattern in medical institution D. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A06: constipation drug; A10: antidiabetics; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; C01: cardiac therapy; C09: renin-angiotensin system agent; C10: lipid-modifying agent; J01: systemic antibacterials; M01: anti-inflammatories and antirheumatics; M03: muscle relaxant; N01: anesthetics; N02: analgesics; R03: obstructive airway disease drug; R05: cough and cold preparation; R06: systemic antihistamine; WHO: World Health Organization.



Concomitant Drug Use Patterns

Following the examination of single therapeutic subgroup patterns, concomitantly used therapeutic subgroups were analyzed to determine the trends in combination drug use in patients with polypharmacy. At medical institution A, the most frequently prescribed therapeutic subgroup combination for patients of the nonpolypharmacy level was B01 and N01. For patients in the polypharmacy level, A02-C01-N02-N05-V04 was the most frequently prescribed combination. Patients in the major polypharmacy level were most frequently prescribed with the therapeutic subgroup combination C01-H02-J01-M09-N01-N02-S01, while patients in the excessive polypharmacy group were most frequently prescribed the combination A02-A03-C01-H02-J01-M09-N01-N02-S01 (Figure 7). The full names for each of the therapeutic subgroups denoted as WHO ATC second-level codes are in Multimedia Appendix 1.

Results for medical institution B showed that patients in the nonpolypharmacy level were most frequently prescribed therapeutic subgroup combination N02-N05-V06, followed by combination H02-N01. The most frequent combination prescribed to patients at the polypharmacy level was B01-C01-C08-V08, followed by A02-G02-J01-M01-N01-V06 and A02-J01-N02. Patients in the major polypharmacy level were also most frequently prescribed the combination B01-C01-C08-V08. The second most frequently prescribed

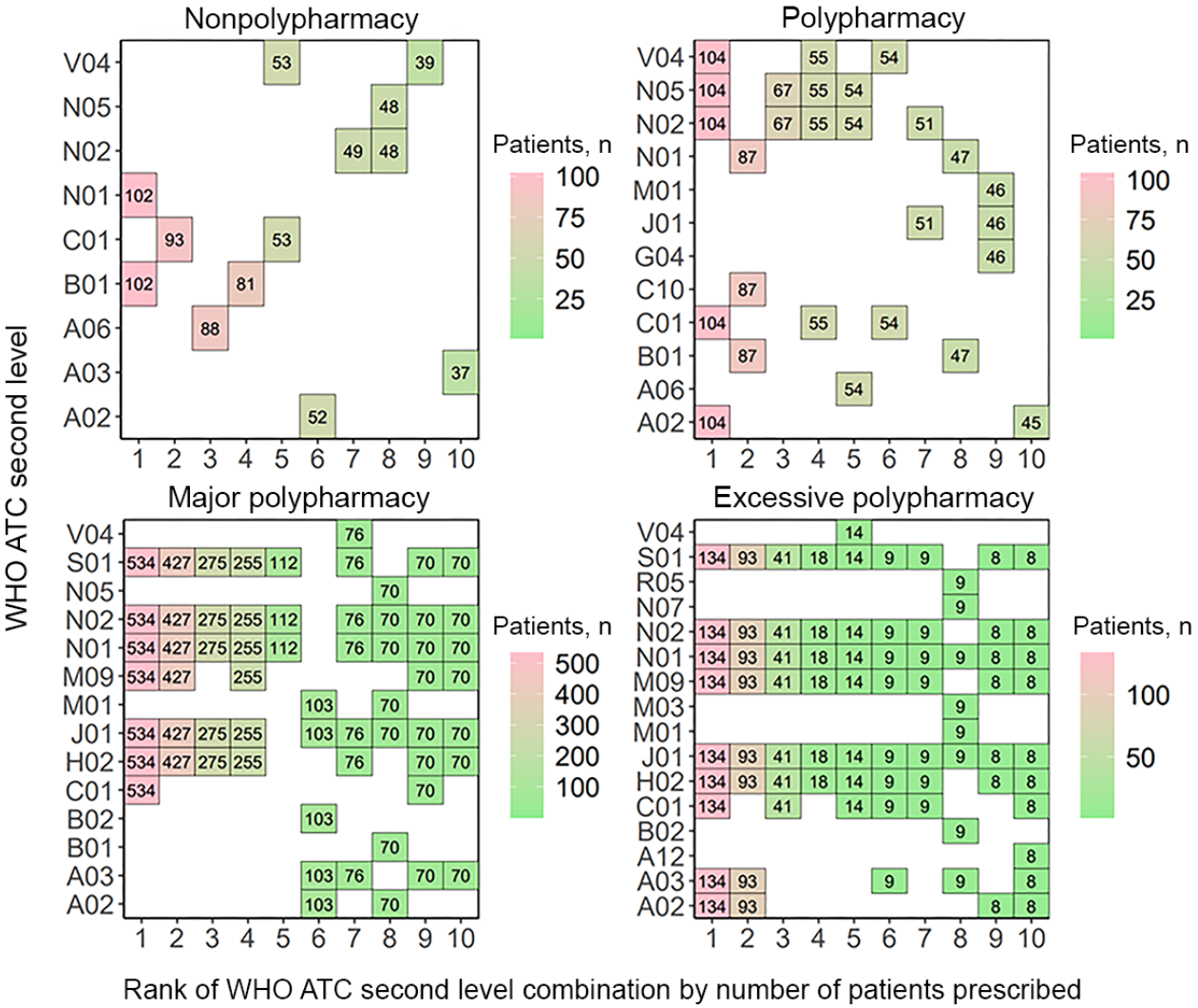
XSL•FO
RenderX

combination at this level was A03-A06-N02-N05-R06-V06, and the third was A12-C01-J01-M09-N02-N05-V06. The most frequently prescribed subgroup combination for patients in the excessive polypharmacy level was A02-A03-B03-D10-N02-N05-V03, followed by A02-A12-B03-D10-N02-N05-V03 and A12-C01-J01-M09-N02-N05-S01-V06 (Figure 8).

In medical institution C, the 3 most frequent therapeutic subgroup combinations prescribed to patients at the nonpolypharmacy level each involved 1 subgroup: R05, N02, and J01. Patients in the polypharmacy level were most frequently prescribed therapeutic subgroup combination A03-M01-N01-S01. The most frequent subgroup combination prescribed for patients in the major polypharmacy level was A03-C01-M01-N01-S01, followed by A03-M01-N01-S01 and A02-A03-C01-N02-N05-V03. Patients in the excessive polypharmacy level were prescribed A02-A04-C01-J01-M03-N01-N02-R03-R05 most frequently. The second frequent combination was the same as the most frequent combination with the addition of subgroups A03 and A07. The third most frequent combination matched the second, excluding subgroup A02 (Figure 9).

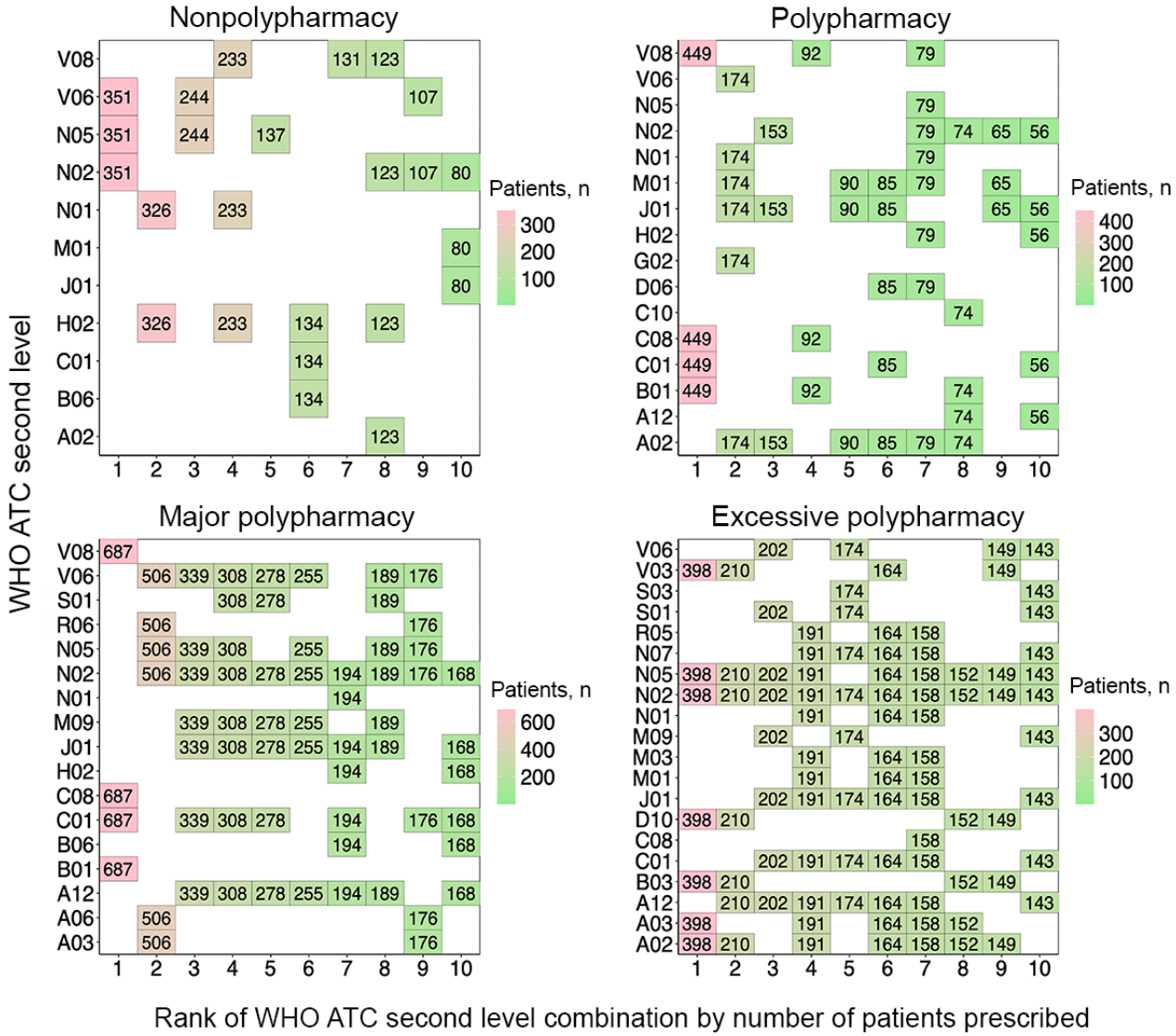
In medical institution D, patients in the nonpolypharmacy level were prescribed therapeutic subgroup J01 most frequently, followed by individual subgroups A02, N02, and A06. Patients in the polypharmacy level were mostly prescribed the combination J01-N01-N02. Those in major polypharmacy level were prescribed B06-C01-H02-J01-N01-N02-S01, followed by J01-N02-R05-R06 and A02-J01-M01-N01-N05. Patients in the excessive polypharmacy level were prescribed combination A03-A04-B02-J01-M01-M03-N01-N02-N07-R05 most frequently, and the second and third most frequent combinations involved the addition of subgroup R03 and the removal of subgroup M01, respectively (Figure 10).

Figure 7. The most frequently prescribed 10 concomitant drug therapeutic subgroups (WHO ATC second-level combinations) used in each polypharmacy level at medical institution A. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A06: constipation drug; A12: mineral supplement; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; C01: cardiac therapy; C10: lipid-modifying agent; G04: urologicals; H02: systemic corticosteroid; J01: systemic antibacterials; M01: anti-inflammatories and antirheumatics; M03: muscle relaxant; M09: other musculo-skeletal system drug; N01: anesthetics; N02: analgesics; N05: psycholeptics; N07: other nervous system drug; R05: cough and cold preparation; S01: ophthalmologicals; V04: diagnostic agent; WHO: World Health Organization.



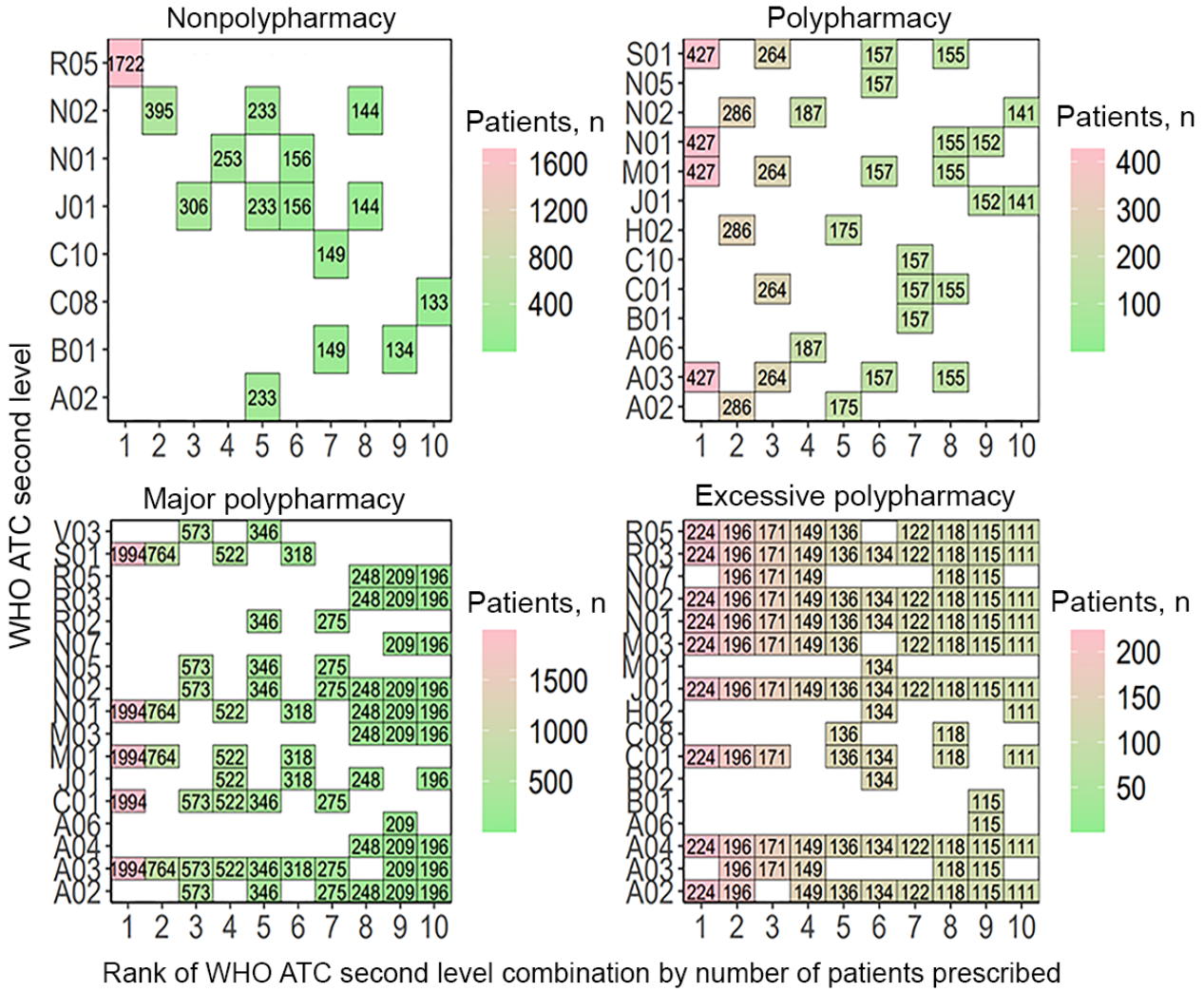
XSL•FO
RenderX

Figure 8. The most frequently prescribed 10 concomitant drug therapeutic subgroups (WHO ATC second-level combinations) used in each polypharmacy level at medical institution B. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A06: constipation drug; A12: mineral supplement; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B03: antianemics; B06: other hematological agent; C01: cardiac therapy; C08: calcium channel blocker; C10: lipid-modifying agent; D06: dermatologic antibiotics and chemotherapeutics; D10: antiacne preparation; G02: other gynecologicals; H02: systemic corticosteroid; J01: systemic antibacterials; M01: anti-inflammatories and antirheumatics; M03: muscle relaxant; M09: other musculo-skeletal system drug; N01: anesthetics; N02: analgesics; N05: psycholeptics; N07: other nervous system drug; R05: cough and cold preparation; R06: systemic antihistamine; S01: ophthalmologicals; S03: ophthalmological and otological preparation; V03: other therapeutic product; V06: general nutrient; V08: contrast media; WHO: World Health Organization.



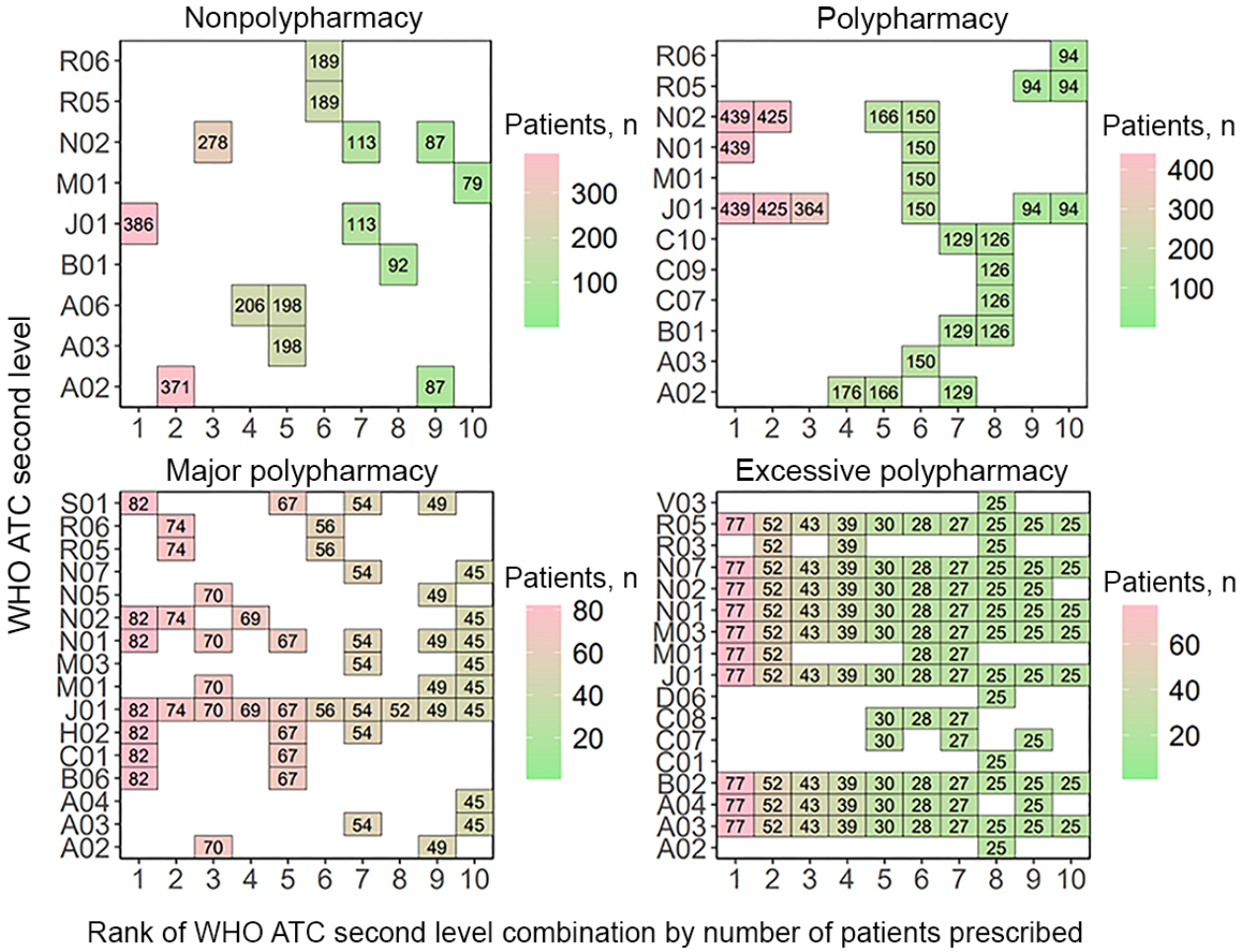
XSL•FO
RenderX

Figure 9. The most frequently prescribed 10 concomitant drug therapeutic subgroups (WHO ATC second-level combinations) used in each polypharmacy level at medical institution C. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A04: antiemetics and antinauseants; A06: constipation drug; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; C01: cardiac therapy; C08: calcium channel blocker; C10: lipid-modifying agent; H02: systemic corticosteroid; J01: systemic antibacterials; M01: anti-inflammatories and antirheumatics; M03: muscle relaxant; N01: anesthetics; N02: analgesics; N05: psycholeptics; N07: other nervous system drug; R02: throat preparation; R03: obstructive airway disease drug; R05: cough and cold preparation; S01: ophthalmologicals; V03: other therapeutic product; WHO: World Health Organization.



XSL•FO
RenderX

Figure 10. The most frequently prescribed 10 concomitant drug therapeutic subgroups (WHO ATC second-level combinations) used in each polypharmacy level at medical institution D. A02: acid-related disorder drug; A03: functional gastrointestinal disorder drug; A04: antiemetics and antinauseants; A06: constipation drug; ATC: Anatomical Therapeutic Chemical; B01: antithrombotics; B02: antihemorrhagics; B06: other hematological agent; C01: cardiac therapy; C07: beta-blocking agent; C08: calcium channel blocker; C09: renin-angiotensin system agent; C10: lipid-modifying agent; D06: dermatologic antibiotics and chemotherapeutics; H02: systemic corticosteriod; J01: systemic antibacterials; M01: anti-inflammatories and antirheumatics; M03: muscle relaxant; N01: anesthetics; N02: analgesics; N05: psycholeptics; N07: other nervous system drug; R03: obstructive airway disease drug; R05: cough and cold preparation; R06: systemic antihistamine; S01: ophthalmologicals; V03: other therapeutic product; WHO: World Health Organization.



Discussion

Principal Findings

This study used EHR data converted to the OMOP CDM from 4 medical institutions for geriatric inpatients aged 65 years and older admitted over 9 years. Our study is the first to collect the EHR of patients and retrospectively study polypharmacy patterns over multiple institutions to determine a polypharmacy threshold in hospitalized older patients. The authors used data from 2 of the largest hospitals in the South Korean metropolitan area and 2 hospitals in the regional area to represent patients from distinct settings.

The use of OMOP CDM in all 4 study settings allowed the use of similar programming languages and software, thereby increasing the efficiency of the study. Although there were differences in the specific columns customized for each institution, the uniformity of the basic OMOP CDM tables and columns across institutions allowed for the distribution of a single analysis code customized for each institution, and the collection of analysis resulted in a standardized format.

We offer a possible definition of polypharmacy in hospitalized older adults. This was obtained by consulting clinical professionals to define the inpatient observation periods and reviewing previous studies [25,26,30] to determine the time point of drug exposure for polypharmacy level determination. We filtered out drugs that did not contribute to adverse events from polypharmacy (eg, WHO ATC code B05 and its lower-level codes), which improved the quality of our data.

The 4 health care institutions (A, B, C, and D) included in this study showed distinct differences in patient scale, proportion of older patients, sex and age distribution, and medication use patterns. Institution C had the largest patient population with 3,349,795 patients and demonstrated high proportions of older patients and excessive polypharmacy levels. Institution B had the next highest patient population with 1,944,140 patients and the highest number of hospitalization events with 1,092,936 events. This trend reflects that institutions located in metropolitan, urban settings had a higher patient population size and hospitalization event count than their rural counterparts. Institution A, despite having a relatively smaller scale with 602,483 patients, was characterized by the highest average

XSL•FO
RenderX

number of medications per age group and the longest hospital stays.

In terms of sex distribution, institution A showed higher male ratios in patients aged 5-60 years, but female patients were more prevalent in those aged 70 years and older. Institution C displayed higher female ratios in patients in their 50s and 80s or older, while male patients were more common in the 60- to 70-year age group. Institution D exhibited a unique pattern with higher female ratios in patients aged 80 years and older and higher male ratios in those aged 5-70 years. Overall, the proportion of female patients tended to increase in the older population, with sex ratios varying across institutions and age groups.

A tendency was observed for hospital length of stay to increase with advancing age across all institutions. Regarding polypharmacy prescriptions, institution B demonstrated the highest "excessive polypharmacy level" ratio across all age groups.

Comparison to Prior Work

The pattern of drug use among inpatients in the 4 medical institutions showed that patients in the major or excessive polypharmacy levels were most prevalent. This trend is consistent with the results of previous studies reporting 5-10 drugs or more as the polypharmacy threshold. A study using nationwide health care claims data found that hospital admission was significantly associated with polypharmacy (use of 5 or more drugs) and hyper-polypharmacy (use of 10 or more drugs) [39]. Another study reported an increase from 5.6 drugs taken per patient on average at admission to 7.6 drugs at discharge [40]. Considering that the number of drugs prescribed tended to increase during hospital admission, the use of 10-19 or more drugs (major polypharmacy or higher) as the threshold for inpatient polypharmacy was considered appropriate. All data were obtained and analyzed within the securely blocked environment of each medical institution.

The mean number of drugs used in medical institution B was higher than that used in other institutions by 3-10 drugs. The prescription frequency of drugs in the cardiac therapy (C01) WHO ATC therapeutic subgroups tended to increase with the severity of polypharmacy in all medical institutions, and the anti-inflammatory and antirheumatic product (M01) therapeutic subgroups increased across all institutions, excluding institution B. These results correspond with those of a study in which hypertension, osteoarthritis, and rheumatoid arthritis were reported to be prevalent in older individuals aged 65 years or older [41].

The patterns of concomitant therapeutic subgroups differed among the medical institutions. In the major polypharmacy category, medical institution A patients were frequently prescribed ophthalmologicals (WHO ATC second-level code S01), along with other therapeutic subgroups. This drug class has been consistently used in most prescribed combinations of therapeutic subgroups. Medical institution B prescribed drugs in various therapeutic categories including acid-related disorders (A02), functional gastrointestinal disorders (A03), antianemic preparations (B03), antithrombotics (B01), and cardiac therapies (C01) to patients in the major and excessive polypharmacy levels. Medical institution C, which had the highest number of study patients, prescribed drugs for WHO ATC second-level functional gastrointestinal disorders (A03), anti-inflammatory and antirheumatic products (M01), and ophthalmologicals (S01). These drug classes were in the 1-3 highest rank of concomitant drugs prescribed.

Distinct polypharmacy patterns in the polypharmacy group were that patients in all 4 institutions were more frequently prescribed lipid-modifying agents (C10), and in 3 institutions, antidiabetics (A10) were more frequently prescribed, compared to other polypharmacy groups (Figures 3-6). This reflects the increased prevalence of chronic diseases in the older population and their demand for multiple medications to treat the disease [42]. Furthermore, the result is consistent with a finding that older patients with diabetes have a high prevalence of polypharmacy, as these patients commonly have multiple concomitant diseases including cardiovascular, renal, and respiratory disorders [43]. The higher prescription frequencies of ophthalmic agents (S01) and anti-inflammatory and analgesic drugs (M01) observed in institution C reflect that this institution is highly specialized in ophthalmology or rheumatology. The major polypharmacy group was frequently prescribed cough and cold preparations (R05) compared to other polypharmacy groups (Figures 3-6). This trend was consistent with results of a previous study, which reported patients on 5 or more drugs having an odds ratio of 3.18 and 6 or more drugs having an odds ratio of 4.57 to receive prescriptions for cough and cold drugs [44].

Distinctive polypharmacy patterns observed in the excessive polypharmacy group were that cardiac drugs (C01) were frequently prescribed in all 4 institutions (Figures 3-6). This observation suggests that patients in the excessive polypharmacy group may have a higher burden of cardiovascular diseases or more severe cardiac comorbidities requiring complex medication regimens [45]. The frequent prescription of cardiac drugs in this group indicates that managing heart-related conditions—such as heart failure, arrhythmia, or ischemic heart disease—contributes substantially to the overall medication load. This reflects that the patients in the excessive polypharmacy group are clinically more fragile, thus requiring multiple medications for symptom control, prevention of complications, and comorbidity management. This group was also frequently prescribed functional gastrointestinal disorder drugs (A03), antihemorrhagics (B02), and muscle relaxants (M03; Figures 7-10). The reason for the high prevalence of antihemorrhagics prescribing indicates frequent surgeries or procedures in this population [46,47]. Muscle relaxants were found to be frequently prescribed in older polypharmacy patients and coprescribed with opioids [48]. This medication was shown to increase the risk of adverse outcomes including all-cause and cardiovascular mortality, warranting enhanced monitoring of drug therapy [48]. Gastrointestinal drugs may be used for the gastric protection caused by polypharmacy or due to the decreased physiological function of the gastrointestinal system [49].

Although each institution had a distinctive pattern, we observed that drugs related to surgeries, including those needed after the operation (eg, WHO ATC second-level codes C01, H02, J01,

XSL•FO
RenderX

N01, and N02), were commonly prescribed across institutions (Figures 7-10). This reflects the trend from 2011 to 2015, where 3,146,452 of 8,497,443 (37%) surgeries were performed in patients aged 65 years or older, and the surgeries were for high-risk conditions, such as glaucoma [50]. We observed that older inpatients frequently underwent surgeries and were prescribed several drugs during admission for these procedures. The high frequency of antibacterials (WHO ATC second-level code J01), anesthetics (N01), and analgesics (N02) in all polypharmacy levels in the older population reflects the complex clinical conditions of hospitalized older patients, who often require management of infections, surgeries or procedures, and pain control [51]. This population has a high prevalence of osteoarthritis, fractures, and cancer-related pain, increasing the risk of anesthetic and analgesic use [52].

Limitations

This study has several limitations. First, we included 4 medical institutions, which restricts the generalizability of the study results to the institutions of interest. Second, drugs without a mapped WHO ATC code were excluded; therefore, the prevalence of polypharmacy might have been underestimated. Third, we analyzed the pattern of polypharmacy using the second-level ATC code; thus, polypharmacy patterns at the ingredient level were not determined. Fourth, although we examined outpatient drug prescriptions up to 7 days after discharge, polypharmacy use in older outpatients was not comprehensively analyzed. This is a topic of interest for future studies. Fifth, we did not assess the appropriateness of polypharmacy based on linking prescribed drugs and their indications. Although an estimate of the drug-indication linkage is possible, the precise linkage is challenging, as the indications for the drugs were not explicitly recorded in the OMOP CDM database. Finally, the tables used for the analyses did not include the note table, which includes clinical results or interpretations in free-text format, and is a potential resource for detecting polypharmacy.

Conclusions

Based on well-established EHR-based CDM datasets in South Korea, we suggest a polypharmacy definition for older inpatients that is different from the commonly used definition of the concurrent prescription of 5 or more drugs. This is because, based on the severity of the patients' condition and the treatment they require, which includes major surgeries, the number of concomitant drugs used was higher than the commonly used definition of polypharmacy in the majority of patients in this study. Therefore, we suggest that 10 or more drugs is a more suitable definition of polypharmacy for geriatric inpatients. As the geriatric population increases, evaluating the appropriateness of existing polypharmacy definitions and serious outcomes of polypharmacy in this population, including renal failure, falls, and fractures, is necessary for future studies.

Acknowledgments

The authors appreciate the financial support provided by the Korea Health Industry Development Institute, Asan Medical Center, and the Korea Institute of Drug Safety and Risk Management. The authors thank Editage [53] for the English language editing.

Funding

This work was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (grant HR21C0198 and RS-2022-KH125153), the Asan Institute for Life Sciences of Asan Medical Center (grant 2023IP0060), and the Korea Institute of Drug Safety and Risk Management in 2022.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy protection policies but are available after approval or exemption by the institutional review boards of the study institutions.

Authors' Contributions

Conceptualization: GJK, JHK, SL
Methodology: CCL, JYH, WMH, SL
Formal analysis: CCL, GJK, MK, SK
Data curation: CCL, GJK, MK, SK
Visualization: CCL, GJK, MK, SK
Writing—original draft preparation: CCL, KHL, GJK
Writing—review and editing: CCL, KHL, GJK, JYH, JYK, KK, SK, MK, WMH, JHK, SL

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

World Health Organization Anatomical Therapeutic Chemical second-level codes and names.
[DOCX File , 31 KB - medinform_v14i1e78353_app1.docx]

References

1. Population projections for Korea (2020~2070). Ministry of Data and Statistics. 2021. URL: https://mods.go.kr/board.es?mid=a20108080000&bid=11748&act=view&list_no=416209 [accessed 2025-12-31]
2. Kim DS, Yun SH, Cho HJ, Chae J. Management of polypharmacy in the elderly. Health Insurance Review and Assessment Service Repository. 2022 Jan. URL: https://repository.hira.or.kr/handle/2019.oak/3039 [accessed 2025-12-31]
3. Faquetti ML, la Torre AM, Burkard T, Obozinski G, Burden AM. Identification of polypharmacy patterns in new-users of metformin using the Apriori algorithm: a novel framework for investigating concomitant drug utilization through association rule mining. Pharmacoepidemiol Drug Saf 2023 Mar;32(3):366-381. [doi: 10.1002/pds.5583] [Medline: 36579709]
4. Management of medications for elderly patients on polypharmacy. Seoul Metropolitan Council Secretariat. 2021. URL: https://books.google.co.kr/books?id=mdOgzwEACAAJ [accessed 2025-12-31]
5. Drug utilization review system of South Korea. Health Insurance Review & Assessment Service. 2022. URL: https://www.hira.or.kr/dummy.do?pgmid=HIRAA990001000330 [accessed 2025-12-31]
6. Ishizaki T, Mitsutake S, Hamada S, Teramoto C, Shimizu S, Akishita M, et al. Drug prescription patterns and factors associated with polypharmacy in >1 million older adults in Tokyo. Geriatr Gerontol Int 2020 Apr;20(4):304-311. [doi: 10.1111/ggi.13880] [Medline: 32048453]
7. George CJ, Verghese J, Izzetoglu M, Wang C, Holtzer R. The effect of polypharmacy on prefrontal cortex activation during single and dual task walking in community dwelling older adults. Pharmacol Res 2019 Jan;139:113-119 [FREE Full text] [doi: 10.1016/j.phrs.2018.11.007] [Medline: 30408573]
8. Jyrkkä J, Enlund H, Korhonen MJ, Sulkava R, Hartikainen S. Patterns of drug use and factors associated with polypharmacy and excessive polypharmacy in elderly persons: results of the Kuopio 75+ study: a cross-sectional analysis. Drugs Aging 2009;26(6):493-503. [doi: 10.2165/00002512-200926060-00006] [Medline: 19591524]
9. Aparasu RR, Mort JR, Brandt H. Polypharmacy trends in office visits by the elderly in the United States, 1990 and 2000. Res Social Adm Pharm 2005 Sep;1(3):446-459. [doi: 10.1016/j.sapharm.2005.06.004] [Medline: 17138489]
10. Li M, Wei N, Shi H, Jing X, Kan X, Gao H, et al. Prevalence and clinical implications of polypharmacy and potentially inappropriate medication in elderly patients with heart failure: results of six months' follow-up. J Geriatr Cardiol 2023 Jul 28;20(7):495-508 [FREE Full text] [doi: 10.26599/1671-5411.2023.07.002] [Medline: 37576481]
11. Zhao Y, Wang J, Zhu X, Zhang X, Zhang Y, Zhang W, et al. Multimorbidity and polypharmacy in hospitalized older patients: a cross-sectional study. BMC Geriatr 2023 Jul 11;23(1):423 [FREE Full text] [doi: 10.1186/s12877-023-04109-4] [Medline: 37434147]
12. Chavarro-Carvajal DA, Sánchez DC, Vargas-Beltran MP, Venegas-Sanabria LC, Muñoz OM. Clinical value of Hospital Admission Risk Profile (HARP) and the Identification of Seniors at Risk (ISAR) scales to predict hospital-associated functional decline in an acute geriatric unit in Colombia. Colomb Med (Cali) 2023;54(1):e2005304 [FREE Full text] [doi: 10.25100/cm.v54i1.5304] [Medline: 37440979]
13. Al Farsi RS, Al Alawi AM, Al Huraizi AR, Al-Saadi T, Al-Hamadani N, Al Zeedy K, et al. Delirium in medically hospitalized patients: prevalence, recognition and risk factors: a prospective cohort study. J Clin Med 2023 Jun 07;12(12):3897 [FREE Full text] [doi: 10.3390/jcm12123897] [Medline: 37373591]
14. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. BMC Geriatr 2017 Oct 10;17(1):230 [FREE Full text] [doi: 10.1186/s12877-017-0621-2] [Medline: 29017448]
15. Cresswell KM, Fernando B, McKinstry B, Sheikh A. Adverse drug events in the elderly. Br Med Bull 2007;83:259-274. [doi: 10.1093/bmb/ldm016] [Medline: 17580312]
16. Hanlon JT, Schmader KE, Koronkowski MJ, Weinberger M, Landsman PB, Samsa GP, et al. Adverse drug events in high risk older outpatients. J Am Geriatr Soc 1997 Aug;45(8):945-948. [doi: 10.1111/j.1532-5415.1997.tb02964.x] [Medline: 9256846]
17. Mabuchi T, Hosomi K, Yokoyama S, Takada M. Polypharmacy in elderly patients in Japan: analysis of Japanese real-world databases. J Clin Pharm Ther 2020 Oct;45(5):991-996. [doi: 10.1111/jcpt.13122] [Medline: 31986233]
18. Standardized data: the OMOP Common Data Model. Observational Health Data Science and Informatics. URL: https://www.ohdsi.org/data-standardization/the-common-data-model [accessed 2024-03-31]
19. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
20. ATC/DDD Index 2025. Norwegian Institute of Public Health. URL: https://atcddd.fhi.no/atc_ddd_index [accessed 2022-10-12]
21. Mair A. Medication safety in polypharmacy: technical report. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/medication-safety-in-polypharmacy-technical-report [accessed 2025-11-01]

XSL•FO
RenderX

22. van Poelgeest E, Seppala L, Bahat G, Ilhan B, Mair A, van Marum R, et al. Optimizing pharmacotherapy and deprescribing strategies in older adults living with multimorbidity and polypharmacy: EuGMS SIG on pharmacology position paper. Eur Geriatr Med 2023 Dec;14(6):1195-1209 [FREE Full text] [doi: 10.1007/s41999-023-00872-0] [Medline: 37812379]
23. Multimorbidity: clinical assessment and management. National Institute for Health and Care Excellence. 2016. URL: https://www.nice.org.uk/guidance/ng56 [accessed 2025-11-01]
24. Young EH, Pan S, Yap AG, Reveles KR, Bhakta K. Polypharmacy prevalence in older adults seen in United States physician offices from 2009 to 2016. PLoS One 2021;16(8):e0255642 [FREE Full text] [doi: 10.1371/journal.pone.0255642] [Medline: 34343225]
25. Vetrano DL, Landi F, De Buyser SL, Carfì A, Zuccalà G, Petrovic M, et al. Predictors of length of hospital stay among older adults admitted to acute care wards: a multicentre observational study. Eur J Intern Med 2014 Jan;25(1):56-62. [doi: 10.1016/j.ejim.2013.08.709] [Medline: 24054859]
26. Sganga F, Vetrano DL, Volpato S, Cherubini A, Ruggiero C, Corsonello A, et al. Physical performance measures and polypharmacy among hospitalized older adults: results from the CRIME study. J Nutr Health Aging 2014;18(6):616-621 [FREE Full text] [doi: 10.1007/s12603-014-0029-z] [Medline: 24950153]
27. Onder G, Liperoti R, Fialova D, Topinkova E, Tosato M, Danese P, et al. Polypharmacy in nursing home in Europe: results from the SHELTER study. J Gerontol A Biol Sci Med Sci 2012 Jun;67(6):698-704. [doi: 10.1093/gerona/glr233] [Medline: 22219520]
28. Chan DD, Hao Y, Wu S. Polypharmacy among disabled Taiwanese elderly: a longitudinal observational study. Drugs Aging 2009;26(4):345-354. [doi: 10.2165/00002512-200926040-00005] [Medline: 19476401]
29. Jyrkkä J, Enlund H, Korhonen MJ, Sulkava R, Hartikainen S. Polypharmacy status as an indicator of mortality in an elderly population. Drugs Aging 2009;26(12):1039-1048. [doi: 10.2165/11319530-000000000-00000] [Medline: 19929031]
30. Kim HA, Shin JY, Kim MH, Park BJ. Prevalence and predictors of polypharmacy among Korean elderly. PLoS One 2014;9(6):e98043 [FREE Full text] [doi: 10.1371/journal.pone.0098043] [Medline: 24915073]
31. Conway J, Eddelbuettel D, Nishiyama T, Prayaga S, Tiffin N. RPostgreSQL: R Interface to the 'PostgreSQL' Database System. 2025. URL: https://cran.r-project.org/web/packages/RPostgreSQL/index.html [accessed 2026-01-02]
32. Grothendieck G. sqldf: Manipulate R Data Frames Using SQL. 2017. URL: https://cran.r-project.org/web/packages/sqldf/index.html/ [accessed 2026-01-02]
33. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. 2025. URL: https://cran.r-project.org/web/packages/dplyr/index.html/ [accessed 2026-01-02]
34. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T. data.table: Extension of 'data.frame'. 2025. URL: https://cran.r-project.org/web/packages/data.table/index.html [accessed 2026-01-02]
35. Moon K. R statistics and graphs for medical papers. 2015. URL: https://cran.r-project.org/web/packages/moonBook/index.html [accessed 2026-01-02]
36. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag New York; 2016. ISBN: 978-3-319-24277-4.
37. Drug code mapping information. Korea Pharmaceutical Information Service. 2019. URL: https://biz.kpis.or.kr/index.html [accessed 2022-10-01]
38. Ataei SM, Golestani A, Khosravi S, Tabatabaei-Malazy O, Malekpour M, Ebrahimpur M, et al. Medication use patterns and polypharmacy among elderly in Iran: a cross-sectional study using national health insurance claims data. BMJ Open 2025 Aug 21;15(8):e097863 [FREE Full text] [doi: 10.1136/bmjopen-2024-097863] [Medline: 40840984]
39. Cho HJ, Chae J, Yoon S, Kim D. Factors related to polypharmacy and hyper-polypharmacy for the elderly: a nationwide cohort study using National Health Insurance data in South Korea. Clin Transl Sci 2023;16(2):193-205 [FREE Full text] [doi: 10.1111/cts.13438] [Medline: 36401587]
40. Viktil KK, Blix HS, Eek AK, Davies MN, Moger TA, Reikvam A. How are drug regimen changes during hospitalisation handled after discharge: a cohort study. BMJ Open 2012;2(6):e001461 [FREE Full text] [doi: 10.1136/bmjopen-2012-001461] [Medline: 23166124]
41. Oh YH. The health status of older Koreans and policy considerations. Health and Welfare Policy Forum 2015;223:29-39. [doi: 10.23062/2015.05.4]
42. Kuribayashi R, Hasebe S, Ishida D, Hirose S, Mimori S, Takizawa T, et al. An age-stratified cross-sectional study of antidiabetic and non-antidiabetic drugs prescribed to Japanese outpatients with diabetes. Glob Health Med 2023 Oct 31;5(5):294-300 [FREE Full text] [doi: 10.35772/ghm.2023.01059] [Medline: 37908508]
43. Remelli F, Ceresini MG, Trevisan C, Noale M, Volpato S. Prevalence and impact of polypharmacy in older patients with type 2 diabetes. Aging Clin Exp Res 2022 Sep;34(9):1969-1983 [FREE Full text] [doi: 10.1007/s40520-022-02165-1] [Medline: 35723858]
44. Komagamine J. Characteristics of the use of cold combination products among older ambulatory patients at the National Hospital Organization Tochigi Medical Center in Japan: a retrospective single-center observational study. BMC Res Notes 2017 Dec 08;10(1):728 [FREE Full text] [doi: 10.1186/s13104-017-3070-2] [Medline: 29221489]

XSL•FO
RenderX

45. Sefera B, Sheleme T, Dechasa M, Ararame GG. Determinants of polypharmacy among ambulatory cardiovascular disease patients in Iluababora and Buno Bedele Zones, Ethiopia: prospective observational study. BMC Cardiovasc Disord 2025 Jul 04;25(1):468 [FREE Full text] [doi: 10.1186/s12872-025-04929-9] [Medline: 40615951]
46. Ockerman A, Vanassche T, Garip M, Vandenbriele C, Engelen MM, Martens J, et al. Tranexamic acid for the prevention and treatment of bleeding in surgery, trauma and bleeding disorders: a narrative review. Thromb J 2021 Aug 11;19(1):54 [FREE Full text] [doi: 10.1186/s12959-021-00303-9] [Medline: 34380507]
47. Gibbs VN, Champaneria R, Sandercock J, Welton NJ, Geneen LJ, Brunskill SJ, et al. Pharmacological interventions for the prevention of bleeding in people undergoing elective hip or knee surgery: a systematic review and network meta-analysis. Cochrane Database Syst Rev 2024 Jan 16;1(1):CD013295 [FREE Full text] [doi: 10.1002/14651858.CD013295.pub2] [Medline: 38226724]
48. Huang Y, Steptoe A, Wei L, Zaninotto P. The impact of high-risk medications on mortality risk among older adults with polypharmacy: evidence from the English Longitudinal Study of Ageing. BMC Med 2021 Dec 16;19(1):321 [FREE Full text] [doi: 10.1186/s12916-021-02192-1] [Medline: 34911547]
49. Pereira LB, Gonçalves AMRF, Fernandes CSE, Fontanella AT, Francisco PMSB, Mengue SS, et al. Use of drugs for gastrointestinal disorders: evidence from National Survey on Access, Use and Promotion of Rational Use of Medicines. Einstein (Sao Paulo) 2020;18:eAO5314 [FREE Full text] [doi: 10.31744/einstein_journal/2020ao5314] [Medline: 32785451]
50. Korean Statistical Information Service. Main surgery statistical yearbook. Statistical Database. 2016. URL: https://kosis.kr/eng/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ETITLE&parmTabId=M_01_01&parentId=F.1;F_10.2;#content-group [accessed 2025-12-24]
51. Al-Qurain AA, Gebremichael LG, Khan MS, Williams DB, Mackenzie L, Phillips C, et al. Prevalence and factors associated with analgesic prescribing in poly-medicated elderly patients. Drugs Aging 2020 Apr;37(4):291-300. [doi: 10.1007/s40266-019-00742-0] [Medline: 32016823]
52. Al-Qurain AA, Gebremichael LG, Khan MS, Williams DB, Mackenzie L, Phillips C, et al. Opioid prescribing and risk of drug-opioid interactions in older discharged patients with polypharmacy in Australia. Int J Clin Pharm 2021 Apr;43(2):365-374. [doi: 10.1007/s11096-020-01191-1] [Medline: 33206289]
53. Editage. URL: https://www.editage.co.kr/ [accessed 2024-04-10]

Abbreviations

ADE: adverse drug event
ATC: Anatomical Therapeutic Chemical
CDM: Common Data Model
EDI: electronic data interchange
EHR: electronic health record
OMOP: Observational Medical Outcomes Partnership
OS: operating system
WHO: World Health Organization

Edited by A Benis; submitted 01.Jun.2025; peer-reviewed by H Demir, YS Lo; comments to author 06.Oct.2025; revised version received 28.Nov.2025; accepted 17.Dec.2025; published 29.Jan.2026.

Please cite as:
Lee CC, Kim GJ, Kim S, Hong JY, Hwang WM, Kim JY, Lee KH, Kim K, Kang M, Kim JH, Lee S
Multi-Institutional Drug Use Patterns in Hospitalized Older Patients: Retrospective Cross-Sectional Study
JMIR Med Inform 2026;14:e78353
URL: https://medinform.jmir.org/2026/1/e78353
doi:10.2196/78353
PMID:

©Chung Chun Lee, Grace Juyun Kim, Suhyun Kim, Jee Young Hong, Won Min Hwang, Jong-Yeup Kim, Kye Hwa Lee, Kwangsoo Kim, Mingyu Kang, Ju Han Kim, Suehyun Lee. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Harmonizing Logical Observation Identifiers Names and Codes (LOINC) Codes and Units in Real-World Oncology Data: Method Development and Evaluation

Parvati Naliyatthaliyazchayil[1], PharmD, MS; Travis Stenerson[1], MHI, MD

ConcertAI, LLC, Cambridge, MA, United States

Corresponding Author:
Parvati Naliyatthaliyazchayil, PharmD, MS
ConcertAI, LLC
1120 Massachusetts Ave.
Cambridge, MA, 02138
United States
Phone: 1 317 985 7429
Email: parumenon.pm@gmail.com

Abstract

Background: The expanding use of multisource real-world electronic health record (EHR) and claims data offers major opportunities for research, drug discovery, and clinical decision support. While standards such as Logical Observation Identifiers Names and Codes (LOINC) can ensure semantic interoperability for laboratory observations, clinical documents, and other clinical terms, properly assigning these concepts remains a challenge. Studies show that 6% to 19% of laboratory tests cannot be accurately mapped to LOINC. Existing systems try to address this challenge but often depend on source data strings and other input features that may be absent, null, or incorrect. This underscores the need for a scalable approach to correct LOINC code assignments, standardize units, and ensure data integrity across multisource laboratory data.

Objective: This paper presents a universally applicable framework that identifies and corrects observable errors in quantitative laboratory results coded to LOINC and the Systematized Nomenclature of Medicine for the unit of measure without relying on raw source data strings. The process seeks to improve the accuracy, conformance, consistency, and completeness of laboratory data while maintaining complete provenance.

Methods: The proposed framework uses a 2-step process. First, LOINC codes are corrected using the associated unit of measure. Second, units are adjusted or populated to match a preferred unit for that LOINC code. In both steps, the quantitative result is checked against a predefined acceptable range to determine validity. The process is driven by 3 knowledge tables. The framework is applied to datasets derived from the ConcertAI database of approximately 10 million patients with cancer, evaluating improvements in LOINC code–unit conformance and unit completeness. Analyses are performed on 4 independently LOINC-coded datasets: the full ConcertAI dataset and 3 high-volume diverse subsets grouped by data source or EHR vendor.

Results: A total of 428 LOINC codes were observed across 6.34 billion records in the ConcertAI database. All 4 datasets were processed using the proposed framework. Before applying the framework, 73.1% (4,634,610,173/6,337,101,453) of records in the ConcertAI dataset had correctly assigned units based on the laboratory reasonable range table; after application, this increased to 99.7% (6,322,375,200/6,341,230,213). Similar improvements were observed across the 3 EHR-specific datasets, increasing from 78.5% (691,315,390/880,250,137) to 99.8% (879,626,472/881,157,852; source 1), 71.4% (2,132,455,936/2,985,465,124) to 99.8% (2,982,319,644/2,988,173,959; source 2), and 63.3% (2,936,710,502/4,640,432,294) to 99.6% (4,618,714,114/4,638,862,412; source 3). Unit completeness also improved substantially, increasing from 92.7% (5,879,071,858/6,341,230,213) to 99.8% (6,331,923,060/6,341,230,213) in the ConcertAI dataset and from 92.5% (814,867,241/881,157,852) to 100% (880,816,133/881,157,852), 94.4% (2,822,107,252/2,988,173,959) to 99.9% (2,986,624,027/2,988,173,959), and 91.7% (4,254,054,966/4,638,862,412) to 99.8% (4,632,935,919/4,638,862,412) in sources 1 to 3, respectively.

Conclusions: Laboratory data quality is crucial in oncology systems for therapy selection, monitoring, and disease progression assessment. This proposed solution is a first-of-its-kind, system-agnostic, and scalable normalization process that addresses key gaps in laboratory data quality across multiple dimensions.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e81254) doi:10.2196/81254

KEYWORDS

real-world data; laboratory data; harmonization; standardization; oncology; health care system; electronic health record; multisourced data; structured data; mapping; Fast Healthcare Interoperability Resources; FHIR; Logical Observation Identifiers Names and Codes; LOINC; Systematized Nomenclature of Medicine; SNOMED; unit conversion

Introduction

Background: Real-World Data in Health Care

Multisourced, comprehensive patient health data have seen a recent rise in prominence and utility in supporting research, drug discovery, pharmacovigilance, and clinical decision support [1]. These real-world data (RWD) are collected for individual patient care and reimbursement in often unconnected systems such as electronic health records (EHRs) and claims clearing houses [2]. Secondary use of RWD holds immense potential for improving health due to its volume, variety, and ubiquity [3].

The different source systems producing RWD bring individual challenges and opportunities, and it is often necessary to stitch together disparate sources to cover a complete patient journey [3]. For instance, medical claims may contain procedures and tests performed on a patient and their associated costs, but the results of those tests are only present in the EHR [4].

Data Quality and Semantic Interoperability Challenges

The path to a usable RWD dataset is fraught with challenges [2]. Errors can arise from any system or actor that generates, transmits, transforms, or persists the data during its lifetime [5]. Errors can be discernible from patterns observed in the data or can be hidden from downstream users when information is lost [6]. Information loss can result in incorrectness or missingness of variables, records, tables, or complete data sources [7]. Monitoring and minimizing information loss is a critical requirement of an RWD pipeline.

Semantic interoperability (SI), or the ability to ensure the same meaning and interpretation of data shared between systems [8], is a well-established and ongoing challenge in RWD [5]. It can manifest as redundant representations of the same concept, such as one EHR using the *International Classification of Diseases, 10th Revision, Clinical Modification*, and another using the Systematized Nomenclature of Medicine (SNOMED) to represent the same disease. Use of data standards and common ontologies can alleviate SI challenges but can also introduce errors if used incorrectly [5], which can commonly occur when a concept is assigned to a string by either a human or machine [9]. As health care data systems continue to evolve, the importance of SI is underscored by the increasing adoption of standards such as Fast Healthcare Interoperability Resources (FHIR) [10] with its associated code system bindings supported by the Food and Drug Administration, which are designed to ensure consistent data interpretation and improve interoperability across systems. Even FHIR-based representations may fall short in fully addressing semantic mismatches, particularly for complex domains such as laboratory data, in which variability in coding, units, and local conventions persists despite standardization efforts [11].

Laboratory Data and Logical Observation Identifiers Names and Codes Mapping Challenges

Laboratory result data are both valuable and particularly challenging in RWD [9]. To address SI, laboratory data primarily rely on codes from the Logical Observation Identifiers Names and Codes (LOINC) code system [12,13]. LOINC is the dominant global coding system for laboratory concepts but also extends to a broader range of clinical terms, including clinical questions and documents [12-15]. LOINC consists of concepts that have 6 attributes each. These include the analyte tested, how it is observed, the duration of observation, and the sample type that the observation is performed on [13]. Precoordinating that much information into a single concept makes it convenient for transmission by reducing the potential for misinterpretation but also makes it difficult to use when mapping to LOINC [14,15]. The complete information for LOINC mapping is rarely present in a single string, column, or record [15]. Relevant information may be spread across one or more fields in a record or laboratory information system, such as the observation's display string (eg, glucose), the data type of the result (numeric or categorical), the unit of measure (mass over volume or molarity), the associated order name or panel name (complete blood count or urinalysis), the associated sample (urine or blood), the device or equipment used to make the observation (urine dipstick or continuous glucose monitor), or other observations that occurred as part of the same panel of tests (Is the patient fasting?). Despite the abundance of resources, accurately translating these fields to the correct LOINC code remains a challenge.

This semantic and translational complexity means that laboratory results in RWD can frequently present with errors introduced during terminology binding to LOINC [16]. Prior studies have shown that LOINC codes could not be assigned for 6% to 19% of laboratory tests due to incomplete or incorrect information [17]. Systems have been proposed to reduce this error by coding via automated means in RWD [9,16,18] and clinical trial data [19]. These algorithms rely on source data strings and other input features that may be absent, null, or observably incorrect.

Unit of Measure and Normalization Challenges

The diversity of information required for LOINC mapping increases the likelihood of mapping difficulty [20] and significant information loss. If a result's unit of measure is absent or recorded in the wrong place or the associated panel name, order name, or sample type is missing from the data, it can mean insufficient information to properly assign the LOINC code [15]. Contextually appropriate filling in of absent units of measure has been shown to be effective in a system that extracts logical expressions from clinical trial inclusion criteria pertaining to quantitative laboratory results but not within RWD [21].

XSL•FO
RenderX

A laboratory result may present with a correct concept code, but it is possible to represent the same result using multiple scales (g or mg), and certain units are used interchangeably between sources (mg per day or mg per 24 hours). Unit normalization is necessary to ensure comparable results during data analysis and aggregation, and the variety of representation, the missingness, and the level of incorrectness can all lead to loss of utility of RWD. Even with a correct LOINC code, missing or inconsistent units have been shown to affect as much as 14% of laboratory records in some systems [22,23]. Mechanisms to normalize synonymous units exist, but these techniques do not handle situations in which the unit is incomplete, incorrect, or completely absent [24].

Proposed Framework and Objective: Correcting Quantitative Laboratory Results

In this paper, we present a framework to identify and correct observable errors in quantitative laboratory results that have already been coded to LOINC for the observation and SNOMED for the unit of measure without reliance on source data strings. This 2-step process first adjusts the LOINC code by using information contained in the associated unit of measure. Next, we use the quantitative result to inform whether we can safely populate a missing unit of measure or correct an erroneous unit of measure. Finally, all synonymous or related units are normalized to a single preselected unit for each LOINC code. We characterize the prevalence of these errors and then use this framework to correct laboratory results in a multisource RWD pipeline of oncology patient data from EHRs, laboratory information systems, and medical claims.

Methods

This section outlines the data collection and study design used to develop this framework.

Data Collection

Laboratory result data were extracted from the ConcertAI database, a US-based, deidentified, patient-level dataset from approximately 10 million patients with cancer aged ≥18 years during the period from 2015 to 2025. The data used in this study were sourced from EHRs, including oncology practices, hospitals, and academic medical centers. To evaluate the framework, we analyzed 4 laboratory datasets: the full ConcertAI database and 3 subsets corresponding to the individual EHR sources (represented as source 1, source 2, and source 3) that transmitted data to ConcertAI. These 3 subsets were selected because they were the top contributors of laboratory data by volume and represented a broad diversity of laboratory tests. Specifically, the ConcertAI database accounted for 6.34 billion records, source 1 included 880 million records, source 2 included 2.98 billion records, and source 3 included 4.64 billion records.

Data were contributed by multiple clinical sites and laboratories across the United States, representing a diverse group of care settings. Each dataset was standardized using LOINC codes for laboratory observations and SNOMED codes for units of measure. For records from the full ConcertAI database, LOINC codification was performed using ConcertAI's internal system; for EHR or data source–derived subsets, the LOINC codes that were delivered with those data were used. Only records with a numeric value were used in this study. All datasets used ConcertAI's unit codification that maps unit strings to SNOMED.

System Design

The laboratory correction framework was applied after standardization to LOINC and SNOMED. The laboratory correction system relies on 3 knowledge tables that have been manually created after careful examination of the ConcertAI laboratory data, as outlined below. Their role in supporting the logical framework is discussed later in this section. Below, we provide the formal definitions of these tables, including their constituent fields and primary keys. The procedures used to create each table and the column definitions are documented in Multimedia Appendix 1.

The LOINC conversion map (LCM) relates a LOINC concept and unit concept that are incongruous to a new LOINC code that is compatible with the unit of measure. The table structure incorporates the following computable fields: "old_loinc_code," "old_loinc_display," "old_loinc_system," "unit_code," "unit_display," "unit_system," "new_loinc_code," "new_loinc_display," and "new_loinc_system." The primary key for the table is defined as the composite of "old_loinc_code" and "unit_code." Examples are shown in Table 1, which has been formatted for readability. The complete computable version is available in Multimedia Appendix 1 and in the accompanying GitHub repository to support transparency and reproducibility.

Table 1. Logical Observation Identifiers Names and Codes (LOINC) conversion map.

LOINC code	Unit (SNOMED[a])	Corrected LOINC code
Platelets (#/volume) in blood (26515-7)	fL (258775009)	Platelet (entitic mean volume) in blood (28542-9)
Glucose (mass/volume) in blood (26515-7)	mEq/L (258865000)	Glucose (moles/volume) in blood (15074-8)
Glucose (mass/volume) in blood (26515-7)	mmol/L (258813002)	Glucose (moles/volume) in blood (15074-8)
Kappa light chains (mass/volume) in serum or plasma (11050-2)	Percentage (118582008)	Kappa lymphocytes/lymphocytes in blood (17096-9)

[a]SNOMED: Systematized Nomenclature of Medicine.

The reasonable range map defines metadata about a single LOINC code, including its assigned correct unit of measure, a minimum and maximum value that the laboratory test can reasonably take, and the mean and median of records with that LOINC code and unit across all 4 datasets. The correct unit of measure for each LOINC code was determined using the

XSL•FO
RenderX

"example_unit" supplied in the official LOINC distribution. These units were mapped to the corresponding SNOMED codes. For LOINC codes that listed more than one example unit, we examined the empirical distribution of units observed in our dataset and selected the unit with the highest frequency. This made the choice of the correct unit data driven and ensured that the process could be consistently reproduced in any system. If Unified Code for Units of Measure (UCUM) units are preferred, LOINC also provides "example_ucum_units," which can be incorporated similarly. Minimum and maximum reasonable values for each LOINC code were defined by estimating the empirical distribution of the data through quantiles of 0.005, 0.025, 0.16, 0.5 (median), 0.84, 0.975, and 0.995. The range was set at or near the highest and lowest of those quantiles, providing broad yet data-anchored thresholds for plausibility checking.

This knowledge table includes the following computable fields: "loinc_code, loinc_name," "unit_code," "unit_name," "min_reasonable," "max_reasonable," "mean," and "median." The primary key for this table is the "loinc_code." A readability-optimized version is shown in Table 2, whereas the complete computable version is provided in Multimedia Appendix 1 and the GitHub repository.

Table 2. Reasonable range map.

LOINC[a] code	Unit (SNOMED[b])	Reasonable range	Median	Mean
Platelets (#/volume) in blood (26515-7)	×10(3)/mcL (1287856009)	1-800	223.5	215.0
Glucose (mass/volume) in blood (2339-0)	mg/dL (258797006)	40-400	138.8	121
Kappa light chains (mass/volume) in serum or plasma (11050-2)	mg/L (258796002)	0.1-5000	68.7	25.9

[a]LOINC: Logical Observation Identifiers Names and Codes.

[b]SNOMED: Systematized Nomenclature of Medicine.

The unit multiplier map relates a correct unit to a synonymous, convertible, or incorrect unit along with a multiplier value, an additive scalar value, and a description of the relationship between the 2 units (eg, incorrect, synonym, or convertible). The scalar value is only used for temperature conversions. An empty "incorrect unit" represents a null unit field and is used to populate absent units. This knowledge table includes the following computable fields: "incorrect_unit_code," "incorrect_unit_name," "correct_unit_code," "correct_unit_name," "multiplier," "scalar_constant," and "multiplier_type." The primary key is the composite of "incorrect_unit_code," "correct_unit_code," and "multiplier_type." A readability-optimized version is shown in Table 3, with the full computable version available in Multimedia Appendix 1 and the GitHub repository.

Table 3. Unit multiplier map.

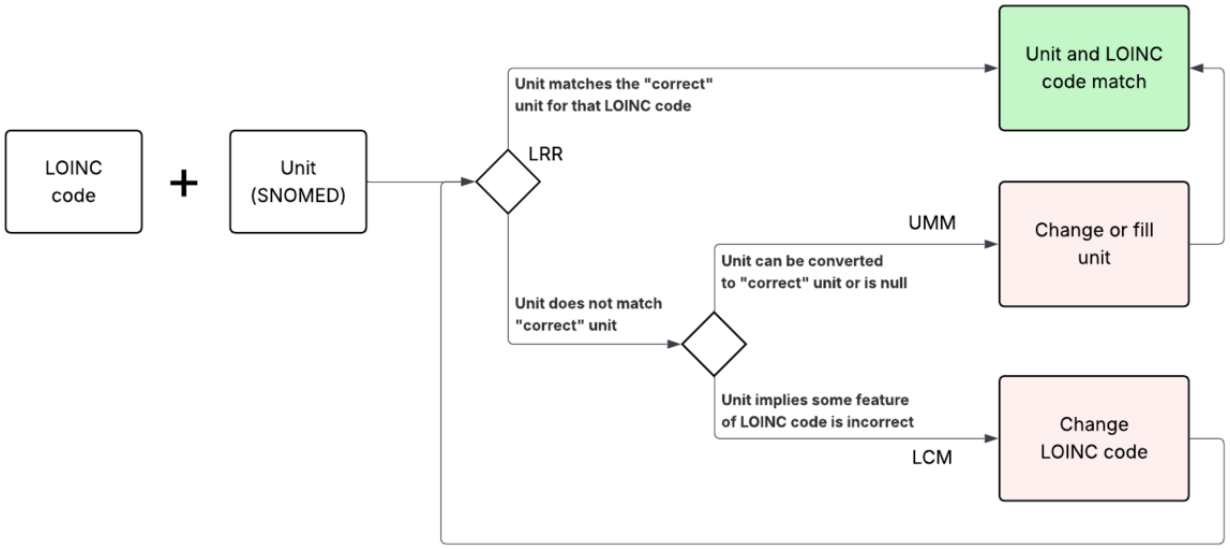
Incorrect unit (SNOMED[a])	Correct unit (SNOMED)	Multiplier	Scalar	Multiplier type (description)
<null>	×10(3)/mcL (1287856009)	1	0	"nullfill" (unit absent)
10*9/L (277288007)	×10(3)/mcL (1287856009)	1	0	"synonym"
cells/µL (258878000)	×10(3)/mcL (1287856009)	0.001	0	"unit_conversion" (convertible unit)
cells/µL (258878000)	×10(3)/mcL (1287856009)	1	0	"unit_typo" (known unit error)
g/dL (258795003)	mg/dL (258797006)	1000	0	"unit_conversion" (convertible unit)
g/dL (258795003)	mg/dL (258797006)	1	0	"unit_typo" (known unit error)
g/L (258794004)	mg/dL (258797006)	1	0	"unit_conversion" (convertible unit)

[a]SNOMED: Systematized Nomenclature of Medicine.

The framework logic is shown in Figure 1. Conformance to the selected "correct" unit is checked against the laboratory reasonable range (LRR) table entry. If the unit does not conform, the LCM provides a better LOINC code if one is known for the unit that is present in the record. In the second transformation, the unit is altered according to the unit multiplier map, normalized if it is a synonym, converted using a multiplier value, or swapped for the correct unit if it is a known error pattern or absent altogether. For example, the LOINC code 1920-8 (glucose [mass/volume] in blood) has the correct unit "mg/dL." If a record with this LOINC code has the unit "mmol/L," the LOINC code is changed using the LCM to 15074-8 (glucose [moles/volume] in blood). The process is only 2 steps, although not all records require both. Records may begin with LOINC code–unit conformance and skip the process altogether. Some require only LOINC conversion, some require only unit normalization, and many require LOINC conversion and then unit normalization.

XSL•FO
RenderX

Figure 1. Framework logic and knowledge table use. LCM: Logical Observation Identifiers Names and Codes (LOINC) conversion map; LRR: laboratory reasonable range; SNOMED: Systematized Nomenclature of Medicine; UMM: unit multiplier map.



Each of these transformations only proceeds when the record's result value falls within the reasonable range associated with the LOINC code–unit pair that would appear in the transformed record. These ranges were selected using known correct data and designed to capture most possible values in that laboratory's value distribution.

In rare instances, the system must select between 2 possible transformations for the same record. This might occur in the situation visible in rows 5 and 6 in Table 3 when trying to convert from g/dL to mg/dL. We identified a common error in many records in which a unit's prefix was absent. A g/dL unit might be a simple data entry error at the source, or it may, in fact, require conversion by multiplying the result by 1000. The system uses the reasonable ranges to determine which of these 2 situations is more likely. If the reasonable ranges are not disparate enough to discern which transformation is required, proximity to the median of the target LOINC code–unit's distribution is compared. The transformation resulting in a record closer to that median is selected. If the compared laboratory results have medians of 0, distance from the mean is used.

To evaluate the success of each transformation, the value distribution of the output for any individual transformation was compared with the value distribution of known correct data with that LOINC code–unit pair. If the distribution of the numeric result of the transformed data matched that of known correct data, it was considered successful. If the distribution did not match, the discrepancy was investigated. These investigations led to the discovery of common error patterns that could then be accounted for in the next iteration by adding to the 3 knowledge tables.

To summarize, the precedence of the rules described above is as follows:

1. LRR conformance check—if the LOINC code–unit pair conforms to the LRR table, no transformation is applied.
2. LOINC conversion (LCM)—if the unit is incompatible with the LOINC code, the framework first selects the appropriate LOINC code.
3. Unit normalization or conversion (unit multiplier map)—after the LOINC code is finalized, unit normalization or conversion is applied.
4. Range validation—a transformation is only accepted if the resulting value falls within the target LOINC code's reasonable range.
5. Conflict resolution—when more than one unit transformation falls within the reasonable range, the system resolves this by checking proximity to the median of the known correct distribution or proximity to the mean when the distribution median is 0.

A few examples of LOINC code–unit incongruencies and framework processing are outlined in Multimedia Appendix 1.

Provenance is maintained within each transformed record in an attribute that acts as a ledger of each transformation that alters that record. This attribute is an array of objects that includes a descriptive string that details what features of the record were adjusted and what features were used to justify that adjustment for human readability. For computability, "object" includes the adjusted attribute names and preadjustment values for those attributes, ensuring transparency to downstream users and reversibility if necessary.

A summary of the steps for this framework is illustrated in Figure 2.

Figure 3 shows an example of LOINC code–unit pairs before transformation, each step of the transformation, and the final output after transformation.

XSL•FO
RenderX

Figure 2. Summary of the framework. LOINC: Logical Observation Identifiers Names and Codes; SNOMED: Systematized Nomenclature of Medicine.

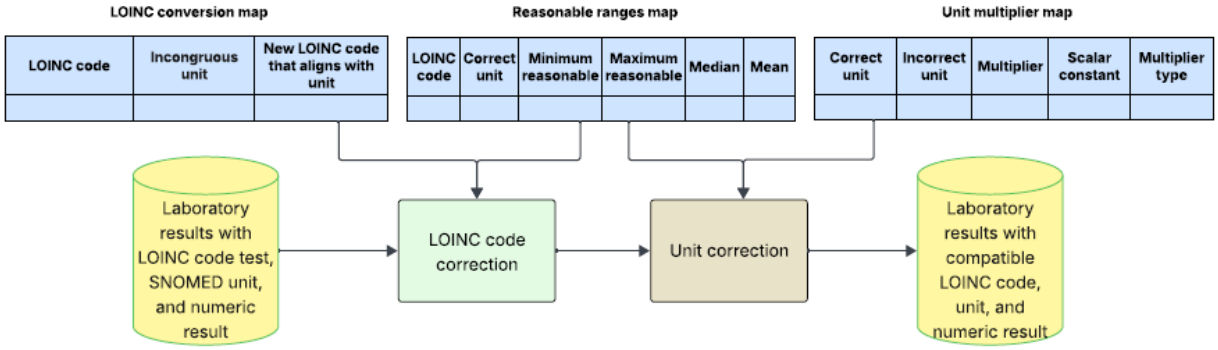
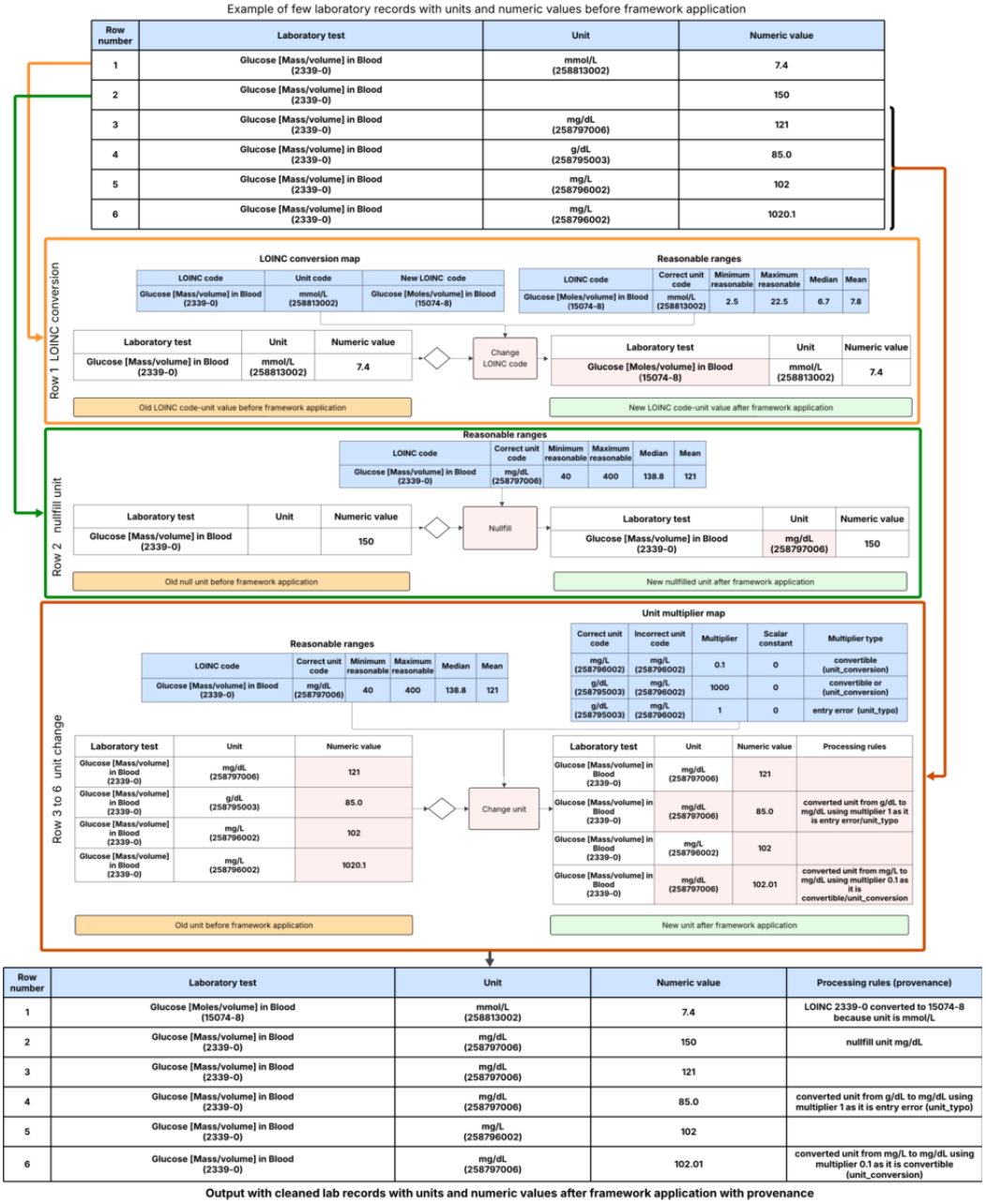


Figure 3. Example of Logical Observation Identifiers Names and Codes (LOINC) code–unit pairs before and after transformation.



XSL•FO
RenderX

Ethical Considerations

This study used retrospective, deidentified data collected for clinical purposes. All data were fully anonymized before analysis, and unique patient identifiers were removed. Analysis of deidentified secondary data does not constitute human subjects research and does not require formal review or approval by an institutional review board, and thus, no institutional review board approval number was sought. Because only deidentified data were used, informed consent from individual patients was not required. Privacy and confidentiality were maintained throughout the study in accordance with all applicable regulations.

Results

Dataset Collection: Overview and LOINC Coverage

All 4 sets of laboratory data underwent the LOINC code and unit correction process, and the proportion of records exhibiting each LOINC code–unit pattern was calculated both before and after the framework execution. A total of 428 LOINC codes were included in this study, chosen based on their prevalence in the ConcertAI database. Not all datasets contained all 428 LOINC codes. These LOINC codes accounted for 6.34 billion records in the ConcertAI database and 880 million, 2.98 billion, and 4.64 billion records in the source-specific datasets from sources 1, 2, and 3, respectively.

System Design

Impact of Framework on Unit Correction

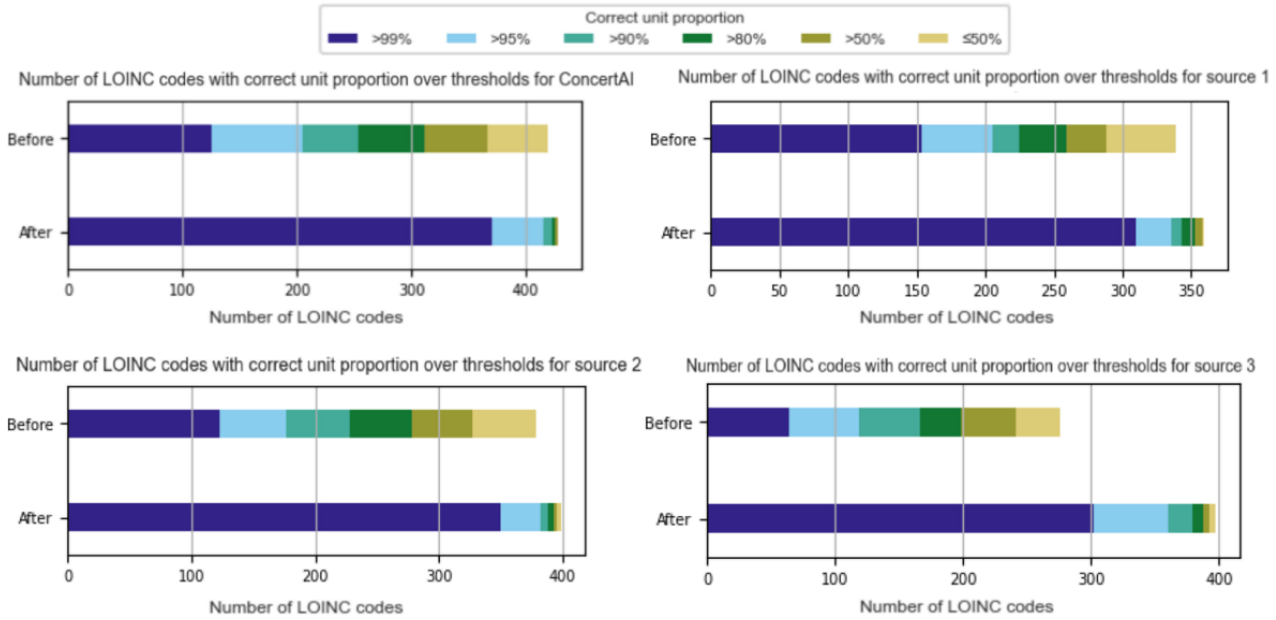
Before the application of the framework, 73.1% (4,634,610,173/6,337,101,453) records in the ConcertAI dataset with these LOINC codes had the correct unit as assigned by the LRR table. Following the application, this increased to 99.7% (6,322,375,200/6,341,230,213) of records. A similar trend was observed in each of the EHR-specific datasets: 78.5% (691,315,390/880,250,137) to 99.8% (879,626,472/881,157,852), 71.4% (2,132,455,936/2,985,465,124) to 99.8% (2,982,319,644/2,988,173,959), and 63.3% (2,936,710,502/4,640,432,294) to 99.6% (4,618,714,114/4,638,862,412) for sources 1, 2, and 3, respectively. The system increased the completion rate of units, defined as the proportion of records with nonmissing units, from 92.7% (5,879,071,858/6,341,230,213) to 99.8% (6,331,923,060/6,341,230,213) in the ConcertAI dataset and similarly in the other 3 datasets: 92.5% (814,867,241/881,157,852) to 100.0% (880,816,133/881,157,852), 94.4% (2,822,107,252/2,988,173,959) to 99.9% (2,986,624,027/2,988,173,959), and 91.7% (4,254,054,966/4,638,862,412) to 99.9% (4,632,935,919/4,638,862,412) for sources 1, 2, and 3, respectively. Because the framework assigns validated units to previously missing values, improvements in the correct unit rate reflect both corrections to misassigned units and the assignment of correct units to previously missing entries. As a result, the correct unit and completion rate metrics are related but not redundant: the correct unit rate captures accuracy among all populated units, whereas the completion rate captures overall presence of any unit.

LOINC-Level Quality Threshold

Not all LOINC codes start with the same degree of incorrectness. To evaluate the proportion of LOINC codes that reached specific data quality thresholds before and after framework application, records with each LOINC code were evaluated for unit correctness, and the number of LOINC codes passing certain unit correctness proportion thresholds was counted. These counts for each dataset are shown in Figure 4. Of the 428 LOINC codes, the number meeting the 99% threshold of unit correctness in the full ConcertAI dataset increased from 125 (29.2%) to 370 (86.4%). Similar improvements were observed across the 3 source datasets, which contained different total proportions of evaluated LOINC codes: 83.9% (359/428) of LOINC codes in source 1, a total of 93.2% (399/428) of LOINC codes in source 2, and 88.6% (379/428) of LOINC codes in source 3. In source 1, the proportion meeting the 99% threshold increased from 42.9% (154/359) to 86.4% (310/359); in source 2, it increased from 30.8% (123/399) to 87.7% (350/399); and in source 3, it increased from 16.9% (64/379) to 79.7% (302/379).

XSL•FO
RenderX

Figure 4. Number of Logical Observation Identifiers Names and Codes (LOINC)–coded records meeting unit correctness thresholds.



Illustrative Subset Analysis

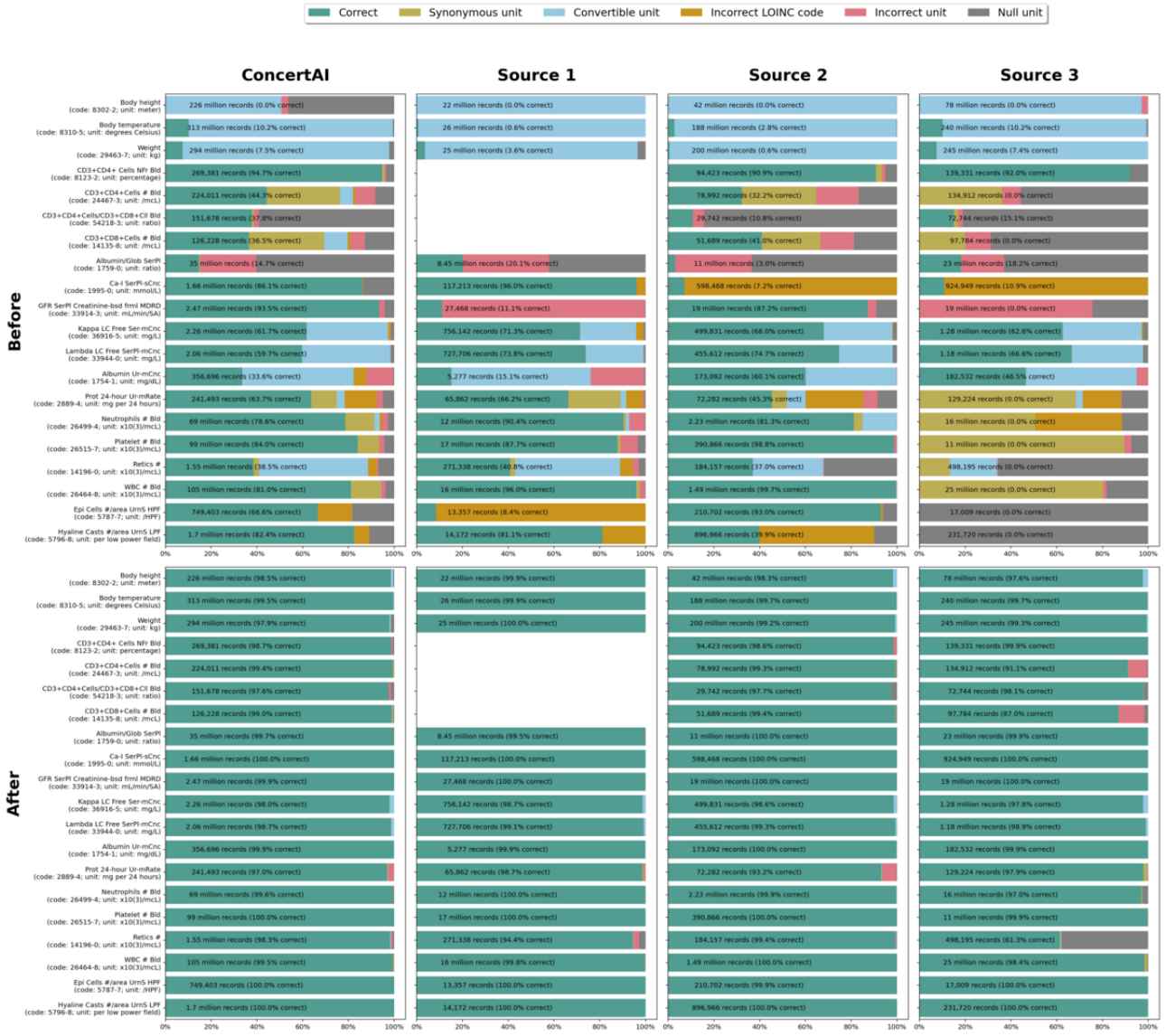
To provide a more detailed view of the framework's impact, Figure 5 presents a closer look at 20 specific LOINC codes. These codes were selected for their varied domains and types of incongruencies in the starting data.

Records with each of these 20 LOINC codes had their unit classified as correct, synonymous to correct, absent, or incorrect. Incorrect units were subclassified as incorrect and suggestive of a LOINC change, incorrect but convertible to the correct unit through a multiplier, or incorrect but not convertible. These unit categories represent the types of errors that this system is designed to correct. The proportions in each category were calculated before and after the framework application and are shown in Figure 5 for each system.

A detailed report showing the proportion of correctness for each LOINC code across all 4 datasets before and after the framework application is available in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 5. Error type proportions for specific Logical Observation Identifiers Names and Codes (LOINC) codes. GFR: glomerular filtration rate; HPF: high power field; MDRD: Modification of Diet in Renal Disease; SA: surface area; WBC: white blood cell.



Discussion

Principal Findings

The proposed normalization framework significantly improved LOINC code–unit congruence and unit absentia across all evaluated datasets, demonstrating generalizability to any clinical data system using standard medical terminologies for observation and unit representation. The method improved LOINC code–unit congruence overall in the ConcertAI laboratory dataset from 73.1% (4,634,610,173/6,337,101,453) to 99.7% (6,322,375,200/6,341,230,213), with similar trends in the other 3 data sources. The method and knowledge tables were developed using the ConcertAI data and applied to these independent datasets to demonstrate generalizability. To the best of our knowledge, this framework is the first of its kind to address 2 key gaps: the absence of a system-agnostic process for cleaning multisource laboratory data for secondary use [18] and an approach tailored to RWD. RWD, where data quality is both challenging [25] and critical to data usability, may lack source-specific semantics such as the local observation concept display name. This would impede the use of a standardization process designed to take advantage of those strings [18].

Mechanism of Framework

The process, driven by 3 small, shareable knowledge tables, uses the observation concept (LOINC), the unit concept (SNOMED), and the numeric result as input. It first adjusts the LOINC code if the unit suggests a better LOINC code. It then normalizes the unit to a predetermined preferred unit for that LOINC code by filling null units, converting related units, or correcting units with overt errors. All transformations only proceed when the numeric result aligns with the predetermined distribution for the target LOINC code–unit pair. This “reasonable range” constraint addresses a shortcoming in existing proposed methodologies [19,21] that handle units without consideration of the numeric result. This range can be used to flag records that fall outside the reasonable distribution for exclusion from analysis or closer review.

XSL•FO
RenderX

Impact on Data Quality Dimensions

Ensuring that laboratory values fall within a reasonable distribution contributes to accuracy, or the degree to which the data represent the true value of what is intended to be measured, and plausibility, 2 prominent dimensions [25] of RWD quality. The results demonstrate improvements across a number of established quality dimensions [25] other than accuracy and plausibility, including conformance, consistency, completeness, and provenance [25-29]. The technique primarily addresses conformance to the approved LOINC code–unit pair set by the LRR table and consistency of semantic representation regardless of data source. Completeness is improved by populating the unit when null. This contextually appropriate population of absent units has not been addressed in prior studies of similar techniques [19,21].

Data Provenance

Data provenance is recognized as critical metadata for systems transmitting, storing, or using health data [30]. Records in RWD pipelines may undergo many transformations on their path to integration into a single dataset, and these transformations must all be tracked to enable error correction and improve trustworthiness with downstream consumers. Data provenance has been defined in many ways [31], and there are established specifications for recording provenance in health care data, such as the FHIR Provenance resource [32] based on the World Wide Web Consortium provenance specification [33]. Establishing the best method for provenance tracking within a system requires an assessment of the stakeholders for those data. The stakeholders for the transformations described in this paper were internal data analysts, informaticists, and quality assurance engineers. Provenance is maintained within each transformed record in an attribute that acts as a ledger of each transformation that altered that record. This attribute is an array of structures that include a descriptive, human-readable string detailing the transformation, as well as the altered attributes and preadjustment values that those attributes took. By consulting this array, transformations are fully reversible, original values are preserved, and a complete audit trail is available to support transparency and verification by downstream users.

Clinical and Research Relevance

Laboratory data quality and usability are critical in systems handling oncology data, where these results are used to select and monitor therapies and assess disease severity and progression [34]. LOINC code–unit incongruence, unit absentia, and lack of unit standardization put an undue burden on any analyst or application attempting to use laboratory data for clinical decision support, retrospective studies, or clinical trial eligibility [35,36]. Improving unit completeness at the source is challenging because units are often inconsistently documented or transmitted in EHR systems [37,38]. Potential upstream improvements could include enforcing unit capture at the point of entry, not allowing free-text entry of units, ensuring consistent transmission in Health Level Seven and FHIR messages, or applying unit standardization at the health system level before data export. However, such approaches depend on local workflows and vendor configurations and cannot be guaranteed for multisource RWD. Consequently, postcodification harmonization frameworks such as the one described in this paper provide a practical and reproducible solution. The technique outlined in this paper presents a minimal set of logical steps to correct these problems within the pipeline of a multisourced data system handling RWD. Enforcing semantic consistency between LOINC and unit improves data usability [24] but can also foster better data exchange [39] between systems by reducing semantic differences. Ultimately, this framework ensures that laboratory data are both accurate and complete, enabling faster data-driven decisions that can enhance patient care, data exchange, clinical trials, and RWD analysis.

Limitations and Future Work

The technique is presently limited by its capacity to correct the body system, which will reduce effectiveness for laboratory tests in which the potential body systems are relatively evenly distributed and share similar units when tested in either sample type. Information related to the body system can be present in the observation name but also in the panel name, the order name, discretely in another attribute in the laboratory result, or in other observations related to the result in question. Efforts are currently underway to evolve the process by incorporating panel or order-related information. The choice of SNOMED for unit may have been less ideal than UCUM [40,41], which is more flexible and more closely aligned with LOINC [42], which suggests a UCUM unit for some of its concepts. The FHIR [43] and Interoperability Standards Platform [44] also suggest that UCUM be used to represent units. However, the framework is easily adaptable by simply replacing the SNOMED codes used in the knowledge tables with UCUM codes without requiring any further changes in script. In addition, the steps described in this paper specifically address quantitative LOINC codes and require modification to be applicable to qualitative scale–type LOINC codes. Another limitation is that the creation of the knowledge tables requires clinical and terminologist expertise to recognize errors and prepare the tables, as well as a large dataset to prepare laboratory value distributions. This framework is designed for secondary use in RWD, where inconsistencies, missing values, and heterogeneous coding practices complicate analysis, and is not intended for integration into direct point-of-care clinical decision-making. Therefore, its use does not pose any direct safety risk to patients. The harmonization of LOINC code–unit combinations and the potential flagging of outlier values can influence downstream analyses such as cohort definitions or outlier detection. It enables data analysts to use laboratory data that would otherwise have required a significant time investment to clean. The primary aim of these corrections is to reduce inconsistencies and facilitate reliable data aggregation for secondary research purposes. Future work will focus on integrating the framework with emerging technologies such as artificial intelligence.

Practical Implementation

To support practical implementation, we have released a starter kit supporting normalization of 146 LOINC codes from blood or serum, covering the “Hematology and cell counts” class and “Chemistry” glucose codes. It has 3 knowledge tables and an accompanying Python script available via a public GitHub repository [45]. The knowledge tables incorporate LOINC and

XSL•FO
RenderX

SNOMED, which are maintained in multiple languages, allowing the tables to be used with datasets in different linguistic contexts. The numeric values in the tables are language independent and do not require conversion, supporting their generalizability to different languages. The Python script executing the rules takes FHIR Observation resources as input, and 6 example resources are provided.

Currently, the framework leverages SNOMED codes for units. However, the design is adaptable: implementers preferring UCUM codes can replace SNOMED codes with UCUM codes in the tables without any modification to the transformation script, allowing for seamless adoption of alternative unit standards. This flexibility ensures that the framework can be applied in diverse institutional or research contexts with minimal overhead.

A key component of the framework is its structured provenance tracking, which records all transformations applied to each laboratory record in a way that allows the original data to be reconstructed if needed. For each modification, information such as the applied rule, details of the transformation, which fields were changed, the original values, and the reference values used in the conversion is systematically captured. Collectively, these elements enable full traceability, providing a clear lineage from the transformed value back to the original measurement. This approach not only supports reproducibility and auditing but also ensures transparency, allowing researchers and data analysts to confidently interpret and validate the standardized data while retaining the ability to recover the source information.

By combining reproducible tables, adaptable unit standards, and robust provenance tracking, the framework provides a practical pathway for implementing consistent laboratory data harmonization in diverse settings while maintaining the transparency and traceability of each transformation.

Conclusions

As the demand for multisourced RWD datasets grows, ensuring data quality and semantic consistency becomes increasingly vital. Persistent challenges regarding SI in laboratory data continue to limit the full potential of these data assets. Common issues include missing or incorrect units and misassigned LOINC codes, all of which can hinder integration and analysis. This framework addresses these challenges by systematically identifying and correcting discrepancies in quantitative laboratory data and has been shown to improve unit accuracy to above 99% for all evaluated datasets. It enables scalable, system-agnostic normalization of laboratory data, addressing critical gaps in several data quality dimensions. By improving laboratory data quality at the foundational level, it strengthens the reliability of RWD for data analysis, insight generation, and utility in software applications.

Acknowledgments

The authors would like to express their sincere gratitude to the ConcertAI reviewers Sheenu Chandwani, Jannette Hanna, Jennifer Rider, and Jocelyn Benson for their constructive feedback and insightful comments, which greatly enhanced the quality of this paper.

Funding

ConcertAI, LLC, sponsored this study and provided financial support for the conduct of this research and preparation of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to proprietary and licensing restrictions but may be available from ConcertAI, LLC, upon reasonable request and subject to applicable data use agreements and approval processes [46].

Conflicts of Interest

PN and TS are employed by ConcertAI, LLC.

Multimedia Appendix 1
Detailed process for building the 3 knowledge tables described in system design, that is, the reasonable range map, unit multiplier map, and Logical Observation Identifiers Names and Codes conversion map.
[DOCX File , 91 KB - medinform_v14i1e81254_app1.docx]

Multimedia Appendix 2
Detailed report on the proportion of correct units before and after application of the framework, stratified by Logical Observation Identifiers Names and Codes code and by data source.
[XLSX File (Microsoft Excel File), 112 KB - medinform_v14i1e81254_app2.xlsx]

References

XSL•FO
RenderX

1. Hao Y, Cheng C, Li J, Li H, Di X, Zeng X, et al. Multimodal integration in health care: development with applications in disease management. J Med Internet Res 2025 Aug 21;27:e76557 [FREE Full text] [doi: 10.2196/76557] [Medline: 40840463]
2. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 05;22(1):287 [FREE Full text] [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
3. Dang A. Real-world evidence: a primer. Pharmaceut Med 2023 Jan 05;37(1):25-36 [FREE Full text] [doi: 10.1007/s40290-022-00456-6] [Medline: 36604368]
4. Wilson J, Bock A. The benefit of using both claims data and electronic medical record data in health care analysis. OPTUM. 2025. URL: https://www.optum.com/content/dam/optum/resources/whitePapers/Benefits-of-using-both-claims-and-EMR-data-in-HC-analysis-WhitePaper-ACS.pdf [accessed 2025-04-29]
5. An D, Lim M, Lee S. Challenges for data quality in the clinical data life cycle: systematic review. J Med Internet Res 2025 Apr 23;27:e60709 [FREE Full text] [doi: 10.2196/60709] [Medline: 40266662]
6. Abedjan Z, Chu X, Deng D, Fernandez RC, Ilyas IF, Ouzzani M, et al. Detecting data errors: where are we and what needs to be done? Proc VLDB Endow 2016 Aug 1;9(12):993-1004. [doi: 10.14778/2994509.2994518]
7. Munappy AR, Bosch J, Olsson HH. Data pipeline management in practice: challenges and opportunities. In: Morisio M, Torchiano M, Jedlitschka A, editors. Product-Focused Software Process Improvement. Cham, Switzerland: Springer; 2020:168-184.
8. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
9. Kelly J, Wang C, Zhang J, Das S, Ren A, Warnekar P. Automated mapping of real-world oncology laboratory data to LOINC. AMIA Annu Symp Proc 2021;2021:611-620 [FREE Full text] [Medline: 35308998]
10. FHIR® - Fast Healthcare Interoperability Resources®. eCQI Resource Center. URL: https://ecqi.healthit.gov/fhir?qt-tabs_fhir=about [accessed 2025-04-30]
11. Vorisek CN, Lehne M, Klopfenstein SA, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
12. Richesson RL, Campion TR, Knosp BM, Hanauer DA. LOINC implementation approaches in academic medical research centers - results from a survey of CTSA sites. J Clin Transl Sci 2025;9(1):e223. [doi: 10.1017/cts.2025.10151] [Medline: 41111944]
13. LOINC. URL: https://loinc.org [accessed 2025-04-29]
14. Drenkhahn C, Ingenerf J. The LOINC content model and its limitations of usage in the laboratory domain. Stud Health Technol Inform 2020 Jun 16;270:437-442. [doi: 10.3233/SHTI200198] [Medline: 32570422]
15. Lin MC, Vreeman DJ, Huff SM. Investigating the semantic interoperability of laboratory data exchanged using LOINC codes in three large institutions. AMIA Annu Symp Proc 2011;2011:805-814 [FREE Full text] [Medline: 22195138]
16. Fidahussein M, Vreeman DJ. A corpus-based approach for automated LOINC mapping. J Am Med Inform Assoc 2014;21(1):64-72 [FREE Full text] [doi: 10.1136/amiajnl-2012-001159] [Medline: 23676247]
17. Khan AN, Griffith SP, Moore C, Russell D, Rosario ACJ, Bertolli J. Standardizing laboratory data by mapping to LOINC. J Am Med Inform Assoc 2006;13(3):353-355 [FREE Full text] [doi: 10.1197/jamia.M1935] [Medline: 16501183]
18. Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. J Am Med Inform Assoc 2018 Oct 01;25(10):1292-1300 [FREE Full text] [doi: 10.1093/jamia/ocy110] [Medline: 30137378]
19. Wu CZ, Wales DL. Laboratory data standardization with SAS. PharmaSUG. 2017. URL: https://pharmasug.org/proceedings/2017/AD/PharmaSUG-2017-AD10.pdf [accessed 2025-04-30]
20. Vreeman DJ, Hook J, Dixon BE. Learning from the crowd while mapping to LOINC. J Am Med Inform Assoc 2015 Nov;22(6):1205-1211 [FREE Full text] [doi: 10.1093/jamia/ocv098] [Medline: 26224334]
21. Hao T, Liu H, Weng C. Valx: a system for extracting and structuring numeric lab test comparison statements from text. Methods Inf Med 2016 May 17;55(3):266-275 [FREE Full text] [doi: 10.3414/ME15-01-0112] [Medline: 26940748]
22. Results final report. AHRQ HCUP. URL: https://hcup-us.ahrq.gov/datainnovations/clinicaldata/3MSummaryResultsReportFinal.jsp [accessed 2025-04-29]
23. Abhyankar S, Demner-Fushman D, McDonald CJ. Standardizing clinical laboratory data for secondary use. J Biomed Inform 2012 Aug;45(4):642-650 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.012] [Medline: 22561944]
24. Hauser RG, Quine DB, Ryder A, Campbell S. Unit conversions between LOINC codes. J Am Med Inform Assoc 2018 Feb 01;25(2):192-196 [FREE Full text] [doi: 10.1093/jamia/ocx056] [Medline: 28637208]
25. Castellanos EH, Wittmershaus BK, Chandwani S. Raising the bar for real-world data in oncology: approaches to quality across multiple dimensions. JCO Clin Cancer Inform 2024 Jan;8:e2300046 [FREE Full text] [doi: 10.1200/CCI.23.00046] [Medline: 38241599]
26. Data quality framework for EU medicines regulation. European Medicines Agency. 2023. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf [accessed 2025-04-30]

XSL•FO
RenderX

27. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. U.S. Food & Drug Administration. 2024 Jul. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory [accessed 2025-04-30]
28. NICE real-world evidence framework. National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/corporate/ecd9/chapter/assessing-data-suitability#assessing-data-suitability [accessed 2025-04-30]
29. Daniel G, Silcox C, Bryan J. Characterizing RWD quality and relevancy for regulatory purposes. Duke-Margolis Center for Health Policy. 2018 Oct 1. URL: https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf [accessed 2025-04-28]
30. Ahmed M, Dar AR, Helfert M, Khan A, Kim J. Data provenance in healthcare: approaches, challenges, and future directions. Sensors (Basel) 2023 Jul 18;23(14):6495 [FREE Full text] [doi: 10.3390/s23146495] [Medline: 37514788]
31. Ram S, Liu J. A new perspective on semantics of data provenance. In: Proceedings of the First International Conference on Semantic Web in Provenance Management. 2009 Presented at: SWPM'09; October 25, 2009; Washington, DC URL: https://ceur-ws.org/Vol-526/InvitedPaper_1.pdf
32. Provenance. HL7 FHIR. URL: https://hl7.org/fhir/provenance.html [accessed 2025-11-30]
33. Groth P, Moreau L. PROV-overview: an overview of the PROV family of documents. W3C Working Group Note. 2013. URL: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/ [accessed 2025-11-30]
34. Cabalar I, Le TH, Silber A, O'Hara M, Abdallah B, Parikh M, et al. The role of blood testing in prevention, diagnosis, and management of chronic diseases: a review. Am J Med Sci 2024 Oct;368(4):274-286 [FREE Full text] [doi: 10.1016/j.amjms.2024.04.009] [Medline: 38636653]
35. Vesper HW, Myers GL, Miller WG. Current practices and challenges in the standardization and harmonization of clinical laboratory tests. Am J Clin Nutr 2016 Sep;104 Suppl 3(Suppl 3):907S-912S [FREE Full text] [doi: 10.3945/ajcn.115.110387] [Medline: 27534625]
36. Muñoz Monjas A, Rubio Ruiz D, Pérez Del Rey D, Palchuk MB. Enhancing real world data interoperability in healthcare: a methodological approach to laboratory unit harmonization. Int J Med Inform 2025 Jan;193:105665. [doi: 10.1016/j.ijmedinf.2024.105665] [Medline: 39500036]
37. Hajia M. Secondary use of laboratory data: potentialities and limitations. Iran J Pathol 2019;14(3):188-192 [FREE Full text] [doi: 10.30699/ijp.2019.95692.1942] [Medline: 31582994]
38. Khela H, Khalil J, Daxon N, Neilson Z, Shahrokhi T, Chung P, et al. Real world challenges in maintaining data integrity in electronic health records in a cancer program. Tech Innov Patient Support Radiat Oncol 2024 Mar;29:100233 [FREE Full text] [doi: 10.1016/j.tipsro.2023.100233] [Medline: 38293266]
39. Zayed AM, Sarikakis I, Delvaux N. Automated standardization and harmonization of laboratory units in large-scale clinical data using open-source R functions. Int J Med Inform 2026 Jan;205:106131 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.106131] [Medline: 41061385]
40. Bietenbeck A, Boeker M, Schulz S. NPU, LOINC, and SNOMED CT: a comparison of terminologies for laboratory results reveals individual advantages and a lack of possibilities to encode interpretive comments. J Lab Med 2018;42(6) [FREE Full text] [doi: 10.1515/labmed-2018-0103]
41. Schadow G, McDonald CJ. The unified code for units of measure. UCUM Organization. 2024. URL: https://ucum.org/ucum [accessed 2025-05-29]
42. Vogl K, Ingenerf J, Kramer J, Chantraine C, Drenkhahn C. LUMA: a mapping assistant for standardizing the units of LOINC-coded laboratory tests. Appl Sci 2022 Jun 08;12(12):5848 [FREE Full text] [doi: 10.3390/app12125848]
43. Datatypes. HL7 FHIR. URL: https://hl7.org/fhir/datatypes.html#Quantity [accessed 2025-05-29]
44. Representing units of measure (for use with numerical references and values). Assistant Secretary for Technology Policy. URL: http://www.healthit.gov/isp/representing-units-measure-use-numerical-references-and-values [accessed 2025-05-26]
45. PrecisionHealthIntelligence. GitHub. URL: https://github.com/PrecisionHealthIntelligence/loinc_unit_harmonization [accessed 2026-03-02]
46. ConcerAI. URL: https://www.concertai.com/contact-us/ [accessed 2026-03-02]

Abbreviations

EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
LCM: Logical Observation Identifiers Names and Codes conversion map
LOINC: Logical Observation Identifiers Names and Codes
LRR: laboratory reasonable range
RWD: real-world data
SI: semantic interoperability
SNOMED: Systematized Nomenclature of Medicine
UCUM: Unified Code for Units of Measure

XSL•FO
RenderX

Edited by A Benis; submitted 25.Jul.2025; peer-reviewed by A Famotire, B Dixon; comments to author 11.Nov.2025; revised version received 23.Jan.2026; accepted 17.Feb.2026; published 09.Mar.2026.

Please cite as:
Naliyatthaliyazchayil P, Stenerson T
Harmonizing Logical Observation Identifiers Names and Codes (LOINC) Codes and Units in Real-World Oncology Data: Method Development and Evaluation
JMIR Med Inform 2026;14:e81254
URL: https://medinform.jmir.org/2026/1/e81254
doi:10.2196/81254
PMID:41802234

©Parvati Naliyatthaliyazchayil, Travis Stenerson. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Complete Blood Count–Derived Inflammation Indices to Predict 3-Year All-Cause Mortality in Patients With Diabetes and Acute Myocardial Infarction in Critical Care: Retrospective Cohort Study With Single-Center External Validation

Tong Zhou[1], MS; Kun Yang[1], PhD; Ying Yang[1], PhD; Song-Mei Liu[1], Prof Dr

Zhongnan Hospital of Wuhan University, Wuhan, China

Corresponding Author:
Song-Mei Liu, Prof Dr
Zhongnan Hospital of Wuhan University
No. 169, Donghu Road, Wuhan City, Hubei Province
Wuhan, 430071
China
Phone: 86 027 678 132 33
Email: smliu@whu.edu.cn

Abstract

Background: Inflammation plays a pivotal role in the progression of diabetes and its cardiovascular complications, particularly acute myocardial infarction (AMI). Patients with AMI often face high mortality and morbidity, making accurate prognosis crucial for clinical decision-making and outcome improvement.

Objective: This study aims to develop and validate a stacked predictive model using inflammation-related indices to predict 3-year all-cause mortality among patients with severe diabetes who experience acute myocardial infarction (AMI), aiming to improve patient prognosis.

Methods: We included 833 patients with severe diabetes combined with AMI from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (training/test cohort) and 166 cases from Zhongnan Hospital, China (external validation cohort with 3-year follow-up). A total of 5 inflammation-related indices (lymphocyte-to-monocyte ratio [LMR], neutrophil-to-lymphocyte ratio [NLR], neutrophil-to-platelet ratio [NPR], platelet-to-lymphocyte ratio [PLR], prognostic inflammatory value [PIV]) were analyzed for their association with mortality using Cox proportional hazards models. Kaplan-Meier curves and restricted cubic spline analysis explored survival probabilities and dose-response relationships. A stacked predictive model incorporating these indices was constructed, and its performance was evaluated using the area under the curve.

Results: LMR was a protective factor (hazard ratio [HR] 0.44, 95% CI 0.25-0.77; *P*<.001), while NLR (HR 1.78, 95% CI 1.19-2.65; *P*=.004) and PIV (HR 1.59, 95% CI 1.09-2.30; *P*=.01) were associated with increased mortality risk. Kaplan-Meier analysis showed mortality increased with decreasing LMR quartiles and increasing NLR, NPR, PLR, and PIV quartiles. Restricted cubic spline confirmed that decreasing LMR and increasing NLR, NPR, PLR, and PIV were associated with higher adverse event risk. The predictive model achieved an area under the curve of 0.803 (95% CI 0.736-0.871) in internal testing and 0.781 (95% CI 0.704-0.858) in the external validation cohort.

Conclusions: The stacked predictive model serves as a robust tool for predicting 3-year outcomes in patients with diabetes combined with AMI. Its reliance on routine, low-cost indices highlights its economic viability and potential for widespread clinical implementation, particularly for optimizing resource allocation in resource-limited settings. However, considering the single-center nature of the derivation data, future multicenter validation is essential to verify the model's generalizability across different health care systems before establishing it as a standard policy for risk stratification.

(JMIR Med Inform 2026;14:e83328) doi:10.2196/83328

KEYWORDS

3-year mortality; acute myocardial infarction; diabetes; inflammation-related indices; machine learning; prognosis

XSL•FO
RenderX

Introduction

Diabetes and its complications represent a growing global health challenge, affecting an increasing number of individuals worldwide. According to the International Diabetes Federation, an estimated 537 million individuals were living with diabetes in 2021, and this number is projected to 1.31 billion by 2050, underscoring the urgent need for effective prevention and management strategies [1].

Type 2 diabetes is characterized by systemic inflammation, driven by complex crosstalk between immune cells and β-cells, which plays a pivotal role in disease progression [2]. Importantly, inflammation not only contributes to the pathogenesis of diabetes itself but also plays a critical role in the development of its complications, including cardiovascular diseases (CVDs) such as atherosclerosis and acute myocardial infarction (AMI) [3-5].

CVD remains one of the leading causes of mortality in patients with diabetes, with AMI being one of the most devastating complications [6]. Patients with diabetes not only exhibit a higher incidence of AMI but also experience worse prognoses compared to individuals without diabetes [7]. Hyperglycemia has been implicated in these outcomes by promoting oxidative stress, endothelial dysfunction, and proinflammatory processes, including M1 macrophage polarization and enhanced monocyte adhesion to activated endothelial cells [8]. These inflammatory changes disrupt vascular homeostasis and contribute to microvascular damage [8]. Long-term prognosis in high-risk patients with diabetes combined with AMI plays a critical role in guiding clinical decision-making and improving outcomes. Prognostic studies provide valuable tools for identifying high-risk patients, enabling timely and targeted interventions to reduce mortality and improve survival rates.

Given the pivotal role of inflammation in both diabetes and AMI, there is growing interest in inflammatory markers as prognostic tools. Markers such as the lymphocyte-to-monocyte ratio (LMR), neutrophil-to-platelet ratio (NPR), platelet-to-lymphocyte ratio (PLR), neutrophil-to-lymphocyte ratio (NLR), and prognostic inflammatory value (PIV) have been widely used to reflect systemic inflammation and predict outcomes in CVDs and cancers [9-12]. Importantly, these indices derived from complete blood count (CBC) inherit its advantages of being stable, cost-effective, and widely accessible, making them potentially applicable for routine clinical use and large-scale prognostic studies. However, whether these inflammatory markers could be used for predicting clinical outcomes of patients with diabetes combined with AMI remains unclear. To address this gap, we leveraged data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database and the Department of Critical Care Medicine, Zhongnan Hospital of Wuhan University, China, to systematically compare and evaluate the prognostic utility of LMR, NPR, PLR, NLR, and PIV in predicting long-term mortality risk and clinical outcomes in this vulnerable population.

Methods

Ethical Considerations

This study adheres to the principles declared in the Declaration of Helsinki and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines and was approved by the Medical Ethics Committee of Zhongnan Hospital of Wuhan University (approval: 2020009). The Institutional Review Board of Beth Israel Deaconess Medical Center (BIDMC) granted a waiver of informed consent and approved the sharing of research resources. The authors obtained access to the MIMIC-IV database (certificate number: 65461627).

Source of Data

This study conducted a retrospective analysis using the large, publicly available intensive care database MIMIC-IV (version 3.0).

For external validation, data were obtained from Zhongnan Hospital of Wuhan University, China. This cohort included 166 patients admitted between January 1, 2021, and June 30, 2022. The final follow-up for outcome ascertainment was completed on July 22, 2025, by hospital record review and telephone interviews. During this period, 56 mortality events were recorded.

Case Inclusion and Exclusion Criteria

Patients diagnosed with both diabetes and AMI were included in this study. Patient identification was based on diagnostic codes from the *ICD-9* (*International Classification of Diseases, Ninth Revision*), and the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*). AMI and diabetes mellitus were identified using *ICD-9* and *ICD-10* codes.

The exclusion criteria were as follows: (1) age <18 years; (2) an intensive care unit (ICU) stay of <24 hours; (3) multiple admissions, with only data from the first hospitalization included; and (4) missing data on lymphocyte, monocyte, neutrophil, or platelet count levels.

Data Extraction

The *psql* tool of PostgreSQL was used to extract the data. Laboratory variables were extracted from the first result recorded within 24 hours of ICU admission. Features with a missing data rate exceeding 30% were excluded, while multiple imputation was applied to those with a missing data rate below 30%. Missing data were handled using multiple imputation by chained equations. We used the predictive mean matching method with 5 imputations (m=5) and set the random seed to 1 for reproducibility. All predictor variables were included in the imputation model.

The LMR was calculated as lymphocyte count (10^9/L) / monocyte count (10^9/L), NPR as neutrophil count (10^9/L) / platelet count (10^9/L), the PLR as platelet count (10^9/L) / lymphocyte count (10^9/L), the NLR as neutrophil count (10^9/L) / lymphocyte count (10^9/L), and the PIV as lymphocyte count

XSL•FO
RenderX

(10^9/L) × platelet count (10^9/L) × monocyte count (10^9/L)/ lymphocyte count (10^9/L) [10-12].

Time-Zero Alignment

All predictor variables (including CBC components and inflammatory indices) were extracted as the first recorded value on admission to prevent any look-ahead bias. To ensure transportability, this identical protocol—using the admission baseline—was applied to the external validation cohort.

Outcomes

The start date of follow-up was defined as the date of hospital admission, and the study outcome was 3-year mortality.

Statistical Analysis

As this study is a retrospective analysis, sample size calculations were not performed. For categorical variables, the chi-square test or Fisher exact test was performed for analysis. Student 2-tailed *t* test and the Wilcoxon rank-sum test were used for continuous variables. Kaplan-Meier (K-M) curves and Cox proportional hazards models were used to determine the relationships between LMR, NPR, PLR, NLR, PIV, and the risk of mortality. In Model 1, adjustments were made for age, gender, BMI. In Model 2, further adjustments were made for age, gender, BMI, troponin T, creatine kinase isoenzymes (CK-MB), insulin, history of revascularization (percutaneous coronary intervention and coronary artery bypass grafting), hyperlipidemia, atrial fibrillation, and stroke. Subgroup analysis was performed to examine the relationship between the continuous inflammation index and 3-year mortality across different subgroups. Restricted cubic spline (RCS) analysis was conducted to explore the associations between the 5 indices and mortality in a dose–response manner.

Feature Selection

The variance inflation factor (VIF) was used to assess multicollinearity among features. Features with a VIF greater than 5 were removed due to concerns about multicollinearity (Table S1 in Multimedia Appendix 1).

To identify the most important features associated with prognosis, the Boruta (version 9.0.0; R packages) algorithm was used. As a robust feature selection method, Boruta identifies all relevant predictors by iteratively comparing the importance of real features with random shadow features. Features marked as important (green) were selected for subsequent machine learning model construction.

Construction and Validation of Predictive Models

To construct the predictive model, the important features were identified using the Boruta algorithm. Subsequently, the dataset was randomly split into a training set (666/833, 80%) and a testing set (176/833, 20%) with a fixed random seed of 123. Machine learning models were then developed using the H2O AutoML platform (version 3.44.0.3) in R (R Foundation for Statistical Computing). The AutoML process was configured with the following hyperparameters: maximum number of models = 200, maximum runtime = 24 hours, and class balancing enabled (balance_classes = TRUE) to address outcome imbalance. Finally, models were ranked based on the area under the curve (AUC), and a stacked ensemble model was generated using the top-performing base models.

Each base model was trained on prognostic outcomes derived from the training data, and prognostic scores for each patient were calculated and scaled to a range of 0 to 1, with higher scores indicated worse prognosis. An ensemble method was then applied by compiling the top 5 base models, ranked according to their area under the receiver operating characteristic curve, into a matrix of prognostic scores. A second-layer generalized linear model algorithm was then used to create the final model. Five-fold cross-validation was performed during model training to minimize overfitting and ensure the robustness of the results. A probability threshold of .5 was used to classify patients into high- and low-risk groups for the confusion matrix analysis.

Here, 5 inflammation-related indices—LMR, NPR, PLR, NLR, and PIV—were also selected as features for model construction. Separate models were built for each individual feature following the methodology described above. Additionally, a stacked ensemble model incorporating all 5 indices was constructed to explore their combined predictive performance.

Results

Baseline Characteristics

Following the inclusion and exclusion criteria, 833 patients with diabetes combined with AMI from the MIMIC-IV database were used as the training and test cohorts, and 166 cases from Zhongnan Hospital of Wuhan University, China, were enrolled as an external validation cohort. Figure 1 illustrates the flow of patient selection for this study.

XSL•FO
RenderX

Figure 1. Patient screening and feature selection. Inclusion and exclusion criteria for patients from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. AMI: acute myocardial infarction; ICU: intensive care unit.

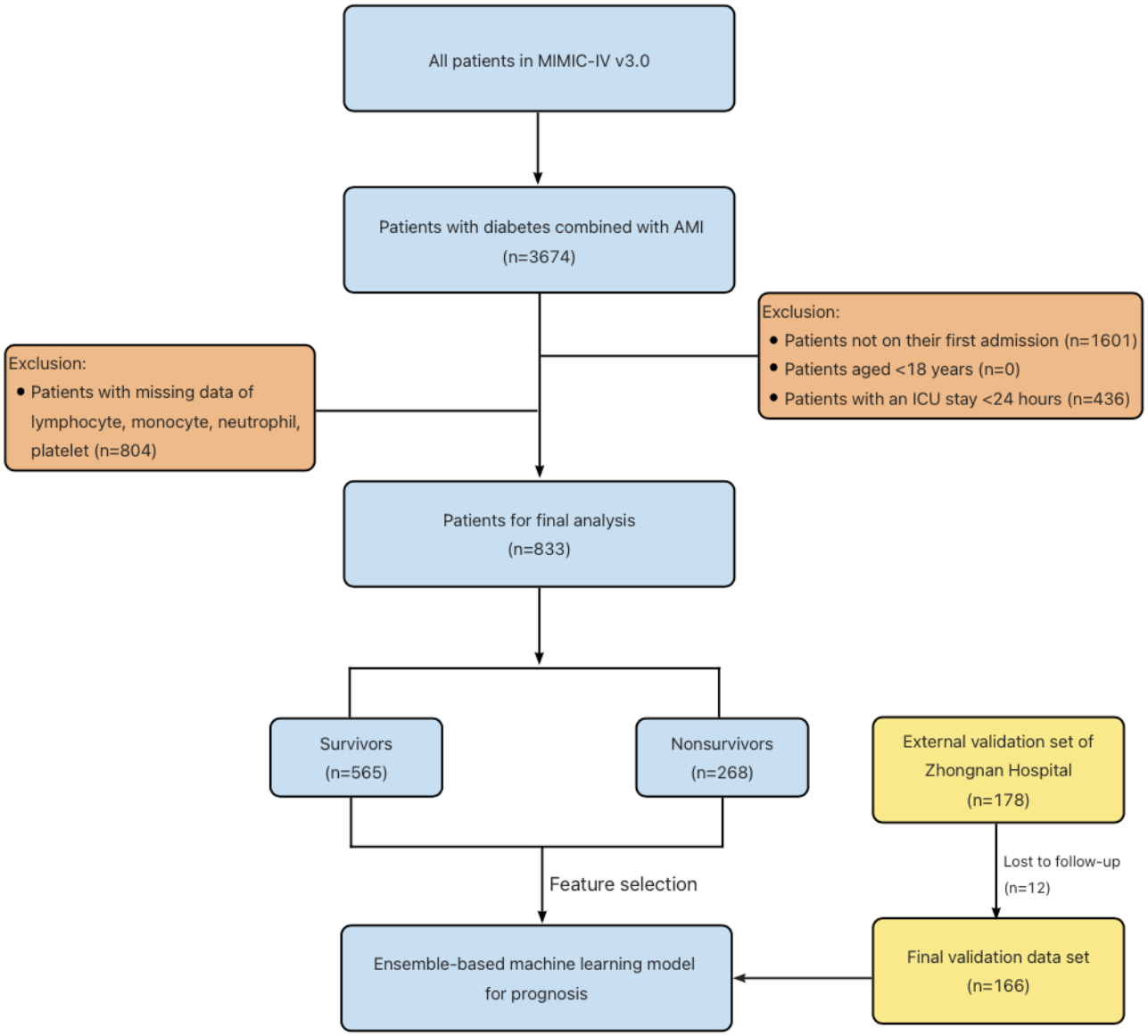


Table 1 provides a detailed comparison of baseline characteristics between survivors and nonsurvivors in the MIMIC-IV cohort (n=833), based on the 3-year mortality outcomes. Overall, the nonsurvivors were older and had a lower proportion of males compared with survivors. Among the comorbidities, the nonsurvivors exhibited significantly higher incidences of stroke, atrial fibrillation, and heart failure. Notably, the proportion of patients undergoing coronary artery bypass grafting was significantly lower in the nonsurvivor group. The nonsurvivors also exhibited significant differences in vital signs, including higher heart rate and respiratory rate, as well as higher diastolic blood pressure. As for laboratory parameters, the nonsurvivors had markedly elevated levels of mean corpuscular volume, red cell distribution width, white blood cells, phosphate, potassium, magnesium, alkaline phosphatase, creatinine, and urea nitrogen, suggesting metabolic and immune disturbance. Furthermore, inflammation-related indices also differed significantly between the 2 groups, with nonsurvivors exhibiting higher levels of NPR, PLR, NLR, and PIV, while LMR was lower. However, there was no significant change in troponin T or CK-MB between the survivors and nonsurvivors, demonstrating that these 2 markers typically show a significant increase during the acute phase of myocardial infarction and decrease rapidly during the recovery phase. These findings suggest that CBC-derived inflammation-related indices hold promise for predicting long-term prognosis in patients with diabetes combined with AMI.

XSL•FO
RenderX

Table 1. Baseline characteristics between survivors and nonsurvivors.

Characteristics	Total (n=833)	Survivors (n=565)	Nonsurvivors (n=268)	*P* value
Demographics				
Male, n (%)	576 (69.1)	408 (72.5)	168 (62.7)	.07
Age (years), mean (SD)	68.99 (11.10)	66.92 (10.64)	73.36 (10.81)	<.001
Weight (kg), mean (SD)	86.65 (23.09)	87.31 (22.07)	85.26 (25.11)	.23
BMI (kg/m^2)	29.87 (7.01)	29.85 (6.46)	29.93 (8.06)	.88
Vital signs, mean (SD)				
Heart rate (bpm)	86.82 (18.70)	83.80 (15.87)	93.21 (22.30)	<.001
Respiratory rate (breaths/min)	18.76 (6.86)	17.50 (5.75)	21.44 (8.14)	<.001
Systolic blood pressure (mm Hg)	118.67 (23.51)	117.55 (21.82)	120.93 (26.61)	.05
Diastolic blood pressure (mm Hg)	66.63 (15.75)	65.74 (14.51)	68.49 (17.90)	.02
Comorbidities, n (%)				
Hypertension	293 (35.2)	245 (43.4)	48 (17.9)	<.001
Stroke	71 (9)	27 (5)	44 (16)	<.001
Atrial fibrillation	61 (7)	33 (6)	28 (10)	<.001
Heart failure	457 (54.9)	271 (48)	186 (69.4)	<.001
Hyperlipidemia	601 (72.1)	438 (77.5)	163 (60.8)	<.001
Treatment, n (%)				
Insulin	809 (97.1)	552 (97.7)	257 (95.9)	.22
Percutaneous coronary intervention	94 (11)	53 (9)	41 (15)	.02
Coronary artery bypass grafting	439 (52.7)	396 (70.1)	43 (16)	<.001
Complete blood count, mean (SD)				
White blood cells (10^9/L)	11.66 (7.42)	10.60 (5.33)	13.89 (10.21)	<.001
Red cell distribution width (%)	14.41 (2.16)	13.95 (1.76)	15.39 (2.56)	<.001
Hemoglobin (g/dL)	11.61 (2.36)	12.13 (2.19)	10.51 (2.35)	<.001
Hematocrit (%)	35.71 (6.69)	37.00 (6.16)	33.01 (6.95)	<.001
Platelet count (10^9/L)	216.84 (81.65)	217.75 (77.27)	214.93 (90.31)	.64
Neutrophil count (10^9/L)	10.50 (5.78)	9.93 (5.12)	11.70 (6.82)	<.001
Lymphocyte count (10^9/L)	1.92 (4.34)	1.98 (1.20)	1.81 (7.46)	.61
Monocyte count (10^9/L)	0.78 (0.54)	0.67 (0.48)	0.99 (0.59)	<.001
Eosinophil count (10^9/L)	0.13 (0.19)	0.14 (0.18)	0.11 (0.20)	.02
Basophil Count (10^9/L)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	.25
Mean corpuscular volume (fL)	90.43 (6.79)	89.84 (6.46)	91.68 (7.30)	.001
Mean corpuscular hemoglobin (pg)	29.32 (2.56)	29.42 (2.52)	29.11 (2.64)	.10
Mean corpuscular hemoglobin concentration (g/dL)	32.43 (1.50)	32.74 (1.43)	31.77 (1.43)	<.001
Immature granulocytes (%)	1.04 (0.92)	1.01 (0.90)	1.10 (0.95)	.17
Metabolic, mean (SD)				
Bicarbonate (mEq/L)	22.44 (4.40)	23.02 (3.94)	21.22 (5.04)	<.001
Phosphate (mg/dL)	3.95 (1.32)	3.75 (1.15)	4.38 (1.52)	<.001
Chloride (mEq/L)	101.03 (5.42)	100.99 (4.56)	101.13 (6.89)	.71
Potassium (mEq/L)	4.38 (0.73)	4.31 (0.65)	4.52 (0.84)	<.001

XSL•FO
RenderX

Characteristics	Total (n=833)	Survivors (n=565)	Nonsurvivors (n=268)	*P* value
Magnesium (mg/dL)	1.98 (0.36)	1.94 (0.34)	2.06 (0.40)	<.001
Calcium total (mg/dL)	8.70 (0.74)	8.76 (0.69)	8.58 (0.83)	.001
Aspartate aminotransferase (IU/L)	80.46 (131.84)	69.50 (117.58)	103.85 (155.73)	.001
Alanine aminotransferase (IU/L)	51.68 (94.95)	42.40 (71.36)	71.30 (129.58)	<.001
Albumin (g/dL)	3.45 (0.58)	3.61 (0.50)	3.13 (0.61)	<.001
Alkaline phosphatase (IU/L)	101.09 (104.40)	90.53 (52.22)	123.37 (165.74)	<.001
Glucose (mg/dL)	203.05 (105.71)	200.08 (102.18)	209.30 (112.72)	.24
Urea nitrogen (mg/dL)	31.57 (22.82)	25.91 (17.38)	43.49 (27.82)	<.001
Creatinine (mg/dL)	1.77 (1.68)	1.50 (1.44)	2.33 (1.99)	<.001
Troponin T (ng/mL)	1.18 (2.57)	1.08 (2.42)	1.39 (2.85)	.10
Creatine kinase isoenzymes (ng/mL)	28.73 (64.48)	28.48 (60.27)	29.26 (72.68)	.87
Coagulation, mean (SD)				
Partial thromboplastin time (sec)	48.09 (32.14)	48.73 (31.37)	46.73 (33.73)	.40
Inflammation-related indices, mean (SD)				
Lymphocyte-to-monocyte ratio	3.74 (4.28)	4.62 (4.76)	1.93 (2.11)	<.001
Neutrophil-to-platelet ratio	0.05 (0.05)	0.05 (0.05)	0.06 (0.05)	.001
Platelet-to-lymphocyte ratio	187.02 (184.20)	164.53 (169.65)	234.37 (203.96)	<.001
Neutrophil-to-lymphocyte ratio	9.55 (11.84)	7.63 (9.23)	13.61 (15.26)	<.001
Prognostic inflammatory value	2291.85 (3761.61)	1808.40 (2741.59)	3291.33 (5143.52)	<.001

For the external validation cohort (n=166), baseline analysis focused specifically on CBC data and indices in the prognostic model (LMR, NLR, NPR, PIV, and PLR). The detailed comparison between survivors and nonsurvivors based on the 3-year mortality outcomes are presented in Table S2 in Multimedia Appendix 1.

Associations Between the Inflammation-Related Indices and Mortality Risk

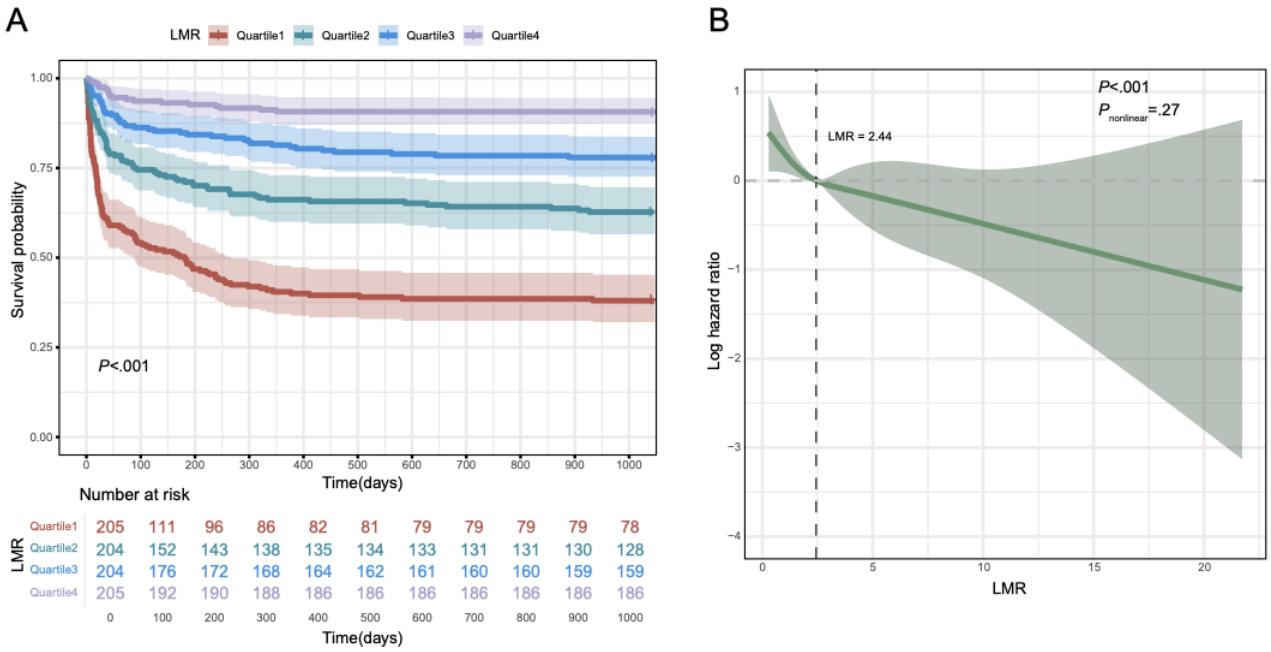
Cox proportional hazards analysis identified significant associations between inflammation-related indices (LMR, NPR, PLR, NLR, and PIV) and 3-year mortality risk (Table S3 in Multimedia Appendix 1). LMR was determined to be a protective factor, while NPR, PLR, NLR, and PIV were associated with an increased risk of mortality. These findings underscore the prognostic value of systemic inflammation markers in predicting long-term outcomes in this population.

LMR and Mortality Risk

Patients were stratified into quartiles based on LMR values: quartile 1 (LMR≤1.17; n=205), quartile 2 (1.17<LMR≤2.48; n=204), quartile 3 (2.48<LMR≤4.60; n=204), and quartile 4 (LMR>4.60; n=205). In Cox proportional hazards models, quartiles 2, 3, and 4 were significantly associated with reduced 3-year mortality risk compared to quartile 1, in both unadjusted and adjusted models (Table S3 in Multimedia Appendix 1). K-M survival analysis further demonstrated that patients in quartile 1 had the lowest 3-year survival probability, indicating the detrimental impact of low LMR (Figure 2A). RCS analysis revealed a decreasing trend between LMR and mortality risk, with a critical inflection point at LMR=2.44 (Figure 2B). Below this threshold, lower LMR values were strongly associated with increased mortality risk.

XSL•FO
RenderX

Figure 2. Kaplan Meier (K-M) survival analysis curves and restricted cubic spline (RCS) regression analysis of lymphocyte-to-monocyte ratio (LMR) for 3-year mortality. (A) K-M curves of LMR. (B) RCS analysis of LMR.

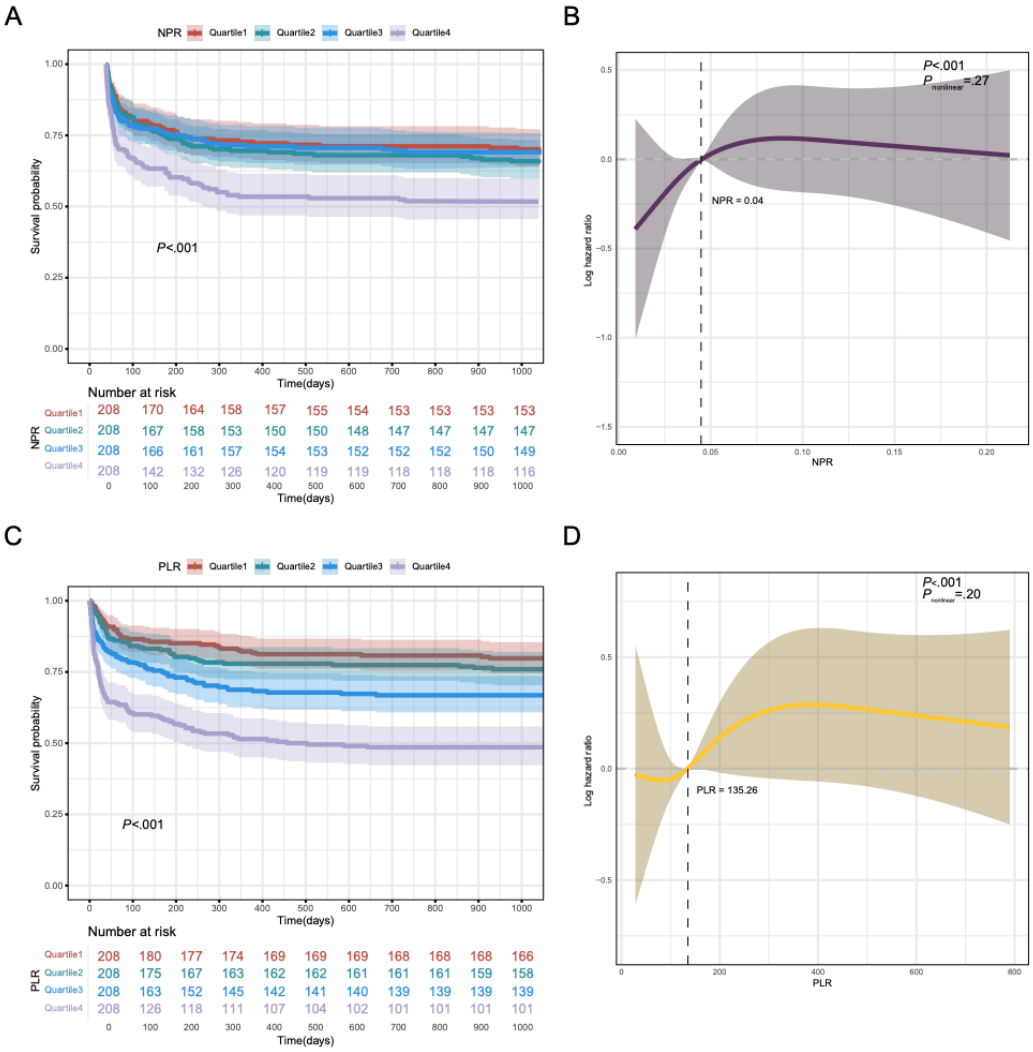


NPR and Mortality Risk

For NPR, patients were divided into 4 quartiles: quartile 1 (NPR≤0.033), quartile 2 (0.033<NPR≤0.046), quartile 3 (0.046<NPR≤0.064), and quartile 4 (NPR>0.064). In the unadjusted model, higher NPR values were significantly associated with increased mortality risk. However, this association was attenuated and lost statistical significance after adjustment for potential confounders (Table S3 in Multimedia Appendix 1). K-M survival analysis demonstrated that patients in quartile 4 (highest NPR) had the lowest 3-year survival probability compared with other quartiles, underscoring the detrimental impact of elevated NPR values (Figure 3A). RCS analysis revealed a nonlinear trend in the association between NPR and mortality risk, with a potential threshold observed around NPR=0.04. However, the test for nonlinearity did not reach statistical significance ($P_{\text{nonlinearity}}$=.27), suggesting that a linear relationship between these variables cannot be excluded (Figure 3B).

XSL•FO
RenderX

Figure 3. Kaplan-Meier (KM) survival analysis curves and restricted cubic spline (RCS) regression analysis of neutrophil-to-platelet ratio (NPR) and platelet-to-lymphocyte ratio (PLR) for 3-year mortality. (A) K-M curves of NPR. (B) RCS analysis of NPR. (C) K-M curves of PLR. (D) RCS analysis of PLR.



PLR and Mortality Risk

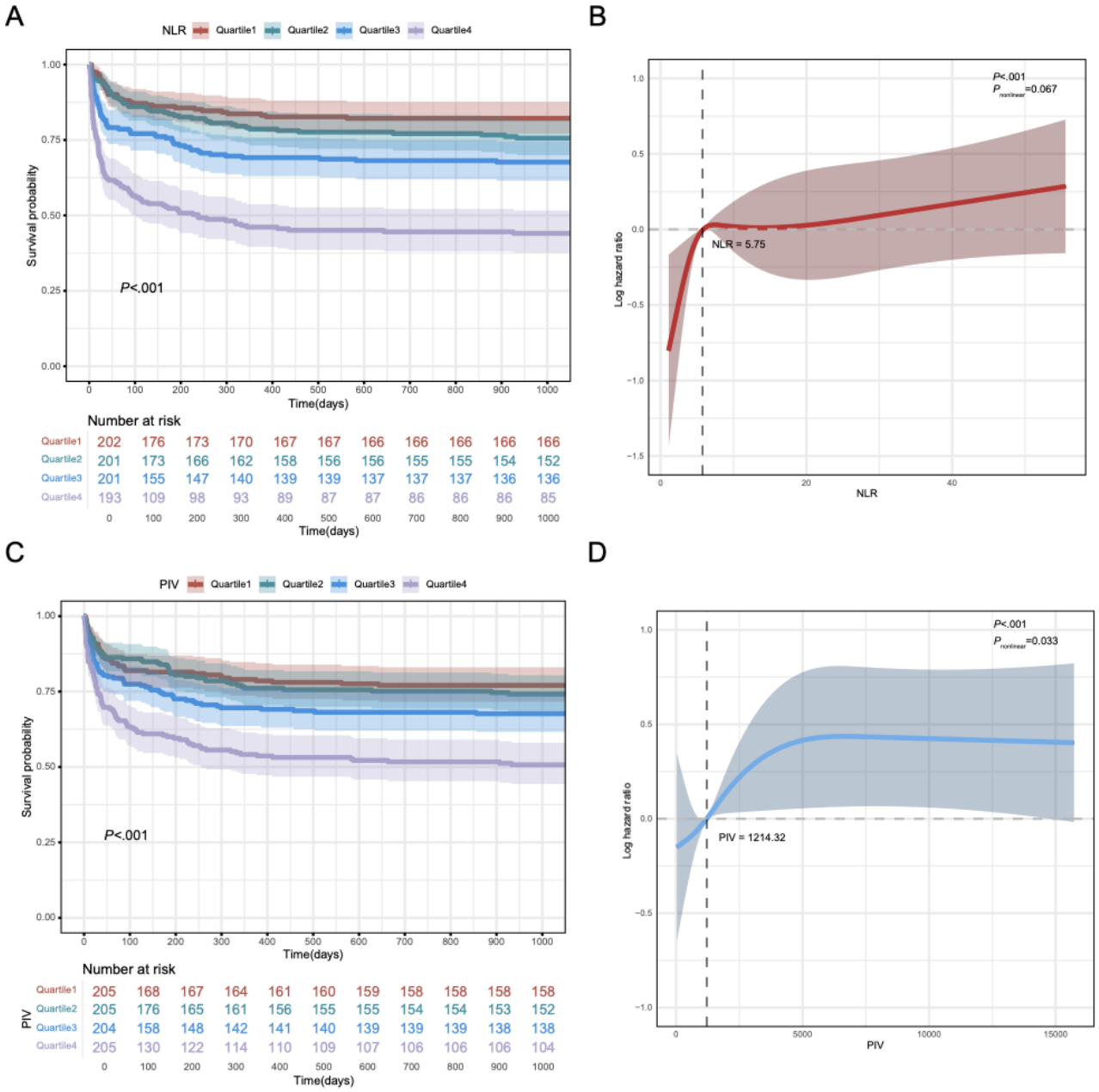
PLR was categorized into quartile 1 (PLR≤83.50), quartile 2 (83.50<PLR≤134.37), quartile 3 (134.37<PLR≤255.65), and quartile 4 (PLR>255.65). In the unadjusted model, higher PLR values were associated with increased mortality risk. However, after adjustment for confounding variables, this association was no longer statistically significant, suggesting potential confounding effects (Table S3 in Multimedia Appendix 1). K-M survival analysis demonstrated that patients in quartile 4 (highest PLR) had the lowest 3-year survival probability, consistent with the hypothesis that elevated PLR reflects a heightened inflammatory state (Figure 3C). RCS analysis indicated that as PLR increased, mortality risk also increased, with a notable value at PLR=135.26 (Figure 3D).

NLR and Mortality Risk

NLR quartiles were defined as follows: quartile 1 (NLR≤3.50), quartile 2 (3.50<NLR≤5.86), quartile 3 (5.86<NLR≤10.62), and quartile 4 (NLR>10.62). In both unadjusted and adjusted Cox proportional hazards models, patients in quartile 4 (highest NLR) exhibited a significantly increased risk of 3-year mortality compared with those in lower quartiles (Table S3 in Multimedia Appendix 1). K-M survival analysis further demonstrated that quartile 4 had the lowest 3-year survival probability, indicating the adverse prognostic impact of elevated NLR (Figure 4A). RCS analysis indicated that lower NLR values were associated with lower mortality risk, whereas higher NLR values significantly increased mortality risk, especially beyond a threshold of 5.75. Although the *P*=.07 suggests the nonlinear relationship was not fully significant, the trend between NLR and mortality risk is noteworthy (Figure 4B).

XSL•FO
RenderX

Figure 4. Kaplan-Meier (KM) survival analysis curves and restricted cubic spline (RCS) regression analysis of neutrophil-to-lymphocyte ratio (NLR) and prognostic inflammatory value (PIV) for 3-year mortality. (A) K-M curves of NLR. (B) RCS analysis of NLR. (C) K-M curves of PIV. (D) RCS analysis of PIV.



PIV and Mortality Risk

For PIV, the quartiles were defined as quartile 1 (PIV≤550.46), quartile 2 (550.46<PIV≤1207.40), quartile 3 (1207.40<PIV≤2619.32), and quartile 4 (PIV>2619.32). Similarly, in both unadjusted and adjusted Cox models, quartile 4 (highest PIV) was significantly associated with increased 3-year mortality risk compared with lower quartiles (Table S3 in Multimedia Appendix 1). K-M survival analysis demonstrated that patients in quartile 4 had the lowest 3-year survival probability (Figure 4C), suggesting an association between elevated PIV and poor outcomes. The RCS curve shows that 1214.32 is a risk turning point; beyond this value, risk increased significantly (Figure 4D).

Stratified Analyses

To further explore whether the relationship between inflammation-related indices and 3-year all-cause mortality persists under different conditions, subgroup analyses were conducted based on gender, age, BMI, hypertension, and hyperlipidemia. In the subgroup of patients aged ≥60 years, the hazard ratios (HRs) for all indices related to 3-year all-cause mortality were significant (Figures S1 and S2 in Multimedia Appendix 2). However, in the subgroup of patients aged <60 years, the HRs for NPR, PLR, and PIV did not show statistical significance. Furthermore, the HR for NPR was not statistically significant in the subgroup of patients with BMI<18 kg/m^2 and the subgroup of patients with hypertension. Similarly, the HRs for NPR and PIV were not statistically significant in the

XSL•FO
RenderX

subgroup of patients with BMI>28 kg/m^2 (Figures S1 and S2 in Multimedia Appendix 2).

Feature Selection

Initially, we compared the prognostic performance of individual CBC counts (lymphocyte, monocyte, neutrophil, and platelet counts) with inflammation-related indices using Cox regression models. The C-index values for individual cell counts ranged from 0.62 to 0.68, significantly lower than those of composite indices (0.72-0.79; Figure S3 in Multimedia Appendix 2). This disparity highlights the limited utility of isolated cell counts in capturing systemic inflammatory states. Given that the inflammation-related indices are derived from these individual cell counts, they inherently exhibit a higher degree of collinearity. Therefore, to focus more on the inflammation-related features themselves and minimize the impact of collinearity in the analysis, we excluded the individual immune cell counts before performing the VIF calculation.

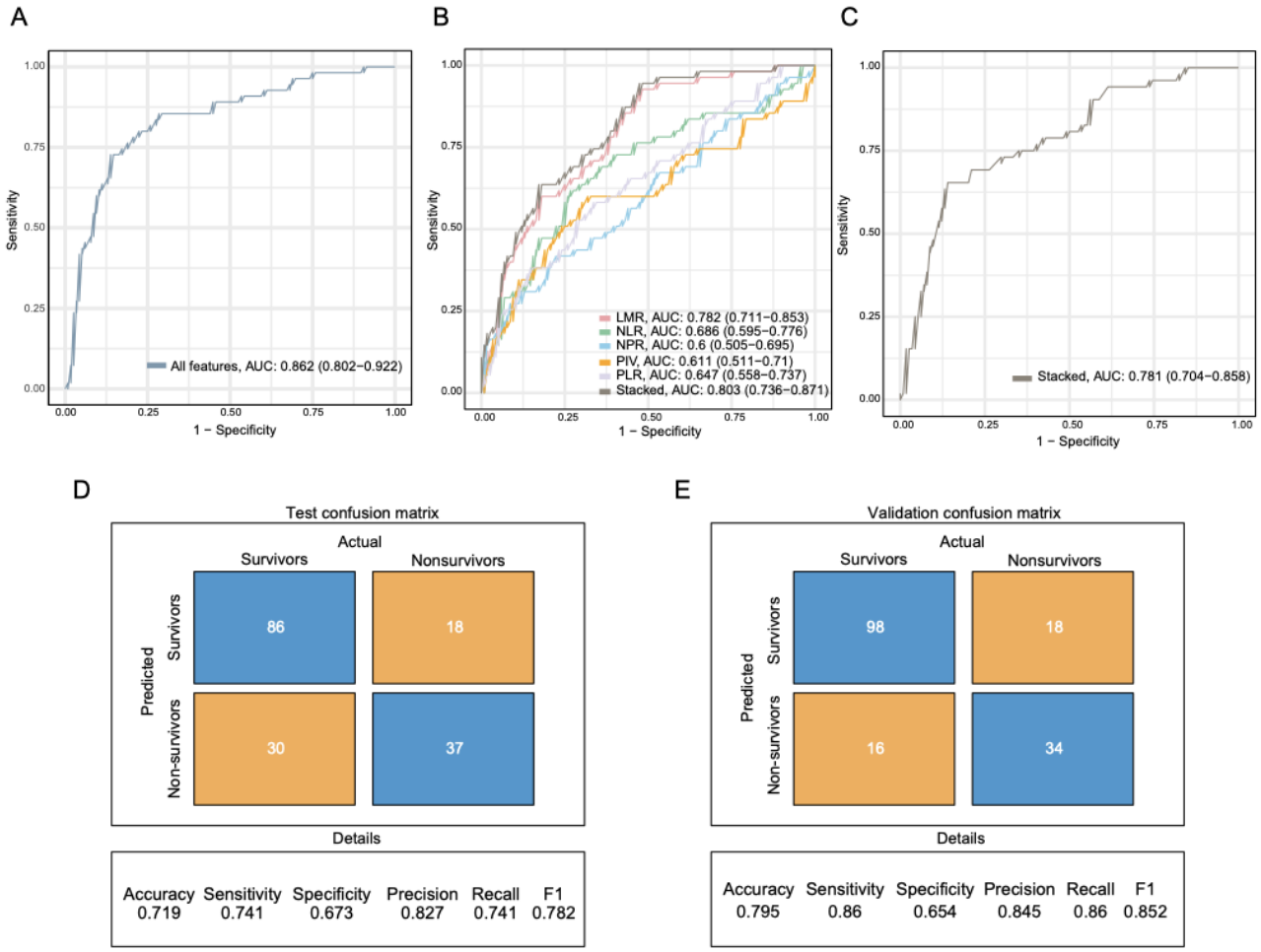
We then examined the VIF to detect multicollinearity among variables, and features with a VIF>5 were removed (Table S1 in Multimedia Appendix 1). Subsequently, feature selection was performed using the Boruta algorithm, in which variables in the green area were identified as important features, whereas those in the red area were deemed unimportant (Figure S4 in Multimedia Appendix 2). Both CK-MB and troponin T were classified as unimportant (red box in Figure S4 in Multimedia Appendix 2), corroborating their limited prognostic value observed in baseline and Cox analysis (Table 1 and Figure S3 in Multimedia Appendix 2). This exclusion by the algorithm suggests that, although CK-MB and troponin T are critical for diagnosing acute myocardial injury, their transient elevation may not reflect the chronic inflammatory processes that drive long-term mortality in patients with diabetes combined with AMI.

Development and Validation of an Ensemble Machine Learning Model for Prognostic Prediction

After feature selection, we constructed a stacked model using all the important features, achieving an AUC of 0.862 in the test cohort (Figure 5A). To further explore the effectiveness of inflammation-related indices in predicting mortality risk in patients with diabetes combined with AMI, we built models based on individual inflammation-related indices, with AUCs of 0.782, 0.686, 0.598, 0.611, and 0.648 for LMR, NLR, NPR, PIV, and PLR, respectively, in the test cohort. We also combined these 5 indices into a stacked model, which achieved an AUC of 0.803 in the test cohort (Figure 5B; detailed parameters are shown in Figure 5D). Importantly, this combined inflammation-related indices model demonstrated robust generalizability, achieving an AUC of 0.781 in the external validation cohort (Figure 5C), with its predictive performance detailed in the confusion matrix (Figure 5E).

XSL•FO
RenderX

Figure 5. Receiver operating characteristic (ROC) curves of the machine learning prognostic models. (A) ROC curve for the model constructed using all important features in the test cohort. (B) ROC curves for models based on individual inflammation-related indices (lymphocyte-to-monocyte ratio [LMR], neutrophil-to-lymphocyte ratio [NLR], neutrophil-to-platelet ratio [NPR], prognostic inflammatory value [PIV], platelet-to-lymphocyte ratio [PLR]) and their combined model in the test cohort. (C) ROC curve for the combined inflammation-related indices model in the external validation cohort (area under the curve [AUC]=0.781). (D) Confusion matrix for the combined inflammation-related indices model in the test cohort. (E) Confusion matrix for the combined inflammation-related indices model in the external validation cohort.



Although the combined model performed slightly worse than the model using all important features in the test cohort, it demonstrated comparable and generalizable prognostic ability. The combined inflammation-related indices model may contribute to future prognostic prediction by offering a simplified and cost-effective alternative for clinical applications.

Beyond discrimination, we evaluated the calibration and clinical utility of the model. The calibration plots displayed good agreement between predicted probabilities and observed outcomes, with Brier scores of 0.146 in the validation set and 0.155 in the external cohort (Figure S5 in Multimedia Appendix 2). Furthermore, decision curve analysis (DCA) indicated that the model provided a substantial net benefit over the "treat-all" or "treat-none" strategies across a wide range of threshold probabilities in both cohorts (Figure S5 in Multimedia Appendix 2).

In the external validation cohort, we also evaluated an enhanced model incorporating readily available vitals (age, heart rate, blood pressure, respiration rate). This "Indices + Vitals" model showed marginal improvement (AUC=0.788) and overlapping net benefit curves in DCA (Figure S6 in Multimedia Appendix 2), supporting the use of the 5-index model as a parsimonious stand-alone tool.

We further compared the stacked model with the best-performing single index LMR (Figure S8 in Multimedia Appendix 2). Although the improvements in net reclassification improvement and discrimination improvement were not statistically significant, the stacked model demonstrated superior overall performance. Specifically, it achieved a higher AUC (0.781 vs 0.771 in external validation) and, more importantly, better calibration (Brier score: 0.155 vs 0.183; calibration slope: 1.589 vs 3.147), suggesting that the multi-index approach provides more robust and accurate risk estimates than any single biomarker.

Application of the Model

Using the above model, we developed a web tool to predict the prognosis of patients with diabetes and acute myocardial infarction (Figure S9 in Multimedia Appendix 2). Users can get prediction results by directly inputting features on the web page [13].

XSL•FO
RenderX

Discussion

Overview

To the best of our knowledge, this study is the first to systematically compare 5 inflammation-related indices (LMR, NLR, NPR, PLR, and PIV) derived from CBC in relation to the 3-year mortality risk of patients with diabetes combined with AMI, and to explore their potential application in predicting long-term mortality risk. The findings indicate that inflammation may play a critical role in the long-term prognosis of patients with diabetes combined with AMI. By constructing both single-indicator models and combined predictive models, we demonstrated that these inflammation indices can effectively assess mortality risk while also providing a reference for rapid and straightforward clinical risk assessment methods. Interestingly, adding readily available vital signs to the inflammation-related indices provided limited incremental value in the external cohort, with minimal improvement. This suggests that the 5-index model may offer a streamlined tool for risk stratification (Figure S6 in Multimedia Appendix 2).

From a clinical perspective, the proposed model primarily relies on routine CBC components, which are widely available, rapid, and inexpensive. This accessibility could make the model a potentially useful tool for early risk stratification in settings such as emergency departments and ICUs, potentially helping clinicians to prioritize resource allocation.

The prognostic value of these CBC-derived inflammation-related indices likely stems from the synergistic pathophysiology of diabetes and AMI. The chronic hyperglycemia and insulin resistance unique to patients with diabetes induce low-grade chronic inflammation, which not only exacerbates cardiovascular damage but also drives disease progression through various mechanisms [14]. When AMI occurs, it triggers an acute stress response involving the sympathetic nervous system and the hypothalamic–pituitary–adrenal axis [15]. Both chronic metabolic inflammation and acute ischemic injury leads to a profound dysregulation of hematopoiesis, manifesting as the specific alterations in leukocytes and platelets observed in this study [16].

The elevated risk associated with high NLR, PLR, and PIV can be mechanically explained by the exacerbated proinflammatory and prothrombotic phenotype in patients with diabetes. Hyperglycemia and insulin resistance prime neutrophils to overreact during myocardial ischemia. A critical mechanism linking these cells to poor prognosis is the formation of neutrophil extracellular traps. In a diabetic environment, activated neutrophils release excessive neutrophil extracellular traps, which not only directly damage cardiomyocytes but also serve as a scaffold for platelet aggregation and thrombosis [17,18]. This interaction is particularly detrimental in AMI, as it contributes to microvascular obstruction and extends the infarct size [19]. Furthermore, diabetic platelets exhibit surface protein glycation and increased reactivity, making them more prone to adhering to leukocytes [20]. PIV (the systemic immune-inflammation index), which integrates neutrophils, platelets, and monocytes, essentially captures this extensive interplay between coagulation and inflammation, explaining its strong association with mortality in our cohort of patients with diabetes.

Conversely, the protective role of LMR highlights the importance of adaptive immunity in healing. Lymphocytes, particularly regulatory T cells, are essential for transitioning the immune response from the inflammatory phase to the reparative phase [21]. However, in patients with diabetes complicated by AMI, lymphocytes face dual suppression: chronic oxidative stress from hyperglycemia induces lymphocyte apoptosis, while the acute cortisol surge from AMI causes massive lymphopenia [22,23]. This severe depletion of lymphocytes impairs the containment of the inflammatory storm. Simultaneously, hyperglycemia promotes differentiation of monocytes into the proinflammatory M1 phenotype rather than the reparative M2 phenotype, further delaying scar healing and promoting adverse ventricular remodeling [24,25]. Therefore, a reduced LMR serves as a composite marker indicating an uncontrolled inflammatory response correlating with worse long-term survival.

The elevated levels of NLR, PLR, NPR, and PIV may reflect a proinflammatory state driven by interactions among platelets, neutrophils, and lymphocytes, which could be a critical driver of poor prognosis in patients with diabetes and CVDs. For instance, PIV, as an integrative inflammatory marker, captures the interplay between platelets, neutrophils, and lymphocytes. Studies have shown that elevated PIV levels are associated with worse outcomes in patients with hypertension [20]. Additionally, previous research has demonstrated that high NLR and PLR values are not only independent predictors of in-hospital mortality in patients with heart failure but are also closely associated with the risk of cardiac death [26,27]. Elevated NLR may indicate enhanced proinflammatory neutrophil activity and impaired lymphocyte function, whereas increased PLR may reflect both platelet activation and reduced lymphocyte-mediated immune function. In contrast, NPR primarily reflects the ratio between neutrophils and platelets, and its elevation has been associated with worse outcomes in patients with myocardial infarction with nonobstructive coronary arteries [28].

The protective role of LMR and the risk-associated roles of NLR, NPR, PLR, and PIV further underscore the critical importance of systemic inflammation and immune balance in the prognosis of patients with diabetes combined with AMI. These mechanisms suggest that protective and risk-related inflammatory markers may influence the prognosis of patients with diabetes combined with AMI through distinct pathways, providing a theoretical basis for future research on underlying mechanisms and advancements in diagnostic and therapeutic strategies.

In addition, the findings of this study highlight the significant value of inflammation-related indices in predicting the long-term prognosis of patients with diabetes combined with AMI.

Compared with traditional complex biomarkers or imaging techniques, these inflammation indices are derived from CBC, which is a widely available and affordable diagnostic tool. Furthermore, CBC results are highly stable and reproducible, making them reliable for both research and clinical practice [29]. These advantages highlight their practicality and feasibility

XSL•FO
RenderX

for preliminary prognostic assessment in patients with diabetes combined with AMI. The combined predictive model constructed based on these 5 indices (AUC=0.803) demonstrated a predictive ability comparable to that of the full importance features model (AUC=0.862). Moreover, our DCA confirmed that using this model to guide clinical decisions provides a higher net benefit than default strategies across a wide range of threshold probabilities (Figures 5A and B in Multimedia Appendix 2). By identifying high-risk patients, such prognostic models may facilitate early clinical interventions and potentially improve survival rates.

Limitations

Despite confirming the potential role of inflammation-related indices in predicting the long-term prognosis of patients with diabetes combined with AMI, this study has several limitations.

First, the regulatory mechanisms of inflammation in diabetes and AMI are highly complex, with different types of inflammatory cells potentially exerting bidirectional effects during disease progression. This intricate cellular regulation may not be fully captured by simplified inflammation-related indices.

Second, although 5 commonly used inflammation-related indices were analyzed, other potentially important inflammatory markers (eg, interleukin 6 and tumor necrosis factor α) were not included, which may underestimate the comprehensive impact of inflammation. Additionally, these indices primarily reflect systemic inflammation and may not adequately capture the influence of local immune microenvironments.

Third, a direct comparison with established scoring systems, such as the Global Registry of Acute Coronary Events or the Thrombolysis in Myocardial Infarction score, was not feasible in this study. Critical variables required for these scores—specifically cardiac arrest at admission, Killip classification, and coronary artery stenosis severity—were not systematically recorded or exhibited high levels of missingness in the structured electronic health records of the MIMIC-IV database. Regarding clinical utility, the DCA results suggest that using our model to guide clinical decisions would yield a net benefit over default strategies within a broad threshold range. However, the observed calibration drift in the external validation cohort highlights the heterogeneity between populations (Figures 5A and B in Multimedia Appendix 2). While the model's discriminative power remains robust, this finding emphasizes the necessity of model recalibration to adjust for baseline risk differences before local clinical implementation.

Fourth, the external validation was asymmetrical. While the primary 5-index model was successfully validated externally, we could not validate the full model containing ancillary covariates because these specific variables were not available in the external center's dataset. This limitation restricts the generalizability of the covariate-adjusted analysis to the derivation cohort.

Fifth, while MIMIC-IV provides a large-scale, high-quality dataset, it represents a single-center experience in the United States. Consequently, practice patterns, patient demographics, and admission criteria may differ from those in other institutions or countries, potentially limiting the direct generalizability of our findings. Although external validation was performed to assess transportability, future multicenter prospective studies are widely recommended.

Finally, regarding model transparency, while our model demonstrated excellent calibration in the internal validation, calibration drift was observed in the external validation cohort (Figures 5C and D in Multimedia Appendix 2), showing a tendency to underestimate risk in the new population. This suggests that while the model is effective for risk ranking, it may require recalibration using local data to optimize absolute probability estimates for different clinical settings.

Acknowledgments

During the preparation of this manuscript, the authors used DeepSeek (DeepSeek Inc) for grammar and language checking.

Funding

This work was supported by the National Natural Science Foundation of China (82172359, 82372338 and 81972009 to SML; 82402724 to KY; and 82202619 to YY) and the Translational Medicine and Interdisciplinary Research Joint Fund of Zhongnan Hospital of Wuhan University (ZNJC202302).

Data Availability

The datasets generated and/or analyzed during this study are available from the following sources. The Medical Information Mart for Intensive Care IV (MIMIC-IV) database is publicly available via PhysioNet [30], with access requiring completion of the Collaborative Institutional Training Initiative (CITI) Data or Specimens Only Research training and approval by the review board. The dataset from Zhongnan Hospital is not publicly available due to patient privacy and ethical restrictions; however, deidentified data may be available from the corresponding author upon reasonable request and with permission from the Ethics Committee of Zhongnan Hospital. To facilitate reproducibility, the code has been deposited in a public repository and is available at gitee [31].

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File , 31 KB - medinform_v14i1e83328_app1.docx]

Multimedia Appendix 2
Additional figures.
[DOCX File , 3211 KB - medinform_v14i1e83328_app2.docx]

References

1. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023;402(10397):203-234 [FREE Full text] [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
2. Lee YS, Olefsky J. Chronic tissue inflammation and metabolic disease. Genes Dev 2021;35(5-6):307-328 [FREE Full text] [doi: 10.1101/gad.346312.120] [Medline: 33649162]
3. Wei X, Song H, Yin L, Rizzo MG, Sidhu R, Covey DF. Fatty acid synthesis configures the plasma membrane for inflammation in diabetes. Nature 2016;539(7628):294-298 [FREE Full text] [doi: 10.1038/nature20117] [Medline: 27806377]
4. Yang K, Sun J, Zhang Z, Xiao M, Ren D, Liu SM. Reduction of mRNA mA associates with glucose metabolism via YTHDC1 in human and mice. Diabetes Res Clin Pract 2023;198:110607. [doi: 10.1016/j.diabres.2023.110607] [Medline: 36878322]
5. Qin S, Yang K, Han S, Yuan Y, Mo J, Xiao M. N -methyladenine profiling of low-input multiplex clinical samples on transcriptome reveals RNA modifications implicated in type 2 diabetes and acute myocardial infarction. Clin Transl Med 2023;13(1):e1165. [doi: 10.1002/ctm2.1165] [Medline: 36631931]
6. Wong ND, Sattar N. Cardiovascular risk in diabetes mellitus: epidemiology, assessment and prevention. Nat Rev Cardiol 2023;20(10):685-695. [doi: 10.1038/s41569-023-00877-z] [Medline: 37193856]
7. Xia JG, Li B, Zhang H, Li QX, Lam SM, Yin CL. Precise metabolomics defines systemic metabolic dysregulation distinct to acute myocardial infarction associated with diabetes. Arterioscler Thromb Vasc Biol 2023;43(4):581-596. [doi: 10.1161/ATVBAHA.122.318871] [Medline: 36727520]
8. Edgar L, Akbar N, Braithwaite AT, Krausgruber T, Gallart-Ayala H, Bailey J. Hyperglycemia induces trained immunity in macrophages and their precursors and promotes atherosclerosis. Circulation 2021;144(12):961-982 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.046464] [Medline: 34255973]
9. Dentali F, Nigro O, Squizzato A, Gianni M, Zuretti F, Grandi AM. Impact of neutrophils to lymphocytes ratio on major clinical outcomes in patients with acute coronary syndromes: a systematic review and meta-analysis of the literature. Int J Cardiol 2018;266:31-37. [doi: 10.1016/j.ijcard.2018.02.116] [Medline: 29887466]
10. Trinh H, Dzul SP, Hyder J, Jang H, Kim S, Flowers J. Prognostic value of changes in neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR) and lymphocyte-to-monocyte ratio (LMR) for patients with cervical cancer undergoing definitive chemoradiotherapy (dCRT). Clin Chim Acta 2020;510:711-716. [doi: 10.1016/j.cca.2020.09.008] [Medline: 32919942]
11. Lin Y, Dai W, Chen Y, He X, Xu Y. Neutrophil-to-platelet ratio predicts mortality following percutaneous coronary intervention in patients with acute ST-elevation myocardial infarction. Front Physiol 2022;13:1011048. [doi: 10.3389/fphys.2022.1011048] [Medline: 36200052]
12. Guven DC, Sahin TK, Erul E, Kilickap S, Gambichler T, Aksoy S. The association between the pan-immune-inflammation value and cancer prognosis: a systematic review and meta-analysis. Cancers (Basel) 2022;14(11):2675. [doi: 10.3390/cancers14112675] [Medline: 35681656]
13. Prognostic tool for diabetic patients with acute myocardial infarction. URL: https://vigbly.shinyapps.io/shiny/ [accessed 2026-02-02]
14. Pradhan AD, Ridker PM. Do atherosclerosis and type 2 diabetes share a common inflammatory basis? Eur Heart J 2002;23(11):831-834. [doi: 10.1053/euhj.2001.3052] [Medline: 12042000]
15. Kun W, Jie Z, Shuai C, Xin WU, Guoqi Z, Shengbing WU, et al. Electroacupuncture ameliorates cardiac dysfunction in myocardial ischemia model rats: a potential role of the hypothalamic-pituitary-adrenal axis. J Tradit Chin Med 2023;43(5):944-954 [FREE Full text] [doi: 10.19852/j.cnki.jtcm.20230727.001] [Medline: 37679982]
16. Yamagishi S, Maeda S, Matsui T, Ueda S, Fukami K, Okuda S. Role of advanced glycation end products (AGEs) and oxidative stress in vascular complications in diabetes. Biochim Biophys Acta 2012;1820(5):663-671. [doi: 10.1016/j.bbagen.2011.03.014] [Medline: 21440603]
17. Döring Y, Soehnlein O, Weber C. Neutrophil extracellular traps in atherosclerosis and atherothrombosis. Circ Res 2017;120(4):736-743 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.309692] [Medline: 28209798]

XSL•FO
RenderX

18. Menegazzo L, Ciciliot S, Poncina N, Mazzucato M, Persano M, Bonora B. NETosis is induced by high glucose and associated with type 2 diabetes. Acta Diabetol 2015;52(3):497-503. [doi: 10.1007/s00592-014-0676-x] [Medline: 25387570]
19. Sreeramkumar V, Adrover JM, Ballesteros I, Cuartero MI, Rossaint J, Bilbao I. Neutrophils scan for activated platelets to initiate inflammation. Science 2014;346(6214):1234-1238 [FREE Full text] [doi: 10.1126/science.1256478] [Medline: 25477463]
20. Wu B, Zhang C, Lin S, Zhang Y, Ding S, Song W. The relationship between the pan-immune-inflammation value and long-term prognoses in patients with hypertension: National Health and Nutrition Examination Study, 1999-2018. Front Cardiovasc Med 2023;10:1099427 [FREE Full text] [doi: 10.3389/fcvm.2023.1099427] [Medline: 36937901]
21. Simões FC, Riley PR. Immune cells in cardiac repair and regeneration. Development 2022;149(8) [FREE Full text] [doi: 10.1242/dev.199906] [Medline: 35502777]
22. Gubser PM, Vasanthakumar A. Going sugar free: Treg cells avoid glucose to maintain functional fitness. Immunol Cell Biol 2021;99(6):558-560. [doi: 10.1111/imcb.12461] [Medline: 34013589]
23. Núñez J, Núñez E, Bodí V, Sanchis J, Mainar L, Miñana G. Low lymphocyte count in acute phase of ST-segment elevation myocardial infarction predicts long-term recurrent myocardial infarction. Coron Artery Dis 2010;21(1):1-7. [doi: 10.1097/mca.0b013e328332ee15] [Medline: 20050312]
24. Tapp LD, Shantsila E, Wrigley BJ, Pamukcu B, Lip GYH. The CD14++CD16+ monocyte subset and monocyte-platelet interactions in patients with ST-elevation myocardial infarction. J Thromb Haemost 2012;10(7):1231-1241 [FREE Full text] [doi: 10.1111/j.1538-7836.2011.04603.x] [Medline: 22212813]
25. van der Laan AM, Hirsch A, Robbers LF, Nijveldt R, Lommerse I, Delewi R. A proinflammatory monocyte response is associated with myocardial injury and impaired functional outcome in patients with ST-segment elevation myocardial infarction: monocytes and myocardial infarction. Am Heart J 2012;163(1):57-65.e2 [FREE Full text] [doi: 10.1016/j.ahj.2011.09.002] [Medline: 22172437]
26. Liu T, Wang B, Xiao S, Sun L, Zhu Z, Wang S. Correlation analysis between the static and the changed neutrophil-to-lymphocyte ratio and in-hospital mortality in critical patients with acute heart failure. Postgrad Med 2023;135(1):50-57. [doi: 10.1080/00325481.2022.2129177] [Medline: 36154549]
27. Tamaki S, Nagai Y, Shutta R, Masuda D, Yamashita S, Seo M. Combination of neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios as a novel predictor of cardiac death in patients with acute decompensated heart failure with preserved left ventricular ejection fraction: a multicenter study. J Am Heart Assoc 2023;12(1):e026326 [FREE Full text] [doi: 10.1161/JAHA.122.026326] [Medline: 36565197]
28. Mohammed AA, Liu L, Mareai RM, Mohammed A, Yin G, Singh S. Combination of white blood cell count to mean platelet volume ratio and neutrophil-to-platelet ratio predicts long-term adverse events in patients with MINOCA. Mediators Inflamm 2022;2022:5642406 [FREE Full text] [doi: 10.1155/2022/5642406] [Medline: 36016662]
29. Doeleman MJH, Esseveld A, Huisman A, de Roock S, Tiel Groenestege WM. Stability and comparison of complete blood count parameters between capillary and venous blood samples. Int J Lab Hematol 2023;45(5):659-667. [doi: 10.1111/ijlh.14080] [Medline: 37128804]
30. MIMIC-IV. URL: https://physionet.org/content/mimiciv [accessed 2026-02-02]
31. ZhouTong202103/MIMIC. URL: https://gitee.com/zhoutong202103/mimic [accessed 2026-02-02]

Abbreviations

AMI: acute myocardial infarction
AUC: area under the curve
BIDMC: Beth Israel Deaconess Medical Center
CBC: complete blood count
CK-MB: creatine kinase isoenzymes
CVD: cardiovascular disease
DCA: decision curve analysis
HR: hazard ratio
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-9: International Classification of Diseases, Ninth Revision
ICU: intensive care unit
K-M: Kaplan-Meier
LMR: lymphocyte-to-monocyte ratio
MIMIC-IV: Medical Information Mart for Intensive Care IV
NLR: neutrophil-to-lymphocyte ratio
NPR: neutrophil-to-platelet ratio
PIV: prognostic inflammatory value
PLR: platelet-to-lymphocyte ratio
RCS: restricted cubic spline

XSL•FO
RenderX

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factor

Edited by A Benis; submitted 31.Aug.2025; peer-reviewed by M Al-Agil, Y Zhu, Y Chen; comments to author 11.Nov.2025; revised version received 06.Jan.2026; accepted 08.Jan.2026; published 12.Mar.2026.

Please cite as:
Zhou T, Yang K, Yang Y, Liu SM
Complete Blood Count–Derived Inflammation Indices to Predict 3-Year All-Cause Mortality in Patients With Diabetes and Acute Myocardial Infarction in Critical Care: Retrospective Cohort Study With Single-Center External Validation
JMIR Med Inform 2026;14:e83328
URL: https://medinform.jmir.org/2026/1/e83328
doi:10.2196/83328
PMID:

©Tong Zhou, Kun Yang, Ying Yang, Song-Mei Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Neutrophil Percentage–to-Albumin Ratio as a Novel Prognostic Biomarker in Adult Diffuse Gliomas: Retrospective Study Integrating 3 Machine Learning Models and Cox Regression

Congcong Zhu[1,2], MD; Jiyang An[3], PhD; Lili Zhou[1], PhD

[1]Department of Oncology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

[2]Department of Oncology, The First Affiliated Hospital, and College of Clinical Medicine of Henan University of Science and Technology, Luoyang, Henan, China

[3]Department of Neurosurgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

Corresponding Author:
Lili Zhou, PhD
Department of Oncology
The First Affiliated Hospital of Zhengzhou University
No. 1, East Jian She Road, Er Qi District
Zhengzhou, Henan, 450052
China
Phone: 86 176371002322
Email: zhoulili0@126.com

Abstract

Background: Adult-type diffuse glioma (ADG) is the most common primary malignant tumor of the central nervous system. Its highly invasive nature, marked heterogeneity, and resistance to therapy contribute to a high risk of recurrence and poor prognosis. At present, the lack of reliable prognostic tools poses a significant barrier to the development of individualized treatment strategies.

Objective: This study aimed to develop an effective prognostic model for ADG by integrating multiple machine learning algorithms, in order to enhance the precision of individualized clinical decision-making.

Methods: In this retrospective study, 160 newly diagnosed patients with ADG who underwent surgical resection and histopathological confirmation at our institution between June 2019 and September 2021 were included. A total of 32 variables, including clinical characteristics, molecular biomarkers, and preoperative hematological indicators, were collected. Overall survival (OS) and progression-free survival (PFS) were defined as the study endpoints. Feature selection was performed using least absolute shrinkage and selection operator regression, extreme gradient boosting, and random forest algorithms. Kaplan-Meier survival curves and log-rank tests were used for survival analysis. Multivariate Cox proportional hazards models were constructed to identify independent prognostic factors, and nomograms were developed accordingly. The model's discriminative ability, calibration, and clinical utility were evaluated using the concordance index, area under the receiver operating characteristic curve (area under the curve), calibration plots, and Kaplan-Meier analysis.

Results: Age, neutrophil percentage–to-albumin ratio (NPAR), and platelet-to-mean platelet volume ratio were identified as independent prognostic factors for OS, while age and NPAR were independent predictors for PFS (all $P<.001$). The prognostic models based on these variables demonstrated good predictive performance, with concordance index values of 0.731 and 0.763 for the training and validation cohorts in the OS model, respectively. The PFS model also showed robust performance. Area under the curve values and calibration curves further supported the models' accuracy and stability. Risk stratification analysis revealed clear survival differences between risk groups (all $P<.05$), indicating strong clinical applicability.

Conclusions: This study is the first to identify preoperative NPAR as a significant prognostic biomarker for ADG using machine learning approaches. The prognostic model incorporating NPAR, platelet-to-mean platelet volume ratio, and age demonstrated favorable predictive performance, offering a novel perspective for accurate risk stratification and personalized treatment in patients with ADG.

(JMIR Med Inform 2026;14:e79945) doi:10.2196/79945

XSL•FO
RenderX

KEYWORDS

adult-type diffuse glioma; machine learning; neutrophil percentage-to-albumin ratio; NPAR; prediction model; prognosis

Introduction

Glioma is the most common primary malignant tumor of the central nervous system, with an annual incidence of approximately 4.5 per 100,000 individuals [1]. Among its subtypes, adult-type diffuse glioma (ADG) constitutes the predominant pathological category, characterized by high invasiveness, frequent postoperative recurrence, and poor overall survival (OS) [2]. Epidemiological data indicate that ADG incidence exhibits age and sex specificity, with the highest prevalence observed in individuals aged 45-70 years (peak incidence at age 65-75 years) and a male predominance (male-to-female ratio: 1.3-1.6:1) [3].

Despite advances in multimodal treatment—including maximal surgical resection combined with radiochemotherapy—ADG survival outcomes have shown limited improvement over the past decade [4,5]. Novel therapies such as immune checkpoint inhibitors (eg, anti–PD-1 monoclonal antibodies) and targeted agents (eg, epidermal growth factor receptor inhibitors) offer promise, but their clinical benefits are restricted to specific molecular subgroups and are frequently accompanied by immune-related adverse events, such as pneumonitis (reported in 15%-20% of cases) and acquired resistance [6]. As such, there remains an urgent need for precise prognostic tools to identify treatment-sensitive subpopulations and guide individualized therapeutic decision-making.

Currently, prognostic evaluation of ADG largely depends on traditional clinicopathological parameters, including age, isocitrate dehydrogenase 1 (IDH1) mutation status, and O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation [7,8]. However, these markers alone exhibit limited predictive accuracy for treatment response or survival. Emerging evidence suggests that inflammation within the tumor microenvironment (TME) plays a critical role in glioma progression by promoting angiogenesis and suppressing antitumor immune activity [9]. Peripheral blood–derived inflammatory biomarkers, such as the neutrophil-to-lymphocyte ratio (NLR), lymphocyte-to-monocyte ratio (LMR), and platelet-to-lymphocyte ratio (PLR), have demonstrated prognostic relevance in several malignancies, including breast and colorectal cancer [10,11]. However, their utility in ADG remains underinvestigated.

Recently, the neutrophil percentage-to-albumin ratio (NPAR)—a composite marker integrating systemic inflammation and nutritional status—has been proposed as a novel prognostic indicator in solid tumors [12]. Nevertheless, its prognostic value in ADG, particularly in the context of long-term survival and treatment response, remains unclear.

Moreover, most existing prognostic models in gliomas are based on conventional Cox regression analysis, which is limited in handling high-dimensional data and complex nonlinear interactions [13,14]. In contrast, machine learning algorithms such as least absolute shrinkage and selection operator (LASSO) regression, extreme gradient boosting (XGBoost), and random forest (RF) have shown superior performance in variable selection and pattern recognition [15-18]. However, no studies to date have integrated these 3 algorithms to enhance the robustness of prognostic modeling in ADG.

Therefore, this study proposes an innovative machine learning–based approach that integrates LASSO regression (for feature selection), XGBoost (a gradient-boosted decision tree method), and RF (an ensemble learning model) to develop a multidimensional prognostic tool based on clinical, molecular, and inflammatory parameters. The objective is to improve predictive accuracy and provide refined support for clinical decision-making in patients with ADG.

Methods

Ethical Considerations

This retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of The First Affiliated Hospital of Zhengzhou University (Approval No 2023-KY-0223-002). Given the retrospective nature of this study, the requirement for informed consent was waived by the ethics committee. All patient data were anonymized and deidentified prior to analysis. Personal identifiers were removed, and each patient was assigned a unique study code to ensure confidentiality throughout the research process. No financial compensation was provided to the participants.

Study Population

This retrospective study enrolled 160 treatment-naïve patients with ADG who underwent maximal safe surgical resection with histopathological confirmation at our institution between June 2019 and September 2021. The postoperative treatment regimen followed the standard protocol, consisting of concurrent chemoradiotherapy followed by adjuvant temozolomide chemotherapy. In cases of disease progression, targeted therapies such as bevacizumab or BRAF (B-Raf proto-oncogene, serine/threonine kinase) inhibitors (eg, for tumors harboring the BRAF V600E mutation) were administered based on clinical indications [19].

Inclusion criteria were as follows: (1) age ≥18 years, diagnosis of ADG in accordance with the 5th edition of the *WHO Classification of Tumors of the Central Nervous System* (World Health Organization [WHO] grade 3-4); (2) completion of preoperative magnetic resonance imaging or computed tomography evaluation; (3) no prior antitumor therapy; (4) receipt of standard treatment protocol; (5) availability of complete clinical data, including general characteristics, preoperative blood tests, cardiac, hepatic, and renal function assessments, as well as postoperative immunohistochemistry and molecular profiling; and (6) complete follow-up records.

Exclusion criteria were (1) postoperative mortality within 30 days due to surgical complications or nonneoplastic causes; (2) presence of severe systemic disease or active infection

XSL•FO
RenderX

potentially affecting study outcomes; (3) diagnosis of other concurrent malignancies; (4) significant cardiac, hepatic, or renal dysfunction; and (5) history of prior cranial surgery.

Postoperative follow-up included clinical and radiological evaluations every 3 months during the first year and every 6 months thereafter, until October 1, 2024. Progression-free survival (PFS) was defined as the time from diagnosis to radiographic or clinical disease progression or death from any cause. OS was defined as the time from diagnosis to death or the last follow-up.

Study Variables

Clinical data were collected from patients with ADG who met the inclusion and exclusion criteria. The variables analyzed in this study included: clinical characteristics, molecular biomarkers, genetic alteration markers, preoperative peripheral hematological indicators, and inflammation-related ratios.

Clinical Characteristics

Age, gender, preoperative tumor location, extent of tumor resection, maximum tumor diameter, and midline shift.

Molecular Biomarkers

Immunohistochemical markers included P53 protein (P53), oligodendrocyte transcription factor 2 (Olig-2), S100 calcium-binding protein (S100), epithelial membrane antigen (EMA), synuclein (SYN), alpha-thalassemia/mental retardation syndrome X-linked (ATRX), cluster of differentiation 34 (CD34), neuronal nuclei (NeuN), glial fibrillary acidic protein (GFAP), and proliferation cell nuclear antigen-67 (Ki-67).

Genetic Alteration Markers

IDH1 immunophenotype, telomerase reverse transcriptase (TERT) promoter region mutation, BRAF V600E point mutation, epidermal growth factor receptor (EGFR) amplification, MGMT promoter methylation status, 1p loss of heterozygosity (1pLOH), and 19q loss of heterozygosity (19qLOH).

Preoperative Peripheral Hematological Indicators

White blood cell (WBC), platelet (PLT), neutrophil (Neut), monocyte (Mono), lymphocyte (Lym), prothrombin time (PT), activated partial thromboplastin time (APTT), lactate dehydrogenase (LDH), and serum albumin (ALB). All preoperative peripheral hematological indicators, including NPAR, were measured from blood samples collected within 1 week prior to surgery, before the administration of any corticosteroids or other interventions that could significantly alter these values.

Inflammation-Related Ratios

According to international standards for inflammatory biomarkers, the following parameters were calculated and defined: PLR = PLT/Neut, NLR = Neut/Lym, LMR = Lym/Mono, monocyte-to-lymphocyte ratio (MLR) = Mono/Lym, prognostic nutritional index (PNI) = ALB + 5 × Lym, systemic inflammation response index (SIRI) = (Neut × Mono)/Lym, systemic immune-inflammation index (SII) = (PLT × Neut)/Lym, aggregate index of systemic inflammation (AISI) = (Neut × PLT × Mono)/Lym, NPAR = (Neut/WBC%)/ALB, neutrophil-lymphocyte-platelet ratio (NLPR) = Neut/(Lym × PLT).

Study Design and Statistical Analysis

The study design is illustrated in Figure 1.

A random seed was set to divide the 160 cases into a training cohort (n=112) and a validation cohort (n=48) at a 7:3 ratio. The primary endpoint was OS, and the secondary endpoint was PFS. A total of 32 variables were included, encompassing clinical characteristics (age, gender, tumor location, etc), molecular biomarkers (Ki-67, IDH1, R132H, etc), and inflammatory indicators (NPAR, NLR, etc).

Continuous variables were dichotomized using the median as the cutoff. This approach was chosen to simplify the clinical interpretation of the model and to facilitate risk stratification, while also mitigating the impact of potential outliers and meeting the model's assumption of proportional hazards for categorical predictors. Categorical variables were expressed as frequencies (percentages), and group comparisons were conducted using the chi-square test (expected frequency ≥5) or Fisher exact test (expected frequency <5). The Shapiro-Wilk test was used to assess the normality of continuous variables. For normally distributed data, comparisons between groups were performed using the 2-tailed independent-samples *t* test (2 groups) or 1-way ANOVA (3 or more groups), with data presented as mean (SD). For nonnormally distributed data, comparisons were performed using the Mann-Whitney *U* test (2 groups) or the Kruskal-Wallis H test (3 or more groups), with post hoc pairwise comparisons adjusted using the Bonferroni method. These data were reported as median (IQR). A 2-tailed *P* value <.05 was considered statistically significant for all analyses.

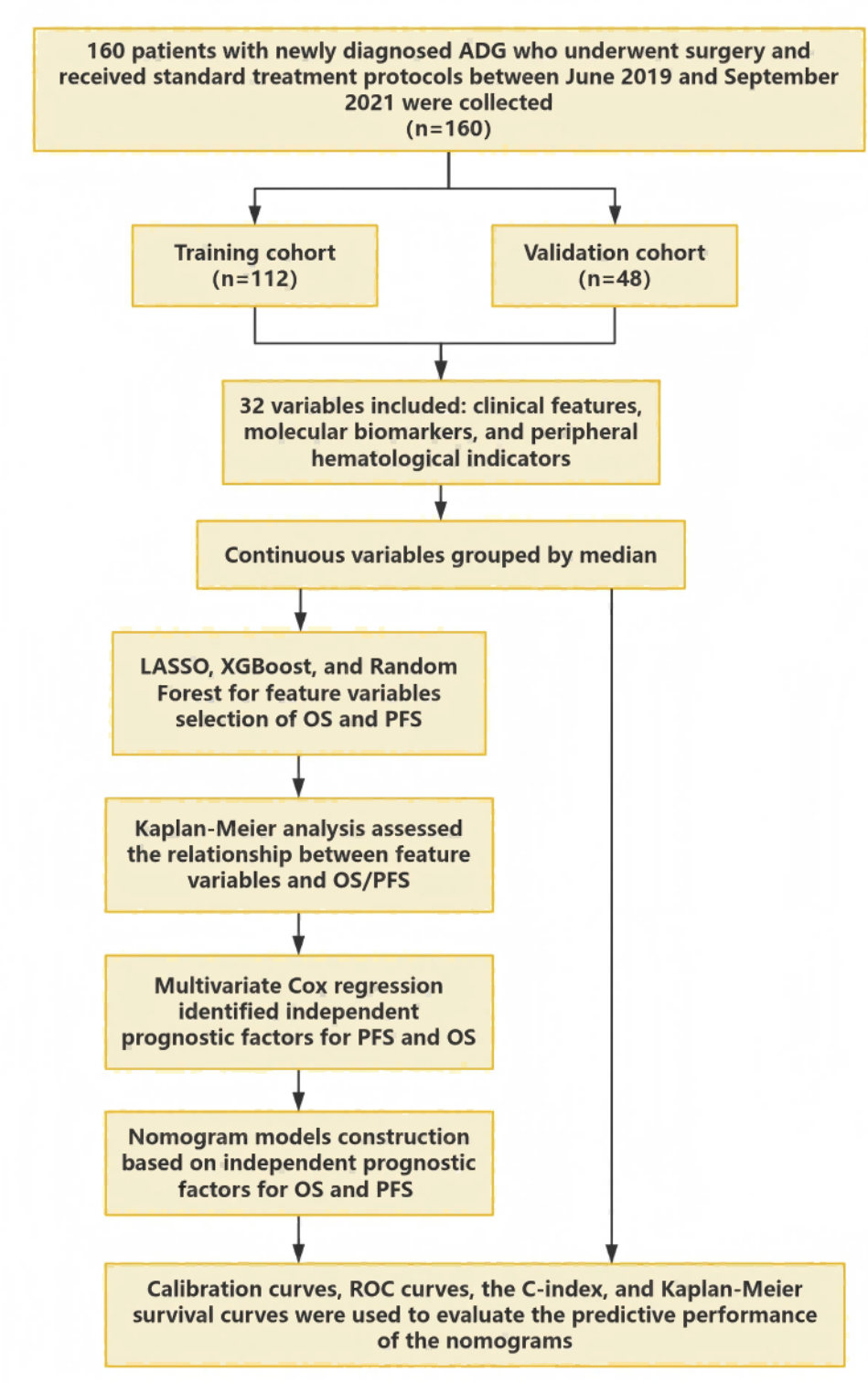
A total of 3 machine learning algorithms were applied: LASSO regression (using the *glmnet* package with 10-fold cross-validation to optimize λ), XGBoost (using the *xgboost* package with max depth=6 and learning rate=0.01), and RF (using the *randomForest* package with *ntree*=1000). Venn diagrams were used to identify overlapping important variables selected by all 3 algorithms. Survival differences between groups were visualized using Kaplan-Meier (KM) curves and compared using the log-rank test.

A Cox proportional hazards regression model was constructed to estimate hazard ratios (HRs) and 95% CIs. Based on multivariate Cox results, a nomogram was developed using the *rms* package. Time-dependent receiver operating characteristic (ROC) curves were generated, and the area under the curve (AUC) at 1, 3, and 5 years was calculated. The concordance index (C-index) was used to evaluate the model's discrimination ability. Calibration was assessed using calibration curves based on 1000 bootstrap resamples. KM survival analysis based on the median nomogram score was performed to evaluate clinical utility. The validation cohort was used to assess the model's predictive performance.

All statistical analyses were conducted using R software (version 4.4.1; R Foundation for Statistical Computing). A 2-sided *P* value <.05 was considered statistically significant.

XSL•FO
RenderX

Figure 1. Flowchart of the study design.



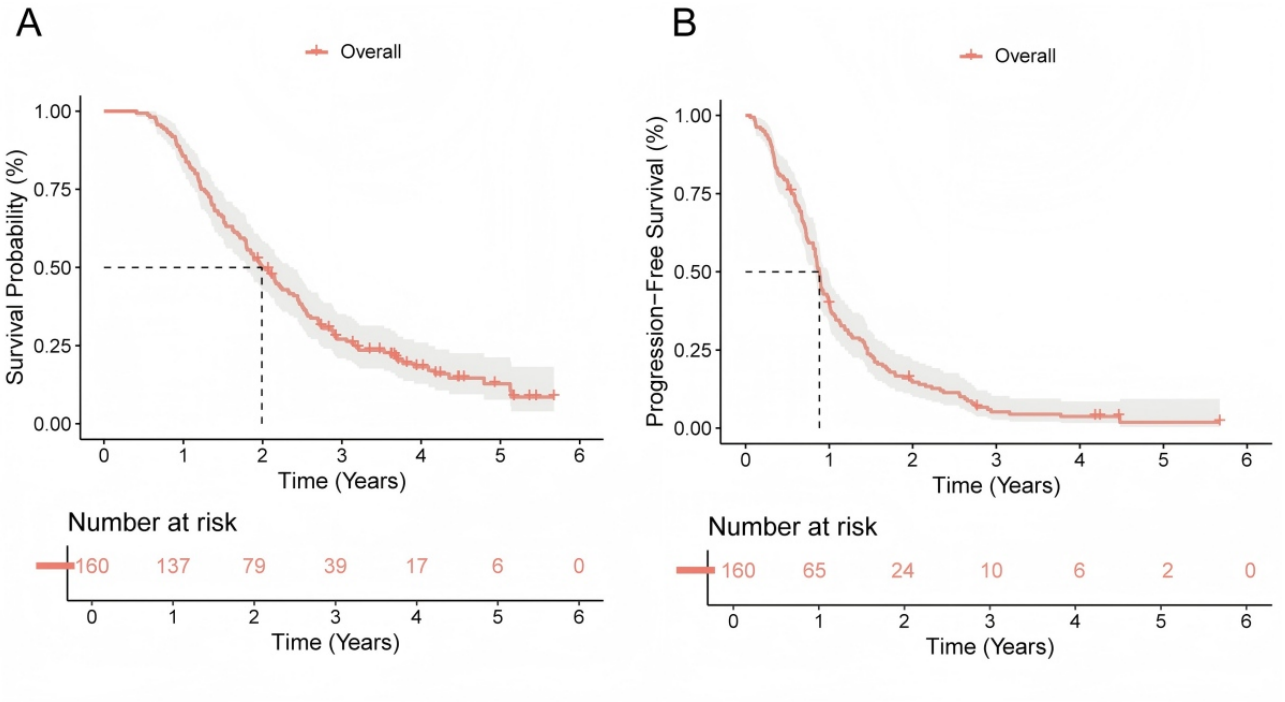
Results

Survival Outcomes

Among the 160 patients included in this study, a total of 132 (82.5%) experienced death events, and 152 (95%) experienced disease progression. The median OS was 23.0 months (IQR 15.0-38.3 months), while the median PFS was 10.5 months (IQR 6.8-17.5 months). The 1-, 3-, and 5-year OS rates were 85.6%, 27.1%, and 12.8%, respectively. The corresponding 1-, 3-, and 5-year PFS rates were 40.0%, 6.5%, and 3.4%, respectively. KM survival curves are presented in Figure 2.

XSL•FO
RenderX

Figure 2. Postoperative survival curves of 160 patients with adult-type diffuse glioma: (A) overall survival (OS) and (B) progression-free survival (PFS).



Baseline Characteristics

A total of 160 patients were included in this study, with a male-to-female ratio of approximately 1.4:1. The mean age was 51.3 (SD 13.6) years, and the median age was 53 (IQR 44-61) years. The maximum tumor diameter ranged from 1.0 to 9.5 cm, with a mean of 4.7 (SD 1.7) cm and a median of 5.0 cm (IQR 3.4-5.9 cm). Tumors were predominantly unilateral, with 76 cases (47.5%) on the left, 80 (50%) on the right, and 2 (2.5%) involving both sides. Midline shift was observed in 59.4% (95/160) of patients. All cases were classified as WHO grade 3-4. No significant differences in baseline characteristics were found between the training and validation cohorts (all $P>.05$). Detailed information is provided in Table 1.

XSL•FO
RenderX

Table 1. Baseline characteristics of 160 patients with adult-type diffuse glioma (ADG) stratified by group.

Variable	Training group, n (%)	Validation group, n (%)	*P* value
Age (years)			.58
<53	56 (50)	21 (44)	
≥53	56 (50)	27 (56)	
Sex			.06
Male	71 (63.4)	22 (46)	
Female	41 (36.6)	26 (54)	
Maximum tumor diameter (cm)			.89[a]
<5.0	55 (49.1)	23 (47)	
≥5.0	57 (50.9)	25 (52)	
Location			.74[b]
Left	54 (48.2)	22 (46)	
Right	56 (50.0)	24 (50)	
Double	2 (1.8)	2 (4)	
Midline shift			.16
No	41 (36.6)	24 (50)	
Yes	71 (63.4)	24 (50)	
Neutrophil percentage–to-albumin ratio			.92[b]
<1.578	55 (49.1)	24 (50)	
≥1.578	57 (50.9)	24 (50)	
Aggregate index of systemic inflammation			.60
<272.854	58 (51.8)	22 (46)	
≥272.854	54 (48.2)	26 (54)	
Immune-inflammation index			.86
<146.649	55 (49.1)	25 (52)	
≥146.649	57 (50.9)	23 (48)	
Lymphocyte-to-monocyte ratio			.39
<3.750	59 (52.7)	21 (44)	
≥3.750	53 (47.3)	27 (56)	
Neutrophil-to-lymphocyte ratio			.12
<2.677	51 (45.5)	29 (60)	
≥2.677	61 (54.5)	19 (40)	
Platelet-to-lymphocyte ratio			.23
<138.251	52 (46.4)	28 (58)	
≥138.251	60 (53.6)	20 (42)	
Monocyte-to-lymphocyte ratio			.39
<0.267	53 (47.3)	27 (56)	
≥0.267	59 (52.7)	21 (44)	
Prognostic nutritional index			.47
<50.55	52 (46.4)	26 (54)	
≥50.55	60 (53.6)	22 (46)	
Systemic inflammation response index			.86

XSL•FO
RenderX

Variable	Training group, n (%)	Validation group, n (%)	*P* value
<1.166	55 (49.1)	25 (52)	
≥1.166	57 (50.9)	23 (48)	
L actate dehydrogenase (IU/L)			.54
<174	53 (47.3)	26 (54)	
≥174	59 (52.7)	22 (46)	
Albumin (g/L)			.12
<42.35	51 (45.5)	29 (60)	
≥42.35	61 (54.5)	19 (40)	
Prothrombin time/Activated partial thromboplastin time			.39
<0.372	53 (47.3)	27 (56)	
≥0.372	59 (52.7)	21 (44)	
Platelet-to-mean platelet volume ratio			.12
<25.322	61 (54.5)	19 (40)	
≥25.322	51 (45.5)	29 (60)	
Proliferation cell nuclear antigen-67			.16
<30	28 (25)	18 (37)	
≥30	84 (75)	30 (63)	
Glial fibrillary acidic protein			>.99[b]
Positive	108 (96.4)	47 (98)	
Negative	4 (3.6)	1 (2)	
S100 calcium-binding protein			.74[b]
Positive	105 (93.8)	44 (92)	
Negative	7 (6.2)	4 (8)	
Oligodendrocyte transcription factor 2			>.99[b]
Positive	107 (95.5)	46 (96)	
Negative	5 (4.5)	2 (4)	
P53 protein			.09
Positive	103 (92)	39 (81)	
Negative	9 (8)	9 (19)	
Alpha-thalassemia/mental retardation syndrome X-linked			.27
Positive	96 (85.7)	37 (77)	
Negative	16 (14.3)	11 (23)	
Cluster of differentiation 34			.66
Positive	34 (30.4)	17 (35)	
Negative	78 (69.6)	31 (65)	
Epithelial membrane antigen			.56[b]
Positive	12 (10.7)	3 (6)	
Negative	100 (89.2)	45 (94)	
Synuclein			.40
Positive	22 (19.6)	13 (27)	
Negative	90 (80.4)	35 (73)	
Neuronal nuclei			>.99[b]

XSL•FO
RenderX

Variable	Training group, n (%)	Validation group, n (%)	*P* value
Positive	2 (1.9)	1 (2)	
Negative	110 (58.1)	47 (98)	
O6-methylguanine-DNA methyltransferase promoter methylation			>.99[b]
Methylated	6 (5.4)	2 (4)	
Unmethylated	106 (94.6)	46 (96)	
Isocitrate dehydrogenase 1 R132			.78[b]
Positive	12 (10.7)	4 (8)	
Negative	100 (89.2)	44 (92)	
1p loss of heterozygosity			.07[b]
Positive	2 (1.9)	4 (8)	
Negative	110 (58.1)	44 (92)	
19q loss of heterozygosity			.74[b]
Positive	7 (6.2)	4 (8)	
Negative	105 (93.8)	44 (92)	

[a]Indicates without continuity correction.

[b]Indicates Fisher exact test.

Differences in PFS and OS Between Groups

Since the measurement data did not meet the assumption of normality according to the Shapiro-Wilk test, the nonparametric Wilcoxon rank-sum test was applied to evaluate the impact of different clinical characteristics on PFS and OS. Data are presented as median (IQR).

Using PFS as the outcome variable, significant differences (all $P<.05$) were observed in the following 10 variables: age ($P<.001$), tumor location ($P=.03$), NPAR ($P=.002$), NLR ($P=.002$), PNI ($P=.01$), SYN ($P=.003$), MGMT promoter methylation ($P=.004$), IDH1 R132 ($P<.001$), 1pLOH ($P=.04$), and 19qLOH ($P<.001$).

Using OS as the outcome variable, a total of 12 variables showed significant associations (all $P<.05$): age ($P<.001$), NPAR ($P<.001$), LMR ($P=.03$), NLR ($P=.02$), MLR ($P=.03$), PNI ($P=.01$), LDH ($P=.04$), SYN ($P=.009$), MGMT ($P=.04$), IDH1 R132 ($P<.001$), 1pLOH ($P=.003$), and 19qLOH ($P<.001$). Detailed results are presented in Table 2.

XSL•FO
RenderX

Table 2. Comparison of progression-free survival (PFS) and overall survival (OS) among patients with glioblastoma between groups, median (IQR). The z and H values denote the test statistics for the Wilcoxon rank-sum test used in 2-group comparisons and the Kruskal-Wallis test used in multigroup comparisons, respectively.

Variable	Frequency, n	PFS (year), median (IQR)	z or H	*P* value	OS (year), median (IQR)	z or H	*P* value
Age (years)			4.42[a]	<.001		6.67[a]	<.001
<53	77	1.02 (0.74-2.37)			2.76 (1.96-3.67)		
≥53	83	0.73 (0.36-1.07)			1.36 (1.03-2.07)		
Sex			0.49[a]	.63		0.16[a]	.88
Male	93	0.90 (0.63-1.43)			2.03 (1.22-3.06)		
Female	67	0.84 (0.52-1.61)			1.95 (1.29-2.90)		
Maximum tumor diameter (cm)			–1.14[a]	.25		–1.04[a]	.30
<5.0	78	0.85 (0.59-1.45)			1.87 (1.21-2.91)		
≥5.0	82	0.92 (0.54-1.49)			2.13 (1.33-3.02)		
Location			7.050[b]	.03		3.26 [b]	.20
Left	76	0.90 (0.65-1.54)			2.05 (1.35-2.89)		
Right	80	0.86 (0.59-1.46)			1.98 (1.21-3.21)		
Double	4	0.33 (0.27-0.41)			1.13 (0.82-1.62)		
Midline-shift			–0.54[a]	.59		–0.86[a]	.39
No	65	0.85 (0.55-1.52)			1.80 (1.21-2.91)		
Yes	95	0.90 (0.58-1.45)			2.09 (1.33-3.00)		
Neutrophil percentage–to-albumin ratio			2.86[a]	.004		3.73[a]	<.001
<1.578	79	0.99 (0.67-1.74)			2.47 (1.53-3.58)		
≥1.578	81	0.83 (0.36-1.18)			1.72 (1.03-2.44)		
Aggregate index of systemic inflammation			1.35[a]	.18		1.28[a]	.20
<272.854	80	0.89 (0.67-1.47)			2.13 (1.32-3.38)		
≥272.854	80	0.88 (0.39-1.46)			1.91 (1.26-2.71)		
Immune-inflammation index			1.69[a]	.09		0.68[a]	.498
<146.649	80	0.89 (0.66-1.72)			2.07 (1.22-2.99)		
≥146.649	80	0.83 (0.50-1.24)			1.95 (1.30-2.90)		
Lymphocyte-to-monocyte ratio			–0.59[a]	.56		–2.13[a]	.03
<3.750	80	0.85 (0.53-1.43)			1.85 (1.22-2.52)		
≥3.750	80	0.89 (0.59-1.54)			2.45 (1.33-3.49)		
Neutrophil-to-lymphocyte ratio			2.83[a]	.005		2.41[a]	.02
<2.677	80	0.94 (0.67-1.81)			2.33 (1.35-3.56)		
≥2.677	80	0.83 (0.35-1.24)			1.85 (1.15-2.52)		
Platelet-to-lymphocyte ratio			0.85[a]	.40		0.47[a]	.64
<138.251	81	0.88 (0.61-1.54)			2.07 (1.22-3.13)		
≥138.251	79	0.88 (0.48-1.44)			1.94 (1.30-2.92)		
Monocyte-to-lymphocyte ratio			0.59[a]	.56		2.13[a]	.03
<0.267	80	0.89 (0.59-1.54)			2.45 (1.33-3.49)		
≥0.267	80	0.85 (0.53-1.43)			1.85 (1.22-2.52)		
Prognostic nutritional index			–2.26[a]	.02		–3.27[a]	.001

XSL•FO
RenderX

Variable	Frequency, n	PFS (year), median (IQR)	z or H	*P* value	OS (year), median (IQR)	z or H	*P* value
<50.55	78	0.83 (0.37-1.24)			1.79 (1.04-2.49)		
≥50.55	82	0.98 (0.64-1.61)			2.35 (1.51-3.21)		
Systemic inflammation response index			1.63[a]	.10		1.82[a]	.07
<1.166	80	0.90 (0.67-1.53)			2.39 (1.33-3.49)		
≥1.166	80	0.84 (0.38-1.46)			1.87 (1.22-2.56)		
Lactate dehydrogenase			1.04[a]	.30		2.11[a]	.04
<174	79	0.90 (0.51-1.79)			2.21 (1.34-3.21)		
≥174	81	0.85 (0.60-1.19)			1.80 (1.22-2.61)		
Albumin			–0.62[a]	.54		–1.61[a]	.11
<42.35	80	0.85 (0.64-1.27)			1.80 (1.21-2.63)		
≥42.35	80	0.92 (0.50-1.70)			2.15 (1.38-3.07)		
Prothrombin time/activated partial thromboplastin time			–0.12[a]	.91		0.52[a]	.60
<0.372	80	0.87 (0.51-1.47)			1.98 (1.21-3.56)		
≥0.372	80	0.88 (0.61-1.41)			2.00 (1.38-2.75)		
Platelet-to-mean platelet volume ratio			1.03[a]	.30		0.10[a]	.92
<25.322	80	0.88 (0.66-1.68)			1.96 (1.29-2.92)		
≥25.322	80	0.87 (0.50-1.31)			2.01 (1.26-2.97)		
Proliferation cell nuclear antigen-67			1.08[a]	.28		1.33[a]	.19
<30	46	0.92 (0.62-1.72)			2.36 (1.37-3.79)		
≥30	114	0.87 (0.52-1.43)			1.92 (1.24-2.74)		
Glial fibrillary acidic protein			0.21[a]	.84		1.09[a]	.28
Positive	155	0.88 (0.58-1.46)			1.96 (1.25-2.92)		
Negative	5	1.28 (0.21-2.76)			2.55 (2.22-4.35)		
S100 calcium-binding protein			–0.57[a]	.57		0.37[a]	.71
Positive	149	0.88 (0.59-1.46)			1.98 (1.22-2.92)		
Negative	11	0.88 (0.35-1.29)			2.55 (1.42-3.00)		
Oligodendrocyte transcription factor 2			–0.74[a]	.46		0.15[a]	.89
Positive	153	0.88 (0.58-1.46)			2.03 (1.22-2.92)		
Negative	7	0.83 (0.55-0.94)			1.89 (1.76-2.54)		
P53 protein			1.08[a]	.28		0.58[a]	.57
Positive	142	0.87 (0.58-1.43)			1.98 (1.22-2.89)		
Negative	18	1.13 (0.55-2.36)			1.97 (1.33-3.18)		
Alpha-thalassemia/mental retardation syndrome X-linked			1.71[a]	.09		1.48[a]	.14
Positive	133	0.85 (0.54-1.43)			1.94 (1.22-2.76)		
Negative	27	1.13 (0.70-2.35)			2.73 (1.40-3.55)		
Cluster of differentiation 34			–0.39[a]	.70		–0.47[a]	.64
Positive	51	0.85 (0.60-1.57)			2.07 (1.34-3.03)		
Negative	109	0.89 (0.51-1.43)			1.95 (1.22-2.89)		
Epithelial membrane antigen			–1.27[a]	.20		–0.51[a]	.61
Positive	15	1.02 (0.85-1.59)			1.81 (1.45-3.29)		

XSL•FO
RenderX

Variable	Frequency, n	PFS (year), median (IQR)	z or H	*P* value	OS (year), median (IQR)	z or H	*P* value
Negative	145	0.86 (0.54-1.46)			2.03 (1.22-2.92)		
Synuclein			2.86[a]	.004		2.63[a]	.009
Positive	35	0.67 (0.35-0.92)			1.44 (1.01-2.39)		
Negative	125	0.91 (0.64-1.54)			2.15 (1.35-3.14)		
Neuronal nuclei			–1.11[a]	.27		–0.35[a]	.73
Positive	3	1.72 (1.27-1.83)			1.96 (1.95-2.35)		
Negative	157	0.88 (0.55-1.46)			1.99 (1.22-2.94)		
O6-methylguanine-DNA methyltransferase promoter methylation			–2.86[a]	.004		–2.09[a]	.04
Positive	8	1.96 (1.25-2.44)			2.87 (2.16-3.63)		
Negative	152	0.86 (0.54-1.43)			1.94 (1.22-2.90)		
Isocitrate dehydrogenase 1 R132			–4.57[a]	<.001		–4.36[a]	<.001
Positive	16	1.88 (1.39-3.10)			3.34 (2.84-3.82)		
Negative	144	0.85 (0.51-1.23)			1.85 (1.21-2.72)		
1p loss of heterozygosity			–1.98[a]	.049		–3.00[a]	.003
positive	6	2.08 (1.19-2.52)			3.69 (3.31-4.27)		
negative	154	0.87 (0.56-1.43)			1.94 (1.22-2.89)		
19q loss of heterozygosity			–3.76[a]	<.001		–3.69[a]	<.001
positive	11	1.79 (1.32-3.73)			3.68 (2.63-4.19)		
negative	149	0.85 (0.53-1.42)			1.89 (1.21-2.84)		

[a]z values determined using the Wilcoxon rank sum test with two-group comparisons.

[b]H values determined using the Kruskal-Wallis test with multiple-group comparisons.

Construction and Validation of an OS-Based Prediction Model

Feature Variable Selection

Using OS as the outcome variable, a total of 3 machine learning algorithms—LASSO regression, XGBoost, and RF—were used to screen feature variables.

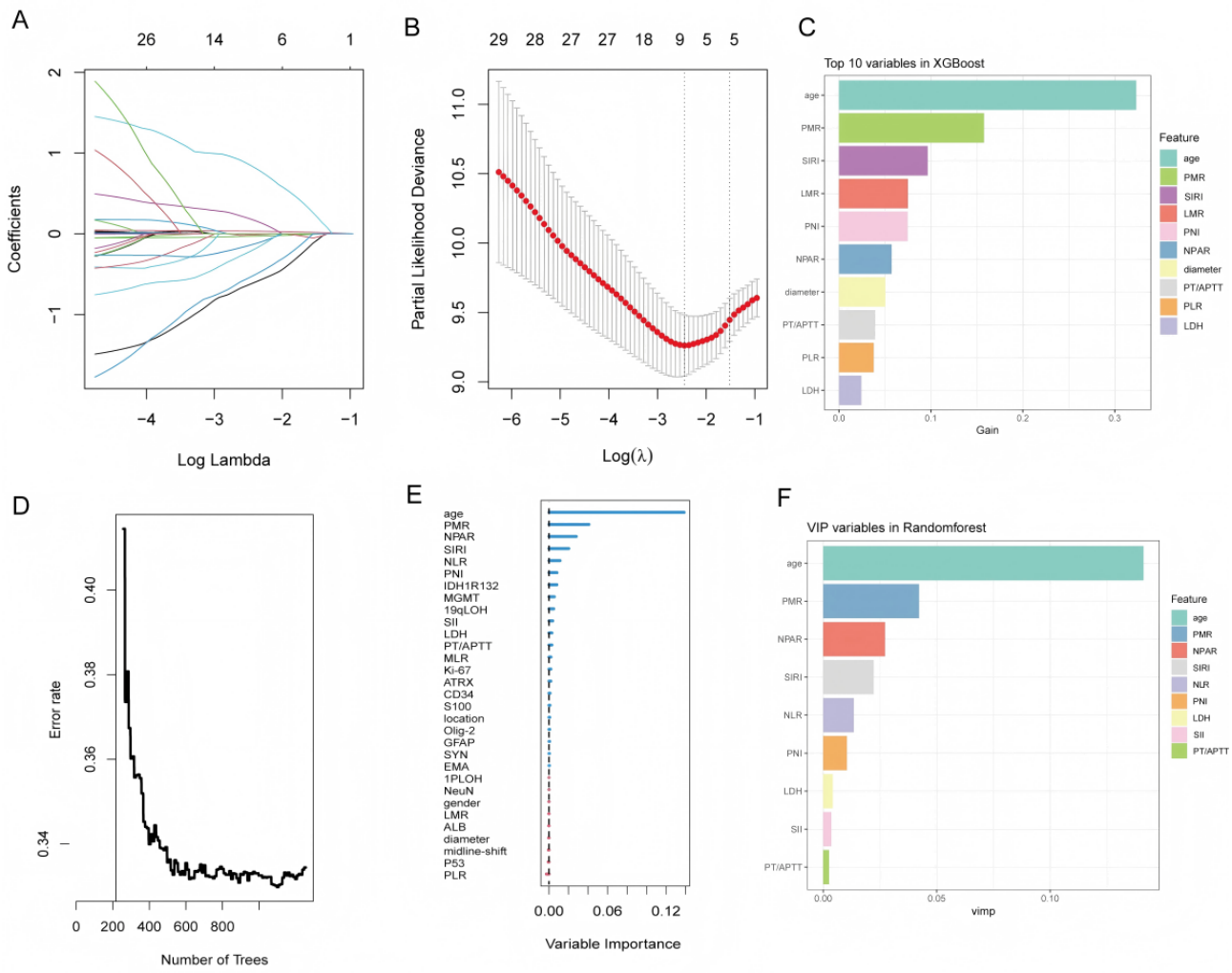
LASSO regression: based on the minimum λ value (λ_min), a total of 9 variables with nonzero coefficients were selected, including age, NPAR, platelet-to-mean platelet volume ratio (PMR), CD34 expression, SYN expression, EMA expression, IDH1 R132, 19qLOH, and MGMT promoter methylation (Figures 3A and B).

XGBoost model: according to feature importance scores, a total of 10 variables were selected, including age, PMR, SIRI, LMR, PNI, NPAR, maximum tumor diameter, PT/APTT, PLR, and LDH (Figure 3C).

RF: based on the mean decrease in Gini index, a total of 9 variables were selected, including age, PMR, NPAR, SIRI, NLR, PNI, LDH, SII, and PT/APTT (Figures 3D-F).

XSL•FO
RenderX

Figure 3. Feature variable selection based on 3 machine learning methods: (A) least absolute shrinkage and selection operator (LASSO) regression coefficient path plot: variable coefficients as a function of λ; (B) LASSO regression λ value tuning: 10-fold cross-validation to select the minimum λ value (λ_min); (C) extreme gradient boosting (XGBoost) feature importance: top 10 variables; (D) random forest (RF) error rate change: error rate as a function of the number of decision trees; (E) feature importance: assessed by Gini index; and (F) RF selected variables: top 9 variables. 19qLOH: 19q loss of heterozygosity; 1pLOH: 1p loss of heterozygosity; ALB: albumin; ATRX: alpha-thalassemia/mental retardation syndrome X-linked; CD34: cluster of differentiation 34; EMA: epithelial membrane antigen; GFAP: glial fibrillary acidic protein; IDH1R132: isocitrate dehydrogenase 1 R132; Ki-67: proliferation cell nuclear antigen-67; LDH: lactate dehydrogenase; LMR: lymphocyte-to-monocyte ratio; MGMT: O6-methylguanine-DNA methyltransferase; MLR: monocyte-to-lymphocyte ratio; NeuN: neuronal nuclei; NLR: neutrophil-to-lymphocyte ratio; NPAR: neutrophil percentage-to-albumin ratio; Olig-2: oligodendrocyte transcription factor 2; P53: P53 protein; PLR: platelet-to-lymphocyte ratio; PMR: platelet-to-mean platelet volume ratio; PNI: prognostic nutritional index; PT/APTT: prothrombin time/activated partial thromboplastin time; S100: S100 calcium-binding protein; SII: immune-inflammation index; SIRI: systemic inflammation response index; SYN: synuclein.



Survival Analysis

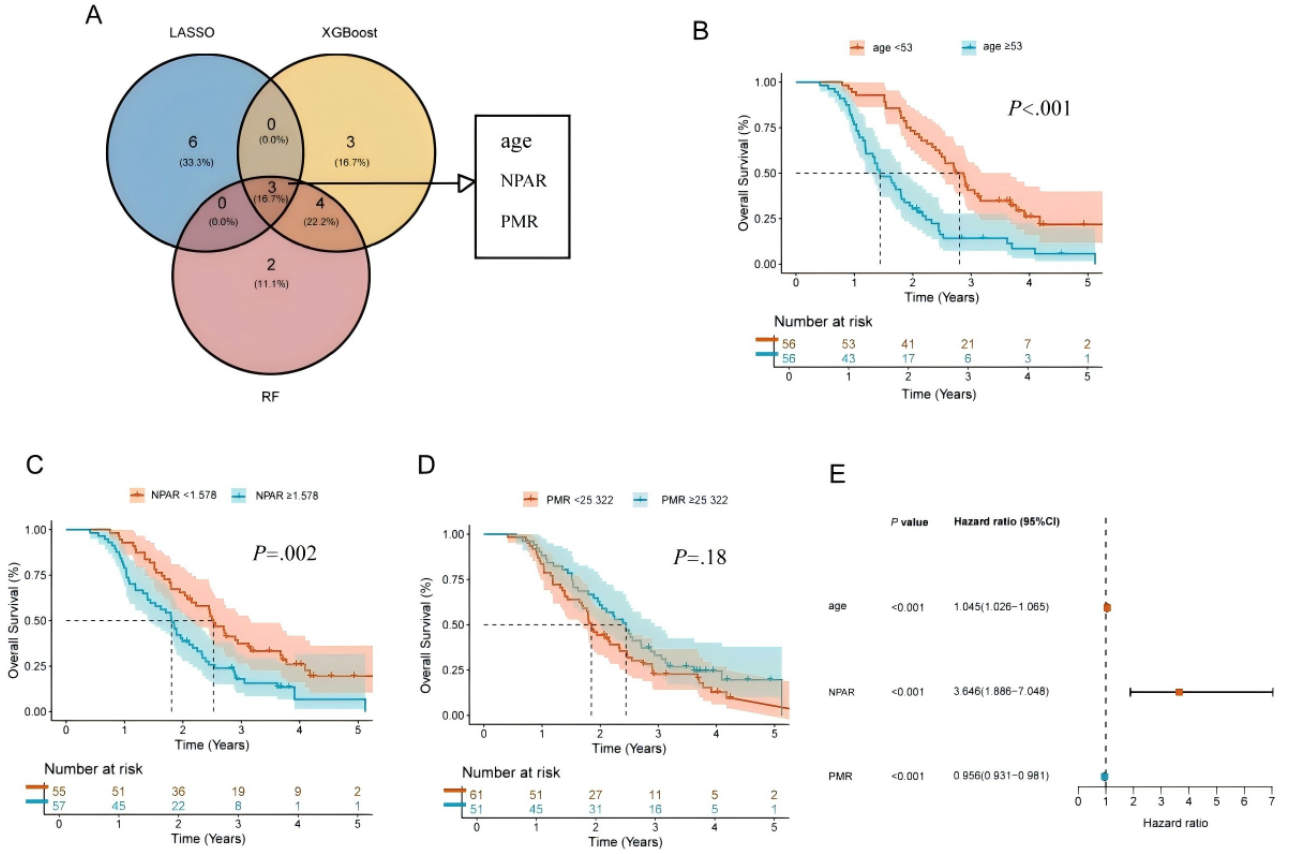
Based on the combined screening results from LASSO, XGBoost, and RF, Venn diagram intersection analysis identified age, NPAR, and PMR as key factors influencing OS (Figure 4A).

The prognostic impact of each individual variable was assessed using KM survival analysis (Figures 4B-D). Log-rank test results indicated that patients aged <53 years had significantly longer OS compared to those aged ≥53 years (P<.001), patients with NPAR <1.578 had significantly better OS than those with NPAR ≥1.578 (P<.001), while no significant difference in OS was observed between patients with PMR <25.322 and those with PMR ≥25.322 (P=.18).

Multivariate Cox proportional hazards regression analysis identified age (HR 1.05, 95% CI 1.03-1.07; P<.001), NPAR (HR 3.65, 95% CI 1.89-7.05; P<.001), and PMR (HR 0.96, 95% CI 0.93-0.98; P<.001) as independent prognostic factors for OS.

XSL•FO
RenderX

Figure 4. Identification of independent prognostic factors for overall survival (OS): (A) intersection analysis of feature variables from 3 models, (B) impact of age group on patient OS, (C) impact of neutrophil percentage-to-albumin ratio (NPAR) level on patient OS, (D) impact of platelet-to-mean platelet volume ratio (PMR) level on patient OS, and (E) forest plot of multivariate Cox regression analysis for OS. LASSO: least absolute shrinkage and selection operator.



Construction and Validation of the OS Nomogram

Based on the multivariate Cox regression analysis, age, NPAR, and PMR were identified as independent prognostic factors for OS. A nomogram was constructed using these variables to predict 1-, 3-, and 5-year OS probabilities. Each variable was assigned a corresponding risk score, and the total score was used to estimate individual survival probability (Figure 5A). The calibration curve showed good agreement between the predicted and observed 1-, 3-, and 5-year OS rates (Figure 5B). Time-dependent ROC analysis demonstrated that the AUCs for 2-, 3-, and 4-year OS in the training cohort were 0.836, 0.820, and 0.801, respectively, indicating strong predictive performance across multiple timepoints (Figure 5C).

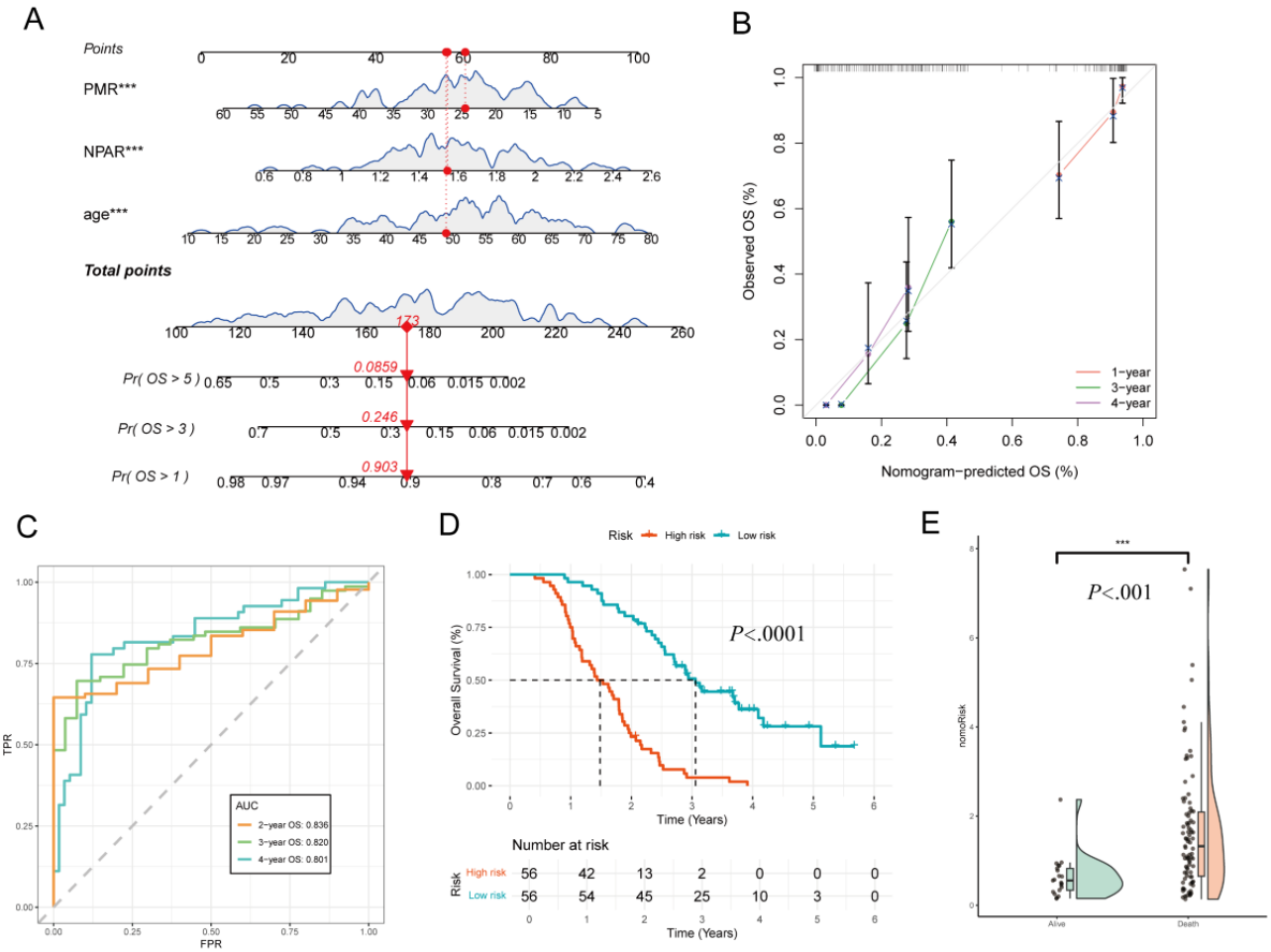
Patients were stratified into high- and low-risk groups based on the median nomogram score. KM survival analysis revealed that the high-risk group had significantly poorer OS compared with the low-risk group (P<.001), supporting the nomogram's utility in risk stratification (Figure 5D).

To further evaluate the discriminative ability of the nomogram score, a raincloud plot was used to visualize the score distribution between the survival and death groups. This plot, which integrates a density plot, box plot, and scatter plot, clearly illustrated the differences in score distribution, providing additional evidence of the nomogram's effectiveness in prognostic assessment (Figure 5E).

The calibration curve in the validation group indicated good predictive accuracy of the model for 1-, 2-, and 3-year survival in patients with ADG (Figure 6A). Time-dependent ROC analysis showed that the AUCs for 1-, 3-, and 4-year OS were 0.822, 0.797, and 0.722, respectively (Figure 6B). Risk stratification based on the nomogram score revealed a significant difference in score distribution between survivors and nonsurvivors. Patients in the high-risk group had significantly poorer OS compared with those in the low-risk group (P<.01), supporting the prognostic discriminatory power of the nomogram (Figure 6C). The C-index remained high in both the training and validation sets, suggesting robust and consistent predictive performance (Figure 6D).

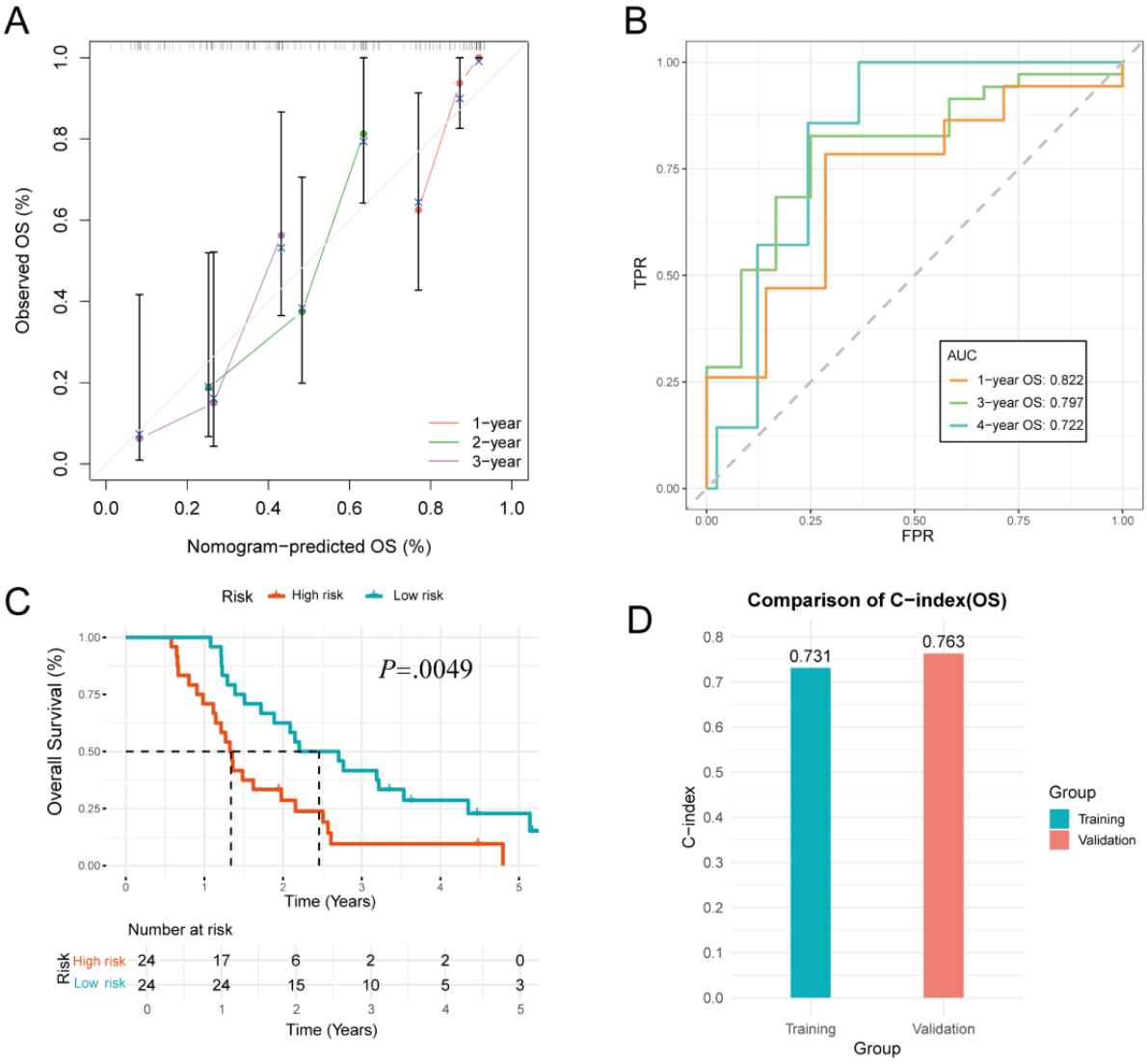
XSL•FO
RenderX

Figure 5. Predictive performance of the overall survival (OS) nomogram: (A) prognostic nomogram integrating age, neutrophil percentage-to-albumin ratio (NPAR), and platelet-to-mean platelet volume ratio (PMR) to predict 1-, 3-, and 5-year OS probabilities through score conversion. For individual prediction, assign points for each variable, sum them to obtain a total score, and project this score downward to the bottom probability axes; (B) the calibration curve demonstrates the agreement between predicted and observed 1-, 3-, and 4-year OS probabilities in the training group; (C) time-dependent receiver operating characteristic (ROC) curves show area under the curve (AUC) values of 0.836, 0.820, and 0.801 for 2-, 3-, and 4-year OS, respectively; (D) survival analysis based on the median nomogram score reveals significant differences in OS between the high-risk and low-risk groups (P<.001); and (E) raincloud plot comparing the distribution of nomogram scores between the survival and death groups.



XSL•FO
RenderX

Figure 6. Validation of the overall survival (OS) nomogram: (A) the calibration curve demonstrates the agreement between predicted and observed 1-, 2-, and 3-year OS probabilities in the validation group; (B) time-dependent receiver operating characteristic (ROC) curves show area under the curve (AUC) values of 0.822, 0.797, and 0.722 for 1-, 3-, and 4-year OS, respectively; (C) survival analysis based on the median nomogram score reveals significant differences in OS between the high-risk and low-risk groups; and (D) comparison of the concordance index (C-index) between the training and validation groups.



Construction and Validation of the PFS Prediction Model for ADG

Feature Variable Selection

Using PFS as the outcome variable, 3 machine learning algorithms—LASSO regression, XGBoost, and RF—were used to screen feature variables.

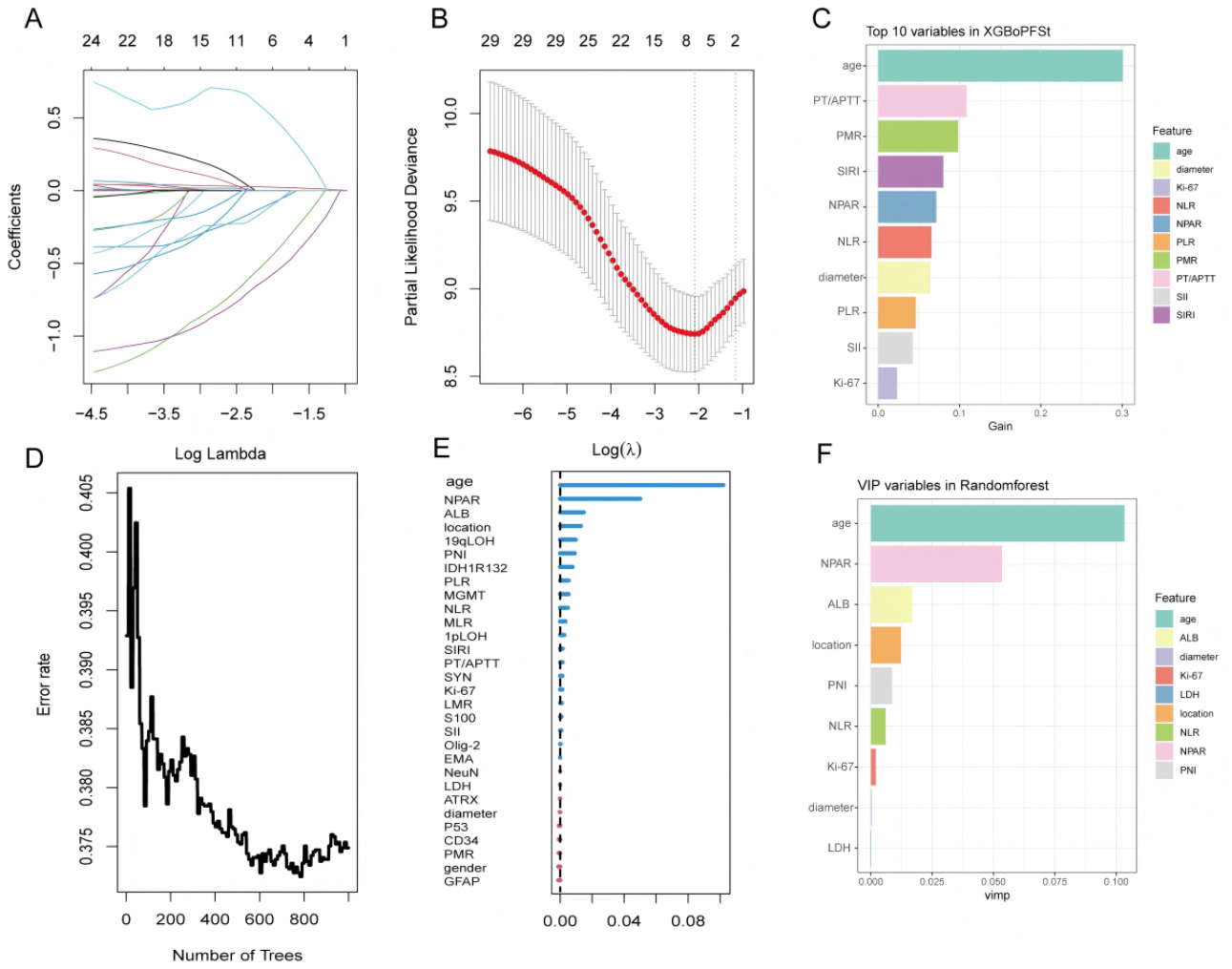
LASSO regression: based on the minimum λ value (λ_min), a total of 6 nonzero coefficient variables were selected, including age, NPAR, MLR, MGMT methylation, 1pLOH, and 19qLOH (Figure 7A and B).

XGBoost model: based on feature importance scores, a total of 10 variables were selected, including age, PT/APTT, PMR, SIRI, NPAR, NLR, maximum tumor diameter, PLR, SII, and Ki-67 expression (Figure 7C).

RF: based on the mean decrease in Gini index, a total of 9 variables were selected, including age, NPAR, ALB, tumor location, PNI, NLR, Ki-67, maximum tumor diameter, and LDH (Figures 7D and E).

XSL•FO
RenderX

Figure 7. Feature variable selection based on 3 machine learning methods: (A) least absolute shrinkage and selection operator (LASSO) regression coefficient path plot: variable coefficients as a function of λ, (B) LASSO regression parameter λ tuning based on the minimum criterion, (C) top 10 variables ranked by importance in the extreme gradient boosting (XGBoost) model, (D) change in error rate as a function of the number of decision trees during random forest (RF) model training, (E) RF feature importance: assessed by Gini index, and (F) RF selected variables: top 9 variables. 19qLOH: 19q loss of heterozygosity; 1pLOH: 1p loss of heterozygosity; ALB: albumin; ATRX: alpha-thalassemia/mental retardation syndrome X-linked; CD34: cluster of differentiation 34; EMA: epithelial membrane antigen; GFAP: glial fibrillary acidic protein; IDH1R132: isocitrate dehydrogenase 1 R132; Ki-67: proliferation cell nuclear antigen-67; LDH: lactate dehydrogenase; LMR: lymphocyte-to-monocyte ratio; MGMT: O6-methylguanine-DNA methyltransferase; MLR: monocyte-to-lymphocyte ratio; NeuN: neuronal nuclei; NLR: neutrophil-to-lymphocyte ratio; NPAR: neutrophil percentage-to-albumin ratio; Olig-2: oligodendrocyte transcription factor 2; P53: P53 protein; PLR: platelet-to-lymphocyte ratio; PMR: platelet-to-mean platelet volume ratio; PNI: prognostic nutritional index; PT/APTT: prothrombin time/activated partial thromboplastin time; S100: S100 calcium-binding protein; SII: immune-inflammation index; SIRI: systemic inflammation response index; SYN: synuclein.

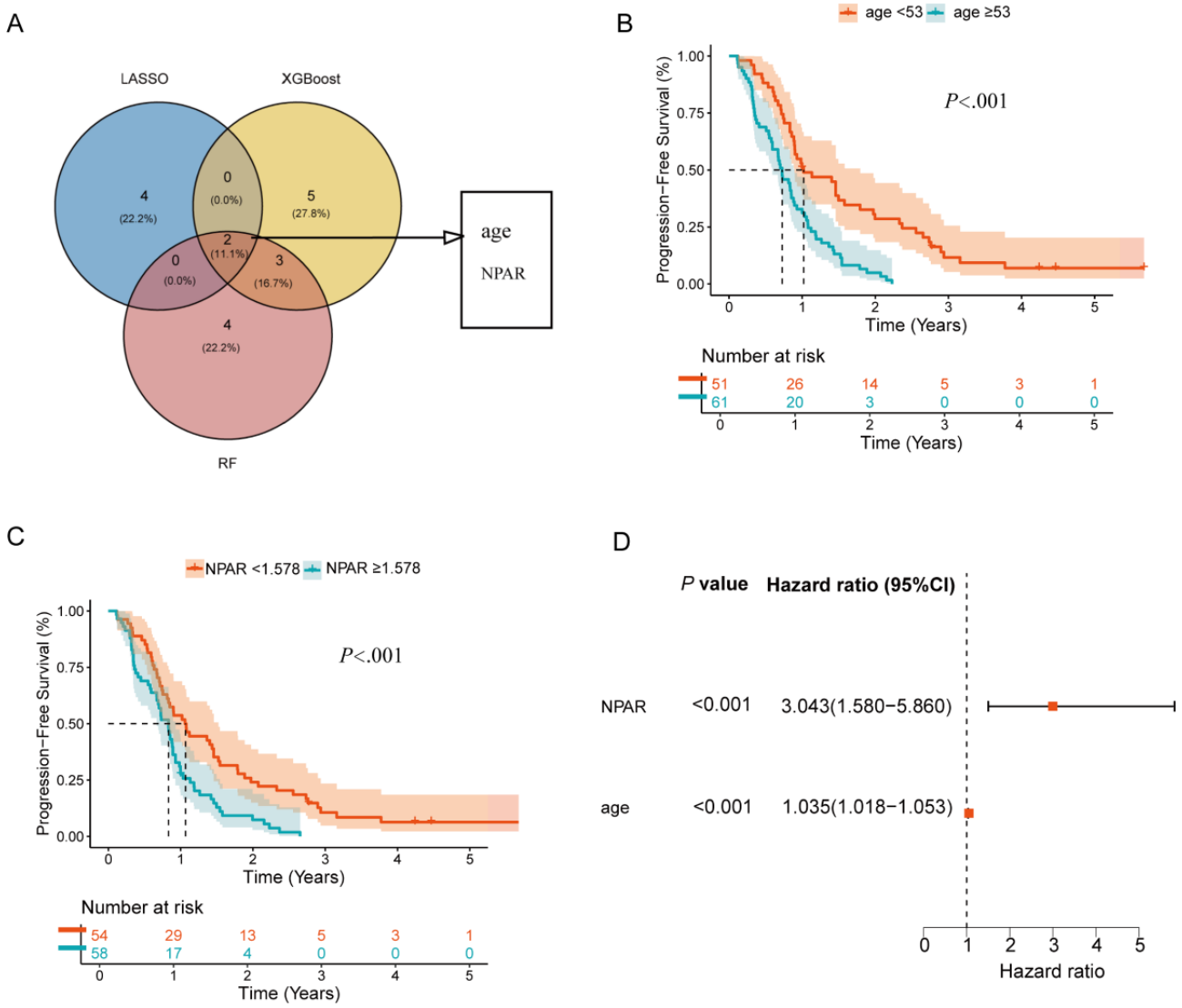


Survival Analysis

Based on the important variables screened by LASSO, XGBoost, and RF, Venn diagram intersection analysis identified age and NPAR as key factors influencing PFS (Figure 8A). The impact of each single factor on prognosis was assessed using KM survival analysis (Figures 8B and C). Log-rank test results showed that patients in the age <53 years group had significantly longer PFS than those in the age ≥53 years group (P<.001); patients in the NPAR <1.578 group had significantly longer PFS than those in the NPAR ≥1.578 group (P<.001). Multivariate Cox proportional hazards regression analysis revealed that age (HR 1.04, 95% CI 1.02-1.05; P<.001) and NPAR (HR 3.04, 95% CI 1.58-5.86; P<.001) were independent prognostic factors for PFS (Figure 8D).

XSL•FO
RenderX

Figure 8. Identification of independent prognostic factors for progression-free survival (PFS): (A) intersection analysis of feature variables from 3 models, (B) impact of age group on patient PFS, (C) impact of neutrophil percentage-to-albumin ratio (NPAR) level on patient PFS, and (D) forest plot of multivariate Cox regression analysis for PFS. LASSO: least absolute shrinkage and selection operator.



Construction and Validation of the PFS Nomogram

Based on multivariate Cox regression analysis, age and NPAR were identified as independent prognostic factors for PFS. A nomogram was constructed using these variables to predict 1-, 3-, and 5-year PFS rates. Each predictive variable was assigned a corresponding risk score, and the total score estimated the patient's probability of disease progression (Figure 9A).

The calibration curve demonstrated good agreement between the model-predicted and observed 12-, 24-, and 30-month PFS rates (Figure 9B). Time-dependent ROC curve analysis showed that the AUC for 1-, 2-, and 3-year PFS in the training group was 0.817, 0.864, and 0.873, respectively, indicating high predictive accuracy of the model at different timepoints (Figure 9C). Patients were divided into high-risk and low-risk groups based on the median nomogram score. KM survival curves revealed that the high-risk group had significantly lower PFS than the low-risk group (P<.001), demonstrating that the nomogram score effectively stratified patients' prognostic risk (Figure 9D).

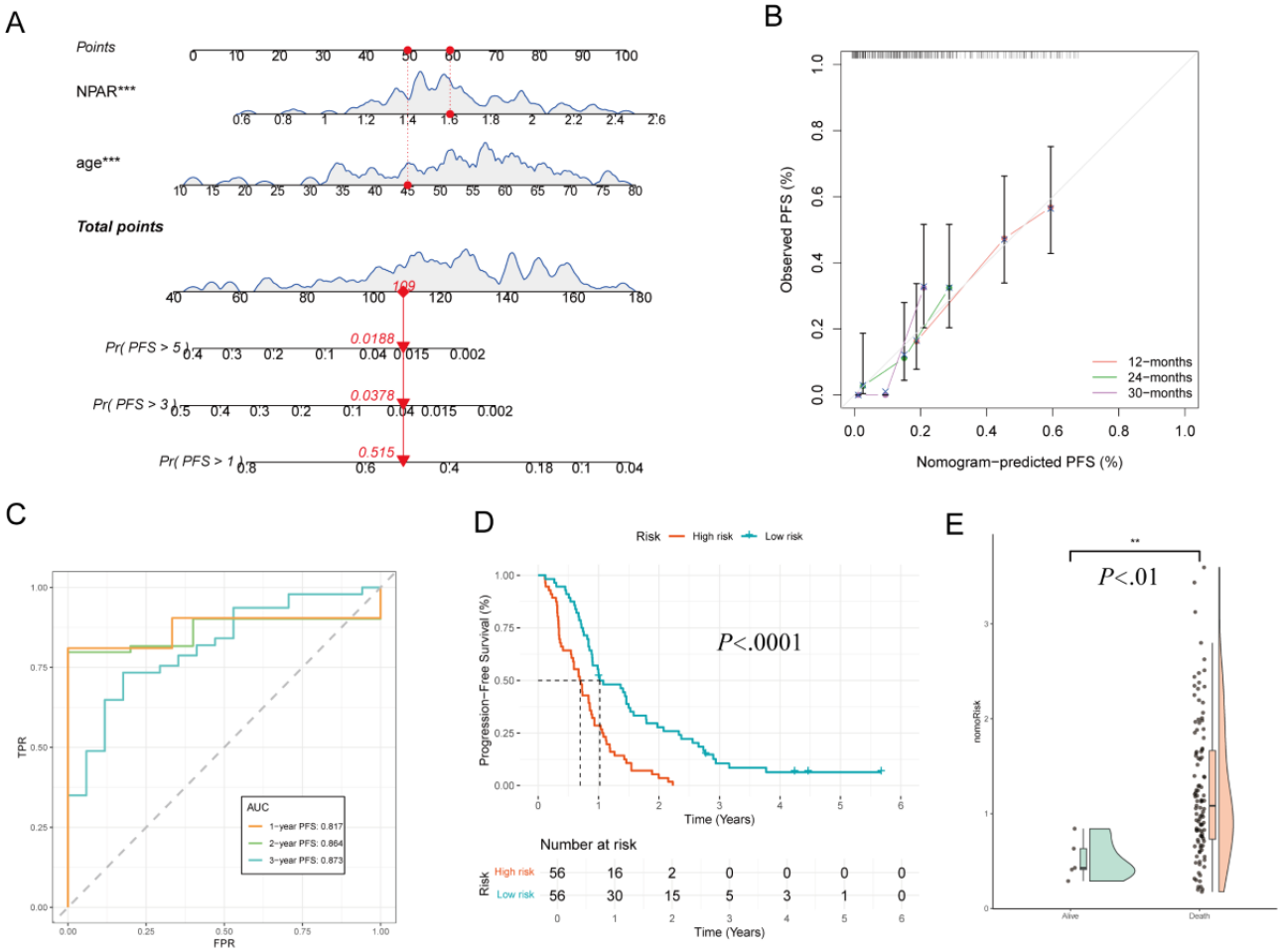
The raincloud plot provides a visual representation of the score distribution differences between the 2 groups and further supporting the effectiveness of the nomogram score in prognostic assessment (Figure 9E).

Due to the high risk of patient recurrence, the calibration curve for the validation group demonstrated good predictive performance for 6-, 12-, and 18-month disease progression risk in patients with ADG (Figure 10A). Time-dependent ROC curve analysis showed that the AUC for 1-, 3-, and 5-year PFS in the validation group was 0.709, 0.843, and 0.953, respectively (Figure 10B). Based on the median nomogram score, patients in the validation group were divided into high-risk and low-risk groups. KM survival curves revealed significant differences in OS between the 2 groups (P<.05), further validating the prognostic discriminative ability of the nomogram score (Figure 10C).

The C-index of the model was excellent in both the training and validation groups, indicating stable predictive performance across different datasets (Figure 10D).

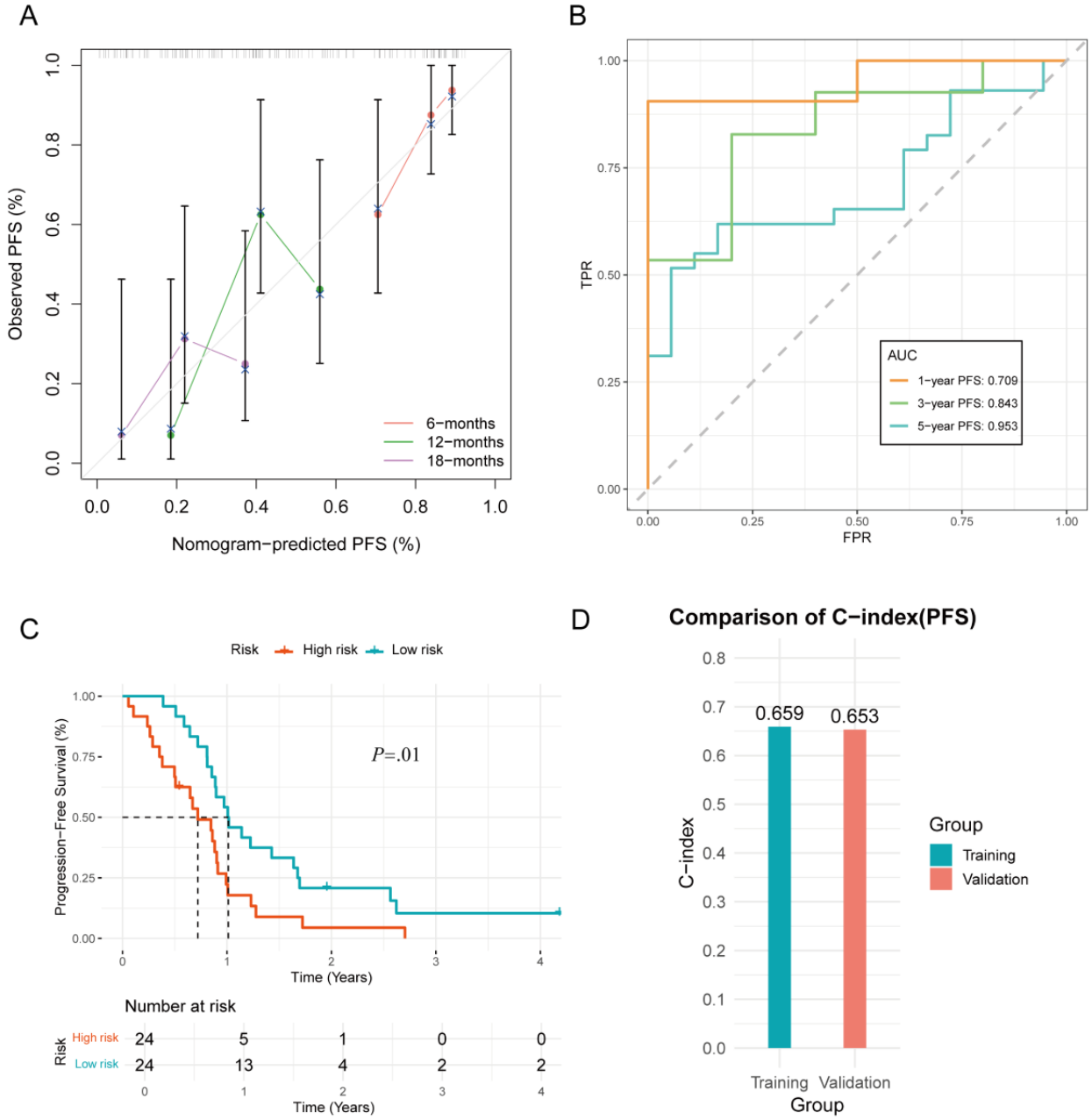
XSL•FO
RenderX

Figure 9. Predictive performance of the progression-free survival (PFS) nomogram: (A) prognostic nomogram integrating age and NPAR to predict 1-, 3-, and 5-year PFS probabilities through score conversion. For individual prediction, assign points for each variable, sum them to obtain a total score, and project this score downward to the bottom probability axes. (B) The calibration curve demonstrates the agreement between predicted and observed 12-, 24-, and 30-month PFS probabilities in the training group. (C) Time-dependent receiver operating characteristic (ROC) curves show area under the curve (AUC) values of 0.817, 0.864, and 0.873 for 1-, 2-, and 3-year PFS, respectively. (D) Survival analysis based on the median nomogram score reveals significant differences in PFS between the high-risk and low-risk groups (P<.001). (E) Raincloud plot comparing the distribution of nomogram scores between the survival (alive) and death (death) groups.



XSL•FO
RenderX

Figure 10. Validation of the progression-free survival (PFS) nomogram: (A) the calibration curve demonstrates the agreement between predicted and observed 6-, 12-, and 18-month PFS probabilities in the validation group; (B) time-dependent receiver operating characteristic (ROC) curves show area under the curve (AUC) values of 0.709, 0.843, and 0.953 for 1-, 3-, and 5-year PFS, respectively; (C) survival analysis based on the median nomogram score reveals significant differences in PFS between the high-risk and low-risk groups; and (D) comparison of the concordance index (C-index) between the training and validation groups.



Discussion

Principal Results

ADG, characterized by high malignancy and strong invasiveness, accounts for 45% of all gliomas [20,21]. The tumor heterogeneity and infiltrative growth patterns of ADG often lead to progressive neurological impairment, posing significant challenges for clinical management [9,22]. This study demonstrated that standardized comprehensive treatment (surgery combined with concurrent chemoradiotherapy and temozolomide) extended the median overall survival (mOS) to 23.0 (IQR 15.0-38.3) months, which is superior to the approximately 15 months reported in previous studies [2,23,24]. However, 95% (152/160) of patients experienced postoperative recurrence, with a median progression-free survival (mPFS) of only 10.8 months and a 5-year PFS rate of merely 1.9% (3/160). Moreover, the long-term prognosis remains poor, with a 5-year OS rate of just 12.8% (20/160) [25]. Therefore, it is of great significance to further explore the key prognostic factors affecting patients with ADG, establish novel prognostic models, and develop effective preventive strategies.

The key methodological advance of our work lies in the robust feature selection achieved by integrating 3 distinct machine learning algorithms (LASSO, XGBoost, and RF) [26]. This integrative approach circumvents the inherent pitfalls of relying on a single method, be it traditional Cox regression or an

XSL•FO
RenderX

individual machine learning model [13,14]. Consequently, it ensures that the identified prognostic factors (age, NPAR, and PMR) are consistently salient across different modeling paradigms, thereby significantly strengthening the credibility of our results.

Compared to existing prognostic models for ADG, our model demonstrates superior predictive performance. While Kim et al [27] reported a C-index of 0.70 using a single machine learning approach, the OS model in this study achieved C-indices of 0.731 and 0.763 in the training and validation cohorts, respectively. Furthermore, the model maintained high time-dependent AUC values at multiple timepoints (AUCs of 0.836, 0.820, and 0.801 for 2-, 3-, and 4-year OS, respectively) along with favorable calibration. Importantly, our model innovatively incorporates readily accessible hematological biomarkers (NPAR and PMR). This provides a more cost-effective and clinically feasible prognostic assessment tool compared to traditional models that rely solely on clinical and molecular parameters [7,8].

Analysis of 32 variables showed that longer survival was observed in patients with canonical molecular markers such as IDH1 mutation and MGMT promoter methylation, implicating them as potential prognostic factors. IDH1 mutations play a critical role in suppressing tumor metabolism and proliferation [28,29]. Patients with IDH1-mutant astrocytoma included had an mOS of 40.0 (IQR 33.6–NA, NA is censored data—some patients remained alive at follow-up cutoff, so the upper quartile was not observed) months, markedly better than the 22.2 (IQR 14.4–34.7) months observed in IDH1 wild-type patients. Similarly, MGMT methylation enhances sensitivity of tumor cells to alkylating agents such as temozolomide by inhibiting MGMT transcription and reducing DNA repair capacity, thereby significantly prolonging both PFS and OS [29-31]. In this study, a 30-year-old patient with IDH-mutant astrocytoma and MGMT methylation achieved a survival period exceeding 64 months, demonstrating significant survival benefit [32].

Furthermore, by integrating multiple machine learning methods, the study systematically identified age, the inflammation-nutrition marker NPAR, and the platelet quantity and function marker PMR as independent prognostic factors for patients with ADG [16]. These variables demonstrated significant predictive power in forecasting the prognosis of patients with ADG [33,34].

Age was confirmed as a key prognostic factor in our model. Patients aged ≥53 years had a higher risk of disease progression and significantly shorter survival time compared with those aged <53 years. This finding aligns with previous research trends regarding the prognosis of older adults with ADG, which may be related to reduced treatment tolerance and alterations in tumor biological characteristics in older individuals. It suggests that clinicians should pay attention to age stratification and develop personalized treatment plans accordingly [28].

This study is the first to identify NPAR as an independent prognostic factor for ADG. Patients with high NPAR levels had significantly shorter mOS and mPFS, along with a markedly increased risk of disease recurrence and worse prognosis. As a novel indicator that comprehensively reflects systemic inflammation and nutritional status, the NPAR has increasingly been recognized for its prognostic value in various cancers such as pancreatic and urological cancers in recent years. The NPAR integrates both systemic inflammation (reflected by neutrophil percentage) and nutritional status (reflected by serum ALB), 2 key pathophysiological processes known to influence cancer prognosis. Neutrophils can impair immune surveillance by suppressing natural killer cell function and promote tumor cell proliferation and metastasis [12,35]. Serum ALB levels not only reflect the body's protein metabolic status but are also closely associated with immune dysfunction [36]. The prognostic value of NPAR may stem from its correlation with systemic inflammatory responses mediated by proinflammatory cytokines in the TME. This pathophysiological cascade can accelerate the progression of cachexia and facilitate the evolution of malignant tumor phenotypes by modulating the TME [37,38].

The prognostic value of NPAR stems from its integration of 2 pivotal pathophysiological pathways: protumor inflammation and host nutritional status. The neutrophil component reflects systemic inflammation. Within the glioma TME, neutrophils are polarized into tumor-associated neutrophils (TANs), which facilitate tumor progression by secreting matrix metalloproteinases to enhance invasion, releasing proangiogenic factors like vascular endothelial growth factor, and contributing to T-cell dysfunction [39]. Emerging evidence specifically links this protumorigenic neutrophil activity to glioma progression [40]. Conversely, hypoalbuminemia, indicative of poor nutritional status and diminished physiological reserve, is a well-established marker of impaired immune competence and poorer outcomes in patients with cancer [41]. Thus, NPAR comprehensively captures this detrimental interplay between a proinflammatory TME and a compromised host state.

Furthermore, this study identified a higher PMR as an independent prognostic factor for prolonged overall survival in patients with ADG. This finding is consistent with a study on angiosarcoma, which also reported that a high mean platelet volume–to–platelet count ratio served as an independent risk factor for patient mortality [42]. This consistent pattern across malignancies highlights the biological significance of composite indices based on platelet count and volume.

While studies have established that platelets can facilitate glioma progression by releasing factors like platelet-derived growth factor to stimulate proliferation [43] and by modulating the immune microenvironment to promote an M2 phenotype and immune escape [44], our discovery of PMR as a protective factor suggests a more nuanced role. We hypothesize that an elevated PMR reflects a distinct platelet phenotype characterized by a lower mean platelet volume which is generally associated with diminished platelet reactivity. Consequently, a high PMR may indicate a reduced capacity of platelets to engage in these documented protumorigenic processes, thereby explaining its association with improved survival in ADG.

The prognostic nomogram developed from the aforementioned factors demonstrates significant translational potential. This tool can be integrated into electronic medical record systems to enable automated risk stratification at the time of diagnosis. Consequently, high-risk patients could be considered for more

XSL•FO
RenderX

aggressive treatment strategies—such as intensified chemoradiation following maximal safe resection—or more intensive follow-up (eg, neuroimaging every 2-3 months), and even enrollment in novel clinical trials. Conversely, low-risk patients may be suitable candidates for treatment de-escalation, such as shortened adjuvant chemotherapy cycles or reduced radiation doses, aiming to preserve oncological control while enhancing quality of life. This nomogram-based decision-making model represents a constructive step toward precision neuro-oncology.

Limitations

This study is subject to several limitations, including its single-center retrospective design, which may limit generalizability, and a variable-to-sample size ratio that raises concerns about overfitting despite internal validation. These limitations underscore the necessity for external validation in independent multicenter cohorts as an essential next step. Moreover, the molecular mechanisms linking NPAR and PMR to glioma prognosis are not fully understood. To address these points and advance this research, future work must focus on multicenter prospective validation, augmenting the model's power through the integration of radiomics and genomics, and leveraging single-cell and multi-omics technologies to decipher the functional roles of these circulating biomarkers in glioma pathogenesis.

Conclusion

In conclusion, this study significantly improved the robustness of prognostic biomarkers through a multimodel cross-validation strategy. Among the numerous factors, age, the preoperative inflammation-nutrition composite indicator NPAR, and platelet quantity and function marker PMR were key prognostic factors. Patients with ADG with younger age and lower NPAR values had a lower risk of disease progression and better prognosis; meanwhile, patients with higher PMR exhibited longer survival. These biomarkers are not only easily accessible and cost-effective but also highly beneficial, offering a new perspective for personalized treatment of ADG.

Acknowledgments

We gratefully acknowledge the use of R statistical software and thank the funding agencies for their support of this research.

Authors' Contributions

CZ conceived the study concept, CZ and JA designed the study protocol and performed statistical analysis and interpretation, JA and LZ retrieved and selected the data, CZ and JA completed the draft writing and manuscript revision, and LZ oversaw the integration of the entire study and approved the final version. All authors have approved the submission of the final manuscript.

Conflicts of Interest

None declared.

References

1. Chen W, Guo S, Wang Y, Shi Y, Guo X, Liu D, et al. Novel insight into histological and molecular astrocytoma, IDH-mutant, grade 4 by the updated WHO classification of central nervous system tumors. Cancer Med 2023;12(18):18666-18678 [FREE Full text] [doi: 10.1002/cam4.6476] [Medline: 37667984]
2. Shaalan F, Ballout N, Chamoun WT. Insights into the role of Bmi-1 deregulation in promoting stemness and therapy resistance in glioblastoma: a narrative review. Cancer Med 2025;14(1):e70566 [FREE Full text] [doi: 10.1002/cam4.70566] [Medline: 39791545]
3. Sakly H, Said M, Seekins J, Guetari R, Kraiem N, Marzougui M. Brain tumor radiogenomic classification of O-methylguanine-DNA methyltransferase promoter methylation in malignant gliomas-based transfer learning. Cancer Control 2023;30:10732748231169149 [FREE Full text] [doi: 10.1177/10732748231169149] [Medline: 37078100]
4. Cen B, Zhang J, Pan X, Xu Z, Li R, Chen C, et al. Stimuli-responsive peptide/siRNA nanoparticles as a radiation sensitizer for glioblastoma treatment by co-inhibiting RELA/P65 and EGFR. Int J Nanomedicine 2024;19:11517-11537 [FREE Full text] [doi: 10.2147/IJN.S483252] [Medline: 39539970]
5. Schonfeld E, Choi J, Tran A, Kim LH, Lim M. The landscape of immune checkpoint inhibitor clinical trials in glioblastoma: a systematic review. Neurooncol Adv 2024;6(1):vdae174. [doi: 10.1093/noajnl/vdae174] [Medline: 39534539]
6. D'Aprile S, Denaro S, Gervasi A, Vicario N, Parenti R. Targeting metabolic reprogramming in glioblastoma as a new strategy to overcome therapy resistance. Front Cell Dev Biol 2025;13:1535073 [FREE Full text] [doi: 10.3389/fcell.2025.1535073] [Medline: 40078366]
7. Johnstad C, Reinertsen I, Thurin E, Dunås T, Bouget D, Sagberg LM, et al. The prognostic importance of glioblastoma size and shape. Acta Neurochir (Wien) 2024;166(1):450. [doi: 10.1007/s00701-024-06351-0] [Medline: 39531176]
8. Le Rhun E, Bink A, Felsberg J, Gramatzki D, Brandner S, Benhamida JK, et al. The clinical and molecular landscape of diffuse hemispheric glioma, H3 G34-mutant. Neuro Oncol 2025;27(6):1519-1535. [doi: 10.1093/neuonc/noaf015] [Medline: 39842935]
9. Sharma P, Aaroe A, Liang J, Puduvalli VK. Tumor microenvironment in glioblastoma: current and emerging concepts. Neurooncol Adv 2023;5(1):vdad009 [FREE Full text] [doi: 10.1093/noajnl/vdad009] [Medline: 36968288]

XSL•FO
RenderX

10. Attem J, Vemuganti GK. The role of small extracellular vesicles in retinoblastoma development and progression. Curr Eye Res 2025:1-15. [doi: 10.1080/02713683.2025.2457102] [Medline: 39905667]
11. Zhao Y, Wang Y, Jiang Y, Yang J, Zhang Y. The prognostic impact of neutrophil-to-lymphocyte ratio and platelet-to-lymphocyte ratio on patients with small cell lung cancer receiving first-line platinum-based chemotherapy: a systematic review and meta-analysis. BMC Pulm Med 2024;24(1):630 [FREE Full text] [doi: 10.1186/s12890-024-03447-2] [Medline: 39710633]
12. Li X, Wu M, Chen M, Liu R, Tao Q, Hu Y, et al. The association between neutrophil-percentage-to-albumin ratio (NPAR) and mortality among individuals with cancer: insights from National Health and Nutrition Examination Survey. Cancer Med 2025 Jan;14(2):e70527 [FREE Full text] [doi: 10.1002/cam4.70527] [Medline: 39831739]
13. Sun M, Li S, Guo Y, Zhou J, Fu D, Hua X, et al. Clinicopathological, immunohistochemical and therapeutic approaches on survival in patients with epithelioid glioblastoma: institutional experience in the management of 58 patients. Neurosurg Rev 2024;47(1):763. [doi: 10.1007/s10143-024-02957-1] [Medline: 39382734]
14. Zhang Z, Zhao Y, Ma Y, Chen C, Li Z, Wang Y, et al. Prediction of STAS in lung adenocarcinoma with nodules ≤ 2 cm using machine learning: a multicenter retrospective study. BMC Cancer 2025;25(1):417 [FREE Full text] [doi: 10.1186/s12885-025-13783-z] [Medline: 40055661]
15. Kurdi M, Alkhotani A, Alsinani T, Alkhayyat S, Katib Y, Jastaniah Z, et al. Effects of radiotherapy alone versus concomitant radiotherapy with temozolomide chemotherapy on the outcome of IDH-wildtype glioblastoma patients. Clin Oncol (R Coll Radiol) 2025;38:103741 [FREE Full text] [doi: 10.1016/j.clon.2024.103741] [Medline: 39742692]
16. Babajide Mustapha I, Saeed F. Bioactive molecule prediction using extreme gradient boosting. Molecules 2016;21(8):983 [FREE Full text] [doi: 10.3390/molecules21080983] [Medline: 27483216]
17. Hou W, Li C, Wang Z, Wang W, Wan S, Zou B. Predicting lymph node metastasis in rectal cancer: development and validation of a machine learning model using clinical data. JMIR Med Inform 2025;13:e73765 [FREE Full text] [doi: 10.2196/73765] [Medline: 40986886]
18. Alhumaidi NH, Dermawan D, Kamaruzaman HF, Alotaiq N. The use of machine learning for analyzing real-world data in disease prediction and management: systematic review. JMIR Med Inform 2025;13:e68898 [FREE Full text] [doi: 10.2196/68898] [Medline: 40537090]
19. Chu L, Sun Y, Tang X, Duan X, Zhao Y, Xia H, et al. The tumor-derived exosomes enhanced bevacizumab across the blood-brain barrier for antiangiogenesis therapy against glioblastoma. Mol Pharm 2025;22(2):972-983. [doi: 10.1021/acs.molpharmaceut.4c01227] [Medline: 39895311]
20. Incekara F, Smits M, van der Voort SR, Dubbink HJ, Atmodimedjo PN, Kros JM, et al. The association between the extent of glioblastoma resection and survival in light of MGMT promoter methylation in 326 patients with newly diagnosed IDH-wildtype glioblastoma. Front Oncol 2020;10:1087 [FREE Full text] [doi: 10.3389/fonc.2020.01087] [Medline: 32766140]
21. Kirishima M, Yokoyama S, Akahane T, Higa N, Uchida H, Yonezawa H, et al. Prognosis prediction via histological evaluation of cellular heterogeneity in glioblastoma. Sci Rep 2024;14(1):24955 [FREE Full text] [doi: 10.1038/s41598-024-76826-8] [Medline: 39438642]
22. Selvi H, Brüning-Richardson A, Danovi D. Systematic review of pre-clinical systems using artificial microenvironments and anti-migratory drugs to control migration of glioblastoma cells. Expert Rev Mol Med 2025;27:e6. [doi: 10.1017/erm.2024.33] [Medline: 39844377]
23. Bafiti V, Ouzounis S, Chalikiopoulou C, Grigorakou E, Grypari IM, Gregoriou G, et al. A 3-miRNA signature enables risk stratification in glioblastoma multiforme patients with different clinical outcomes. Curr Oncol 2022;29(6):4315-4331 [FREE Full text] [doi: 10.3390/curroncol29060345] [Medline: 35735454]
24. Sloan L, Sen R, Liu C, Doucet M, Blosser L, Katulis L, et al. Radiation immunodynamics in patients with glioblastoma receiving chemoradiation. Front Immunol 2024;15:1438044 [FREE Full text] [doi: 10.3389/fimmu.2024.1438044] [Medline: 39346903]
25. Su H, Peng Y, Wu Y, Zeng X. Overcoming immune evasion with innovative multi-target approaches for glioblastoma. Front Immunol 2025;16:1541467 [FREE Full text] [doi: 10.3389/fimmu.2025.1541467] [Medline: 39911397]
26. Yu C, Niu Z, Fang Q, Lei Q. Development and validation of a predictive model for severe tubular atrophy/interstitial fibrosis in patients with IgA nephropathy: multicenter retrospective study. JMIR Med Inform 2025;13:e78761 [FREE Full text] [doi: 10.2196/78761] [Medline: 41150862]
27. Kim Y, Kim KH, Park J, Yoon HI, Sung W. Prognosis prediction for glioblastoma multiforme patients using machine learning approaches: development of the clinically applicable model. Radiother Oncol 2023;183:109617. [doi: 10.1016/j.radonc.2023.109617] [Medline: 36921767]
28. Jeon YH, Choi KS, Lee KH, Jeong SY, Lee JY, Ham T, et al. Deep learning-based quantification of T2-FLAIR mismatch sign: extending IDH mutation prediction in adult-type diffuse lower-grade glioma. Eur Radiol 2025;35(9):5193-5202. [doi: 10.1007/s00330-025-11475-7] [Medline: 40050456]
29. Alnahhas I. Molecular testing in gliomas: what is necessary in routine clinical practice? Curr Oncol Rep 2024;26(11):1277-1282. [doi: 10.1007/s11912-024-01602-w] [Medline: 39361075]

XSL•FO
RenderX

30. Zhao B, Qi H, Ma W. MGMT promoter methylation and chemotherapy outcomes in low-grade and anaplastic gliomas. JAMA Oncol 2023;9(12):1734-1735. [doi: 10.1001/jamaoncol.2023.4736] [Medline: 37856142]
31. Sareen H, Ma Y, Becker TM, Roberts TL, de Souza P, Powter B. Molecular biomarkers in glioblastoma: a systematic review and meta-analysis. Int J Mol Sci 2022;23(16):8835 [FREE Full text] [doi: 10.3390/ijms23168835] [Medline: 36012105]
32. Butler M, Pongor L, Su Y, Xi L, Raffeld M, Quezado M, et al. MGMT status as a clinical biomarker in glioblastoma. Trends Cancer 2020;6(5):380-391 [FREE Full text] [doi: 10.1016/j.trecan.2020.02.010] [Medline: 32348734]
33. Neil ZD, Pierzchajlo N, Boyett C, Little O, Kuo CC, Brown NJ, et al. Assessing metabolic markers in glioblastoma using machine learning: a systematic review. Metabolites 2023;13(2):161 [FREE Full text] [doi: 10.3390/metabo13020161] [Medline: 36837779]
34. Mohammadzadeh I, Niroomand B, Hajikarimloo B, Habibi MA, Mortezaei A, Behjati J, et al. Can we rely on machine learning algorithms as a trustworthy predictor for recurrence in high-grade glioma? a systematic review and meta-analysis. Clin Neurol Neurosurg 2025;249:108762 [FREE Full text] [doi: 10.1016/j.clineuro.2025.108762] [Medline: 39884144]
35. Tingle SJ, Severs GR, Goodfellow M, Moir JA, White SA. NARCA: a novel prognostic scoring system using neutrophil-albumin ratio and Ca19-9 to predict overall survival in palliative pancreatic cancer. J Surg Oncol 2018;118(4):680-686. [doi: 10.1002/jso.25209] [Medline: 30196571]
36. Sasagasako T, Ueda A, Mineharu Y, Mochizuki Y, Doi S, Park S, et al. Postoperative Karnofsky performance status prediction in patients with IDH wild-type glioblastoma: a multimodal approach integrating clinical and deep imaging features. PLoS One 2024;19(11):e0303002 [FREE Full text] [doi: 10.1371/journal.pone.0303002] [Medline: 39527541]
37. Li R, Sun Z, Song S, He X, Shi X, Li Z, et al. NARFIB: a novel prognostic score based on the neutrophil-to-albumin ratio and fibrinogen can predict the prognosis of gastrointestinal stromal tumors. Cancer Manag Res 2020;12:11183-11190 [FREE Full text] [doi: 10.2147/CMAR.S281375] [Medline: 33177869]
38. Garcia-Martinez R, Andreola F, Mehta G, Poulton K, Oria M, Jover M, et al. Immunomodulatory and antioxidant function of albumin stabilises the endothelium and improves survival in a rodent model of chronic liver failure. J Hepatol 2015;62(4):799-806. [doi: 10.1016/j.jhep.2014.10.031] [Medline: 25450713]
39. Gao X, Xu M, Xiao H, Han Z, Wang Z, Sun G, et al. Tumor-associated neutrophils: a complex role in cancer. Clin Immunol 2025;280:110558. [doi: 10.1016/j.clim.2025.110558] [Medline: 40633742]
40. Sun C, Wang S, Ma Z, Zhou J, Ding Z, Yuan G, et al. Neutrophils in glioma microenvironment: from immune function to immunotherapy. Front Immunol 2024;15:1393173 [FREE Full text] [doi: 10.3389/fimmu.2024.1393173] [Medline: 38779679]
41. Yoo S, Chowell D, Valero C, Morris LGT, Chan TA. Pre-treatment serum albumin and mutational burden as biomarkers of response to immune checkpoint blockade. NPJ Precis Oncol 2022;6(1):23 [FREE Full text] [doi: 10.1038/s41698-022-00267-7] [Medline: 35393553]
42. Horisaki K, Taki T, Mori S, Okumura M, Akiyama M. Peripheral blood mean platelet volume and the ratio of mean platelet volume to platelet count as prognostic biomarkers in patients with cutaneous angiosarcoma of the head and neck: a retrospective cohort study. Acta Derm Venereol 2024;104:adv42227 [FREE Full text] [doi: 10.2340/actadv.v104.42227] [Medline: 39508501]
43. Mayr L, Neyazi S, Schwark K, Trissal M, Beck A, Labelle J, et al. Effective targeting of PDGFRA-altered high-grade glioma with avapritinib. Cancer Cell 2025;43(4):740-756.e8. [doi: 10.1016/j.ccell.2025.02.018] [Medline: 40086436]
44. Li X, Chen X, Gong S, Zhao J, Yao C, Zhu H, et al. Platelets promote CRC by activating the C5a/C5aR1 axis via PSGL-1/JNK/STAT1 signaling in tumor-associated macrophages. Theranostics 2023;13(6):2040-2056 [FREE Full text] [doi: 10.7150/thno.80555] [Medline: 37064877]

Abbreviations

19qLOH: 19q loss of heterozygosity
1pLOH: 1p loss of heterozygosity
ADG: adult-type diffuse glioma
AISI: aggregate index of systemic inflammation
ALB: albumin
APTT: activated partial thromboplastin time
ATRX: alpha-thalassemia/mental retardation syndrome X-linked
AUC: area under the curve
BRAF: B-Raf proto-oncogene, serine/threonine kinase
CD34: cluster of differentiation 34
C-index: concordance index
EGFR: epidermal growth factor receptor
EMA: epithelial membrane antigen
GFAP: glial fibrillary acidic protein

XSL•FO
RenderX

HR: hazard ratio
IDH1: isocitrate dehydrogenase 1
Ki-67: proliferation cell nuclear antigen-67
KM: Kaplan-Meier
LASSO: least absolute shrinkage and selection operator
LDH: lactate dehydrogenase
LMR: lymphocyte-to-monocyte ratio
Lym: lymphocyte
MGMT: O6-methylguanine-DNA methyltransferase
MLR: monocyte-to-lymphocyte ratio
Mono: monocyte
mOS: median overall survival
mPFS: median progression-free survival
NeuN: neuronal nuclei
Neut: neutrophil
NLPR: neutrophil-lymphocyte-platelet ratio
NLR: neutrophil-to-lymphocyte ratio
NPAR: neutrophil percentage–to-albumin ratio
Olig-2: oligodendrocyte transcription factor 2
OS: overall survival
P53: P53 protein
PFS: progression-free survival
PLR: platelet-to-lymphocyte ratio
PLT: platelet
PMR: platelet-to-mean platelet volume ratio
PNI: prognostic nutritional index
PT: prothrombin time
RF: random forest
ROC: receiver operating characteristic
S100: S100 calcium-binding protein
SII: immune-inflammation index
SIRI: systemic inflammation response index
SYN: synuclein
TAN: tumor-associated neutrophil
TERT: telomerase reverse transcriptase
TME: tumor microenvironment
WBC: white blood cell
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by J Klann; submitted 01.Jul.2025; peer-reviewed by Z Zhang, Y Liu, Z Chen; comments to author 03.Nov.2025; accepted 25.Nov.2025; published 13.Jan.2026.

Please cite as:
Zhu C, An J, Zhou L
Neutrophil Percentage–to-Albumin Ratio as a Novel Prognostic Biomarker in Adult Diffuse Gliomas: Retrospective Study Integrating 3 Machine Learning Models and Cox Regression
JMIR Med Inform 2026;14:e79945
URL: https://medinform.jmir.org/2026/1/e79945
doi:10.2196/79945
PMID:

©Congcong Zhu, Jiyang An, Lili Zhou. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting Pediatric Urological Surgery Duration Through Multimodal Patient-Physician Feature Fusion: Deep Learning Framework Incorporating Clinical Text Embedding

Yonggen Zhao[1,2*], BE; Ruoge Lin[3*], MS; Yiying Sun[1,2*], MS; Lingdong Chen[1,2], MS; Jian Huang[1,2], PhD; Guangjie Chen[4], MD; Zhu Zhu[1,2*], PhD; Gang Yu[1,2], PhD

[1]National Clinical Research Center for Children and Adolescents' Health and Diseases, Children's Hospital, Zhejiang University School of Medicine, 3333 Binsheng Rd, Hangzhou, China

[2]Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China

[3]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China

[4]Department of Urology, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Gang Yu, PhD
National Clinical Research Center for Children and Adolescents' Health and Diseases, Children's Hospital, Zhejiang University School of Medicine, 3333 Binsheng Rd, Hangzhou, China

Abstract

Background: Accurate prediction of surgical duration is critical for optimizing operating room scheduling and resource allocation. Existing models, however, exhibit limited applicability in pediatric urology due to the unique anatomical and developmental characteristics of children.

Objective: This study aimed to develop and validate a specialty-tailored prediction framework for estimating the duration of pediatric urological surgeries.

Methods: We integrated multisource heterogeneous data, encompassing patient demographics, surgical details, surgeon-specific features, and electronic medical record narratives, to develop a customized prediction system. Large language model techniques were used to extract semantic representations from unstructured clinical text, while a multihead perceptron architecture enabled the efficient fusion of structured and unstructured features. Pediatric-specific clinical variables, such as developmental stage and the severity of urinary tract malformations, were explicitly modeled to capture their impact on surgical duration.

Results: The proposed approach achieved a mean absolute error of 11.39 minutes and a root mean square error of 15.58 minutes, markedly outperforming existing methods. Comparative analyses demonstrated that the Qwen-based structured preprocessing with text embeddings provided superior feature representation, surpassing both traditional long short-term memory and direct Embedding-3 approaches. Feature importance analysis identified the primary surgical procedure, surgical plan, and preoperative diagnosis as dominant predictive factors.

Conclusions: By combining innovative feature engineering with a tailored model architecture, the proposed framework substantially improves the accuracy of surgical duration prediction in pediatric urology. These findings offer robust technical support for precision operating room scheduling and hold significant clinical value in enhancing the efficiency of surgical resource utilization.

(JMIR Med Inform 2026;14:e82329) doi:10.2196/82329

KEYWORDS

surgical duration prediction; operating room management; deep neural network; multihead perceptron; medical informatics

Introduction

Background and Motivation

As one of the hospital's most resource-intensive departments, the operating room plays a pivotal role in determining the institution's financial health and overall success [1,2]. Optimizing operating room scheduling under constraints of available bed capacity and human resources has become a key lever for improving health care service efficiency. Within this process, accurate prediction of surgical duration is a prerequisite for effective scheduling decisions [3-6]. However, due to the presence of multiple sources of uncertainty, such as patient heterogeneity (eg, disease complexity and anatomical variation), intraoperative contingencies (instrument failure and unexpected

XSL•FO
RenderX

bleeding), and dynamic adjustments to the surgical plan [7,8], predicting surgical duration remains a major challenge in perioperative management [9,10]. Traditional estimation methods largely rely on the surgeon's personal experience or simple historical averages [11,12]. These approaches lack standardized classification of surgical procedures and fail to systematically integrate dynamic variables, such as patient-specific characteristics, team collaboration efficiency, and equipment turnover rates, leading to substantial discrepancies between estimated and actual durations [13,14]. Such inaccuracies can trigger a cascade of operational disruptions. The underestimation of duration can lead to staff fatigue, congestion in emergency channels [15], and prolonged patient waiting and fasting times, thereby increasing the risk of adverse events. Overestimation, on the other hand, results in idle resource time. More critically, errors in estimating a single case can be amplified through the scheduling system, ultimately causing mismatches in staffing, cost, and heightened dissatisfaction among patients and their families [16,17]. Therefore, developing more accurate and reliable methods for surgical duration prediction is both important and urgent.

In recent years, machine learning has made significant advances in surgical duration prediction [18,19]. This research can be broadly categorized by the types of features used.

Patient- and Procedure-Oriented Studies

These methods mainly rely on patient characteristics and surgical information, leveraging demographic data, preoperative physiological indicators, and procedural details for predictive modeling. Commonly used algorithms include multilayer perceptrons (MLPs) and Extreme Gradient Boosting, with some studies incorporating real-time intraoperative parameters, such as heart rate and bispectral index, to enable dynamic predictions [20-23]. Key predictive factors identified in these studies include patient BMI, surgical type, and procedural complexity, with certain models achieving prediction errors—the mean absolute error (MAE) within a clinically acceptable margin of 30 minutes.

Surgeon- and Procedure-Oriented Studies

Such studies emphasize surgeon-related and procedural factors, particularly the impact of operator experience, surgical techniques, and case urgency on operative duration. Ensemble methods, such as bagged decision trees, gradient boosting trees, and Extreme Gradient Boosting, have demonstrated strong performance in this domain, especially for single-specialty or emergency procedures, with reported MAEs as low as 14.9 minutes [24-28]. Feature importance analyses consistently highlight surgical complexity and surgeon proficiency as dominant predictors.

Comprehensive Studies Integrating Patient, Doctor, and Surgical Information

The most comprehensive approaches integrate multimodal clinical data, combining patient-specific variables, surgeon expertise, and procedural factors, to optimize predictive accuracy. Random forests and MLPs are frequently adopted in such frameworks, while some studies have further improved performance through specialized department-specific modeling or 2-stage prediction architectures, achieving MAEs below 16 minutes in targeted clinical settings [29-31]. Emerging techniques, including computer vision–based analysis of surgical videos and large language models (LLMs) (eg, GPT-4), are also gaining traction, demonstrating promising capabilities in processing unstructured clinical narratives for duration estimation [32,33].

Despite the diversity of approaches in existing studies, several critical limitations persist. A primary issue lies in feature engineering, where current methods predominantly rely on structured data fields, such as procedure names and diagnostic codes, while largely neglecting the wealth of information contained in unstructured clinical notes. These free-text entries, which include detailed condition descriptions and laboratory results, often capture patient-specific variations that could significantly enhance prediction accuracy. Furthermore, existing models demonstrate limited applicability to pediatric surgical contexts. Adult-derived models fail to account for pediatric-specific physiological considerations, including developmental variations in organ systems and anatomical structures. This oversight is particularly problematic, given the narrower intraoperative safety margins in pediatric patients due to their reduced physiological reserves [12], which demands greater predictive precision.

This study focuses on pediatric urological surgery and develops a novel specialty-oriented prediction method. Compared with adult procedures, pediatric urology presents unique clinical challenges: the patient age range spans from newborns to adolescents, with marked developmental variations in physiological parameters, and differences in urinary tract development and malformation severity influence the level of operative precision required. These factors make adult-trained models poorly transferable to pediatric cases.

To address these issues, we designed an end-to-end prediction system, which integrates multimodal clinical data, including patient characteristics, surgical information, surgeon attributes, and clinical narratives, to model and accurately predict pediatric urological surgery duration. We innovatively used LLM techniques to automatically process electronic medical record (EMR) text containing detailed case descriptions, examination findings, and physiological features, extracting and embedding unstructured clinical information efficiently. In addition, we proposed a multihead MLP architecture for heterogeneous data fusion, allowing categorical, numerical, and text feature vectors to be processed independently before integration, outperforming traditional single-head MLP approaches. We further applied permutation importance analysis to identify the most influential predictors, including the lead surgeon, primary surgical procedure, preoperative diagnosis, and pediatric-specific disease characteristics (eg, prepuce characteristics and testicular characteristics). Comparative and ablation experiments demonstrated that our method substantially improved prediction accuracy, achieving an MAE of 11.39 minutes and a root mean square error (RMSE) of 15.58 minutes—significantly better than the existing work. Our model exhibits clear advantages in predicting pediatric urological surgery duration, providing reliable support for precise operating room scheduling.

XSL•FO
RenderX

Methods

Data Acquisition and Preprocessing Flow

The data used in this study were obtained from surgical cases performed at the Urology Surgery Center of the Children's Hospital, Zhejiang University School of Medicine, between May 22, 2024, and July 10, 2025. The hospital is an independent tertiary pediatric medical center. All surgical information was extracted from the hospital information system, yielding a total of 4526 cases. These encompassed 4 major categories of pediatric urological procedures: penile lengthening (n=2067, 45.7%), orchidopexy (n=937, 20.7%), laparoscopic high ligation of the processus vaginalis (n=1158, 25.6%), and circumcision (n=364, 8%).

Feature Extraction

In this study, the duration of surgery was defined as the time interval commencing from "anesthesia induction" and concluding when the patient left the operating room. This definition encompasses not only the surgical procedure itself but also key stages, such as anesthesia induction, preoperative preparation, and postoperative transfer prior to the patient's departure from the operating room. As a result, it more comprehensively reflects the actual occupancy period of the operating room. When compared with simplified definitions, such as "from skin incision to suture completion" or "from entry to exit of the operating room," the time frame used in this study more accurately reflects the real-world workflow under multidisciplinary collaboration and holds direct guiding significance for hospital surgical scheduling and resource allocation [34]. For example, precise prediction of this period facilitates the rational arrangement of consecutive surgeries, optimizes the coordination between the anesthesia team and the recovery room, and mitigates idle time or congestion arising from estimation errors. All duration data were obtained from the hospital information system, guaranteeing the consistency and traceability of the definition. Considering that previous research has suggested that the duration of surgery generally follows a log-normal distribution [35,36], we standardized the target variable (with a mean of 0 and an SD of 1) to improve the stability of model training and the balance of feature scales.

Additional predictor variables extracted from the information system included patient age, patient weight, primary surgical procedure, lead surgeon features, surgical plan, preoperative diagnosis, and case-specific characteristics, including preoperative preparation details. All extracted variables were used by the prediction model.

Patient age and weight are continuous variables. Patient age was recorded in the format "x years x months" and was converted to years, followed by standardization. If the age was missing, it was replaced with the age mentioned in the "case characteristics and preoperative preparation" notes. Patient weight, recorded in kilograms, was also standardized. If missing or abnormal, it was imputed using the average weight of patients in the same age group.

Primary surgical procedure and lead surgeon features are categorical variables. The primary surgical procedure is a single-category variable referring to the main pediatric urological surgery performed for the patient, with only 1 such designation per case. This study primarily covered 4 main procedure types: penile lengthening, orchidopexy, laparoscopic high ligation of the processus vaginalis, and circumcision. This variable was one-hot encoded. The lead surgeon data were expanded into 3 distinct features to capture specific physician characteristics: name, years of practice, and annual surgical volume. The surgeon name is a single-category variable representing the primary attending specialist and was processed using one-hot encoding. Years of practice is a continuous numeric variable representing the duration of the surgeon's professional experience. Annual surgical volume is a discrete numeric variable indicating the total number of surgeries performed by that surgeon in the past year. Both numeric features (years of practice and annual surgical volume) were standardized to normalize their scales. Missing values for these features were imputed using the mean value of the respective column.

Preoperative diagnosis and surgical plan are multilabel categorical variables. The preoperative diagnosis refers to the diagnostic categories assigned to a patient before surgery, and each patient may have 1 or multiple diagnoses. For example, patient A's diagnosis might be "hydrocele," whereas patient B's diagnosis could be "hydrocele, phimosis, and incomplete testicular descent." Diagnoses were split by commas, and occurrence counts were tallied for each category. To avoid excessive sparsity, only diagnostic categories appearing ≥10 times were retained; others were grouped into a "rare type" category. Multilabel encoding was then applied. For example, "hydrocele" = [1, 0, 0], "phimosis" = [0, 1, 0], "incomplete testicular descent" = [0, 0, 1], and "hydrocele, phimosis, incomplete testicular descent" = [1,1,1].

The "surgical plan" refers to the preoperative plan drafted by the surgeon, and each patient may have 1 or more planned procedures. For example, patient A's plan could be "bilateral laparoscopic high ligation of the processus vaginalis," whereas patient B's plan might be "bilateral laparoscopic high ligation of the processus vaginalis + orchidopexy + circumcision." Due to differences in writing habits, plans often have inconsistent formatting, with procedure types separated by various delimiters, such as plus signs, commas, or phrases like "simultaneously perform." Naming conventions for the same procedure were also inconsistent; for instance, "right orchidopexy" and "right-sided orchidopexy" refer to the same surgery. For our prediction target, surgical duration, laterality ("left" vs "right") was assumed to have no effect for unilateral procedures, so these were combined into "unilateral" categories, whereas "bilateral" procedures were treated separately due to their significant impact on duration. Initially, manual tokenization and grouping were performed based on delimiters (plus signs, commas, etc) and standardization of terms (eg, merging "circumcision" and "circumcision [stapler/ligation/manual suture]"). This yielded 170 categorized procedure types, of which only those occurring ≥10 times were retained. The rest were grouped as "other surgical procedures." All surgical plans were encoded using the same multilabel approach.

XSL•FO
RenderX

However, due to the high variability in surgical plans, manual tokenization and grouping could not fully resolve inconsistencies, and many equivalent procedures remained uncategorized. To address this, we used an advanced approach using the LLM Qwen3, developed by Alibaba in April 2025. The selection of Qwen over other prominent LLMs (eg, GPT-4 and Llama-3) was driven by its specific alignment with our study's linguistic and security requirements. Unlike Llama-3, which is primarily optimized for English, Qwen demonstrates superior proficiency in understanding specialized Chinese clinical terminology and semantics relevant to our dataset. Furthermore, in contrast to cloud-based services, such as GPT-4, that necessitate external data transmission, Qwen supports local deployment. This feature allows us to process sensitive patient data entirely within the hospital's secure intranet, ensuring strict adherence to data privacy regulations while maintaining reasoning and extraction capabilities compared to top-tier proprietary models.

By calling Qwen3's application programming interface for batch processing, the model summarized all patients' surgical plans into 23 standardized procedure categories (including "other surgical procedures"). This intelligent consolidation reduced unnecessary feature dimensionality and improved downstream analysis. For example, Qwen3 automatically grouped "penile nevus excision," "penile cyst excision," and "penile scar excision" under "penile lesion excision." Similarly, these standardized categories were then multilabel-encoded for each patient.

Case characteristics and preoperative preparation are unstructured free-text variables describing reasons for admission, degree of organ development, precise lesion locations, and other specifics. Due to similar issues with inconsistent formatting, we again leveraged Qwen3 for structured information extraction. We predefined 4 target features for extraction: prepuce characteristics description, testicular characteristics description, scrotal characteristics description, and penile characteristics description. During extraction, Qwen3 was instructed to output both the structured fields and the corresponding original text segments to guard against hallucinations. Since the extracted text lengths varied, we used the Embedding-3 model (256 dimensions) to produce fixed-length semantic vectors for model input. Embedding-3, developed by ZHIPU AI (Tsinghua University), is a text-embedding model capable of capturing deep semantic relationships and outputting high-quality embeddings. It supports multilingual and long-text inputs, making it well-suited for high-precision semantic representation tasks.

Missing category and text vectors were replaced with null vectors. A detailed description of the surgical feature variables is provided in Table 1.

It is worth noting that while "surgeon name" provides specific identity information, it lacks generalizability, as it is strictly bound to the current roster of physicians and cannot accommodate new surgeons joining the department in the future. To address this limitation, we designed 2 distinct experimental configurations. The first configuration includes "surgeon name" to capture individual-specific variance. The second configuration substitutes "surgeon name" with the quantitative metrics "years of practice" and "annual surgical volume" to construct a more robust and generalizable model capable of predicting surgical duration for new or unobserved surgeons.

XSL•FO
RenderX

Table . Description of specific surgical feature variables (N=4526).

Patient variables and type	Values/range and statistics
Age (year)	
Numeric	Float (average 5.3, SD 3.7, max 17.0, min 0.0)
Weight (kg)	
Numeric	Float (average 24.6, SD 14.6, max 140, min 5)
Primary surgical procedure	
Categorical	Penile lengthening surgery (2067, 45.6%), laparoscopic high ligation of the processus vaginalis (1158, 25.6%), orchidopexy (937, 20.7%), circumcision (364, 8%)
Lead surgeon	
Categorical	Name: total =31, including “unknown”
Numeric	Years of practice: float (average 16.1, SD 7.1, max 40.6, min 1.5) (y)
Numeric	Annual surgical volume: integer (average 461.2, SD 223.8, max 863, min 2) (cases/y)
Preoperative diagnosis	
Multicategorical	Concealed penis (1839, 40.6%), hydrocele (1202, 26.6%), and cryptorchidism (733, 16.2%).
Surgical plan	
Multicategorical	Penile lengthening (1956, 43.2%), preputioplasty (1649, 36.4%).
Case characteristics and preoperative preparation	
Free text	Prepuce characteristics description: eg, “Phimosis prepuce, can be manually retracted to expose the glans.”
Free text	Testicular characteristics description: eg, “Both testes are located within the scrotum and are well-developed.”
Free text	Scrotal characteristics description: eg, “Scrotum is well-developed.”
Free text	Penile characteristics description: eg, “The penis appears small, with only the distal half of the glans visible on the body surface in the supine position.”
Surgery duration (min)	
Numeric	Float (average 38.6, SD 24.9, max 539, min 4)

Model Construction

We developed multiple machine-learning methods based on neural networks and traditional models. From the entire dataset of 4526 cases, 15% (n=679 cases) were randomly selected as test data for evaluating the prediction models. The remaining data (referred to as “training data”) were further randomly split into training (n=3168 cases) and validation (n=679 cases) sets at an 85% to 15% ratio. Detailed descriptions of model hyperparameters and validation performance metrics can be found in the Results section. For the finalized model, performance evaluation was conducted only once on the test dataset.

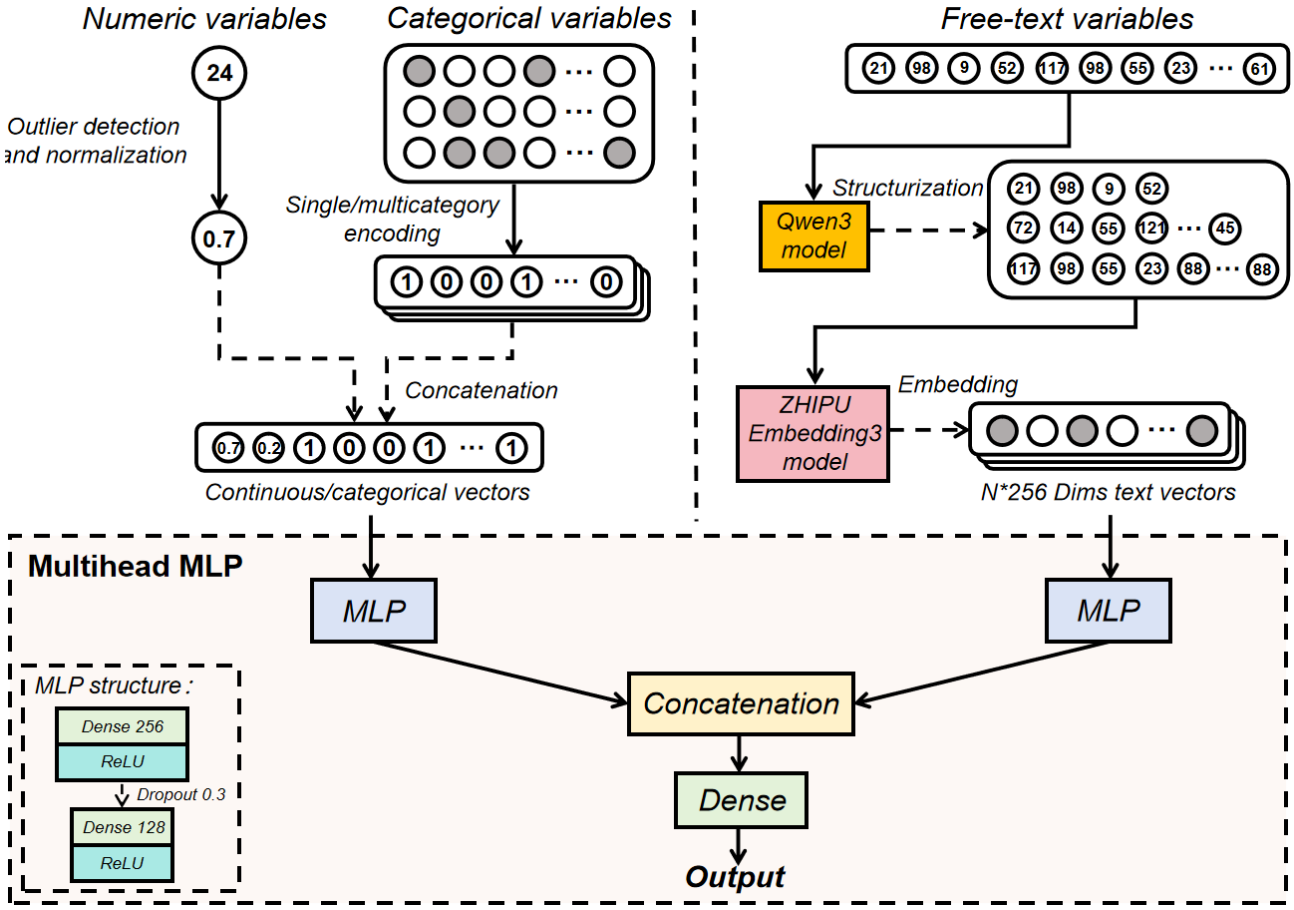
Pediatric Urology Surgical Duration Prediction Model Based on Multihead MLP

In this study, we trained an MLP model with a multihead structure (multihead MLP) to predict surgical duration in pediatric urology. The core idea is to process different types of input features (text vectors and categorical/numerical vectors) separately through independent subnetworks or branches and then concatenate their outputs and feed them into a shared MLP layer for the final prediction.

Traditional single-head MLP concatenates all heterogeneous features into 1 large input vector, which is then passed through a single MLP for prediction. However, due to the severe dimensional imbalance among features in this study (eg, the primary surgery procedure vector is 4-dimensional, while text vectors have dimensions of 4×256), single-head MLP tends to be dominated by features with larger dimensions, drowning out those with smaller dimensions and leading to a tendency for overfitting. To address this, the multihead MLP uses parallel representation learning by establishing independent subnetworks for each data group. This design enables the model to capture the unique patterns and characteristics of each data type separately, thereby reducing interference and ensuring that heterogeneous data are fused within a balanced latent space. To mitigate overfitting, our multihead MLP was trained with early stopping. The architecture of the proposed model is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Architecture of the proposed pediatric urology surgical duration prediction model. MLP: multilayer perceptron; ReLU: rectified linear unit.



The multihead MLP architecture comprises 2 distinct subnetworks designed to process heterogeneous input features separately: one dedicated to textual features and the other to categorical and numerical features. Each subnetwork consists of 2 fully connected (dense) layers with 256 and 128 neurons, respectively, each followed by a dropout layer with a rate of 0.3 to mitigate overfitting. Rectified linear unit activation functions are applied throughout. Outputs from both subnetworks are concatenated in a feature fusion layer, which feeds into a subsequent dense layer of 64 neurons accompanied by another dropout layer (rate 0.3). The final output layer uses a single neuron with a linear activation function to produce continuous predictions of surgical duration. The model is optimized using the Adam optimizer with a learning rate of 0.001, trained over 200 epochs, and uses mean squared error as the loss function.

Model Training and Experimental Settings

This study included a total of 4526 pediatric urology surgical cases with complete recorded surgical times for model development and evaluation. The training dataset (comprising both training and validation sets) contained 3847 cases, including 1730 cases of penile lengthening surgery, 810 cases of orchiopexy, 992 cases of laparoscopic high ligation of the processus vaginalis, and 315 cases of circumcision, accounting for 85% of the total dataset. The test dataset consisted of 679 cases, including 337 penile lengthening surgeries, 127 orchiopexies, 166 laparoscopic high ligations, and 49 circumcisions, representing 15% of the total dataset.

Python 3.8.3 (Python Software Foundation) was used for all feature extraction, algorithm implementation, and performance evaluation. The Alibaba Qwen3 LLM was used for all text analysis and structured field extraction. The ZHIPU AI Embedding-3 large model was used for all text embedding. Google's TensorFlow 2.4.1 (Google Brain) was used to implement the multihead MLP and other neural networks in comparative experiments. The Scikit-Learn library (scikit-learn community and INRIA) was applied for machine-learning methods, such as decision tree regression (DT) in comparative experiments, with default parameters used for all models.

Evaluation Metrics

Model performance was assessed using 3 metrics: MAE, RMSE, and the coefficient of determination (R-squared, R^2).

MAE represents the average of the absolute differences between predicted and true values. It directly reflects the average magnitude of prediction errors and is defined by the following formula:

.(1)MAE=1n∑i=1n|yi−y^i|

RMSE is the square root of the average of the squared differences between predicted and true values. It is a commonly used metric to measure the deviation between model predictions and actual values. The formula is as follows:

.(2)RMSE=1n∑i=1n(yi−y^i)2

R^2 measures the proportion of variance in the dependent variable (true values) that is explained by the model. It represents the

XSL•FO
RenderX

fraction of the total variability in the dependent variable accounted for by the model. An R^2 value of 1 indicates a perfect fit, explaining all variance; R^2=0 means the model explains no more variance than simply predicting the mean; and R^2<0 indicates the model performs worse than predicting the mean. The formula is as follows:

.(3)R2=1−RSSTSS

The formulas for the residual sum of squares and the total sum of squares are as follows:

,(4)RSS=∑i=1n(yi−y^i)2

.(5)TSS=∑i=1n(yi−y¯)2

where n denotes the number of samples, y_i represents the true surgical duration for the ith sample, y¯ is the mean of all true values, and yi^ is the model's predicted surgical duration for the ith sample.

Feature Importance Evaluation

Permutation importance is used to assess the significance of individual features within the model. As a model-agnostic method, its core principle is to measure the impact on model performance by randomly permuting the values of a single feature. Initially, predictions are made on the unshuffled test set to establish a baseline performance score. Then, for each feature in turn, its values in the test set are randomly shuffled while keeping other features unchanged, followed by generating new predictions and recalculating the performance score. The difference between the baseline and the shuffled performance indicates the importance of that feature. A significant drop in performance after shuffling a feature implies that the feature is highly influential to the model's predictive capability.

Ethical Considerations

From an ethical perspective, this study focuses on the preoperative diagnosis and duration of surgery for patients. The information collection methods adopted by the researchers are designed to ensure that the patients cannot be directly identified, nor can they be indirectly identified through related identifiers. In this study, only the surgery-related content was analyzed, and no personally identifiable information was extracted or associated. Additionally, the identities of the hospital staff involved in the surgery were anonymized and represented only by numerical IDs. This study was performed in line with the principles of the Declaration of Helsinki. All data used in this study were obtained from retrospective surgical cases, and approval was granted by the Academic Ethics Committee of Children's Hospital Zhejiang University School of Medicine (IRB No. 2023-IRB-0181-P-01; granted on July 20, 2023). We have applied for an informed consent waiver for our study. No compensation was provided to participants, as this was a retrospective study using existing surgical records without direct patient involvement or additional data collection.

Results

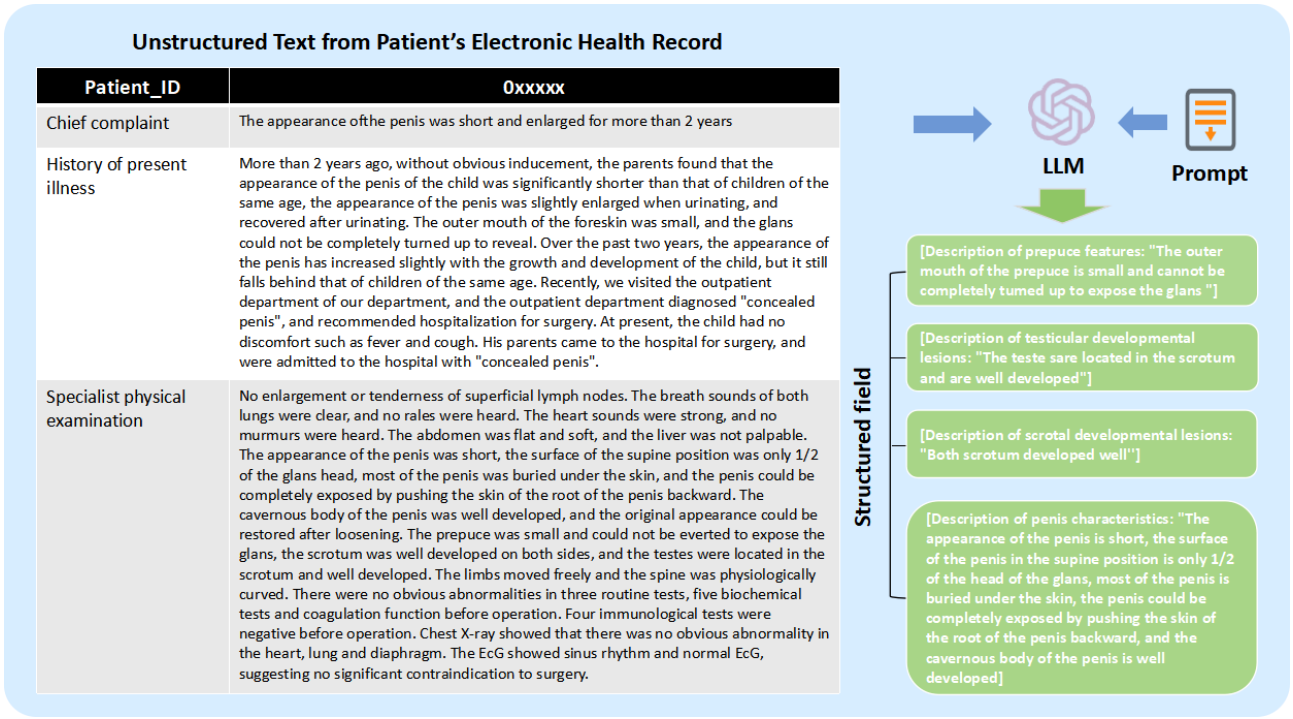
Accuracy Verification of LLM-Based Extraction

To quantitatively evaluate the fidelity of the structured data generated by Qwen-3 and to assess the prevalence of hallucinations, we conducted a manual validation on a randomly selected subset of 500 cases. This evaluation focused on 4 key features extracted from the unstructured "case characteristics and preoperative preparation" text: prepuce characteristics description, testicular characteristics description, scrotal characteristics description, and penile characteristics description. The validation was performed by a senior medical informatics engineer with over 10 years of experience in clinical data processing. A strict scoring metric (ranging from 0% to 100%) was applied to measure the consistency between the original text and the extracted structured data. Specifically, to rigorously penalize hallucinations, any extraction containing factually incorrect or fabricated information was assigned a score of 0%. Incomplete extractions were scored proportionally based on the extent of the missing information.

Figure 2 shows a concrete example of extracting structured fields from patient electronic health records using LLMs and prompt text. The evaluation results demonstrated high reliability in the model's extraction capabilities. The average accuracy scores for the 4 features were 99.6% for prepuce characteristics, 98.94% for testicular characteristics, 93.43% for scrotal characteristics, and 99.18% for penile characteristics. These results confirm that Qwen-3 effectively maintained semantic consistency with the source text while minimizing hallucinations.

XSL•FO
RenderX

Figure 2. A case of extracting structured fields from patients' electronic health records text using LLM and prompting. EcG: electrocardiogram; LLM: large language model.



Model Performance Comparison

We compared the proposed method with 6 outstanding surgical duration prediction approaches in the field, including ridge regression (RR) [37], DT [38], support vector regression (SVR) [39], random forest regression (RF) [40], gradient boosting regression tree (GBR) [41], and the machine learning–mixture density network (ML-MDN) model that combines long short-term memory (LSTM) and MLP for surgical duration prediction [23]. All models were trained and tested using the same datasets and feature categories as the proposed model. The RR, DT, SVR, RF, and GBR models used the feature engineering pipeline proposed in this study, whereas the ML-MDN model was implemented using the feature processing strategy described in its original publication. Specifically, the ML-MDN method integrates both structured and unstructured text features using a neural network architecture. In this method, structured variables (categorical and numerical) are encoded or normalized, whereas unstructured text variables (including primary surgical procedure, surgical plan, preoperative diagnoses, and case characteristics) are tokenized and embedded into vectors fed into an LSTM network. The LSTM captures temporal dependencies and contextual information, outputting semantic vectors that are concatenated with structured features and fed into an MDN. The MDN outputs a probability distribution governed by 3 parameters: mean (μ), SD (σ), and mixture coefficients (α). As the other baseline models in this study are deterministic regressors that output single-point estimates, a direct comparison with a probability distribution is infeasible. To ensure a fair and consistent evaluation framework, we used the expected value (mean, μ) of the MDN's predicted distribution as the representative point prediction. This allows for the standard calculation of MAE, RMSE, and R^2. Additionally, we compared these metrics against a naive baseline ("average") that simply uses the global mean of the surgical duration from the training set for prediction.

In the first experimental configuration using "surgeon name" as a feature, the proposed method achieved the best performance (MAE=10.74 min, RMSE=14.91, R^2=0.58). SVR (MAE=11.74 min), RF (MAE=11.62 min), and GBR (MAE=11.68 min) showed comparable, second-best results. RR (MAE=13.06 min) and DT (MAE=15.07 min) demonstrated relatively lower performance. The ML-MDN model's performance (MAE=16.91 min) was inferior to both our multihead MLP model (MAE=10.74 min) and other traditional machine-learning models. In contrast, the naive "average" baseline, which relied solely on the global mean without utilizing any features or machine-learning algorithms, yielded the poorest performance (MAE=19.06 min). The detailed model evaluation results are shown in Table 2.

XSL•FO
RenderX

Table . Performance comparison of different surgical duration prediction models and methods on the test dataset utilizing specific surgeon identity "surgeon name".

Models	MAE[a]	RMSE[b]	R^2[c]
Average	19.0579	24.8830	0.0000
RR[d] [37]	13.0554	19.3512	0.2910
DT[e] [38]	15.0692	21.8759	0.0939
SVR[f] [39]	11.7400	16.8024	0.4655
RF[g] [40]	11.6238	16.5873	0.4791
GBR[h] [41]	11.6821	16.3646	0.4930
ML-MDN[i] [23]	16.9117	21.4013	0.1328
Multihead MLP[j]	10.7391[k]	14.9134[k]	0.5789[k]

[a]MAE: mean absolute error.

[b]RMSE: root mean squared error.

[c]R^2: coefficient of determination.

[d]RR: ridge regression.

[e]DT: decision tree.

[f]SVR: support vector regression.

[g]RF: random forest.

[h]GBR: gradient boosting regression.

[i]ML-MDN: machine learning–mixed dense network.

[j]MLP: multilayer perceptron.

[k]These values in the last row indicate the best performance among all tested models across each evaluation metric.

In the second configuration, where "surgeon name" was substituted with "years of practice" and "annual surgical volume," all models exhibited a slight increase in overall error rates. However, the overall performance trend remained consistent: the proposed multihead MLP model still secured the best performance (MAE=11.39 min; RMSE=15.58; R^2=0.54). Detailed results for this configuration are shown in Table 3.

XSL•FO
RenderX

Table . Performance comparison of different surgical duration prediction models on the test dataset utilizing generalized surgeon metrics "years of practice" and "annual surgical volume".

Models	MAE[a]	RMSE[b]	R^2[c]
Average	19.0579	24.8830	0.0000
RR[d] [37]	13.6891	20.7138	0.1876
DT[e] [38]	14.6289	21.9035	0.0916
SVR[f] [39]	11.8670	16.9053	0.4589
RF[g] [40]	12.0465	16.8062	0.4652
GBR[h] [41]	11.6706	16.4278	0.4890
ML-MDN[i] [23]	17.2747	21.7401	0.1051
Multihead MLP[j]	11.3889[k]	15.5808[k]	0.5404[k]

[a]MAE: mean absolute error.

[b]RMSE: root mean squared error.

[c]R^2: coefficient of determination.

[d]RR: ridge regression.

[e]DT: decision tree.

[f]SVR: support vector regression.

[g]RF: random forest.

[h]GBR: gradient boosting regression.

[i]ML-MDN: machine learning−mixed dense network.

[j]MLP: multilayer perceptron.

[k]These values in the last row indicate the best performance among all tested models across each evaluation metric.

As previously discussed, relying on specific surgeon names limits the model's applicability to the current roster and fails to accommodate unobserved or newly hired surgeons, thereby reducing generalizability. Therefore, despite the marginal decrease in prediction accuracy, substituting specific identities with quantitative professional metrics (years of practice and annual surgical volume) represents a more robust and scalable solution, ensuring that the model remains effective for future clinical applications. Consequently, all subsequent results and analyses presented in this study utilize "years of practice" and "annual surgical volume" in place of "surgeon name."

To evaluate the proposed model's robustness, 5-fold cross-validation was performed. The model demonstrated remarkable stability in point-wise prediction, achieving an average MAE of 11.61 (0.47) minutes, as shown in Table 4. The low SD (0.47 min) indicates that the model's predictive accuracy is consistent across different data partitions. It is worth noting that while 4 out of 5 folds yielded R^2 scores exceeding 0.50, the second fold showed a performance dip (R^2=0.28; RMSE=24.95). The discrepancy between the relatively stable MAE (12.05 min) and the spiked RMSE in this fold suggests the presence of extreme outliers (eg, rare complex cases with unexpectedly long durations) in that specific test subset, rather than a systematic failure of the model. Overall, the aggregated results (R^2=0.47±.11) confirm the model's generalizability in clinical scenarios.

Table . Performance metrics of the proposed multihead MLP[a] model using 5-fold cross-validation.

Models	1-fold	2-fold	3-fold	4-fold	5-fold	Mean (SD)
MAE[b]	11.9191	12.0483	11.8336	10.9581	11.2793	11.6077 (0.4669)
RMSE[c]	16.0396	24.9463	17.6053	16.0877	15.4815	18.0321 (3.9448)
R^2[d]	0.5291	0.2779	0.5177	0.5396	0.5010	0.4731 (0.1100)

[a]MLP: multilayer perceptron.

[b]MAE: mean absolute error.

[c]RMSE: root mean squared error.

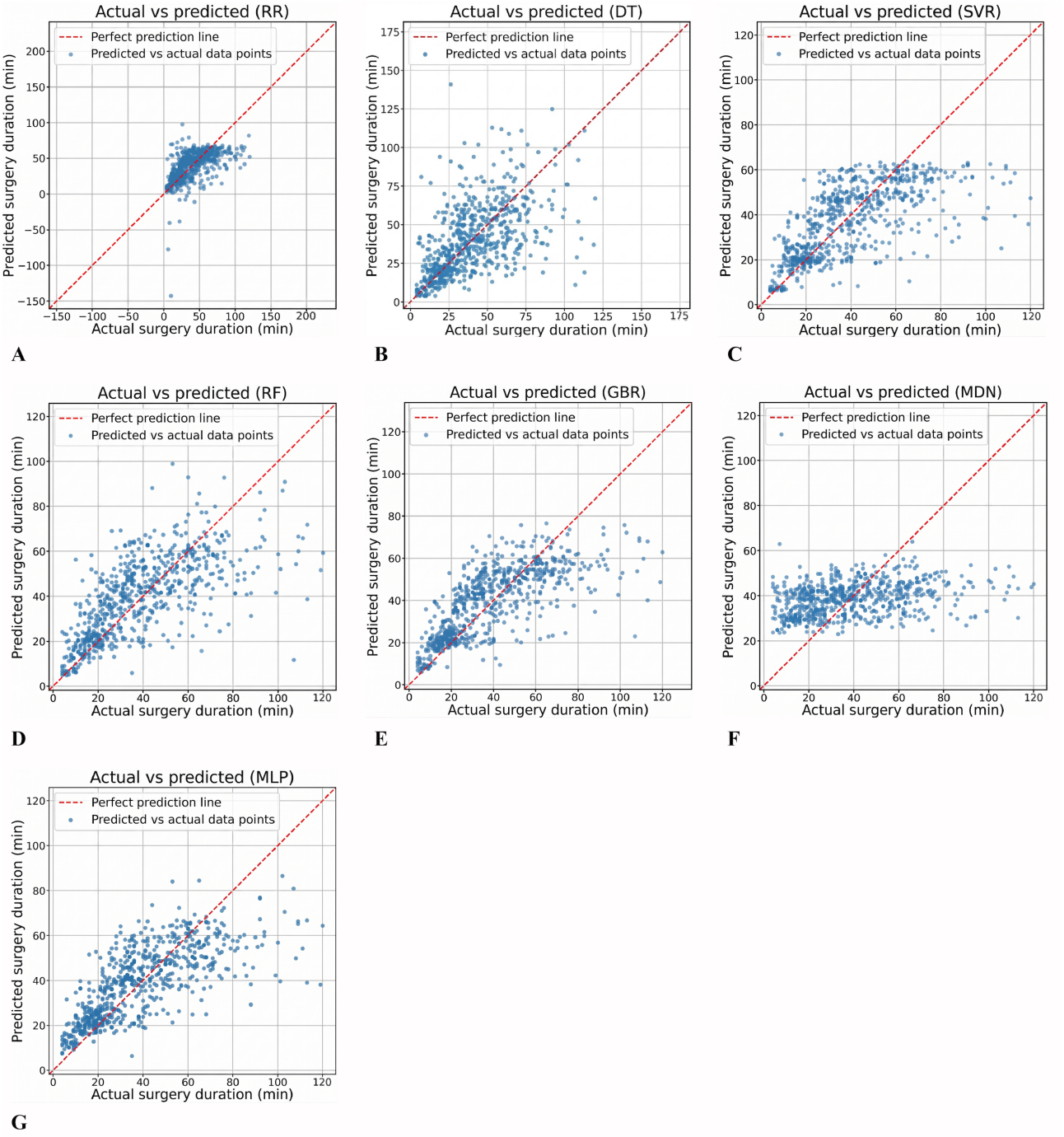
[d]R^2: coefficient of determination.

The scatterplot further visualizes the predictive performance of the models, where the red-dashed line represents the ideal prediction—that is, the predicted values equal the true values. The blue scatter points show the distribution of predicted versus actual values in the test set; the closer the points are to the red line, the more accurate the model's predictions. Scatterplots for

XSL•FO
RenderX

all models are shown in Figure 3. It can be observed that the points predicted by RR and DT are highly dispersed (with the largest range on the horizontal axis). Additionally, the ML-MDN model's predicted points also deviate notably from the red line, consistent with its poorer performance metrics reported earlier. In contrast, the remaining models show less difference, with the multihead MLP exhibiting the most uniform scatter distribution, indicating more stable performance.

Figure 3. Scatterplots of different models on the test dataset. (A) Ridge regression (RR), (B) decision tree (DT), (C) support vector regression (SVR), (D) random forest (RF), (E) gradient boosting regression (GBR), (F) machine learning–mixture density network (ML-MDN), (G) multihead multilayer perceptron (MLP).



Ablation Study

To demonstrate the effectiveness of the proposed multihead MLP architecture, detailed ablation experiments were conducted on its components. Four different model frameworks were tested separately: single-head MLP with categorical or numerical feature vectors, singlehead MLP with text feature vectors, single-head MLP with concatenated categorical or numerical and text feature vectors, and the multihead MLP. The detailed evaluation results are shown in Table 5. The experimental results indicate that directly concatenating categorical or numerical feature vectors with text feature vectors causes the single-head MLP to be biased toward the higher-dimensional text features, resulting in degraded performance, which is worse than using only the categorical or numerical features. In contrast, the multihead MLP effectively integrates heterogeneous data,

XSL•FO
RenderX

leveraging the strengths of features with different dimensions, and achieves the best performance.

Table . Ablation study on the proposed model architectures and feature combinations.

Different model architectures	MAE[a]	RMSE[b]	R^2[c]
Single-head MLP + categorical/numeric vectors	11.9365	16.4067	0.4903
Single-head MLP[d] + text vectors	12.7383	17.0647	0.4486
Single-head MLP + categorical/numeric vectors + text vectors	12.3874	16.9184	0.4581
Multihead MLP + categorical/numeric vectors + text vectors	11.3889	15.5808	0.5404

[a]MAE: mean absolute error.

[b]RMSE: root mean squared error.

[c]R^2: coefficient of determination.

[d]MLP: multilayer perceptron.

Comparison of Performance Using Different Text-Embedding Methods

To further investigate the impact of different text-embedding strategies on model prediction performance, this study designed 2 additional text embedding methods for comparison. The first method uses a traditional tokenization embedding combined with an LSTM network. Specifically, the raw text undergoes preprocessing steps, such as the removal of newline characters and extra whitespace for normalization; the Keras tokenizer layer is then used to tokenize the text and build a vocabulary. The text sequences are padded or truncated to a fixed length and passed through an embedding layer to map tokens into low-dimensional dense vectors. Since the embedding layer only performs initial vector conversion without contextual understanding or complex semantic modeling, the subsequent LSTM layer processes these embedded token vectors to output a semantic vector representing the entire text sequence. The second method directly uses the ZHIPU AI Embedding-3 large model to embed the entire text segment without prior tokenization or Qwen3-based structured preprocessing. This approach treats the text as a whole, aiming to evaluate the large pretrained model's capacity for text comprehension without explicit preprocessing.

In this experiment, the classification or numerical feature branch was removed, and only the text vectors were used as model input to better assess the influence of different text-embedding strategies on model performance. The results, shown in Table 6, indicate that the method based on the ZHIPU AI Embedding-3 large model outperforms the traditional tokenization + LSTM approach. Furthermore, the Qwen3-based structured preprocessing of the text enhances the embedding model's performance compared to the direct embedding of unprocessed full text. These findings strongly support the superiority of our proposed LLM-based unstructured text extraction and embedding method in predictive tasks.

Table . Performance comparison of different text-embedding methods.

Text-embedding methods	MAE[a]	RMSE[b]	R^2[c]
Tokenizer + embedding layer (Keras) + LSTM[d]	18.7119	22.9938	−0.0010
ZHIPU Embedding-3 model (full text)	12.8774	17.2793	0.4347
ZHIPU Embedding-3 model (structured text)	12.6437[e]	17.0410[e]	0.4502[e]

[a]MAE: mean absolute error.

[b]RMSE: root mean squared error.

[c]R^2: coefficient of determination.

[d]LSTM: long short-term memory.

[e]These values in the last row indicate the best performance among all tested methods across each evaluation metric.

Feature Importance Evaluation

Permutation importance was used to assess the contribution of each feature within the multihead MLP model. The ranking of feature importance is presented in Multimedia Appendix 1. It is evident that disrupting the feature order of the primary surgical procedure leads to the largest increase in the model's MAE, indicating that the primary surgical procedure is the most critical feature by a significant margin compared to others. Following this, the surgical plan and preoperative diagnosis rank next in importance, while case characteristics and

XSL•FO
RenderX

preoperative preparation, including the unstructured text of the patient's prepuce description, exhibit the lowest importance.

Discussion

Principal Findings

This study demonstrates that an end-to-end surgical duration prediction model integrating heterogeneous multisource data, including structured patient and surgeon characteristics alongside unstructured EMR text, can significantly improve prediction accuracy for pediatric urology procedures. The proposed multihead MLP architecture, which processes categorical, numerical, and textual features through dedicated input branches, effectively addresses the challenge of heterogeneous data fusion and achieves state-of-the-art performance. In comprehensive comparisons against the "average" baseline and 6 established surgical duration prediction methods [23,37-41], our model attains the best results across all metrics (MAE=11.39 min; RMSE=15.58; R^2=0.54), confirming its advantage in this specialized clinical setting. This improvement is of clear clinical significance. In the dataset used for this study, each operating room typically handles 5 to 8 surgeries per day. Consequently, on a daily cumulative basis, the proposed multihead MLP model reduces the total prediction error by 38.3 to 61.4 minutes compared to the naive "average" baseline. Given that the mean surgical duration is only 38.6 (SD 24.9) minutes, this improvement is substantial, which implies the potential to schedule 1 to 2 additional surgeries, or crucially, to prevent the cancellation of the last patient due to schedule overruns. Furthermore, compared to the traditional ML-MDN model, our approach reduces the daily cumulative error by 29.4 to 47.1 minutes, which is roughly equivalent to the duration of 1 full surgery. Thus, the proposed model directly enhances operating room turnover efficiency and schedule predictability.

Advantages of Patient-Physician Heterogeneous Feature Fusion

Experimental results indicate that our multisource feature fusion strategy yields significant accuracy improvements over single-source models. Compared to models using only categorical or numerical features or only textual features, we achieve MAE reductions of 4.6% and 10.6%, respectively. The multihead MLP also outperforms the conventional single-head MLP approach by reducing MAE by 8.1%. This enhancement stems from our architecture's ability to prevent high-dimensional text vectors from dominating the training process at the expense of lower-dimensional features, such as patient age and weight. By providing separate subnetworks for each feature type, our model enables parallel learning across different dimensions, facilitating deeper feature interactions and synergistic enhancement. This design offers valuable methodological insights for addressing similar heterogeneous data integration challenges in medical applications.

Automated EMR Text Processing Pipeline

Our second major contribution is an automated EMR text processing pipeline leveraging large-scale pretrained models. This approach demonstrates several advantages over traditional tokenization with LSTM networks, including immediate usability, multilingual support, and superior handling of long text inputs. Furthermore, our structured field extraction preprocessing method yields higher-quality embeddings compared to direct processing of raw text segments, confirming the effectiveness of our automated approach in capturing deep semantic relationships within clinical narratives.

Advantages Over Traditional Methods

In the Results section, we observed that the traditional ML-MDN model (LSTM + MLP) performed significantly worse than our proposed multihead MLP model, falling behind even the simpler RR baseline. This performance discrepancy can be attributed to 3 primary factors: first, the difference in text-encoding mechanisms. ML-MDN relies on LSTMs, which process text sequentially and often struggle with long-range dependencies inherent in verbose medical records. In contrast, Qwen uses a self-attention mechanism, allowing the model to capture global semantic relationships simultaneously and ensuring that critical clinical cues are effectively integrated regardless of their position. This superiority of LLM-based encoding is further corroborated by our ablation study in the Results section. The second factor is the divergence in feature processing strategies. ML-MDN applies a uniform LSTM encoding to all unstructured text fields (including surgical plans and case characteristics). Conversely, our approach adopts a hybrid strategy: we reserve LLM-based encoding exclusively for verbose narrative fields (eg, case characteristics) while using precise multilabel encoding for semistructured concepts, such as surgical plans. The fact that the RR model using our proposed hybrid feature set outperformed ML-MDN serves as strong evidence for the effectiveness of this feature engineering strategy. The third factor is architectural differences. While ML-MDN relies on the simple concatenation of all features, our model uses a multihead input structure. As discussed in the Methods section, this design better handles heterogeneous data, further widening the performance gap.

Key Predictors and Clinical Insights

Permutation importance analysis identifies the primary surgical procedure as the most influential predictor, aligning with clinical experience that the specific type of surgery is the fundamental determinant of procedural duration. The surgical plan and preoperative diagnosis demonstrated secondary importance, reflecting their roles as key indicators of surgical complexity. Furthermore, the lead surgeon's years of practice and annual surgical volume also exhibited high predictive value; this finding underscores the necessity of incorporating surgeon-specific metrics into operating room scheduling. Interestingly, textual case descriptions displayed relatively lower importance, likely attributable to the clinical tendency to use standardized language when documenting similar conditions. For instance, descriptions of conditions, such as "small penile appearance" typically adhere to consistent phrasing patterns, limiting their discriminative power. This observation suggests that refining clinical documentation practices could offer opportunities to enhance future predictive modeling.

XSL•FO
RenderX

Limitations and Further Research Directions

A primary limitation of this research stems from its reliance on data obtained from a single tertiary pediatric hospital. While the dataset exhibits strong internal consistency and an adequate sample size, the single-center provenance may nonetheless constrain the model's generalizability across other health care settings, particularly where disparities exist in clinical workflows, surgical practices, or electronic health record documentation standards. Such limitations are frequently encountered in medical artificial intelligence applications. Future efforts could involve validating model robustness through multicenter collaborations and exploring approaches, such as domain adaptation or lightweight transfer learning. For instance, fine-tuning models with limited annotated data from target institutions improve applicability across diverse clinical environments.

Moreover, the current predictive model relies exclusively on preoperative data, whereas actual operative duration is frequently influenced by intraoperative factors that cannot be anticipated. For example, the incidental identification of contralateral cryptorchidism may necessitate a shift from a unilateral to a bilateral procedure, while unforeseen events, such as anatomical anomalies or intraoperative bleeding, can extend surgical time. Such dynamic intraoperative developments fall outside the scope of static preoperative variables and may introduce deviations in predictions. To this end, we plan to build a 2-stage prediction framework in future work: the first stage generates baseline estimates using preoperative data, and the second stage dynamically corrects these estimates by integrating real-time information, such as laparoscopic video streams, anesthesia records, and instrument usage logs, after the operation begins. This approach has the potential to maintain the efficiency of preoperative planning while improving the adaptability and precision of resource allocation during surgery.

Conclusion

This study tackles the clinical challenge of surgical duration prediction in pediatric urology by introducing a specialized predictive model that leverages an innovative multimodal data fusion approach. Our findings demonstrate that combining LLM-based text processing with a multihead MLP architecture successfully integrates structured clinical data with unstructured medical narratives, leading to a significant improvement in prediction accuracy. The model outperforms existing methods across key performance metrics, including MAE and RMSE, confirming the value of heterogeneous feature collaboration in predictive modeling. Feature importance analysis identifies critical predictors, such as the primary surgical procedure, surgical plan, and preoperative diagnosis, offering evidence-based insights for clinical resource allocation. Additionally, sensitivity analysis underscores the importance of pediatric-specific clinical characteristics, including developmental stage variations and urinary tract malformation severity, reinforcing the need for specialty-tailored modeling approaches. These results provide reliable decision-support for operating room scheduling and establish a methodological framework for surgical duration prediction in other specialized fields.

While the study yields promising results, certain limitations must be acknowledged, particularly the reliance on single-institution data. Future research should prioritize multicenter data collection to enhance model generalizability, investigate dynamic prediction methods that incorporate real-time intraoperative data, and develop more interpretable tools for analyzing feature interactions. Such refinements would further strengthen the model's applicability in complex clinical settings. Overall, this work presents a practical and effective technological solution for intelligent surgical management in pediatric subspecialties, with meaningful implications for optimizing operating room efficiency and health care resource use.

Acknowledgments

The authors have not used generative artificial intelligence tools at any stage in the preparation of this manuscript.

Funding

This study was partially supported by the Key R&D Program of Zhejiang (2023C03101) and the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2024KY1176).

Data Availability

Data that support the findings of the study are available upon request from the corresponding author.

Authors' Contributions

Conceptualization: YZ
Methodology: RL, YS, ZZ
Software: RL, ZZ
Validation: LC, JH
Formal analysis: GC
Visualization: RL, YS
Funding acquisition: YZ, GY

XSL•FO
RenderX

Writing—manuscript: YZ, RL, YS, ZZ
Writing—review and editing: YZ, RL, YS, ZZ
Supervision: ZZ, GY
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Feature importance ranking of different variables in the proposed surgical duration prediction model.
[PNG File, 72 KB - medinform_v14i1e82329_app1.png]

References

1. MacMillan L, Madura GM, Elliot M, et al. What affects operating room turnover time? A systematic review and mapping of the evidence. Surgery 2025 May;181:109263. [doi: 10.1016/j.surg.2025.109263] [Medline: 40054053]
2. Wu Q, Xie N, Shao Y. Day surgery appointment scheduling with patient preferences and stochastic operation duration. Technol Health Care 2021;29(4):697-708. [doi: 10.3233/THC-192086] [Medline: 33386830]
3. Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. Anesth Analg 2008 Apr;106(4):1232-1241. [doi: 10.1213/ane.0b013e318164f0d5] [Medline: 18349199]
4. Strum DP, May JH, Vargas LG. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. Anesthesiology 2000 Apr;92(4):1160-1167. [doi: 10.1097/00000542-200004000-00035] [Medline: 10754637]
5. Stepaniak PS, Vrijland WW, de Quelerij M, de Vries G, Heij C. Working with a fixed operating room team on consecutive similar cases and the effect on case duration and turnover time. Arch Surg 2010 Dec;145(12):1165-1170. [doi: 10.1001/archsurg.2010.255] [Medline: 21173290]
6. Gabriel RA, Harjai B, Simpson S, Goldhaber N, Curran BP, Waterman RS. Machine learning-based models predicting outpatient surgery end time and recovery room discharge at an ambulatory surgery center. Anesth Analg 2022 Jul 1;135(1):159-169. [doi: 10.1213/ANE.0000000000006015] [Medline: 35389380]
7. Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 2007 Feb;10(1):13-24. [doi: 10.1007/s10729-006-9005-4] [Medline: 17323652]
8. Pandit JJ. Rational planning of operating lists: a prospective comparison of “booking to the mean” vs. “probabilistic case scheduling” in urology. Anaesthesia 2020 May;75(5):642-647. [doi: 10.1111/anae.14958] [Medline: 31867710]
9. Park HS, Kim SH, Bong MR, et al. Optimization of the operating room scheduling process for improving efficiency in a tertiary hospital. J Med Syst 2020 Aug 15;44(9):1-7. [doi: 10.1007/s10916-020-01644-0] [Medline: 32803733]
10. Zhang B, Su Q. Dynamic surgery scheduling based on an improved genetic algorithm. J Healthc Eng 2021;2021:1559050. [doi: 10.1155/2021/1559050] [Medline: 34853667]
11. Dexter F, Macario A, Epstein RH, Ledolter J. Validity and usefulness of a method to monitor surgical services’ average bias in scheduled case durations. Can J Anaesth 2005 Nov;52(9):935-939. [doi: 10.1007/BF03022054] [Medline: 16251558]
12. Dexter EU, Dexter F, Masursky D, Kasprowicz KA. Prospective trial of thoracic and spine surgeons’ updating of their estimated case durations at the start of cases. Anesth Analg 2010 Apr 1;110(4):1164-1168. [doi: 10.1213/ANE.0b013e3181cd6eb9] [Medline: 20145282]
13. Nasiri MM, Shakouhi F, Jolai F. A fuzzy robust stochastic mathematical programming approach for multi-objective scheduling of the surgical cases. OPSEARCH 2019 Sep;56(3):890-910. [doi: 10.1007/s12597-019-00379-y]
14. Vali-Siar MM, Gholami S, Ramezanian R. Multi-period and multi-resource operating room scheduling under uncertainty: a case study. Comput Ind Eng 2018 Dec;126:549-568. [doi: 10.1016/j.cie.2018.10.014]
15. Erdem E, Qu X, Shi J. Rescheduling of elective patients upon the arrival of emergency patients. Decis Support Syst 2012 Dec;54(1):551-563. [doi: 10.1016/j.dss.2012.08.002]
16. Tuwatananurak JP, Zadeh S, Xu X, et al. Machine learning can improve estimation of surgical case duration: a pilot study. J Med Syst 2019 Jan 17;43(3):1-7. [doi: 10.1007/s10916-019-1160-5] [Medline: 30656433]
17. Tankard K, Acciavatti TD, Vacanti JC, et al. Contributors to operating room underutilization and implications for hospital administrators. Health Care Manag (Frederick) 2018;37(2):118-128. [doi: 10.1097/HCM.0000000000000214] [Medline: 29578931]
18. Abdelrazig Merghani AM, Ahmed Esmail AK, Mubarak Osman AME, Abdelfrag Mohamed NA, Mohamed Ali Shentour SM, Abdelrazig Merghani SM. The role of machine learning in management of operating room: a systematic review. Cureus 2025 Feb;17(2):e79400. [doi: 10.7759/cureus.79400] [Medline: 40125180]
19. Macario A. What does one minute of operating room time cost? J Clin Anesth 2010 Jun;22(4):233-236. [doi: 10.1016/j.jclinane.2010.02.003] [Medline: 20522350]

XSL•FO
RenderX

20. Abbas A, Mosseri J, Lex JR, et al. Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty. Int J Med Inform 2022 Feb;158:104670. [doi: 10.1016/j.ijmedinf.2021.104670] [Medline: 34971918]
21. Gabriel RA, Harjai B, Simpson S, et al. An ensemble learning approach to improving prediction of case duration for spine surgery: algorithm development and validation. JMIR Perioper Med 2023 Jan 26;6:e39650. [doi: 10.2196/39650] [Medline: 36701181]
22. DelliCarpini G, Passano B, Yang J, et al. Utilization of machine learning models to more accurately predict case duration in primary total joint arthroplasty. J Arthroplasty 2025 May;40(5):1185-1191. [doi: 10.1016/j.arth.2024.10.100] [Medline: 39477036]
23. Jiao Y, Xue B, Lu C, Avidan MS, Kannampallil T. Continuous real-time prediction of surgical case duration using a modular artificial neural network. Br J Anaesth 2022 May;128(5):829-837. [doi: 10.1016/j.bja.2021.12.039] [Medline: 35090725]
24. Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. Comput Methods Programs Biomed 2021 Sep;208:106220. [doi: 10.1016/j.cmpb.2021.106220] [Medline: 34161848]
25. Babayoff O, Shehory O, Shahoha M, Sasportas R, Weiss-Meilik A. Surgery duration: optimized prediction and causality analysis. PLoS One 2022;17(8):e0273831. [doi: 10.1371/journal.pone.0273831] [Medline: 36037243]
26. Miller LE, Goedicke W, Crowson MG, Rathi VK, Naunheim MR, Agarwala AV. Using machine learning to predict operating room case duration: a case study in otolaryngology. Otolaryngol Head Neck Surg 2023 Feb;168(2):241-247. [doi: 10.1177/01945998221076480] [Medline: 35133897]
27. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. JMIR AI 2023 Sep 8;2(1):e44909. [doi: 10.2196/44909] [Medline: 38875567]
28. Riahi V, Hassanzadeh H, Khanna S, et al. Improving preoperative prediction of surgery duration. BMC Health Serv Res 2023 Dec 2;23(1):1343. [doi: 10.1186/s12913-023-10264-6] [Medline: 38042831]
29. Strömblad CT, Baxter-King RG, Meisami A, et al. Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. JAMA Surg 2021 Apr 1;156(4):315-321. [doi: 10.1001/jamasurg.2020.6361] [Medline: 33502448]
30. Huang L, Chen X, Liu W, Shih PC, Bao J. Automatic surgery and anesthesia emergence duration prediction using artificial neural networks. J Healthc Eng 2022;2022:2921775. [doi: 10.1155/2022/2921775] [Medline: 35463687]
31. Park JB, Roh GH, Kim K, Kim HS. Development of predictive model of surgical case durations using machine learning approach. J Med Syst 2025 Jan 14;49(1):8. [doi: 10.1007/s10916-025-02141-y] [Medline: 39808376]
32. Frassini E, Vijfvinkel TS, Butler RM, van der Elst M, Hendriks BHW, van den Dobbelsteen JJ. Deep learning methods for clinical workflow phase-based prediction of procedure duration: a benchmark study. Comput Assist Surg (Abingdon) 2025 Dec;30(1):2466426. [doi: 10.1080/24699322.2025.2466426] [Medline: 39992712]
33. Ramamurthi A, Neupane B, Deshpande P, et al. Applying large language models for surgical case length prediction. JAMA Surg 2025 Aug 1;160(8):894-902. [doi: 10.1001/jamasurg.2025.2154] [Medline: 40632526]
34. Dexter F, Epstein RH. Operating room efficiency and scheduling. Curr Opin Anaesthesiol 2005 Apr;18(2):195-198. [doi: 10.1097/01.aco.0000162840.02087.15] [Medline: 16534338]
35. May JH, Spangler WE, Strum DP, Vargas LG. The surgical scheduling problem: current research and future opportunities. Prod Oper Manag 2011 May;20(3):392-405. [doi: 10.1111/j.1937-5956.2011.01221.x]
36. van Eijk RPA, van Veen-Berkx E, Kazemier G, Eijkemans MJC. Effect of individual surgeons and anesthesiologists on operating room time. Anesth Analg 2016 Aug;123(2):445-451. [doi: 10.1213/ANE.0000000000001430] [Medline: 27308953]
37. Zhao B, Waterman RS, Urman RD, Gabriel RA. A machine learning approach to predicting case duration for robot-assisted surgery. J Med Syst 2019 Jan 5;43(2):32. [doi: 10.1007/s10916-018-1151-y] [Medline: 30612192]
38. Master N, Zhou Z, Miller D, Scheinker D, Bambos N, Glynn P. Improving predictions of pediatric surgical durations with supervised learning. Int J Data Sci Anal 2017 Aug;4(1):35-52. [doi: 10.1007/s41060-017-0055-0]
39. Li X, Zhan Y, Ni J, et al. Multimodal feature analysis for surgery duration predication. Presented at: Proceedings of the 2024 16th International Conference on Bioinformatics and Biomedical Technology (ICBBT '24); May 24-26, 2024; Chongqing, China p. 188-195. [doi: 10.1145/3674658.3674688]
40. Bellini V, Guzzon M, Bigliardi B, Mordonini M, Filippelli S, Bignami E. Artificial intelligence: a new tool in operating room management. role of machine learning models in operating room optimization. J Med Syst 2019 Dec 10;44(1):20. [doi: 10.1007/s10916-019-1512-1] [Medline: 31823034]
41. Bartek MA, Saxena RC, Solomon S, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. J Am Coll Surg 2019 Oct;229(4):346-354. [doi: 10.1016/j.jamcollsurg.2019.05.029] [Medline: 31310851]

Abbreviations

DT: decision tree regression
EMR: electronic medical record
GBR: gradient boosting regression tree

XSL•FO
RenderX

LLM: large language model
LSTM: long short-term memory
MAE: mean absolute error
ML-MDN: machine learning–mixture density network
MLP: multilayer perceptron
RF: random forest regression
RMSE: root mean square error
RR: ridge regression
SVR: support vector regression

Edited by A Benis; submitted 13.Aug.2025; peer-reviewed by MH Tsai, T Ashihara; revised version received 06.Mar.2026; accepted 06.Mar.2026; published 28.Apr.2026.

Please cite as:
Zhao Y, Lin R, Sun Y, Chen L, Huang J, Chen G, Zhu Z, Yu G
Predicting Pediatric Urological Surgery Duration Through Multimodal Patient-Physician Feature Fusion: Deep Learning Framework Incorporating Clinical Text Embedding
JMIR Med Inform 2026;14:e82329
URL: https://medinform.jmir.org/2026/1/e82329
doi:10.2196/82329

© Yonggen Zhao, Ruoge Lin, Yiying Sun, Lingdong Chen, Jian Huang, Guangjie Chen, Zhu Zhu, Gang Yu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Data-Efficient Language Model for Assessing Pulmonary Embolism Diagnostic Certainty From Radiology Reports: Model Development and Validation Study

Feifan Liu[1*], PhD; Ruofan Hu[2*], MS; Donghyuk Kim[1,2], MS; Hao Lo[1], MD, MBA; Yevgeniy Kharonov[3], DO; Ansh Johri[1], MD; Ben S Gerber[1], MPH, MD; Elke Rundensteiner[1,2], PhD; Lauren M Westafer[4], DO; Jeroan Allison[1], MD; Catarina Kiefe[1], MD, PhD; Alexander Bankier[1], MD, PhD; Max P Rosen[1], MPH, MD

[1]Chan Medical School, University of Massachusetts, 55 Lake Avenue North, Worcester, MA, United States

[2]Data Science, Worcester Polytechnic Institute, Worcester, MA, United States

[3]Department of Radiology, University of Miami Miller School of Medicine, Miami, FL, United States

[4]Chan Medical School, University of Massachusetts, Baystate, MA, United States

[*]these authors contributed equally

Corresponding Author:
Feifan Liu, PhD
Chan Medical School, University of Massachusetts, 55 Lake Avenue North, Worcester, MA, United States

Abstract

Background: Computed tomography pulmonary angiography (CTPA) is the standard imaging modality for diagnosing pulmonary embolism (PE), but diagnostic uncertainty is common due to technical limitations and vague language, leading to inconsistent interpretation and clinician frustration.

Objective: This study develops a prompt-free, data-efficient method for assessing diagnostic certainty of PE in CTPA reports using small pretrained language models.

Methods: This study examined 173 consecutive CTPA reports from UMass Memorial Health, each annotated by 3 radiologists for PE diagnostic certainty. We developed PECertainty, a lightweight, prompt-free model, and compared it with advanced large language model (LLM)–based methods under limited supervision settings. Baselines included prompt-free methods (support vector machine, random forest, and RoBERTa [Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach]) and prompt-dependent methods (LLM fine-tuning, in-context learning, and ADAPET [A Densely-Supervised Approach to Pattern Exploiting Training]; UNC Chapel Hill) with open-source Gemma3-4B (Google DeepMind) and Llama3.2-3B (Meta), and the proprietary GPT-3.5 (OpenAI). Sensitivity analyses evaluated performance with 1 to 10 training examples per category for the top performer. Model performance was evaluated against radiologist annotations. External validation on 420 CTPA reports from the Baystate Medical Center, with validation limited to distinguishing certain from uncertain reports. Interpretability of the top-performing models (PECertainty and GPT-3.5) was evaluated using integrated gradients and prompt-based explanations reviewed by radiologists.

Results: Among prompt-dependent methods, GPT-3.5 fine-tuning (F_1-score 0.86; 95% CI 0.71 - 1.0) and in-context learning (F_1-score 0.87; 95% CI 0.71 - 1.0) performed best, and the performance of in-context learning consistently outperformed 0-shot learning for Gemma3-4B (F_1-score 0.63, 95% CI 0.56 - 0.79 vs F_1-score 0.45; 95% CI 0.29 - 0.56) and Llama3.2-3B (F_1-score 0.54; 95% CI 0.41 - 0.71 vs F_1-score 0.43, 95% CI 0.28 - 0.62). PECertainty demonstrated numerically better or equivalent performance compared with both the top-performing prompt-dependent methods and all prompt-free baselines. Compared with fine-tuned ClinicalBERT (Bidirectional Encoder Representations From Transformers Pretrained on Clinical Text), PECertainty achieved statistically significant improvements across all metrics (paired bootstrap significance test, $P<.05$). RoBERTa (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach) fine-tuning lagged (F_1-score 0.52; 95% CI 0.35 - 0.71), and simple models such as support vector machine underperformed. In few-shot settings (10 examples/category), PECertainty (F_1-score 0.80; 95% CI 0.59 - 0.94) outperformed both GPT-3.5 fine-tuning (F_1-score 0.74; 95% CI 0.58 - 0.88) and in-context learning (F_1-score 0.65; 95% CI 0.47 - 0.83). External validation on the Baystate dataset showed good generalization for distinguishing certain from uncertain cases (F_1-score 0.77; 95% CI 0.70 - 0.83). Despite its strong performance, PECertainty was rated as less interpretable than fine-tuned GPT-3.5 by radiologists (t test, $P<.05$).

XSL•FO
RenderX

Conclusions: PECertainty enables accurate and data-efficient assessment of diagnostic certainty from free-text CTPA reports in low-resource settings. As an open-source, lightweight alternative to proprietary LLMs, it may support more precise communication between radiologists and referring physicians, with interpretability identified as a key direction for improvement.

(JMIR Med Inform 2026;14:e79972) doi:10.2196/79972

KEYWORDS

pulmonary embolism; radiology reports; natural language processing; machine learning; diagnostic certainty

Introduction

Diagnostic certainty in evaluating pulmonary embolism (PE) is critical, as both missed diagnoses and overdiagnosis can result in serious patient harm, including delayed life-saving treatment or unnecessary anticoagulation with associated bleeding risks. In the emergency department (ED), prompt evaluation of suspected PE is essential for risk assessment and treatment. Computed tomography pulmonary angiography (CTPA) is widely used as the diagnostic standard due to its accessibility and high sensitivity. However, CTPA has limitations, including variability in contrast timing and respiratory motion artifacts. Thus, diagnostic uncertainty in CTPA reports is common and usually conveyed in vague language that can be inconsistently interpreted.

Referring clinicians often express frustrations with ambiguous linguistic hedging, emphasizing the need for clearer and more consistent communication from radiologists as a top priority for improving care quality [1-3]. For example, a reporting statement “No evidence of pulmonary embolism to the lobar level. No airspace or interstitial disease. Mild left axillary lymphadenopathy” appears reassuring at first glance, yet the hedge phrase “to the lobar level” introduces a hedge that indicates reduced diagnostic confidence, as it suggests that while no PE is observed at the lobar level, emboli in the more distal segmental or subsegmental arteries cannot be definitively ruled out. Such subtle expressions of uncertainty can have significant effects on clinical practice. In the high-pressure ED setting, unclear language on diagnostic reporting can give false reassurance or result in overly cautious clinical decisions. Both situations can harm patient outcomes. Thus, clear communication between radiologists and clinicians is essential to ensure accurate diagnosis, improve patient care, reduce mortality, and prevent unnecessary anticoagulation [4,5], especially in the ED setting where clinicians have to make quick and accurate treatment decisions.

Previous efforts to improve report clarity have focused on structured reporting systems that standardize phrases such as “consistent with (>90% certainty)” and “possibly (≈50% certainty)” to improve diagnostic reporting [6]. However, these approaches often focus narrowly on lexicon standardization, overlooking important contextual information such as “rambling” differential diagnoses, using long and vague sentences, and indicating technical limitations. The predefined certainty categories are typically broad and may lack meaningful interpretability for treating clinicians. Although structured reporting is promising, there are recognized limitations. These include interference with radiologist image interpretation, reduced descriptive power by limiting expressive flexibility, added burden of learning and using preferred terminologies, and required availability of lexicons and templates across imaging disciplines (not generalizable) [7,8]. These templates are often institution-specific, further limiting their generalizability across health care settings.

Studies found that a structured report accompanied by a radiologist’s comment is the most preferred style by referring physicians [9]. The use of speech-to-text dictation remains desirable, where radiologists use natural language as though speaking to a human transcriptionist [7]. Successful integration of dictated comments and structured reporting can help bridge the preference gaps in report format and style between radiologists and referring physicians [10] and address issues related to adopting new technological advances [11]. For this integration to be successful, a standardized and calibrated context-aware uncertainty measurement is needed for improved diagnostic uncertainty communication in radiology reporting.

Advances in natural language processing (NLP) with fine-tuning pretrained language models (PLMs) have been successfully applied to different tasks in radiology, including clinical information extraction [12], report generation [13], abnormal sentence classification [14], report coding [15], and diagnostic certainty assessment [16]. However, few studies explored advanced NLP for PE diagnostic certainty assessment, especially when annotated data is limited. Chapman et al [17] developed a rule-based system only to detect uncertainty in PE diagnostic reports (certain vs uncertain), achieving 93% accuracy. However, their approach relied on manually curated rules, which are labor-intensive, nonscalable, and unable to capture probabilistic gradations or longer contextual dependencies. Additionally, the system could not effectively model the interaction between diagnostic findings (positive or negative) and certainty levels (certain or uncertain), limiting its clinical utility.

To address the aforementioned limitations, we developed a lightweight, open-source model, PECertainty, to assess the level of diagnostic certainty expressed in radiology reports for PE. Our goal was to build a solution that performs well even with limited labeled data and does not rely on manually crafted prompts or access to large proprietary models, such as GPT-3 [18]. To this end, we adopted the prompt-free and data-efficient SetFit framework [19] to fine-tune smaller PLMs. We evaluated the performance of PECertainty against a range of strong baselines, including prompt-free methods (SVM [support vector machine], random forest, and RoBERTa) and prompt-dependent approaches, such as large language model (LLM) fine-tuning, in-context learning, and ADAPET [A Densely-Supervised Approach to Pattern Exploiting Training]. The LLMs evaluated include open-source models (Gemma3-4B and Llama3.2-3B)

XSL•FO
RenderX

as well as the proprietary GPT-3.5. We further assess PECertainty under varying training data sizes to evaluate its robustness in low-resource settings. To examine its generalization capability, we test PECertainty on an external dataset. Finally, we compare the interpretability of PECertainty with that of competing models based on radiologist evaluations.

Methods

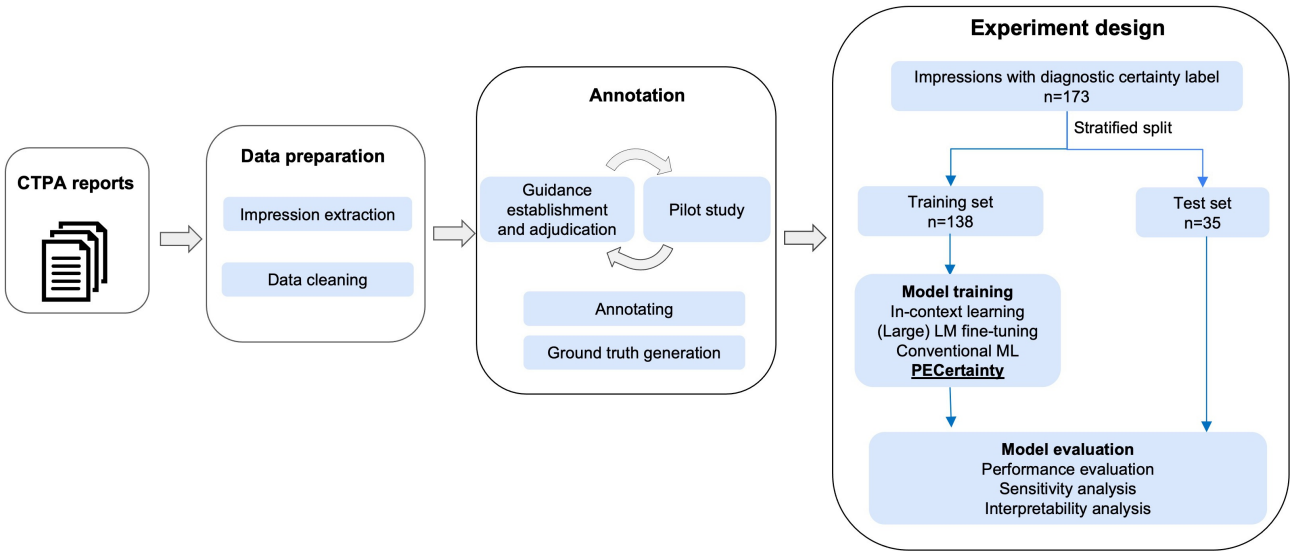
Ethical Considerations

This study was approved by the Institutional Review Board at the University of Massachusetts Chan Medical School (IRB# H00022301). This study involved a retrospective analysis of existing, deidentified clinical text data, and the requirement for informed consent was waived in accordance with applicable federal regulations (45 CFR 46.104[d](4)) [20]. No protected health information or personally identifiable information was accessed or used in this study. All data were deidentified before analysis, and appropriate safeguards were implemented to maintain participant privacy and data confidentiality, including secure data storage and restricted access protocols. Participants received no compensation.

Study Design and Datasets

Figure 1 provides an overview of our study design. The CTPA reports for model training were retrieved from the UMass Memorial Health (UMMH) system using the Nuance M-Power tool and PE-related terms from January 10, 2017, to May 11, 2021 (see Multimedia Appendix 1 for data retrieval details). A total of 205 reports were extracted and converted to plain text. Then, impression sections were extracted using regular expressions. A total of 3 board-certified radiologists independently classified each impression into 1 of 5 certainty categories (Table 1) based on consensus-based annotation guidelines (Multimedia Appendix 1). To ensure high annotation quality, we used an iterative process to refine guidelines and enhance interannotator agreement. Due to class imbalance, we merged the probable positive and probable negative categories into a single “probable” certainty category and excluded nondiagnostic examples. This adjustment resulted in 173 annotated notes for this study: definitive negative (n=106), probable (n=35), and definitive positive (n=32). We performed a stratified data split, using 80% (n=138) for training and 20% (n=35) for testing. Note that in our dataset, each report belongs to a unique patient, so report- and patient-level data splits are equivalent.

Figure 1. Overview of our study design. Data from October 1, 2017, to November 5, 2021, were retrieved, and a total of 205 reports were sampled and preprocessed for this study. This study includes data preparation, annotation, and experiment design. CTPA: Computed tomography pulmonary angiography; LM: language model; ML: machine learning.



XSL•FO
RenderX

Table . Diagnostic certainty of diagnosis in the impression section of a radiology report—categories for annotation.

Certainty categories	Interpretation	Example of impression section (abridged)
Definitive positive	Describing discrete diagnostic positive findings with no hedging (>90% confidence) words.	Bilateral pulmonary embolism originating within the distal main pulmonary arteries and extending up to subsegmental levels…
Definitive negative	Describing discrete diagnostic findings with no hedging (>90% confidence) words.	No evidence of pulmonary embolism. Improving ground glass opacities, likely sequela of prior COVID pneumonia…
Probable positive	Containing contextual information indicating uncertainty with hedging or other differential diagnoses or mentioning suboptimal examination, although "PE"[a] or "evidence of PE" may be mentioned.	There is a low-density filling defect in a segmental pulmonary artery branch, which may represent an embolus but could also be an artifact.
Probable negative	Containing contextual information indicating uncertainty with hedging or other differential diagnoses or mentioning involved suboptimal examination, although "no PE" or "no evidence of PE" may be mentioned.	Mildly limited study due to underopacification of the pulmonary arteries. Within this limitation, there is no evidence for pulmonary embolism up to the lobar branches…
Nondiagnostic	Containing an explicit description of being nondiagnostic due to technical and/or patient-related limitations.	The bolus of contrast is concentrated in the thoracic aorta and not in the pulmonary arteries. The examination is non-diagnostic for the evaluation of emboli of the pulmonary arteries. If there is a concern for pulmonary embolus, consider repeat…

[a]PE: pulmonary embolism.

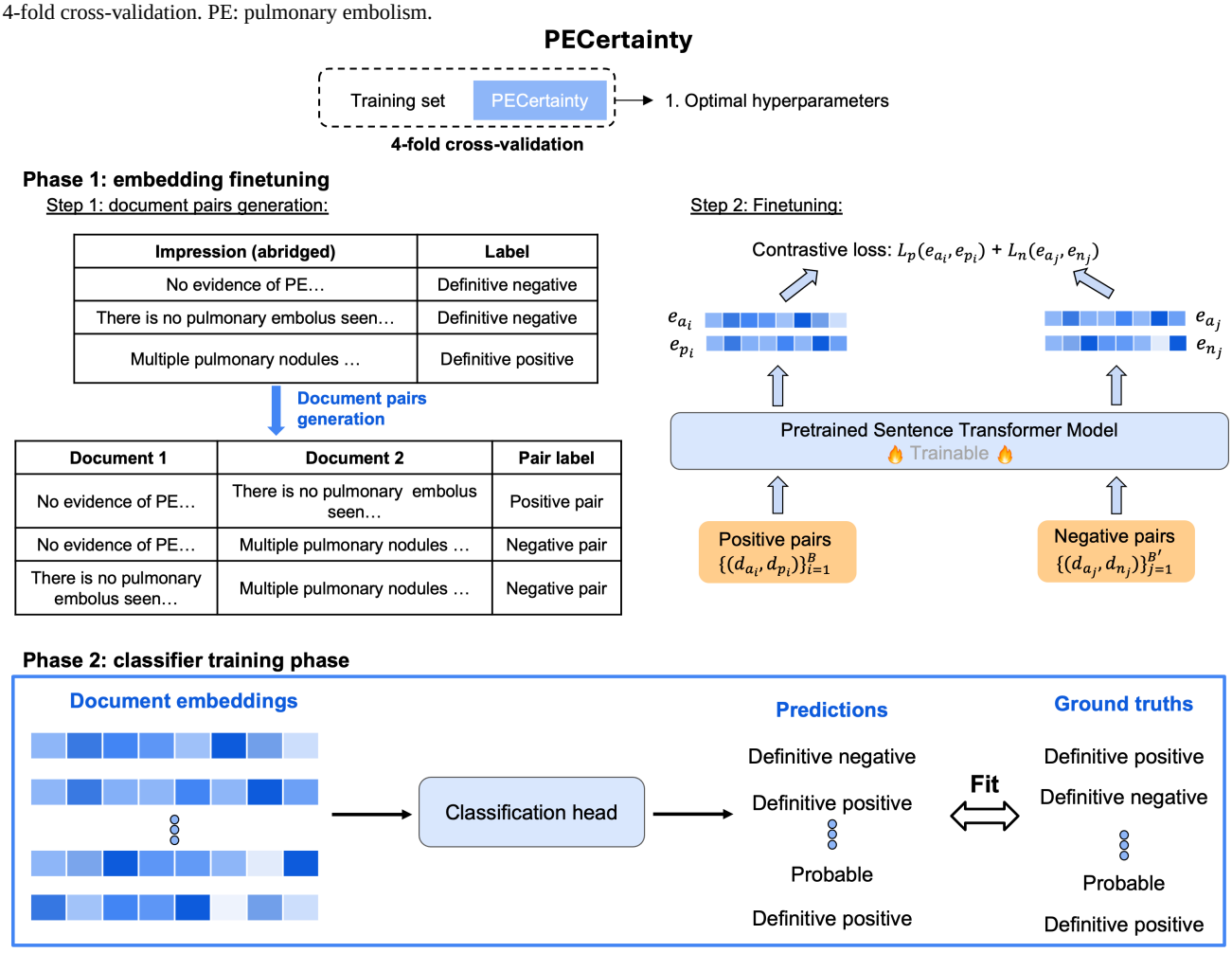
We compared PECertainty against competitive baselines in terms of classification performance and conducted sensitivity analyses across varying training data sizes. To evaluate generalizability, we tested PECertainty on an external dataset of 420 reports from Baystate Medical Center, annotated using a different guideline: definitive (n=356; combining definitive positive and negative) and probable (n=64). Additionally, we performed interpretability analysis to assess the transparency and explainability of PECertainty. The demographic information of the datasets can be found in Multimedia Appendix 1.

PECertainty

We developed PECertainty, a prompt-free data-efficient model for assessing diagnostic certainty of PE findings in radiology reports, leveraging the SetFit framework [19]. Figure 2 illustrates the PECertainty development pipeline, which consists of 2 key phases.

XSL•FO
RenderX

Figure 2. A total of 2-phase training framework of PECertainty. The framework consists of 2 phases. In phase 1, positive and negative document pairs are generated based on diagnostic certainty labels and used to fine-tune a pretrained sentence transformer using contrastive loss. In phase 2, sentence embeddings from the fine-tuned model are used to train a classification head for diagnostic certainty prediction. The hyperparameters are decided based on 4-fold cross-validation. PE: pulmonary embolism.



In phase 1, we generate sentence pairs for contrastive learning by sampling positive pairs from sentences belonging to the same diagnostic certainty class and negative pairs from sentences across different classes. These pairs are used to fine-tune the pretrained sentence transformer model "all-roberta-large-v1," chosen for its strong empirical performance in sentence embedding tasks [21], using a contrastive loss based on cosine similarity. This process yields task-specific embeddings that encode subtle distinctions in diagnostic certainty classes.

In phase 2, we use the learned embeddings as input to a classification head, which is trained to predict 1 of the 3 certainty categories: definitive negative, probable, and definitive positive. This 2-phase training approach allows PECertainty to learn semantically meaningful and diagnostically aligned representations with minimal supervision.

Baselines

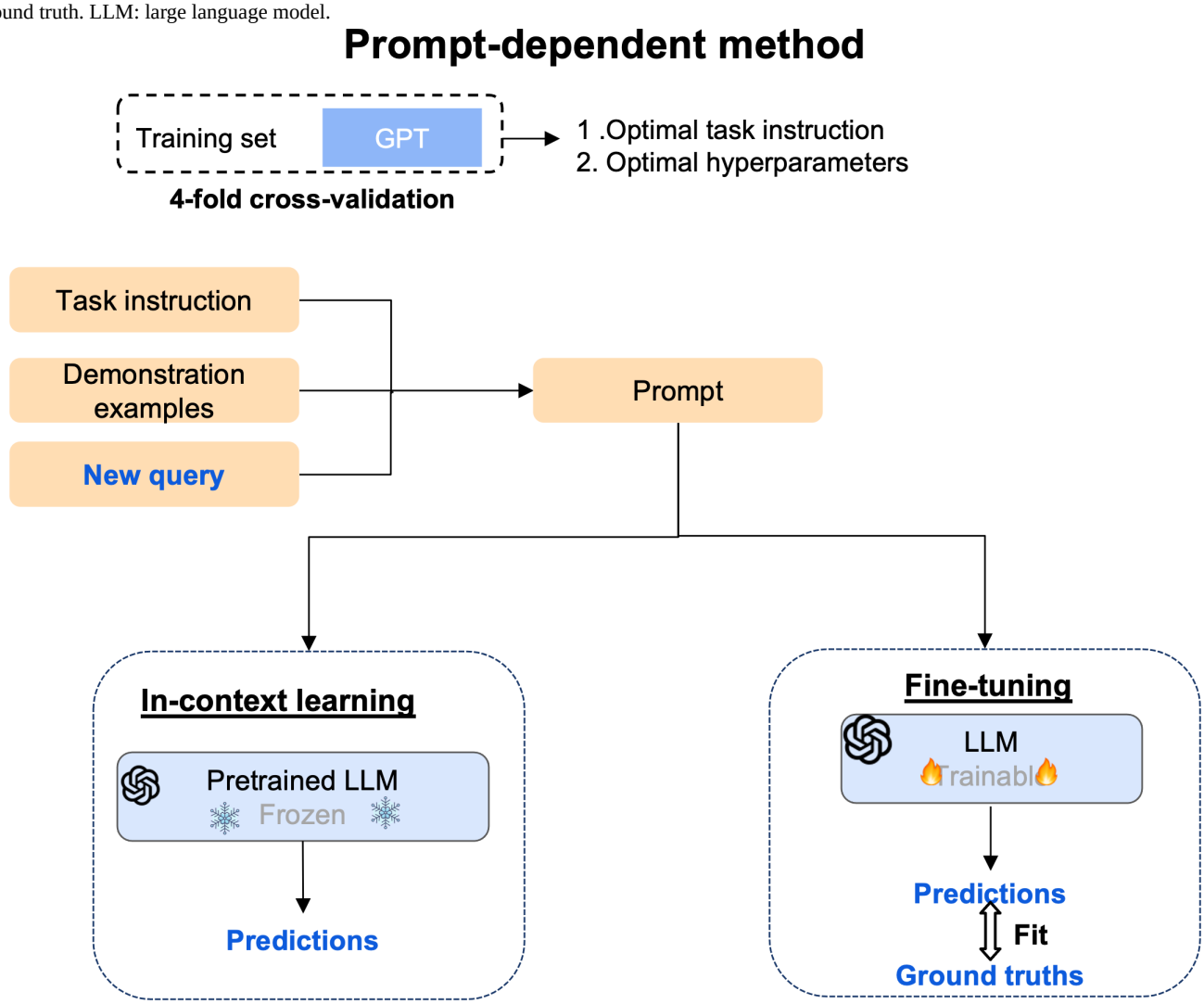
We compare the performance of PECertainty against prompt-free and prompt-dependent methods. Prompt-free methods include conventional machine learning models, SVM [22], and random forest [23] integrated with the BioWordVec embedding model, as well as fine-tuned transformer-based models: RoBERTa-large, BioBERT (Bidirectional Encoder Representations From Transformers for Biomedical Text Mining), and ClinicalBERT (Bidirectional Encoder Representations From Transformers Pretrained on Clinical Text). For prompt-dependent methods, we evaluated state-of-the-art approaches, including ADAPET [24]. We also assessed LLM–based methods, including GPT-3.5 ("gpt-3.5-turbo-0125") fine-tuning and in-context learning [18], as well as Gemma3-4B and Llama3.2-3B in-context learning. For in-context learning, demonstration examples were ranked based on their similarity to the query instance.

Model Implementation

As shown in Figures 2 and 3, hyperparameter optimization was conducted with grid search and 4-fold cross-validation on the training data, which also applied to other methods (Multimedia Appendix 2). The prompt-free methods process the impression section, generate embeddings using various pretrained models, and feed these into a classifier for prediction. In contrast, the prompt-dependent model, such as GPT-3.5, combines task instructions, demonstration examples, and a query impression note to generate a PE diagnostic certainty label (Figure 3). Additional details are in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 3. Pipeline of prompt-dependent method development. We select GPT-3.5 as the example to demonstrate the process. In in-context learning (left), a frozen pretrained LLM receives a prompt composed of task instructions, demonstration examples, and a new query and directly produces a prediction. In fine-tuning (right), an LLM is further trained on labeled data, optimizing model weights using a loss function between predictions and ground truth. LLM: large language model.



Robustness Analysis

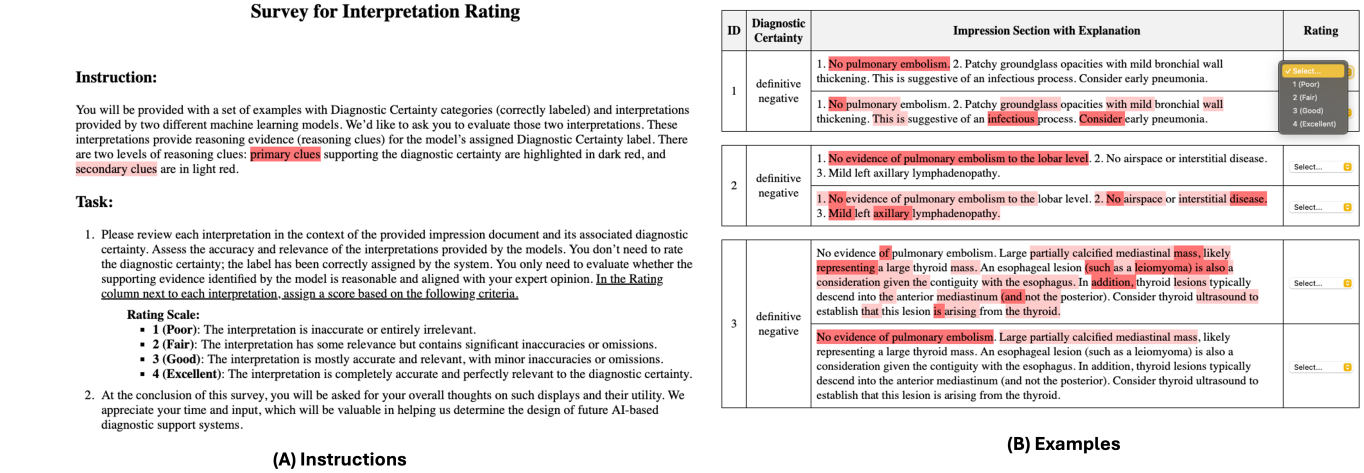
To assess the model's robustness, we applied a k-shot learning setting, where *k* represents the number of labeled demonstration examples per class. We randomly sampled up to 10 examples per category from the training set and systematically varied *k* from 1 to 10. GPT-3.5 fine-tuning was restricted to 5 or more shots per category due to API requirements. Additionally, we evaluate the PECertainty trained on UMMH data on the external data from Baystate Medical Center to evaluate the generalization capability. As the Baystate annotation guideline differs from that of UMMH and defines only definitive and probable classes, the definitive positive and definitive negative predictions were combined into a single definitive category for external validation.

Interpretability Analysis

We also compared the interpretability of the 2 top-performing models, fine-tuned GPT-3.5 and PECertainty. Model interpretation aims to provide transparent reasoning for the model's decision (ie, diagnostic certainty labels), which is important for clinical applications due to its high-stakes nature. For GPT-3.5, 0-shot prompting was used to prompt the model to identify primary and secondary clues based on the impression section and diagnostic certainty label (see Multimedia Appendix 3 for the prompt templates). For PECertainty, the integrated gradient method [21] was applied to calculate word importance scores. Words above the 75th percentile were considered primary clues, and those between the 50th and 75th percentiles were secondary clues. A total of 3 radiologists rated these clues on a 1 - 4 scale (poor to excellent) using the criteria in Figure 4.

XSL•FO
RenderX

Figure 4. Screenshot of the interpretation rating survey. The primary clues supporting the diagnostic certainty are highlighted in dark red, and secondary clues are in light red. AI: artificial intelligence.



Survey for Interpretation Rating

Instruction:

You will be provided with a set of examples with Diagnostic Certainty categories (correctly labeled) and interpretations provided by two different machine learning models. We'd like to ask you to evaluate those two interpretations. These interpretations provide reasoning evidence (reasoning clues) for the model's assigned Diagnostic Certainty label. There are two levels of reasoning clues: primary clues supporting the diagnostic certainty are highlighted in dark red, and secondary clues are in light red.

Task:

1. Please review each interpretation in the context of the provided impression document and its associated diagnostic certainty. Assess the accuracy and relevance of the interpretations provided by the models. You don't need to rate the diagnostic certainty; the label has been correctly assigned by the system. You only need to evaluate whether the supporting evidence identified by the model is reasonable and aligned with your expert opinion. In the Rating column next to each interpretation, assign a score based on the following criteria.

   Rating Scale:
   - **1 (Poor)**: The interpretation is inaccurate or entirely irrelevant.
   - **2 (Fair)**: The interpretation has some relevance but contains significant inaccuracies or omissions.
   - **3 (Good)**: The interpretation is mostly accurate and relevant, with minor inaccuracies or omissions.
   - **4 (Excellent)**: The interpretation is completely accurate and perfectly relevant to the diagnostic certainty.
2. At the conclusion of this survey, you will be asked for your overall thoughts on such displays and their utility. We appreciate your time and input, which will be valuable in helping us determine the design of future AI-based diagnostic support systems.

(A) Instructions

ID	Diagnostic Certainty	Impression Section with Explanation	Rating
1	definitive negative	1. No pulmonary embolism. 2. Patchy groundglass opacities with mild bronchial wall thickening. This is suggestive of an infectious process. Consider early pneumonia.	Select...
		1. No pulmonary embolism. 2. Patchy groundglass opacities with mild bronchial wall thickening. This is suggestive of an infectious process. Consider early pneumonia.	✓ Select... 1 (Poor) 2 (Fair) 3 (Good) 4 (Excellent)
2	definitive negative	1. No evidence of pulmonary embolism to the lobar level. 2. No airspace or interstitial disease. 3. Mild left axillary lymphadenopathy.	Select...
		1. No evidence of pulmonary embolism to the lobar level. 2. No airspace or interstitial disease. 3. Mild left axillary lymphadenopathy.	Select...
3	definitive negative	No evidence of pulmonary embolism. Large partially calcified mediastinal mass, likely representing a large thyroid mass. An esophageal lesion (such as a leiomyoma) is also a consideration given the contiguity with the esophagus. In addition, thyroid lesions typically descend into the anterior mediastinum (and not the posterior). Consider thyroid ultrasound to establish that this lesion is arising from the thyroid.	Select...
		No evidence of pulmonary embolism. Large partially calcified mediastinal mass, likely representing a large thyroid mass. An esophageal lesion (such as a leiomyoma) is also a consideration given the contiguity with the esophagus. In addition, thyroid lesions typically descend into the anterior mediastinum (and not the posterior). Consider thyroid ultrasound to establish that this lesion is arising from the thyroid.	Select...

(B) Examples

Statistical Analysis

To evaluate radiologists' agreement in categorizing impression notes, we used 3 statistics: Cohen [25], Fleiss [26], and Krippendorff [27]. We evaluated model performance using 5 standard metrics: accuracy, sensitivity, specificity, F_1-score, and the area under the receiver operating characteristic curve, with 95% CIs calculated via 2000 times bootstrap resampling. Area under the curve (AUC) was not applicable for GPTs, as they provide token probabilities rather than probabilities for all categories. Overall performance was assessed using the macro-average, which averages metrics across categories.

For the interpretability analysis, we used the 2-tailed Wilcoxon signed-rank test [28] to compare the interpretability ratings of LLM and PECertainty from the same rater. This tests the null hypothesis of no significant difference between the ratings assigned to the 2 models by a given rater. Statistical analyses were conducted in Python (version 3.10; Python Software Foundation) using *scikit-learn*,*krippendorff*,*statsmodels*, and *scipy* packages. To provide a more robust and statistically sound comparison, we employed a non-parametric bootstrap resampling procedure (B=1,000) to estimate confidence intervals for all evaluation metrics. In addition, we conducted paired bootstrap significance tests to assess performance differences between the baseline methods and PECertainty (B = 1,000). Statistical significance was determined using the bootstrap percentile method; results were considered statistically significant at the 0.05 level if the 95% confidence interval did not include zero.

Results

Human Annotator Agreement

Cohen for each pair of annotators is 0.81 (SD 0.038), 0.89 (SD 0.032), and 0.81 (SD 0.039). The mean of pair-wise Cohen is 0.84 (SD 0.030). The Fleiss score is 0.84 (SD 0.030), and the Krippendorff score is 0.85 (SD 0.030). These results collectively indicate a consistent agreement among the annotators with the guideline. We adjudicated all the annotation discrepancies based on the annotators' consensus to serve as ground truth.

Performance

The test set consisted of 35 reports, including 23 definitive negative, 6 probable, and 6 definitive positive. Table 2 summarizes the performance of all models on the withheld test set for PE certainty classification. Among the prompt-dependent methods, GPT-3.5 fine-tuning and in-context learning achieved F_1-scores of 0.86 and 0.87, respectively, with no statistically significant difference between their performances (*t*paired bootstrap significance test, P>.05 for accuracy, recall, and specificity). Its 0-shot performance was lower (F_1-score 0.57, 95% CI 0.40 - 0.75). Overall, GPT-3.5 obtained better results than ADAPET (F_1-score 0.42, 95% CI 0.32 - 50). Gemma3-4B in-context learning (F_1-score 0.63, 95% CI 0.56 - 0.79) yields better performance than both its 0-shot setting (F_1-score 0.45, 95% CI 0.29 - 0.56) and the Llama3.2-3B under both in-context (F_1-score 0.54, 95% CI 0.41 - 0.71) and 0-shot (F_1-score 0.43, 95% CI 0.28 - 0.62) conditions.

XSL•FO
RenderX

Table . Model performance for certainty category prediction. "gpt-3.5-turbo-0125" in the table corresponds to the GPT-3.5 model in the main text. "Biow2v" denotes BioWordVec, which represents biomedical word embeddings generated using fastText. *P* values[a] correspond to paired bootstrap significance test evaluating whether PECertainty's performance differs significantly from that of each comparator model.

	Accuracy	Precision	Recall	F_1-score	Specificity	AUC[b]
Prompt-free method						
SVM[c] (Biow2v)	0.77***[a] (0.63, 0.89)	0.69*** (0.50, 0.86)	0.68*** (0.5, 0.86)	0.67*** (0.49, 0.85)	0.86*** (0.77, 0.95)	0.81*** (0.67, 0.93)
RF[d] (Biow2v)	0.71*** (0.57, 0.83)	0.68*** (0.39, 0.91)	0.57*** (0.40, 0.75)	0.59*** (0.39, 0.78)	0.77*** (0.68, 0.87)	0.79*** (0.64, 0.92)
Fine-tuning RoBERTa-large[e]	0.66*** (0.51, 0.8)	0.53*** (0.35, 0.72)	0.54*** (0.35, 0.73)	0.52*** (0.35, 0.71)	0.80*** (0.71, 0.89)	0.71*** (0.54, 0.85)
Fine-tuning BioBERT[f]	0.66*** (0.57, 0.70)	0.22*** (0.20, 0.25)	0.33*** (0.27, 0.40)	0.26*** (0.23, 0.30)	0.83*** (0.78, 0.90)	0.400*** (0.26, 0.54)
Fine-tuning ClinicalBERT[g]	0.92* (0.83, 1.0)	0.91*** (0.83, 1.0)	0.88*** (0.75, 1.0)	0.87*** (0.71, 1.0)	0.95*** (0.90, 1.0)	0.99*** (0.97, 1.0)
PECertainty	0.94[h] (0.86, 1.0)	0.97[h] (0.94, 1.0)	0.89[h] (0.72, 1.0)	0.92[h] (0.79, 1.0)	0.98[h] (0.95, 1.0)	0.99[h] (0.97, 1.0)
Prompt-dependent method						
ADAPET[i] (RoBERTa-large)	0.71*** (0.66, 0.77)	0.4*** (0.31, 0.47)	0.44*** (0.33, 0.56)	0.42*** (0.32, 0.5)	0.87*** (0.82, 0.91)	0.76*** (0.67, 0.84)
Gemma3-4B Zero-shot	0.71*** (0.66, 0.77)	0.57*** (0.26, 0.60)	0.44*** (0.33, 0.56)	0.45*** (0.29, 0.56)	0.90*** (0.87, 0.92)	N/A[j]
Gemma3-4B In-Context learning	0.68*** (0.54, 0.83)	0.68*** (0.52, 0.82)	0.42*** (0.57, 0.86)	0.63*** (0.56, 0.79)	0.84*** (0.79, 0.89)	N/A
Llama3.2-3B Zero-shot	0.69*** (0.60, 0.77)	0.53*** (0.26, 0.75)	0.43*** (0.30, 0.60)	0.43*** (0.28, 0.62)	0.85*** (0.78, 0.91)	N/A
Llama3.2-3B In-Context learning	0.68*** (0.57, 0.80)	0.61*** (0.43, 0.82)	0.63*** (0.54, 0.76)	0.54*** (0.41, 0.71)	0.84*** (0.79, 0.90)	N/A
gpt-3.5-turbo-0125 Zero-shot	0.77*** (0.66, 0.86)	0.67*** (0.37, 0.85)	0.6* (0.43, 0.76)	0.57*** (0.4, 0.75)	0.9*** (0.82, 0.95)	N/A
gpt-3.5-turbo-0125 In-Context learning	0.91* (0.83, 1.0)	0.90* (0.83, 1.0)	0.87* (0.73, 1.0)	0.87*[h] (0.71, 1.0)	0.95* (0.89, 1.0)	N/A
gpt-3.5-turbo-0125 Fine-tuning	0.92*[h] (0.83, 1.0)	0.91*[h] (0.83, 1.0)	0.88*[h] (0.75, 1.0)	0.86*** (0.71, 1.0)	0.96*[h] (0.90, 1.0)	N/A

[a]*P* value codes: * *P*≥.05; *** *P*<.05.

[b]AUC: area under the receiver operating characteristic curve.

[c]SVM: support vector machine.

[d]RF: random forest classifier

[e]RoBERTa-large: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (Large Model).

[f]BioBERT: Bidirectional Encoder Representations From Transformers for Biomedical Text Mining.

[g]ClinicalBERT: Bidirectional Encoder Representations From Transformers for Clinical Text.

[h]PECertainty and GPT-3.5-turbo-0125 finetuning achieve similar performance and outperform other methods.

[i]ADAPET: A Densely-Supervised Approach to Pattern Exploiting Training.

[j]N/A: area under the receiver operating characteristic curve was not calculated for the large language model–based models because class labels were derived through prompt engineering, and token-level log-probabilities for predefined label strings were not extracted for area under the receiver operating

XSL•FO
RenderX

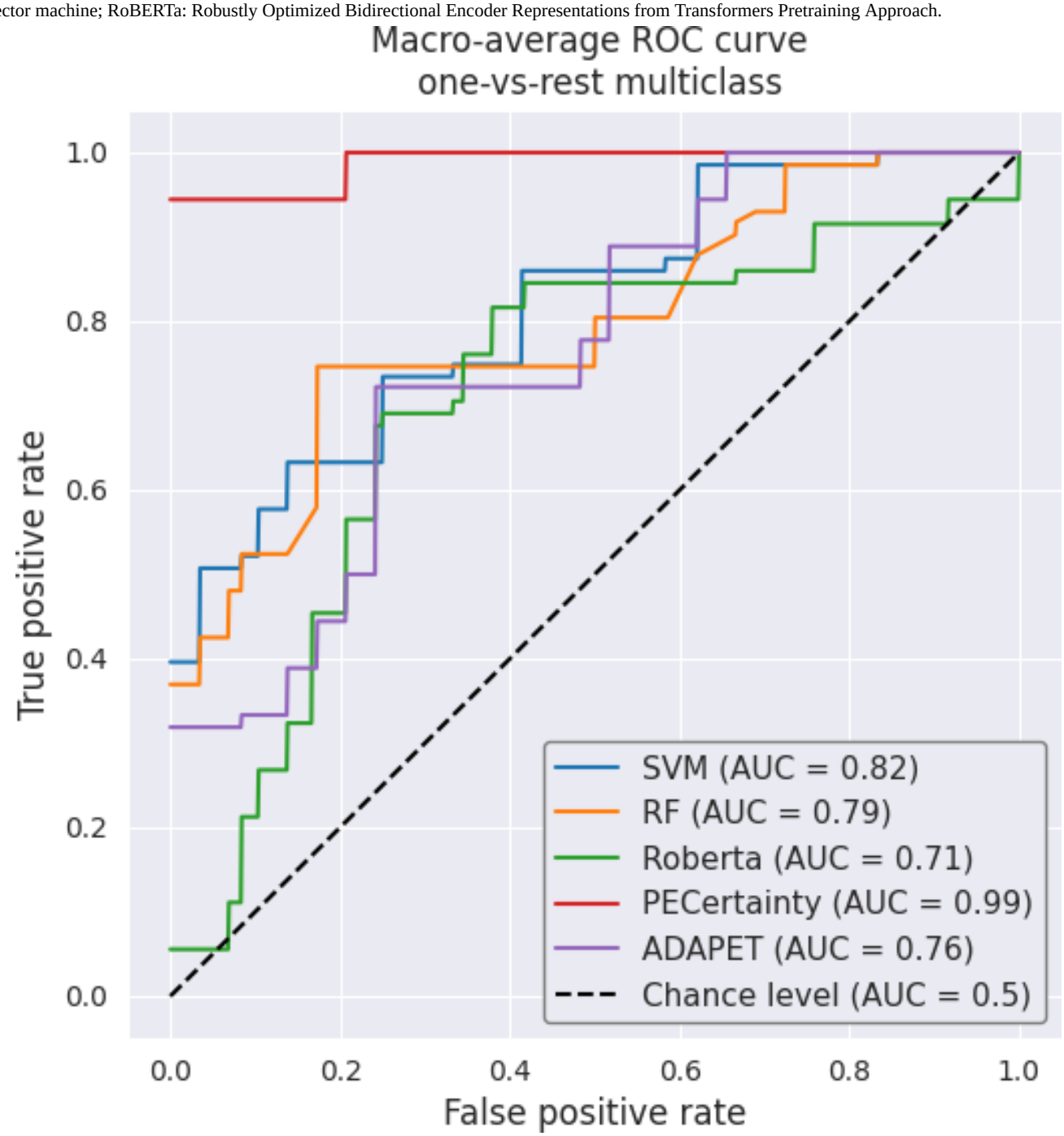
characteristic curve calculation in this study.

PECertainty demonstrates numerically higher performance than all prompt-free methods, including fine-tuned ClinicalBERT (paired bootstrap significance test, $P>.05$ for accuracy and $P<.05$ across all other metrics), and exceeds the prompt-dependent ADAPET by a large margin. Furthermore, it is comparable to GPT-3.5 in both in-context learning and fine-tuning settings (paired bootstrap significance test, $P<.05$ for all metrics). It achieved the best F_1-score of 0.92 (95% CI 0.79 - 1) and the best AUC of 0.99 (95% CI 0.97 - 1), competitively comparable with GPT3.5, yet it has a ≈500 times smaller (355M) parameter size than GPT3.5 (175B). Note that the significance values reported here could be sensitive to small changes in classification outcomes due to limited sample size, so the reported differences may not be definitive statistical evidence.

The misclassified examples and error analysis can be found in Multimedia Appendix 4. While all fine-tuned BERT variants benefit from pretraining, PECertainty consistently yields better performance than ClinicalBERT and BioBERT, suggesting that the additional embedding fine-tuning and task-specific alignment training objective of PECertainty contribute to improved performance. We also found that fine-tuning BioBERT and RoBERTa-large (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach [Large Model]) failed to outperform the conventional machine learning models of SVM and RF (F_1-score of 0.26 - 0.52 vs 0.59 - 0.67 and AUC of 0.40 - 0.71 vs 0.79 - 0.81) when only limited training data is available. In contrast, ClinicalBERT fine-tuning achieved a substantially higher F_1-score of 0.87, outperforming other pretrained models such as RoBERTa-large and BioBERT, thereby highlighting the importance of domain-specific data used during pretraining. RoBERTa-large is not pretrained with biomedical or clinical data. Both BioBERT and ClinicalBERT originate from the BERT architecture but were continually adapted on distinct domain corpora: BioBERT on biomedical literature from PubMed and PubMed Central (PMC), and ClinicalBERT on clinical narratives from the MIMIC-III (Medical Information Mart for Intensive Care III) database, including radiology and discharge notes. As our dataset exhibits linguistic patterns and contextual cues similar to clinical documentation, ClinicalBERT is better aligned with the task, which likely accounts for its superior performance relative to BioBERT. These performance differences can therefore be attributed primarily to variations in the domain-specific data used during pretraining. Figure 5 displays the receiver operating characteristic curves for all methods.

XSL•FO
RenderX

Figure 5. ROC curve of all methods (excluding LLMs). ADAPET: A Densely-Supervised Approach to Pattern Exploiting Training; AUC: area under the receiver operating characteristic curve; LLM: large language model; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine; RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.



Robustness

For the k-shot setting, Table 3 summarizes the performance of the best 3 methods with k examples per category, with k varying from 10 to 1. As the test set is class-imbalanced, the F_1-score is a more informative metric for interpreting the results. For in-context learning, the F_1-score is 0.57 in the 0-shot setting and peaks at 5 shots per category with an F_1-score of 0.85. It then decreases as more shots per category are added, ending at 0.65 for 10 shots per category. The fine-tuned gpt-3.5-turbo-0125 outperforms its in-context learning counterpart with 10 and 7 shots per category, which suggests that while in-context learning can yield rapid gains in performance, fine-tuning can lead to better overall results when there are more examples available. The performance of PECertainty and GPT-3.5 fine-tuning and in-context learning declined in the 10-shot setting, likely reflecting the diminishing utility of additional examples once the most informative demonstrations had been included. As more examples were added, the additional cases were less semantically similar and introduced variability or noise that may have distracted the model from the primary task. Future work should focus on strategies for identifying and selecting high-value demonstrations so that additional examples enhance, rather than degrade, model performance.

XSL•FO
RenderX

Table . Model performance of PE[a] diagnostic certainty classification on the test set under the k-shot setting. Unless otherwise noted, data are mean values, with 95% CIs in parentheses.

k	Accuracy	Precision	Recall	F_1-score	Specificity	AUC[b]
gpt-3.5-turbo-0125 In-context learning						
10	0.77 (0.63, 0.89)	0.72 (0.46, 0.87)	0.72 (0.59, 0.86)	0.65 (0.47, 0.83)	0.88 (0.82, 0.94)	N/A[c]
7	0.8 (0.66, 0.91)	0.82 (0.78, 0.89)	0.81 (0.67, 0.94)	0.76 (0.6, 0.89)	0.88 (0.83, 0.94)	N/A
5	0.91[d] (0.83, 1.0)	0.89 (0.83, 1.0)	0.87[d] (0.73, 1.0)	0.85[d] (0.69, 1.0)	0.95 (0.9, 1.0)	N/A
3	0.89 (0.8, 0.97)	0.86[d] (0.75, 0.95)	0.78 (0.61, 0.94)	0.78 (0.58, 0.94)	0.96[d] (0.92, 0.99)	N/A
1	0.8 (0.71, 0.89)	0.73 (0.45, 0.94)	0.7 (0.61, 0.82)	0.65 (0.5, 0.84)	0.89 (0.81, 0.96)	N/A
0	0.77 (0.66, 0.86)	0.67 (0.37, 0.85)	0.6 (0.43, 0.76)	0.57 (0.4, 0.75)	0.9 (0.82, 0.95)	N/A
gpt-3.5-turbo-0125 Fine-tuning						
10	0.77 (0.63, 0.89)	0.77 (0.63, 0.87)	0.8 (0.66, 0.93)	0.74 (0.58, 0.88)	0.87 (0.82, 0.93)	N/A
7	0.83[d] (0.69, 0.94)	0.84[d] (0.78, 0.92)	0.83[d] (0.69, 0.96)	0.79[d] (0.63, 0.91)	0.9[d] (0.84, 0.96)	N/A
5	0.71 (0.57, 0.86)	0.79 (0.76, 0.85)	0.77 (0.63, 0.9)	0.71 (0.56, 0.84)	0.85 (0.8, 0.9)	N/A
PECertainty						
10	0.89 (0.8, 0.97)	0.87 (0.75, 0.97)	0.78 (0.61, 0.94)	0.80 (0.59, 0.94)	0.92 (0.85, 0.99)	0.91 (0.82, 0.99)
7	0.89 (0.8, 0.97)	0.9 (0.77, 0.99)	0.78 (0.61, 0.94)	0.81 (0.6, 0.96)	0.91 (0.82, 0.98)	0.95[d] (0.88[d], 0.99[d])
5	0.91[d] (0.83, 1.0)	0.96[d] (0.93, 1.0)	0.83 (0.67, 1.0)	0.87[d] (0.73[d], 1.0[d])	0.92 (0.83, 1.0)	0.88 (0.74, 1.0)
3	0.51 (0.37, 0.66)	0.36 (0.27, 0.44)	0.38 (0.23, 0.54)	0.34 (0.24, 0.44)	0.77 (0.69, 0.85)	0.66 (0.51, 0.8)
1	0.52 (0.4, 0.63)	0.23 (0.2, 0.25)	0.26 (0.2, 0.32)	0.24 (0.21, 0.28)	0.68 (0.61, 0.73)	0.59 (0.46, 0.71)

[a]PE: pulmonary embolism.

[b]AUC: area under the receiver operating characteristic curve.

[c]N/A: area under the receiver operating characteristic curve was not calculated for the large language model–based models because class labels were derived through prompt engineering, and token-level log-probabilities for predefined label strings were not extracted for area under the receiver operating characteristic curve calculation in this study.

[d]Best performance for each metric.

On the Baystate Medical Center dataset, PECertainty achieved an accuracy of 0.90 (95% CI 0.87 - 0.92), precision of 0.82 (95% CI 0.76 - 0.89), recall of 0.73 (95% CI 0.67 - 0.79), and an F_1-score of 0.77 (95% CI 0.70 - 0.83). The model also demonstrated strong specificity (0.82, 95% CI 0.76 - 0.89) and AUC of 0.84 (95% CI 0.77 - 0.90). Notably, because the external Baystate dataset merges positive and negative findings into a single definitive category, the reported metrics evaluate the model's ability to distinguish certainty from uncertainty rather than its capacity for diagnostic polarity classification. As the external dataset is highly imbalanced with substantially more definitive cases than probable cases, we also reported class-specific performance to characterize the model

XSL•FO
RenderX

generalization. For the minority "probable" class (n=64), PECertainty achieved a precision of 0.73 (95% CI 0.61 - 0.85), recall of 0.50 (95% CI 0.38 - 0.63), and F_1-score of 0.60 (95% CI 0.48 - 0.80). These results demonstrate that the model identifies a meaningful proportion of uncertain cases and performs substantially better than a naive classifier that predicts only the majority definitive class. For the definitive class (n=356), PECertainty achieved a precision of 0.92 (95% CI 0.90 - 0.94), recall of 0.97 (95% CI 0.95 - 0.98), and F_1-score of 0.94 (95% CI 0.93 - 0.95). While performance naturally varies across classes due to sample size differences, these results indicate that the model generalizes well to the external dataset. Specifically, it maintains strong discriminative ability in distinguishing certain from uncertain cases, even in the presence of class imbalance.

Interpretability

The distributions of the ratings from 3 raters are shown in Figure S1 in Multimedia Appendix 3. The ratings for each impression section are shown in Figure S2 in Multimedia Appendix 3. We examined the differences in ratings provided by the 3 raters, using the average rating value as the overall rating for the interpretations generated by the LLM and the integrated gradient of PECertainty. As shown in Table 4, the Wilcoxon signed-rank test indicates that each of the 3 raters assigned significantly different ratings to LLM and PECertainty. The interpretations generated by LLMs are preferred. Some examples of the interpretations and their corresponding average ratings are shown in Table S4 in Multimedia Appendix 3.

Table . Results of the Wilcoxon signed-rank test. Wilcoxon signed-rank test statistic (W), and a smaller W indicates a larger difference between the 2 samples, suggesting that there is a significant difference between the paired observations.

	Wilcoxon signed-rank test statistic	*P* value
Rater 1	12.50	7.00×10^{-4}
Rater 2	3.50	5.17×10^{-5}
Rater 3	9.00	5.59×10^{-5}

Discussion

Principal Findings

Our study demonstrates that the open-source, prompt-free PECertainty model achieves strong performance compared to both prompt-free and prompt-dependent baselines in assessing PE diagnostic certainty. With approximately 500 times fewer parameters, it yields comparable performance to the closed-source LLM GPT-3.5 when limited annotated data is available. Notably, GPT-3.5 in-context learning performs similarly to fine-tuning while being training-free and more efficient. Our experiments show that PECertainty outperforms ClinicalBERT fine-tuning, but it is noted that the performance comparison between PECertainty (SetFit+ RoBERTa) and the ClinicalBERT baseline may be confounded by differences in training strategy, as ClinicalBERT was not fine-tuned using the SetFit framework. In the k-shot setting, PECertainty benefits from a few examples per category but shows limited improvement with larger datasets. It obtained better performance than GPT-3.5 (gpt-3.5-turbo-0125) in both fine-tuning and in-context learning with more than 5 shots per category, highlighting its strong few-shot learning capability. External validation further supports the model's generalizability for certainty classification, although this evaluation does not assess diagnostic polarity because definitively positive and definitively negative reports are aggregated in the external dataset.

For interpretability analysis, the explanations generated by the LLM were preferred over those produced by PECertainty (*t* test, P<.05). This finding highlights that the quality of model interpretability directly influences the usability of AI tools in clinical practice. Although PECertainty achieves strong classification performance and provides a faithful, model-grounded explanation through integrated gradients, its continuous attribution maps may be less intuitive for radiologists to interpret during routine workflow. While integrating LLM-generated explanations of the PECertainty predictions could improve clinician trust based on our findings, this approach would require deploying both PECertainty and a large LLM simultaneously, thereby increasing computational cost and reducing inference efficiency. Until interpretability is further improved, PECertainty is best deployed as a backend triage and worklist prioritization tool, rather than as a direct clinician-facing decision-support system. In this role, the model can assist by flagging high-risk cases for expedited review while minimizing the risk of clinician confusion or overreliance on the model outputs.

Despite overall strong performance, PECertainty revealed rare but clinically significant failure modes (shown in Multimedia Appendix 4), for example, polarity inversion error by misclassifying a clearly positive statement as "definitive negative" due to contextual ambiguity. This represents a catastrophic negation error with potentially serious patient safety implications for high-risk conditions such as PE, as it could lead to inappropriate clinical reassurance if no safeguards were added. It highlights limitations of end-to-end language models in reliably handling heterogeneous contexts. Therefore, the system should not be used as an autonomous diagnostic tool as the sole basis for ruling out disease. Different mitigation strategies can be explored for future safe deployment. First, hybrid strategies can be implemented to combine data-driven learning with rule-based guardrails, which can identify typical linguistic patterns to help reduce polarity inversion errors. Second, confidence-based thresholds can be used to automatically route ambiguous cases for mandatory radiologist review. Third, a targeted data augmentation in training can be applied to improve the model's sensitivity to this specific type of error. We emphasize that any clinical deployment of this

XSL•FO
RenderX

system would require robust human oversight, with radiologists retaining full responsibility for diagnostic decisions. This human-in-the-loop paradigm helps mitigate the impact of catastrophic errors while preserving the efficiency benefits of automated assistance.

Comparison to Prior Work

The advent of picture archiving and communication system [29] has reduced direct communication between radiologists and referring physicians, making clear diagnostic certainty in reports essential. However, inconsistent interpretation of hedging phrases often leads to ambiguity, delaying treatments or causing unnecessary interventions. Standardized lexicon in hedging [6] fails to capture contextual certainty semantics. Addressing these challenges calls for a standardized, context-aware uncertainty measurement. Yet existing rule-based approaches [17] rely on labor-intensive linguistic indicators, hindering scalability and omitting probabilistic assessments. This method also overlooks nuanced contextual interactions between disease status (eg, positive vs negative) and certainty. In contrast, machine learning techniques, particularly deep learning, can automate and enhance the quantification of diagnostic uncertainty. While traditional models (eg, SVMs and random forests) use word embeddings (eg, BioWordVec), they lack deeper contextual understanding. Fine-tuning PLMs has become a state-of-the-art practice for capturing context, but it usually requires substantial labeled data. When labeled data is scarce, few-shot learning frameworks built on PLMs offer a compelling alternative. The PECertainty was adapted from the SetFit approach and demonstrated its effectiveness and improved data efficiency in assessing diagnostic uncertainty of CTPA reports.

The PECertainty model offers 2 primary advantages. First, unlike prompt-dependent few-shot models, it eliminates the need for manually crafted verbalizers or prompts. Models such as ADAPET require cloze-style phrases and label-word mappings, while GPT-based methods rely on task descriptions, introducing variability from manual engineering. Instead, PECertainty generates rich embeddings from a small set of labeled examples. Second, PECertainty does not require large-scale pretrained models to achieve high performance. While LLMs such as GPT contain billions of parameters, PECertainty leverages smaller models, making training and inference significantly faster. However, interpretability analysis showed that GPT-3.5's explanations were rated higher than PECertainty's gradient-based interpretations.

PECertainty offers a scalable and efficient approach to standardizing diagnostic certainty in radiology reports. By providing referring physicians with clear, structured certainty labels (ie, definitive positive, definitive negative, and probable), the model can reduce ambiguity and support more consistent clinical decision-making. This standardization may help address known issues such as judgment variability among radiologists [30] and improve communication between radiologists and treating physicians. Additionally, once PECertainty is further improved on its interpretability, it could be integrated into clinical workflows as a decision-support tool within picture archiving and communication systems or structured reporting platforms. In this setting, the model could automatically analyze draft or finalized CTPA reports in real time to flag low-certainty impressions for review, promote clarification before report finalization, and alert clinicians to cases that warrant additional follow-up. Such integration could reduce unnecessary hedging, streamline communication, and help avoid treatment delays. In turn, this may help avoid delays in treatment or unnecessary anticoagulation. Future studies can build on this work to evaluate the association between report uncertainty and various clinical outcomes [31] as well as examine variation in report quality by health organizations, radiologists, and patient subgroups.

Limitations

There are limitations to this study. First, although there are a lot of pretrained models in the RoBERTa and BERT families, we did not exhaustively compare all variants. We chose RoBERTa-large because it achieved excellent performance in many NLP benchmarking tasks [32] and has been further pretrained as a sentence transformer, making it more comparable to the model architecture used in PECertainty. In contrast, not all RoBERTa variants have publicly available pretrained sentence transformer versions suitable for direct comparison. BioBERT and ClinicalBERT were included for their domain-specific training on biomedical and clinical texts, respectively. Second**,** the dataset used in this study was modest in size (173 CTPA reports), which may influence model stability and generalizability. Nevertheless, PECertainty was specifically designed to operate under low-resource conditions, and its performance was further evaluated on an external Baystate Medical Center dataset to provide preliminary evidence of generalizability beyond the development cohort. Although the results are encouraging, they are interpreted as preliminary and warrant validation in larger, multi-institutional studies. Third, due to the scarcity of data belonging to the probable negative and probable positive classes, we merged the data from these 2 classes into 1 class labeled as "probable." This consolidation sacrifices the modeling granularity with reduced clinical utility. In other words, the model cannot distinguish cases that convey high diagnostic concern (probable positive) from those reflecting lower suspicion (probable negative), which might lead to ambiguity in downstream decision-making. We plan to explore self-learning and active learning strategies to enrich the dataset with more "probable" examples, thereby enhancing the model's ability to capture subtle clinical nuances in PE diagnostic communication and ultimately advance PE care management.

Fourth**,** the external validation dataset from Baystate Medical Center was annotated using a slightly different guideline, defining only definitive and probable categories, whereas our primary model was trained on 3 classes (definitive positive, definitive negative, and probable). The annotation bias and inconsistencies would be minimal because both teams applied conceptually aligned definitions and guidelines. However, to align these schemes for external validation, we had to combine our 2 "definitive" predictions into 1 category. This limited the scope of our external validation, that is, only validating the model's ability to distinguish "certain" from "uncertain" reports on external data. More validation is warranted when external data with more refined annotations is available to test the model's capacity in separating "definitive positive" reports from "definitive negative" ones. Fifth, the catastrophic negation errors

XSL•FO
RenderX

(Multimedia Appendix 4) noted in the previous discussion indicate potential safety concerns. Therefore, the current iteration of PECertainty must be patched with optimal mitigation strategies before clinical deployment (eg, rule-based guardrails or low-confidence triggered mandatory expert review needs to be implemented and extensively evaluated). Sixth, research findings are time-limited due to the rapid evolution of LLMs and the continuous release of newer models, although newer models do not always mean improvement for a certain task [33]. Future investigators should replicate and extend our study using newer versions of LLMs. To address this limitation, future studies should replicate and extend our findings using newer versions of the LLMs evaluated here. Moreover, PECertainty is designed to be adaptable and modular, allowing integration of newer base models as they become available.

Future Direction

For future work, we would focus on collecting additional cases in the probable category through active learning and developing models that can distinguish between the finer-grained classes (eg, probable negative vs probable positive). Moreover, expanding the evaluation to include a broader range of LLMs and exploring multitask learning (with additional subtasks such as PE acuity and location) may further enhance the model's overall performance. Finally, we will explore effective and innovative ways to improve PECertainty's interpretability that include (1) developing extractive rationalization mechanisms that identify and surface the specific textual evidence spans supporting each certainty classification; (2) exploring hybrid explainer approaches, such as teacher-student frameworks in which a constrained language model generates structured natural-language justifications aligned with PECertainty's outputs; and (3) adopting a human-centered co-design strategy as a core component of future development. We plan to work closely with pulmonary doctors in the ED setting and CTPA radiologists to iteratively evaluate which forms of model evidence are most interpretable and actionable in high-pressure clinical environments.

Conclusion

We developed PECertainty and evaluated baseline methods for assessing diagnostic certainty in CTPA reports under limited annotated data conditions. Our results demonstrate that PECertainty effectively captures the contextual semantics of free-text reporting to assess diagnostic certainty, even with scarce data. This holds the potential to facilitate precise communication of imaging findings between radiologists and referring physicians.

Acknowledgments

This work was supported by the National Institutes of Health (NIH; R21LM014032). The authors attest that generative artificial intelligence technology was not used in the generation of text, figures, or other informational content of this paper.

Funding

This work was supported by NIH grant R21LM014032.

Data Availability

The data underlying this paper cannot be shared publicly due to ethical and privacy concerns. The data can be shared on a reasonable request to the corresponding author under a specific data user agreement. The source code can be accessed online [34].

Authors' Contributions

Conceptualization: FL, RH, DK, HL, YK, JA, BSG, ER, LMW, CK, AB, MPR
Formal analysis: FL, RH
Writing – original draft: FL, RH, DK, HL, YK, JA, BSG, ER, LMW, CK, AB, MPR
Writing – review & editing: FL, RH, DK, HL, YK, JA, BSG, ER, LMW, CK, AB, MPR
FL and MPR were guarantors of integrity of entire study.
FL, RH, and LMW handled literature research.
FL, RH, and DK worked on experimental studies.
HL, YK, and JA did the data annotation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data and annotating.
[DOCX File, 93 KB - medinform_v14i1e79972_app1.docx]

Multimedia Appendix 2
Implementation details.

XSL•FO
RenderX

[DOCX File, 15 KB - medinform_v14i1e79972_app2.docx]

Multimedia Appendix 3
Interpretability analysis.
[DOCX File, 2473 KB - medinform_v14i1e79972_app3.docx]

Multimedia Appendix 4
Error analysis.
[DOCX File, 16 KB - medinform_v14i1e79972_app4.docx]

References

1. Gunn AJ, Tuttle MC, Flores EJ, et al. Differing interpretations of report terminology between primary care physicians and radiologists. J Am Coll Radiol 2016 Dec;13(12 Pt A):1525-1529. [doi: 10.1016/j.jacr.2016.07.016] [Medline: 27595196]
2. Rosenkrantz AB. Differences in perceptions among radiologists, referring physicians, and patients regarding language for incidental findings reporting. AJR Am J Roentgenol 2017 Jan;208(1):140-143. [doi: 10.2214/AJR.16.16633] [Medline: 27657356]
3. Gunn AJ, Sahani DV, Bennett SE, Choy G. Recent measures to improve radiology reporting: perspectives from primary care physicians. J Am Coll Radiol 2013 Feb;10(2):122-127. [doi: 10.1016/j.jacr.2012.08.013] [Medline: 23228373]
4. Goldberg-Stein S, Chernyak V. Adding value in radiology reporting. J Am Coll Radiol 2019 Sep;16(9 Pt B):1292-1298. [doi: 10.1016/j.jacr.2019.05.042] [Medline: 31492407]
5. Lafortune M, Breton G, Baudouin JL. The radiological report: what is useful for the referring physician? Can Assoc Radiol J 1988 Jun;39(2):140-143. [doi: 10.1177/084653718803900214] [Medline: 2967838]
6. Panicek DM, Hricak H. How sure are you, doctor? A standardized lexicon to describe the radiologist's level of certainty. AJR Am J Roentgenol 2016 Jul;207(1):2-3. [doi: 10.2214/AJR.15.15895] [Medline: 27065212]
7. Weiss DL, Langlotz CP. Structured reporting: patient care enhancement or productivity nightmare? Radiology 2008 Dec;249(3):739-747. [doi: 10.1148/radiol.2493080988] [Medline: 19011178]
8. Shea LAG, Towbin AJ. The state of structured reporting: the nuance of standardized language. Pediatr Radiol 2019 Apr;49(4):500-508. [doi: 10.1007/s00247-019-04345-0] [Medline: 30923882]
9. Plumb AAO, Grieve FM, Khan SH. Survey of hospital clinicians' preferences regarding the format of radiology reports. Clin Radiol 2009 Apr;64(4):386-394. [doi: 10.1016/j.crad.2008.11.009] [Medline: 19264183]
10. Bosmans JML, Weyler JJ, De Schepper AM, Parizel PM. The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys. Radiology 2011 Apr;259(1):184-195. [doi: 10.1148/radiol.10101045] [Medline: 21224423]
11. Danton GH. Radiology reporting: changes worth making are never easy. Appl Radiol 2010;39(5):19-23. [doi: 10.37549/AR1751]
12. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
13. Fink MA, Kades K, Bischoff A, et al. Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports. Radiol Artif Intell 2022 Sep;4(5):e220055. [doi: 10.1148/ryai.220055] [Medline: 36204531]
14. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
15. Yan A, McAuley J, Lu X, et al. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell 2022 Jul;4(4):e210258. [doi: 10.1148/ryai.210258] [Medline: 35923376]
16. Liu F, Zhou P, Baccei SJ, et al. Qualifying certainty in radiology reports through deep learning–based natural language processing. AJNR Am J Neuroradiol 2021 Oct;42(10):1755-1761. [doi: 10.3174/ajnr.A7241]
17. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 2011 Oct;44(5):728-737. [doi: 10.1016/j.jbi.2011.03.011] [Medline: 21459155]
18. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
19. Tunstall L, Reimers N, Jo UES, et al. Efficient few-shot learning without prompts. : 36th Conference on Neural Information Processing Systems (NeurIPS 2022); 2022 URL: https://neurips2022-enlsp.github.io/papers/paper_17.pdf [accessed 2026-03-30]
20. 45 CFR part 46. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2026-04-17]
21. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. : Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

XSL•FO
RenderX

Language Pro-cessing (EMNLP-IJCNLP); 2019 URL: https://www.aclweb.org/anthology/D19-1 [accessed 2026-03-30] [doi: 10.18653/v1/D19-1410]
22. Schölkopf B. SVMs - a practical consequence of learning theory. IEEE Intell Syst 1998;13(4):18-21. [doi: 10.1109/5254.708428]
23. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
24. Tam D, Menon RR, Bansal M, Srivastava S, Raffel C. Improving and simplifying pattern exploiting training. Assoc Comput Linguist 2021:4980-4991. [doi: 10.18653/v1/2021.emnlp-main.407]
25. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [doi: 10.11613/bm.2012.031] [Medline: 23092060]
26. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
27. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007 Apr;1(1):77-89. [doi: 10.1080/19312450709336664]
28. Wilcoxon F. Individual comparisons by ranking methods. In: Breakthroughs in Statistics: Springer; 1992:196-202. [doi: 10.1007/978-1-4612-4380-9_16]
29. Reiner B, Siegel E, Protopapas Z, Hooper F, Ghebrekidan H, Scanlon M. Impact of filmless radiology on frequency of clinician consultations with radiologists. Am J Roentgenol 1999 Nov;173(5):1169-1172. [doi: 10.2214/ajr.173.5.10541082]
30. Jaramillo D. Radiologists and their noise: variability in human judgment, fallibility, and strategies to improve accuracy. Radiology 2022 Mar;302(3):511-512. [doi: 10.1148/radiol.211891]
31. Bedayat A, Sewatkar R, Cai T, et al. Association between confidence level of acute pulmonary embolism diagnosis on CTPA images and clinical outcomes. Acad Radiol 2015 Dec;22(12):1555-1561. [doi: 10.1016/j.acra.2015.08.018] [Medline: 26391859]
32. Lan Z, Chen M, Goodman S, et al. ALBERT: a LITE BERT for self-supervised learning of language representations. arXiv. Preprint posted online on Feb 9, 2020. [doi: 10.48550/arXiv.1909.11942]
33. Chen L, Zaharia M, Zou J. How Is ChatGPT's behavior changing over time? Harvard Data Sci Rev 2024;6(2). [doi: 10.1162/99608f92.5317da47]
34. PECertainty. GitHub. URL: https://github.com/ruofanhu/PECertainty [accessed 2026-03-30]

Abbreviations

ADAPET: A Densely-Supervised Approach to Pattern Exploiting Training
AUC: area under the curve
BioBERT: Bidirectional Encoder Representations From Transformers for Biomedical Text Mining
ClinicalBERT: Bidirectional Encoder Representations From Transformers Pretrained on Clinical Text
CTPA: computed tomography pulmonary angiography
ED: emergency department
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care III
NLP: natural language processing
PE: pulmonary embolism
PLM: pretrained language model
RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach
RoBERTa-large: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach (Large Model)
SVM: support vector machine
UMMH: UMass Memorial Health

Edited by A Coristine; submitted 07.Jul.2025; peer-reviewed by M Chukwuka, X Liang, Z Zhong; revised version received 11.Feb.2026; accepted 11.Feb.2026; published 28.Apr.2026.

Please cite as:
Liu F, Hu R, Kim D, Lo H, Kharonov Y, Johri A, Gerber BS, Rundensteiner E, Westafer LM, Allison J, Kiefe C, Bankier A, Rosen MP Data-Efficient Language Model for Assessing Pulmonary Embolism Diagnostic Certainty From Radiology Reports: Model Development and Validation Study
JMIR Med Inform 2026;14:e79972
URL: https://medinform.jmir.org/2026/1/e79972
doi:10.2196/79972

XSL•FO
RenderX

© Feifan Liu, Ruofan Hu, Donghyuk Kim, Hao Lo, Yevgeniy Kharonov, Ansh Johri, Ben S Gerber, Elke Rundensteiner, Lauren M Westafer, Jeroan Allison, Catarina Kiefe, Alexander Bankier, Max P Rosen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Methodological Development and Assessing Prescribing Determinants Through Cumulative Drug Exposure in Hospitalized Patients: Proof-of-Concept Retrospective Study

Mathilde Bories[1], PharmD; Guillaume Bouzillé[1], PhD, HDR; Marc Cuggia[1], PhD, HDR; Pascal Le Corre[2,3,4], PhD, HDR; Aurélie Bannay[5], PhD

[1]Inserm, Laboratoire Traitement du Signal et de l'Image—UMR 1099, Centre Hospitalier Universitaire de Rennes, Université de Rennes, LTSI - Laboratoire Traitement du Signal et de l'Image, Université de Rennes, Campus Santé - Bâtiment 15b, 2 avenue du Professeur Léon Bernard, Rennes, France

[2]Pôle Pharmacie, Secteur Pharmacotechnie et Onco-Pharmacie, Centre Hospitalier Universitaire de Rennes, Rennes, France

[3]Laboratoire de Biopharmacie et Pharmacie Clinique, Faculté de Pharmacie, Université de Rennes, Rennes, France

[4]Centre Hospitalier Universitaire de Rennes, Inserm, Ecole des hautes études en santé publique, Institut de recherche en santé, environnement et travail, UMR_S 1085, Université de Rennes, Rennes, France

[5]Université de Lorraine, Centre Hospitalier Régional Universitaire de Nancy, Centre national de la recherche scientifique, Inria, Laboratoire lorrain de recherche en informatique et ses applications, Nancy, France

Corresponding Author:
Mathilde Bories, PharmD
Inserm, Laboratoire Traitement du Signal et de l'Image—UMR 1099, Centre Hospitalier Universitaire de Rennes, Université de Rennes, LTSI - Laboratoire Traitement du Signal et de l'Image, Université de Rennes, Campus Santé - Bâtiment 15b, 2 avenue du Professeur Léon Bernard, Rennes, France

Abstract

Background: Preventing adverse drug reactions requires accurate monitoring of drug exposure throughout patient care. Conventional metrics, measured at admission or discharge, fail to capture the dynamic and cumulative nature of drug burden during hospitalization. Improving exposure assessment is essential to support clinical decision-making and medication safety. Clinical data warehouses (CDWs), which integrate detailed drug administration records, enable the retrospective reuse of hospital data to develop more granular and dynamic measures of in-hospital drug exposure.

Objective: This exploratory proof-of-concept study aimed to introduce 2 cumulative drug exposure metrics computed from CDW: cumulative drug exposure (CDE) and cumulative drug exposure density (CDED). The study also aimed to compare these metrics with conventional metrics, primarily as a methodological development for characterizing prescribing determinants in hospitalized patients.

Methods: We conducted a retrospective study using the eHOP CDW at Rennes University Hospital. Adults hospitalized for hematological malignancies were included. Four prescribing determinants were analyzed: polypharmacy (PP), hyperpolypharmacy (HPP), drug-drug interactions (DDIs), and potentially inappropriate medications (PIMs). CDE quantified the number of days each determinant was present, while CDED normalized this value to hospital length of stay. Analyses combined descriptive statistics, Spearman correlations, and factorial analysis of mixed data (FAMD).

Results: Mean CDE values were 10.5 days for PP, 5.7 for HPP, 64.7 for DDIs, and 19.0 for PIMs (≥65 years). CDED values ranged from 0.3 to 3.2. Conventional metrics at admission were weakly correlated with cumulative exposure measures (eg, DDIs: r_s=0.04, P=.752; PP: r_s=–0.04, P=.757; HPP: r_s=0.11, P=.364). Stronger, significant correlations emerged at discharge (CDE DDIs: r_s=0.44, CDED: r_s=0.46; both P<.001). PIMs showed strong significant correlations at both time points. FAMD highlighted that cumulative metrics contributed independently to the principal components, capturing dynamics of drug exposure not reflected by conventional indicators.

Conclusions: CDE and CDED, derived from real-time CDW data, offer reproducible and scalable alternatives to conventional metrics for characterizing drug exposure in patients hospitalized for severe conditions. They provide a more accurate characterization of drug burden and hold promise for pharmacoepidemiological research and clinical decision support.

(JMIR Med Inform 2026;14:e76961) doi:10.2196/76961

XSL•FO
RenderX

KEYWORDS

cumulative drug exposure; clinical data warehouse; polypharmacy; hyperpolypharmacy; drug interactions; potentially inappropriate medications

Introduction

Background

Preventing adverse drug reactions (ADRs) is a major challenge for health care systems worldwide, both to improve patient outcomes and to reduce economic burdens [1,2]. Monitoring drug exposure at both the patient and institutional levels is a key strategy to address this issue. Furthermore, information on how drug exposure influences patient response is fundamental to evaluating a medication's safety and effectiveness. Therefore, various tools such as Computerized Provider Order Entry, clinical decision support systems (CDSS), and deprescribing programs have been implemented to limit unnecessary or inappropriate prescriptions [3,4]. Nonetheless, significant challenges remain: generating only relevant alerts to avoid clinician alert fatigue and providing meaningful indicators to assess the effectiveness of deprescribing programs. All of these initiatives depend on accurate assessment of patients' drug exposure across all stages of care [5].

Learning health system (LHS) aims to address this challenge by capturing data from routine practice, transforming them into knowledge, and reintegrating that knowledge to improve care in a continuous cycle. A typical LHS includes three iterative steps:

1. *Practice-to-data*: Data are collected from routine clinical practice, primarily through electronic health records (EHRs) and integrated into appropriate data reuse architecture such as clinical data warehouse (CDW).
2. *Data-to-knowledge*: The available data are analyzed to generate new knowledge, using methods such as statistical modeling, data mining, or machine learning, to identify patterns and predictors.
3. *Knowledge-to-practice*: The newly generated knowledge is translated back into clinical practice via decision support tools, updated protocols, or quality improvement initiatives, creating a continuous feedback loop that promotes better health care delivery and patient outcomes.

Hence, LHSs could contribute to improving current practices in drug exposure measurement by enabling the development of new methods for knowledge extraction from real-world data.

Current Knowledge and Practices on Drug Exposure Measurement

Accurately quantifying drug exposure is fundamental for assessing medication safety and effectiveness. A robust theoretical framework integrates pharmacokinetics, pharmacoepidemiology, and clinical pharmacology, with 4 dimensions: exposure definition, data sources, metrics, and biases [6]. Pharmacokinetic exposure can be described by parameters such as the area under the curve, maximum and minimum concentrations (Cmax, Cmin), or time in therapeutic range; despite being closely related to drug effects, these concentration-based measures are rarely used, even in clinical trials where the administered dose remains the standard [7]. Pharmacoepidemiology defines exposure by drug receipt and refines it with duration, dose, and adherence. Aggregate indicators such as medication possession ratio, proportion of days covered, and the World Health Organization–defined daily dose are useful in chronic care but fail to capture rapid regimen changes or cumulative burden in hospital settings.

Because inpatient treatments vary from day to day, choosing an appropriate observation window is crucial. Assessments limited to admission or discharge underestimate drug burden; a comprehensive evaluation should encompass the entire hospital stay. In addition to quantifying overall exposure, specific prescribing determinants provide complementary insight into the quantity and quality of therapy. Polypharmacy (PP) and hyperpolypharmacy (HPP) reflect the number of concomitant drugs, while potentially inappropriate medications (PIMs) and drug-drug interactions (DDIs) capture the appropriateness of the regimen. These indicators are particularly relevant for older or multimorbid patients.

Whether evaluating overall exposure or these determinants, the same methodological challenges arise. Exposure estimated from prescriptions may overstate administration, while unrecorded medications obtained elsewhere cause underestimation; differences in length of stay introduce further bias, as longer stays inevitably yield more exposure and more determinants. These issues, combined with the dynamic nature of inpatient therapy, expose the limitations of simple point-in-time indicators that record the presence of drugs or determinants at admission, during hospitalization, or at discharge.

To more faithfully characterize drug exposure, especially in contexts of transient high drug burden such as intensive care units or transplantation services, composite measures that integrate both intensity and accumulation are needed. Dose intensity (total dose or number of drugs per unit time), dose density (dosing frequency), and cumulative exposure (sum of doses over the stay) capture both the quantity and the temporal dynamics of medication use. By providing a fuller picture than conventional point-in-time metrics, these cumulative indicators can support more nuanced analyses of prescribing practices and promote safer, more rational medication use.

Date Reuse and the LHS to Improve Measure of Drug Exposure

The reuse of health data has become an essential component of medical research, offering insights into patient care, treatment effectiveness, and disease patterns. Through the integration of EHRs, administrative databases, and insurance claims, researchers can analyze large-scale populations and track health care trends over time [8].

Among the most powerful tools facilitating the practice-to-data step of LHS are CDWs, which aggregate structured and unstructured health information from multiple sources within hospital systems [9]. By consolidating laboratory results, drug

XSL•FO
RenderX

administration records, and patient demographics, CDWs provide invaluable real-world evidence for pharmacovigilance, pharmacoepidemiology, and drug safety research [10]. Moreover, unlike traditional clinical trial data, which often lack generalizability to broader populations, they provide real-world insights into prescribing practices, ADRs, and treatment effectiveness in heterogeneous patient groups [11].

Hospital-based CDWs offer a unique advantage by capturing detailed, real-time data on medication administration, making them particularly valuable for drug exposure. Unlike outpatient prescription databases, which, in most cases, indicate only intended treatments rather than actual drug intake, hospital CDWs provide precise records of administered medications, dosage adjustments, and the timing of drug exposure. Moreover, they facilitate precise evaluation of actual drug administration, offering dynamic analyses over the entire course of hospitalization [12,13].

However, despite their potential, the reuse of health data and CDW-based analyses faces several challenges. Data quality issues, such as missing or inconsistent medication records, variations in documentation practices across hospitals, and incomplete capture of drug administration details, can impact the reliability of findings [8]. Additionally, differences in exposure definitions such as whether drug exposure is measured based on prescriptions, administrations, or estimated treatment durations can lead to variability in results. These discrepancies highlight the need for harmonized methodologies to ensure consistency in developing new drug exposure, particularly in hospitalized patients where medication regimens are frequently modified. Standardized approaches to data integration, exposure quantification, and analytical techniques are crucial to improving the robustness and comparability of real-world evidence on medication use.

Our Contribution

In this context, we introduce two novel cumulative metrics designed to provide a more comprehensive evaluation of drug exposure and prescribing determinants during hospitalization:

1. *Cumulative drug exposure* (*CDE*): It quantifies the total number of days a patient is exposed to PP, HPP, DDIs, or PIMs during hospitalization.
2. *Cumulative drug exposure density* (*CDED*): It normalizes CDE by the length of hospital stay to account for exposure intensity.

These metrics aim to offer a dynamic, patient-centered evaluation of drug exposure, addressing limitations of conventional static indicators and could serve as quality indicators in initiatives promoting safer and more rational use of medications.

The objective of this exploratory study was to develop and validate, as a methodological proof of concept, 2 CDE metrics (CDE and CDED) derived from real-time hospital CDW data. This work should be considered a preliminary methodological step rather than a definitive evaluation. The proposed new indicators were applied to quantify 4 key prescribing determinants (PP, HPP, DDIs, and PIMs) and were compared with traditional point-in-time measures to assess their added value within an LHS for characterizing drug exposure in hospitalized patients undergoing complex and dynamic treatment regimens.

Methods

Data Sources

The data for this study were obtained from three data sources:

1. CDW (eHOP): eHOP, developed at Rennes University Hospital, integrates administrative and clinical data from EHR [14]. It provides structured data, such as drug administration records, and unstructured data, such as clinical notes. For this study, structured data were extracted, including demographic details, drug names, dosages, routes of administration, and administration dates. These data facilitated detailed longitudinal tracking of drug use during hospitalization, which is critical for the computation of cumulative exposure metrics.
2. Theriaque: Theriaque [15] is a drug knowledge database that categorizes medications based on the Anatomical Therapeutic Chemical (ATC) classification system. It offers exhaustive information on drug indications, contraindications, and DDIs. The database categorizes DDIs into 4 severity levels: contraindicated, should be avoided, precaution of use, and to take into account. For this study, these levels were mapped into major (contraindicated, should be avoided) and moderate (precaution of use, to take into account) categories to support analysis [16].
3. GO-PIM list: To identify PIMs, the GO-PIM list [17] a validated reference for geriatric oncology patients, was used.

Prescribing Determinants

PP and HPP were defined based on the daily number of distinct drugs administered during hospitalization. Therefore, the daily number of distinct ATC classes administered during each hospitalization day was computed using structured medication administration data from eHOP. A patient was classified as having PP if the number of ATC classes per day was ≥5, and HPP if the number of ATC classes was ≥10 [18]. Chemotherapy regimens used for transplant conditioning (ATC class L01) administered before the transplantation day were excluded from this calculation to avoid an artificial overestimation of polymedication due to high-dose chemotherapy protocols. This classification was applied daily throughout hospitalization, allowing for the assessment of PP and HPP at admission, discharge, and during the hospital stay.

A DDI refers to a clinically significant unintended modification in drug exposure or response caused by the coadministration of another drug [19]. The detection of DDIs was performed using the Theriaque database, which provides predefined lists of interacting drug pairs (term 1 and term 2) based on ATC codes. A DDI was recorded when both interacting drugs were administered on the same day. DDIs were further categorized into major and moderate severity levels. The daily number of DDIs per patient was computed, and aggregated values were calculated at admission, discharge, and across hospitalization to assess exposure variations over time.

XSL•FO
RenderX

PIMs are drugs for which the risks outweigh the expected benefits in older adult patients, due to an increased risk of ADRs, toxicity, or limited clinical effectiveness [20]. For each patient aged 65 years or older, all administered medications were cross-referenced daily with the GO-PIM ATC-based classification, and any drug listed in GO-PIM was classified as a PIM. The daily number of PIMs per patient was determined, and aggregate values were calculated at admission, discharge, and throughout hospitalization to evaluate temporal patterns of PIMs exposure.

Conventional Metrics of Drug Exposure

To compare cumulative exposure metrics with traditional methods, conventional metrics were assessed at 2 hospitalization time points: admission and discharge. For each prescribing determinant (PP, HPP, DDIs, and PIMs), presence at admission or discharge was evaluated as a binary variable indicating whether the patient met the definition at these time points.

The number of prescribing determinants at each time point was also quantified as follows:

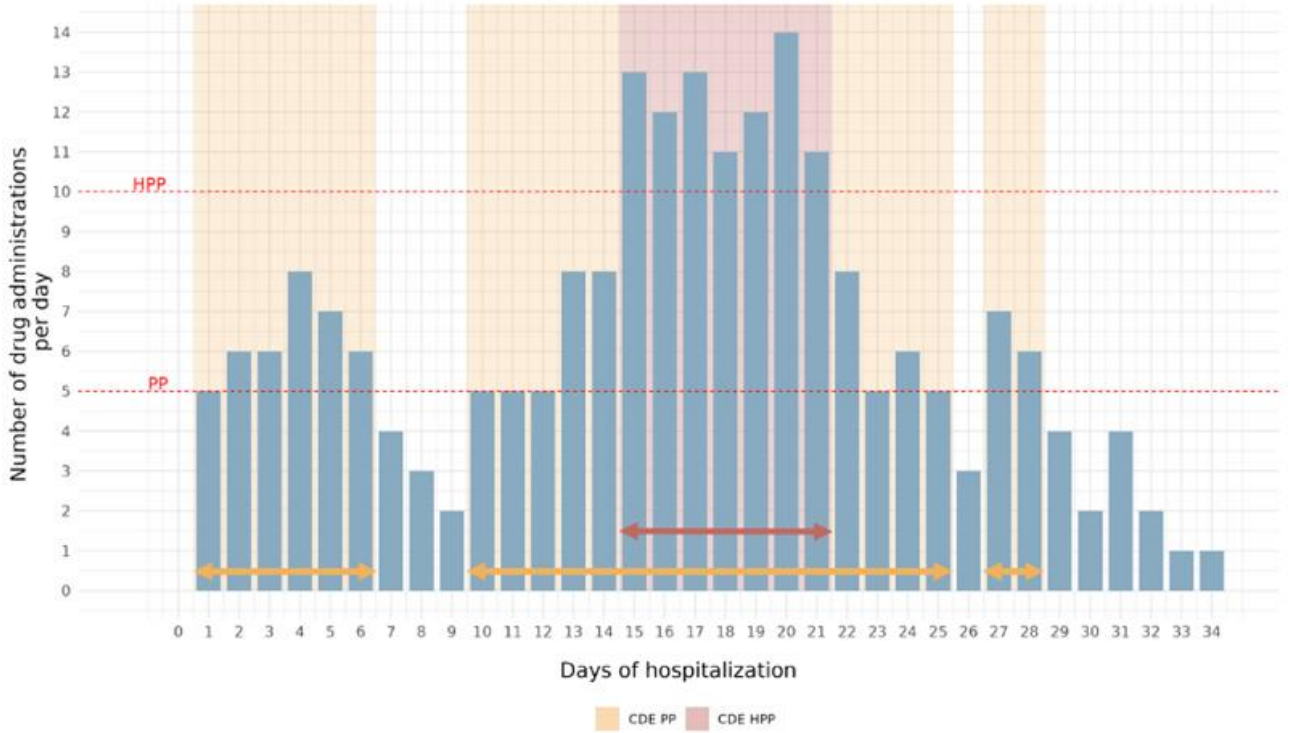
1. *For PP and HPP*: The total number of distinct drugs administered at admission and discharge.
2. *For DDIs*: The total number of DDIs identified at each time point.
3. *For PIMs*: The total number of PIMs administered at admission and discharge.

New Cumulative Exposure Metrics

To quantify the overall exposure to the prescribing determinants throughout hospitalization, 2 cumulative exposure metrics were computed: CDE and cumulative exposure density. CDE was defined as the total number of days a patient was exposed to each prescribing determinant during hospitalization.

For PP and HPP, exposure was recorded as a binary status per day (exposed and nonexposed), and the sum of exposure days was computed across the full hospital stay. However, for DDIs and PIMs, CDE accounted for the number of occurrences per day: if multiple DDIs or PIMs were recorded on a given day, each event contributed separately to the total exposure count. For instance, if a patient experienced 4 distinct DDIs on a single day, this was counted as 4 exposure days in the cumulative measure. To account for variations in hospitalization duration, CDED was calculated by normalizing CDE to the total length of stay, providing an intensity-adjusted measure (Figure 1).

Figure 1. Temporal evolution of daily drug exposure and CDE in a hospitalized patient. This figure presents the daily drug exposure profile of a single patient from the study cohort over a 33-day hospitalization. The *x*-axis represents hospitalization days (days 1-34), while the *y*-axis indicates the number of distinct drugs administered per day. Thresholds for polypharmacy (PP: ≥5 drugs per day) and hyperpolypharmacy (HPP: ≥10 drugs per day) are delineated by shaded bands. CDE to PP and HPP—quantified as the total number of days the patient was exposed to these prescribing determinants—is represented by colored segments. The patient was exposed to PP from admission (day 1, 5 drugs) but was no longer exposed at discharge (day 34, 1 drug), reflecting a progressive reduction in medication burden over the hospital stay. In total, the patient had 24 days of CDE to PP (days 1-6, 10-25, and 27-28). CDE to HPP was observed for 7 days (days 15-21). This representation highlights the dynamic nature of drug exposure throughout hospitalization and underscores the limitations of static assessments at admission or discharge in capturing overall medication burden. CDE: cumulative drug exposure; HPP: hyperpolypharmacy; PP: polypharmacy.



Evaluation: Use Case and Population

Hematopoietic stem cell transplantation (HSCT) patients represent a particularly high-risk group regarding ADRs, as they are exposed to an exceptionally high medication burden throughout hospitalization. Several studies have explored associations between prescribing determinants such as PP, HPP, PIMs, and DDIs, and adverse outcomes such as mortality or

XSL•FO
RenderX

rehospitalization, using conventional metrics [21,22]. While these studies have highlighted potential risks, their findings remain inconsistent, likely due to methodological differences and reliance on conventional, point-in-time metrics. In most cases, drug exposure was assessed only at the time of transplantation, failing to account for the cumulative burden experienced throughout hospitalization. This limitation may lead to an incomplete assessment of medication-related risks, reinforcing the need for more dynamic and temporally sensitive approaches to exposure assessment.

HSCT treatment regimens involve prolonged use of antineoplastics, immunosuppressants, antimicrobials, and supportive therapies, making these patients especially vulnerable to PP, HPP, DDIs, and PIMs [21,23]. Given the duration and complexity of treatment, accurately quantifying CDE of prescribing determinants is essential to identify medication-related risks such as increased toxicity or adverse drug events. It could also help optimize therapeutic strategies by adjusting drug regimens and limiting unnecessary medication burden while refining prescribing guidelines through real-world evidence on exposure patterns and associated risks.

We focused on all adult patients hospitalized at Rennes University Hospital between November 1, 2020, and November 30, 2021, for autologous HSCT. Patients were identified through the diagnosis-related group of their hospital stay (ie, 27Z03) available in the Rennes University Hospital CDW, corresponding to autologous HSCT patients. Comorbidities were assessed using the Charlson Comorbidity Index, calculated from *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnostic codes available in the CDW. For patients with multiple admissions during the study period, only the first hospitalization during which the transplant procedure was performed was retained for analysis, serving as the index stay for the computation of cumulative exposure metrics.

Statistical Analysis

Overview

To evaluate the relevance of cumulative exposure metrics (CDE and CDED) in reflecting overall drug exposure throughout hospitalization, statistical analyses were conducted to examine their relationship with conventional metrics at 2 specific time points: admission (first day of hospitalization) and discharge (last day of hospitalization).

The objective was to determine whether the number of medications, DDIs, and PIMs at admission and discharge, or the presence of these determinants at these time points, were correlated with cumulative exposure throughout hospitalization, or whether these conventional metrics failed to reflect the cumulative and dynamic nature of drug exposure.

The normality of the distribution of CDE, CDED, and the number of medications, DDIs, and PIMs at admission and discharge was assessed using the Shapiro-Wilk test, with a significance level of .05. A *P* value of <.05 indicated a nonnormal distribution, leading to the use of nonparametric statistical methods for subsequent analyses. All statistical analyses were performed using R (version 3.6.0; R Core Team).

Description of CDE and CDED Across Patients' Groups

To describe cumulative exposure to prescribing determinants (PP, HPP, DDIs, and PIMs), patients were categorized into 4 trajectory groups based on their exposure status at admission (first day of hospitalization) and at discharge (last day of hospitalization). For each determinant, the 4 predefined trajectories were: 0→0 (not exposed at admission, nor at discharge), 0→1 (not exposed at admission but exposed at discharge), 1→0 (exposed at admission but not at discharge), and 1→1 (exposed at both admission and discharge).

For each trajectory group, descriptive statistics of cumulative exposure metrics were computed: mean, median, interquartile range, and range of both CDE and CDED. This descriptive approach aimed to characterize cumulative exposure patterns within each trajectory group, providing a detailed overview of drug burden during hospitalization.

Comparison of Cumulative Exposure Metrics With Conventional Metrics

To assess whether cumulative exposure throughout hospitalization was correlated with the number of administered medications, DDIs, and PIMs at admission and discharge, Spearman rank correlation coefficients were calculated. The Spearman correlation test was chosen due to the nonnormal distribution of the variables. This analysis aimed to assess the relationship between the number of medications, DDIs, and PIMs at admission and discharge with their respective CDE and CDED values, and a weak correlation (r_s<0.3 [24]) would suggest that a single time point measure does not adequately reflect overall exposure, highlighting the need for cumulative exposure metrics.

In parallel, we investigated whether the presence of PP, HPP, DDIs, or PIMs at admission or discharge was associated with cumulative exposure metrics. For each determinant, the distributions of CDE and CDED were compared between patients classified as exposed versus nonexposed (eg, PP=0 vs PP=1) at both time points. Boxplots were used to visualize these distributions and examine potential differences in median cumulative exposure across groups. To support this visual comparison, the difference in median values between exposed and nonexposed groups was also calculated (*d*), providing a quantitative summary of potential disparities. This complementary, descriptive analysis provided additional insight into whether conventional binary indicators commonly used could reliably reflect cumulative drug burden in hospitalized patients.

Factorial Analysis of Mixed Data

To further explore the relationship between conventional metrics and cumulative exposure metrics (CDE and CDED), a factorial analysis of mixed data (FAMD) was conducted. This approach was chosen to simultaneously analyze numerical measures of drug exposure (eg, number of drugs administered and cumulative exposure) and categorical classifications of prescribing determinants (eg, presence of PP, HPP, and DDIs). The objective was to assess whether cumulative exposure metrics introduce new information beyond conventional metrics and whether they form an independent dimension in the factorial space, justifying

XSL•FO
RenderX

their use as complementary or alternative measures. Two separate FAMD analyses were performed: one incorporating CDE and another focusing on CDED, each compared with conventional metrics at admission and discharge. The contribution of each variable to the principal dimensions was examined using eigenvalues and variable loadings. Variables strongly associated with a given dimension were identified to determine whether CDE or CDED formed an independent axis of variation. PIMs were excluded from the FAMD analyses as they applied exclusively to patients aged 65 years and older. Including a determinant restricted to a subgroup would have introduced structural heterogeneity and compromised the interpretability of global dimensions across the full cohort.

Ethical Considerations

The study was conducted in accordance with the ethical principles of the 1964 Declaration of Helsinki and its amendments, or equivalent standards. Given its retrospective and noninterventional design based on existing database data, a waiver of informed consent was granted. Confidentiality and anonymity were strictly maintained throughout: the extracted data contained no nominative information and were fully deidentified prior to analysis, in compliance with the institutional data protection procedures. The protocol was approved by the local ethics committee (approval 24.12, May 2024) and complies with French clinical research regulations. The study used routinely collected hospital data. No compensation was provided to participants and no direct contact with patients occurred.

Results

Study Population

This study included 69 adult patients hospitalized at Rennes University Hospital between November 1, 2020, and November 30, 2021, for autologous HSCT. Patients were identified using the diagnosis-related group code (eg, 27Z03). For each patient, only the hospitalization during which the transplant procedure was performed was retained as the index stay, so that each patient contributed a single stay to the analysis. The cohort's median age was 60 years (range 21 - 70 years), with 22 patients out of 69 (31.9%) aged 65 years or older. The median length of hospital stay was 19 days (range 14 - 52 days).

The primary indications for HSCT were multiple myeloma (44/69, 64%), non-Hodgkin lymphoma (19/69, 28%), and Hodgkin lymphoma (6/69, 9%). Patients exhibited a median Charlson Comorbidity Index score of 4, reflecting a significant burden of comorbid conditions.

Prescribing Determinants

Overview

The analysis of prescribing determinants among hospitalized HSCT patients revealed distinct patterns in PP, HPP, DDIs, and PIMs. These results are shown in Table 1 and highlight the dynamic nature of prescribing determinants in hospitalized HSCT patients, demonstrating substantial fluctuations between admission, hospitalization, and discharge.

Nearly all patients (68/69, 99%) experienced at least 1 day of PP during hospitalization, although its prevalence was substantially lower at admission (10/69, 15%) and discharge (39/69, 58%). Similarly, HPP was present in 77% (53/69) of patients at some point during their stay but was rare at admission (1/69, 1%) and remained infrequent at discharge (1/69, 1%). DDIs were highly prevalent throughout hospitalization, with 99% (68/69) of patients experiencing at least 1 interaction, although major DDIs were less common than minor ones. At admission, no patients had a major DDI, while 36% (25/69) had a minor DDI. By discharge, 3% (2/69) of patients had major DDIs and 52% (36/69) still had minor DDIs. Although the proportion of patients with major DDIs at admission or discharge remains low, 20% (14/69) of patients were exposed to at least 1 major DDI at some point during their hospital stay. Among patients aged 65 years or older, all had at least 1 PIM during their hospitalization, with 36% (8/22) presenting with a PIM at admission and 45% (10/22) at discharge.

The mean number of drug administrations per day was 7.7 across the entire hospital stay but was lower at admission (3.0 per day) and at discharge (5.1 per day). The number of DDIs and PIMs administrations followed similar trends: 0.7 DDI per day at admission and 1.5 at discharge and 0.4 PIM per day at admission and 0.7 at discharge.

XSL•FO
RenderX

Table . Conventional metrics and cumulative drug exposure metrics for prescribing determinants during hospitalization[a].

Variable	Hospital stay	Admission	Discharge
Presence of prescribing determinants (prevalence of patients in %)			
PP[b]	99 (68/69)	15 (10/69)	58 (39/69)
HPP[c]	77 (53/69)	1 (1/69)	1 (2/69)
DDIs[d]	99 (68/69)	36 (25/69)	52 (36/69)
DDI major	20 (14/69)	0 (0/69)	3 (2/69)
DDI minor	99 (68/69)	36 (25/69)	52 (36/69)
PIMs[e]	100 (22/22)	36 (8/22)	45 (10/22)
Number of administrations per day (mean)			
Drugs administration	7.7	3.0	5.1
DDI administration	3.2	0.7	1.5
Major	0.1	0.0	0.0
Minor	3.2	0.7	1.5
PIM administration	1.2	0.4	0.7
CDE[f] during hospital stay (days), mean			
PP	16.3	—[g]	—
HPP	5.8	—	—
DDIs	64.7	—	—
Major	2.0	—	—
Minor	62.7	—	—
PIMs	19	—	—
CDED[h], mean			
PP	0.8	—	—
HPP	0.3	—	—
DDIs	3.2	—	—
Major	0.1	—	—
Minor	3.1	—	—
PIMs	1.2	—	—

[a]This table presents the conventional metrics and cumulative drug exposure metrics for prescribing determinants, including polypharmacy (PP), hyperpolypharmacy (HPP), drug-drug interactions (DDIs), and potentially inappropriate medications (PIMs), at 3 key hospitalization time points: admission, discharge, and throughout the hospital stay. The conventional metrics include the presence of each prescribing determinant at admission and discharge, expressed as the percentage of patients who had the determinant at that specific time point. For the presence of prescribing determinants throughout the hospital stay, a patient was considered exposed if they had at least one occurrence of the determinant at any time during hospitalization. Additionally, the number of administrations at admission and discharge is reported. For PP and HPP, this corresponds to the total number of distinct medications administered per day, while for DDIs and PIMs, it represents the total number of interactions or inappropriate medications recorded at each time point. The cumulative drug exposure metrics include cumulative drug exposure (CDE), which quantifies the total number of days a patient was exposed to a prescribing determinant during hospitalization. For DDIs and PIMs, this metric accounts for multiple occurrences per day. Cumulative drug exposure density normalizes CDE by the total length of the hospital stay, providing an intensity-adjusted measure of exposure. The analysis of PIM exposure is restricted to patients aged 65 years and older.

[b]PP: polypharmacy.

[c]HPP: hyperpolypharmacy.

[d]DDIs: drug-drug interactions.

[e]PIMs: potentially inappropriate medications.

[f]CDE: cumulative drug exposure.

[g]Not available.

[h]CDED: cumulative drug exposure density.

XSL•FO
RenderX

Comparison of Cumulative Exposure Metrics Across Patient Groups

To further describe CDE across prescribing determinant trajectories (PP, HPP, DDIs, and PIMs), detailed descriptive statistics were calculated for both CDE and CDED, including mean, median (IQR), and range. Patients classified as exposed at both admission and discharge (1→1) generally showed the highest CDE and CDED values for each prescribing determinant, indicating sustained exposure throughout hospitalization. For example, in this group, median CDE reached 17.5 days (14.5 - 19.8) for PP and 60 days (51-132) for DDIs, corresponding to median CDED values close to 1.0 and 3.5, respectively, indicating sustained exposure throughout hospitalization. However, nonnegligible cumulative exposure was also observed in 0→1 trajectories, particularly for DDIs where the median CDE was 69 days (31-122), highlighting the limitations of conventional point-in-time metrics in capturing dynamic exposure patterns. Even among 0→0 patients, who were not exposed at either time point, cumulative exposure levels remained relatively high. For example, the median CDE was 14 days (10-18) for PP and 31 days (18-63) for DDIs, suggesting that conventional prevalence-based indicators may underestimate transient or intermediate exposure. Full descriptive results are provided in Multimedia Appendix 1.

Comparison of Cumulative Exposure Metrics With Conventional Metrics

Correlation and distribution analyses highlighted heterogeneous relationships between conventional metrics and cumulative exposure metrics (Figures 2 and 3).

For PP, HPP, and DDIs, correlations with CDE and CDED at admission were weak and nonsignificant (r_s<0.3 − P>.05), while stronger significant associations emerged at discharge for PP (r_sCDED=0.43 − P<.001), HPP (r_sCDE=0.44 − P<.001; r_sCDED=0.46 − P<.001) and DDIs (r_s CDE=0.44 − P<.001; r_s CDED=0.46 − P<.001).

In contrast, PIMs showed consistently strong correlations with both CDE and CDED at admission and discharge (r_s CDE=0.73 − P<.001, r_s CDED=0.71 − P<.001 at admission; and r_s CDE=0.52 − P=.012, r_s CDED=0.59 − P=.004 at discharge). Due to the discrete nature of PIM administration (maximum of 1 or 2 drugs at admission or discharge), regression lines were not plotted for PIMs in the scatterplots, as they did not meaningfully represent trends in the data.

Distribution comparisons (boxplots) further confirmed that the presence of PP or DDIs at admission was not associated with noticeably higher cumulative exposure. In contrast, PIM presence at admission consistently corresponded to increased CDE and CDED values. At discharge, only DDIs and PIMs were visibly associated with higher cumulative exposure, while differences for PP remained limited. All results are summarized in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. Comparison of conventional metrics with CDE across prescribing determinants (PP, HPP, DDIs, and PIMs). This figure examines the relationship between conventional metrics at admission and discharge and CDE over the entire hospitalization for each prescribing determinant: PP, HPP, DDIs, and PIMs. Each panel (A, B, C, and D) corresponds to a specific prescribing determinant: (A) PP, (B) HPP, (C) DDIs, and (D) PIMs. For each prescribing determinant, 2 types of analyses were conducted: (1) Correlation analysis (scatter plots, left side): the association between the number of administrations at admission (top left) and discharge (bottom left) and CDE was assessed using Spearman rank correlation coefficient (*r*) and regression lines. (2) Comparison of distributions (boxplots, right side): the exposure status of patients to each prescribing determinant (binary: 0 = not exposed, 1 = exposed) at admission (top right) and discharge (bottom right) was compared against CDE using boxplots. Differences in CDE were evaluated descriptively by examining the distribution and median CDE values between exposed and nonexposed patients. A weak correlation (*r*s <0.3) or minimal differences in CDE distributions between exposed and nonexposed patients at admission or discharge suggest that a single time point measure may not adequately reflect CDE, emphasizing the need for longitudinal assessment. Conversely, stronger correlation (*r*s close to 1) or pronounced distributional differences suggests that conventional metrics at admission or discharge may serve as reliable proxies for CDE. CDE: cumulative drug exposure; DDIs: drug-drug interactions; HPP: hyperpolypharmacy; PIMs: potentially inappropriate medications; PP: polypharmacy.

XSL•FO
RenderX

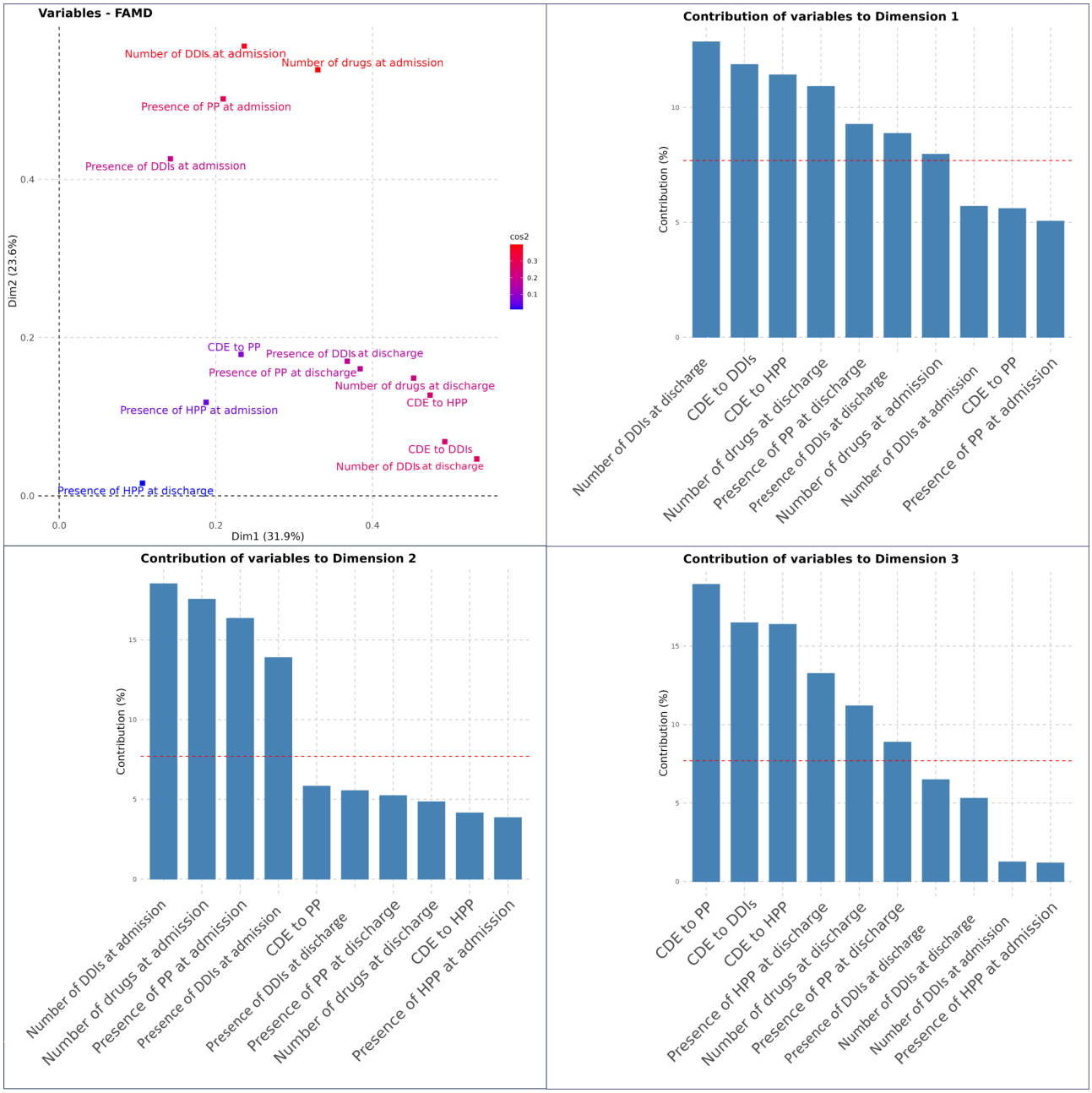
Figure 3. Comparison of conventional metrics with CDED across prescribing determinants (PP, HPP, DDIs, and PIMs). This figure examines the relationship between conventional metrics at admission and discharge and CDED over hospitalization for each prescribing determinant: PP, HPP, DDIs, and PIMs. Each panel (A, B, C, and D) corresponds to a specific prescribing determinant: (A) PP, (B) HPP, (C) DDIs, and (D) PIMs. For each prescribing determinant, two types of analyses were conducted. (1) Correlation analysis (scatter plots, left side): the association between the number of administrations at admission (top left) and discharge (bottom left), and CDED was assessed using Spearman rank correlation coefficient (*r*) and regression lines. (2) Comparison of distributions (boxplots, right side): the exposure status of patients to each prescribing determinant (binary: 0 = not exposed, 1 = exposed) at admission (top right) and discharge (bottom right) was compared against CDED using boxplots. Differences in CDED were evaluated descriptively by examining the distribution and median CDED values between exposed and nonexposed patients. A weak correlation (*r*s<0.3) or minimal differences in CDED distributions between exposed and nonexposed patients at admission or discharge suggest that a single time point measure may not adequately reflect CDED, emphasizing the need for longitudinal assessment. Conversely, stronger correlation (*r*s close to 1) or pronounced distributional differences suggest that conventional metrics at admission or discharge may serve as reliable proxies for CDED. CDED: cumulative drug exposure density; DDIs: drug-drug interactions; HPP: hyperpolypharmacy; PIMs: potentially inappropriate medications; PP: polypharmacy.

XSL•FO
RenderX

Factorial Analysis of Mixed Data

To assess how cumulative exposure measures relate to conventional drug exposure metrics, 2 separate FAMD analyses were conducted: one incorporating CDE and another integrating CDED. Full details, including variable contributions and graphical representations, are provided in Figures 4 and 5 and in Multimedia Appendices 3 and 4.

Figure 4. Representation and contribution of variables in the FAMD on CDE and conventional metrics. This figure presents the results of the FAMD conducted to compare CDE metrics with conventional drug metrics. The top-left panel illustrates the factor space representation of variables, where the color gradient reflects the cos² values, indicating the quality of representation in the first 2 dimensions. The 3 bar plots display the contribution (%) of each variable to the formation of the first 3 principal dimensions in the FAMD performed to compare CDE with conventional drug exposure metrics. Each panel represents 1 of the 3 dimensions. The red dashed line represents the expected average contribution threshold, indicating variables that contribute more than expected to each dimension. CDE: cumulative drug exposure; DDIs: drug-drug interactions; FAMD: factorial analysis of mixed data; HPP: hyperpolypharmacy; PP: polypharmacy.



The FAMD on CDE metrics and conventional metrics shown in Figure 4 identified 3 principal dimensions, explaining 69.2% of the total variance (31.9% for dimension 1, 23.6% for dimension 2, and 13.7% for dimension 3). The first and third dimensions present a mixed structure, integrating both conventional metrics (eg, number of drugs and DDIs at admission and discharge, presence of PP or HPP) and cumulative exposure metrics (CDE to DDIs and HPP for dimension 1; CDE to PP and CDE to DDIs and HPP for dimension 3). In contrast, the second dimension is shaped almost exclusively by conventional metrics, particularly variables describing the patient's medication profile at admission.

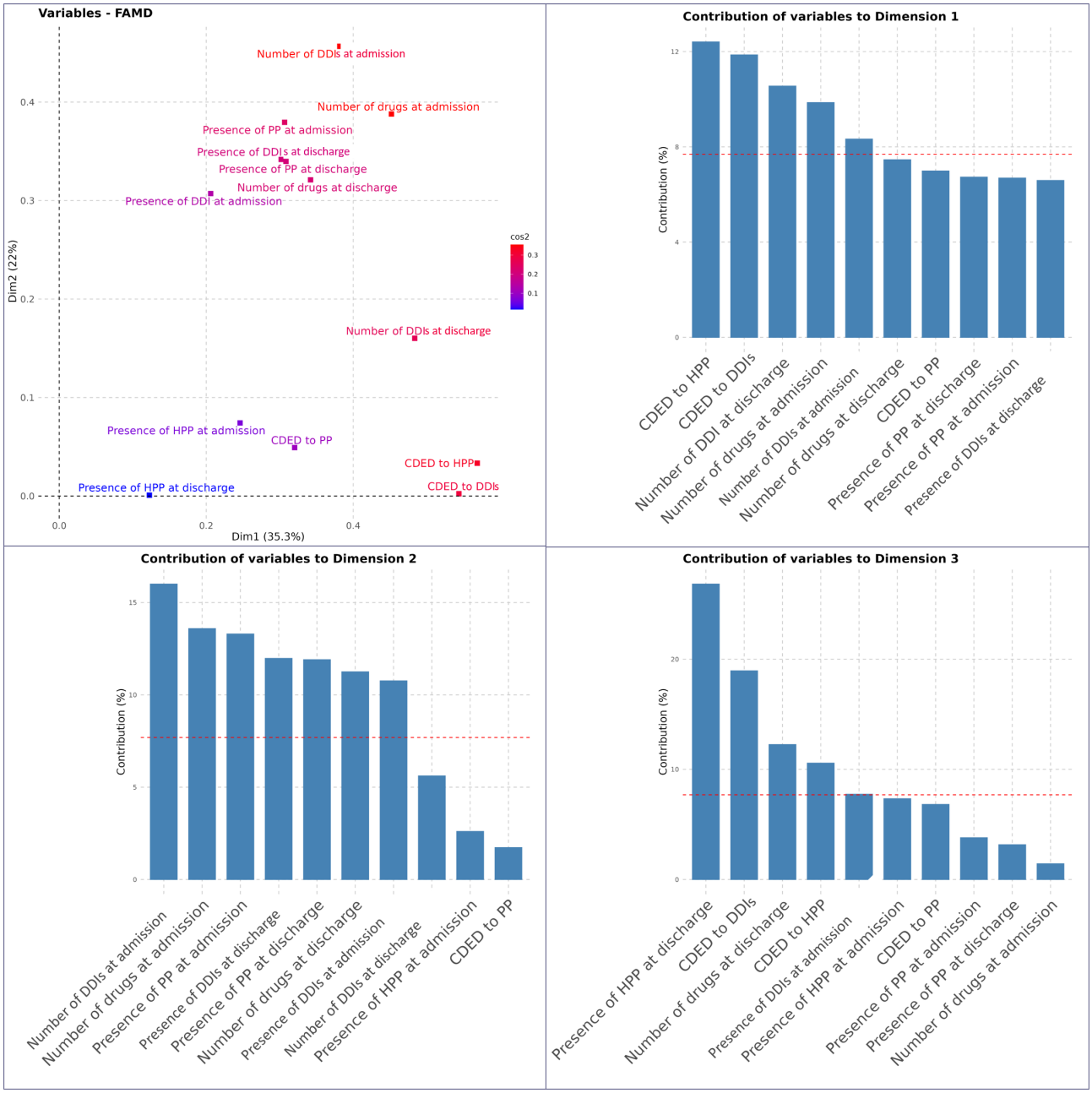
The FAMD on CDED metrics and conventional metrics presented in Figure 5 identified 3 principal dimensions,

XSL•FO
RenderX

explaining 68.3% of the total variance (35.3% for dimension 1, 21.9% for dimension 2, and 11.1% for dimension 3).

Dimension 1 was mainly driven by CDED to HPP and CDED to DDI metrics, alongside conventional metrics such as the number of medications at admission and discharge, as well as the number of DDIs at both time points. This suggests that dimension 1 integrates both conventional metrics and CDE metrics. Dimension 2, by contrast, was solely structured by conventional metrics, with no significant contribution from CDED variables. Dimension 3 reflected a combination of conventional metrics (presence of HPP at discharge and number of drugs at discharge) and CDED metrics related to HPP and DDIs further supporting the complementary role of CDED metrics in capturing dense and variable drug exposure patterns.

Figure 5. Representation and contribution of variables in the FAMD on CDED and conventional metrics. This figure presents the results of the FAMD conducted to compare CDED metrics with conventional drug metrics. The top-left panel illustrates the factor space representation of variables, where the color gradient reflects the cos² values, indicating the quality of representation in the first 2 dimensions. The 3 bar plots display the contribution (%) of each variable to the formation of the first 3 principal dimensions in the FAMD performed to compare CDED with conventional metrics. Each panel represents 1 of the 3 dimensions. The red dashed line represents the expected average contribution threshold, indicating variables that contribute more than expected to each dimension. CDED: cumulative drug exposure density; DDIs: drug-drug interactions; FAMD: factorial analysis of mixed data; HPP: hyperpolypharmacy; PP: polypharmacy.



XSL•FO
RenderX

Discussion

Principal Results

This exploratory study aimed to develop and evaluate 2 CDE metrics: CDE and CDED, and to compare them to conventional, point-in-time indicators in a cohort of autologous HSCT recipients. These metrics were designed to capture the number of days during which patients were exposed to PP, HPP, DDIs, and PIMs, with CDED normalizing this duration by length of stay. Compared with binary measures at admission or discharge, CDE and CDED offered a more nuanced picture of drug burden by incorporating treatment duration and intensity.

Across all prescribing determinants, the number of drugs and DDIs at admission showed no significant correlation with CDE or CDED, and the presence of PP, HPP, or DDIs at admission was not associated with higher CDE or CDED values. These findings suggest that admission-based conventional metrics do not reliably reflect the overall drug burden experienced by patients throughout hospitalization. Conversely, while the number of medications at discharge was not correlated with cumulative exposure to PP, a moderate correlation was observed with cumulative exposure to HPP. Similarly, the number of DDIs at discharge showed a significant positive correlation with cumulative exposure to DDIs.

These findings can be explained by the specific therapeutic trajectory of autologous HSCT. Drug exposure evolves dynamically throughout the hospital stay; it is not determined solely by the patient's status at admission or discharge but by successive clinical phases such as conditioning, aplasia, and management of complications or auto-graft versus host disease [25-27]. Numerous medications (antimicrobials, antivirals, antifungals, immunosuppressants, and supportive care drugs) are initiated to prevent or manage infections, mucositis, nausea, and hematologic toxicities, and many are maintained until discharge and sometimes beyond [23,28]. This progressive accumulation of treatments contributes to cumulative exposure and helps explain why discharge metrics better correlate with cumulative measures.

The heterogeneity of associations reported in the literature between conventional metrics and clinical outcomes in malignant hematological diseases further underscores the limitations of static indicators. Studies have linked PP, PIMs, and DDIs to increased toxicity, reduced survival, or poorer transplantation outcomes, but the strength and consistency of these associations vary widely depending on definitions, study populations, (autologous vs allogeneic HSCT), and the timing of measurement [21,22,29,30]. Such divergences highlight the need for dynamic and integrative exposure measures that better reflect the complexity of pharmacological trajectories.

To further explore whether cumulative metrics capture distinct aspects of drug exposure, we conducted an FAMD. When CDE variables were included, the principal components were largely driven by conventional metrics such as the number of medications and the presence of PP or DDIs at admission and discharge; CDE variables were less strongly represented, suggesting that they capture specific dimensions not fully reflected by static measures. In contrast, when CDED was included, this exposure density measure contributed more substantially to the first dimensions, alongside some conventional metrics, indicating that normalizing cumulative exposure by length of stay captures relevant contrasts between patients.

PIMs were not incorporated into the FAMD because they apply only to patients aged 65 years and older, and including a determinant restricted to a subpopulation would have introduced structural heterogeneity. The dynamics of PIM exposure, often linked to chronic or preexisting treatments, differ from those of other determinants: conventional PIM metrics tend to align closely with cumulative measures because these medications are commonly part of long-term regimens started before hospitalization and continued during and after the inpatient period [31,32].

Limitations

This study has several limitations. First, the relatively small sample size (n=69) and the single-center design inevitably limit the statistical power of the analyses and the generalizability of the findings. Accordingly, this work should be interpreted primarily as an exploratory methodological development rather than as a definitive evaluation of drug exposure in HSCT patients. The study population consisted exclusively of autologous HSCT recipients, a highly specific group treated under standardized therapeutic protocols. This deliberate choice was made because these patients are exposed to dense and complex drug burden, providing a relevant framework to test and refine the proposed metrics. However, the conceptual approach underlying CDE and CDED is not restricted to autologous HSCT recipients: these cumulative measures are designed to be clinically agnostic and can be applied to other hospitalized populations, regardless of their pharmacological profile. While their relevance may be particularly evident in contexts of high therapeutic intensity such as oncology, intensive care, or surgery, they are generalizable to any setting where drug exposure evolves dynamically over time. Future work should therefore focus on validating these metrics in larger, multicenter, and more diverse clinical populations, ideally including outcome analyses to establish their clinical significance.

Second, clinical outcomes such as toxicity, adverse drug events, or hospital readmission were not assessed in this study. While this limits the immediate clinical applicability and interpretation of cumulative exposure, the primary purpose of this work was methodological, aiming to develop and illustrate novel exposure metrics. However, ongoing work is being conducted to evaluate the relationship between these metrics and patient outcomes, in order to support their prognostic and practical value.

Third, although the use of a CDW represents a key strength by offering precise, structured information on actual drug administrations, it is not without limitations. In particular, CDWs do not systematically capture clinical indications, prescriber rationale, or therapeutic intent in a structured format. Retrieving such information would require manual review of unstructured clinical documents such as discharge summaries, without guarantee of consistency or completeness. This limitation can

XSL•FO
RenderX

complicate the interpretation of the clinical appropriateness of certain prescribing patterns such as PP, HPP, or DDIs, which may, in some cases, be justified depending on the clinical context, severity of illness, or treatment objectives. Furthermore, certain contextual elements such as patient-brought medications or off-protocol adaptations may not be consistently documented in structured data fields.

Comparison With Prior Work

The cumulative nature of drug exposure, although conceptually aligned with long-standing practices in public health, remains underexplored in pharmacoepidemiology. Established exposure metrics such as pack-years in tobacco research or cumulative indices in occupational health have long recognized the importance of duration and timing. Conceptually, this approach aligns with the broader exposome framework, which aims to characterize the totality of environmental, biological, and behavioral exposures an individual encounters across the life course [33].

Within this framework, pharmaceuticals represent a major, yet insufficiently characterized component of the chemical exposome, particularly regarding their longitudinal accumulation. Although medications are intentionally administered, their cumulative impact, especially in contexts of PP and prolonged or intensive treatment, can significantly modulate health outcomes beyond the intended therapeutic effect [34].

Recent methodological developments have sought to better account for the temporal dimension of drug exposure. Among them, the weighted cumulative exposure model has extended time-sensitive modeling to pharmacological exposures by estimating the effect of prior doses using flexible, spline-based weight functions over specified exposure windows [35]. However, these approaches rely predominantly on prescription or dispensing data and are implemented within regression-based analytic frameworks, requiring prior assumptions about exposure timing and pharmacological latency. While such models offer valuable insights for studying long-term drug effects in ambulatory care, they are less suited to settings where drug exposure is highly dynamic and context dependent. In this context, CDE and CDED leverage real-time administration data recorded in hospital CDWs, enabling standardized and reproducible quantification of actual in-hospital drug burden.

At the population level, CDE and CDED offer valuable tools to inform hospital-wide strategies for medication management. By providing standardized, temporally resolved indicators of drug exposure, they could be integrated into institutional dashboards to monitor prescribing practices across wards or clinical services. For example, identifying units with persistently high cumulative exposure to PP may help guide resource allocation, staff training, or targeted medication review initiatives. Beyond real-time monitoring, these metrics also allow for longitudinal tracking of drug exposure trends, making them particularly relevant for evaluating the impact of deprescribing policies or institutional changes in prescribing practices. Such applications directly respond to limitations identified in previous work, which highlighted that commonly used point-in-time indicators fail to adequately capture the dynamic nature of drug exposure during hospitalization and emphasized the need for cumulative, temporally informed metrics better suited to quality monitoring and prescribing optimization [36-38].

At the individual level, these new metrics could enhance both the specificity and the clinical relevance of decision support systems. Most CDSS currently rely on static medication lists and rule-based logic applied at fixed points in time, often ignoring how long or how intensively a patient has been exposed to potentially risky prescriptions. This contributes to high rates of nonspecific alerts and clinician alert fatigue [39,40].

By incorporating cumulative exposure into alerting mechanisms, CDSS could move beyond simple presence or absence logic to integrate the duration and intensity of drug exposure, thereby identifying patients with a sustained drug burden more accurately. For example, alerts could be designed to trigger only when exposure to high-risk interactions or inappropriate medications exceeds a predefined cumulative threshold, supporting more clinically meaningful interventions and reducing unnecessary interruptions. Embedding these indicators into patient-level dashboards within EHRs would also provide clinicians with a longitudinal view of drug burden at the bedside, directly informing prescribing and deprescribing decisions.

However, the definition of clinically validated cumulative exposure thresholds remains to be established. Demonstrating associations between these thresholds and meaningful patient outcomes is a critical next step, beyond the methodological proof-of-concept presented here, and would provide the necessary foundation for their integration into CDSS. This represents a relevant direction for future research, particularly for refining risk stratification and alert prioritization within digital clinical tools.

In summary, CDE and CDED address key limitations of existing exposure indicators by leveraging real-world, time-resolved data. Their integration into both institutional governance frameworks and clinical decision tools opens new perspectives for optimizing pharmacotherapy and promoting safer, more appropriate medication use in hospital settings.

Conclusions

This study proposes 2 cumulative exposure metrics, CDE and CDED, derived from daily drug administration data recorded in hospital CDWs. By incorporating the temporal dimension of treatment, these metrics offer a more comprehensive assessment of in-hospital drug exposure than conventional point-in-time metrics. Their reproducibility and adaptability support applications ranging from research to quality monitoring and medication safety. Further analyses are underway to examine how these cumulative profiles relate to patient trajectories and clinical events and to explore their potential contribution to hospital surveillance and prescribing evaluation efforts.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Descriptive analysis of cumulative drug exposure (CDE) and cumulative drug exposure density (CDED) across prescribing determinant trajectories. This table shows descriptive statistics for CDE and CDED across patient groups categorized by their prescribing determinant trajectories: polypharmacy (PP), hyperpolypharmacy (HPP), drug-drug interactions (DDIs), and potentially inappropriate medications (PIMs). Each prescribing determinant was assessed at 2 time points: admission (first day of hospitalization) and discharge (last day of hospitalization). Based on their exposure status at these time points, patients were classified into four trajectory groups: (1) 0→0: Patients not exposed to the determinant at admission nor discharge. (2) 0→1: Patients not exposed to the determinant at admission but exposed at discharge. (3) 1→0: Patients exposed to the determinant at admission but no longer exposed at discharge. (4) 1→1: Patients exposed to the determinant at both admission and discharge. CDE represents the total number of days a patient was exposed to the prescribing determinant throughout the hospital stay (eg, number of days with polypharmacy). CDED corresponds to the ratio of CDE to the length of stay, reflecting the average intensity of exposure over time. For each trajectory group and each prescribing determinant, the table reports the number of patients on each group (n), mean, median with interquartile range (IQR), and minimum-maximum values for both CDE and CDED.
[DOCX File, 5192 KB - medinform_v14i1e76961_app1.docx]

Multimedia Appendix 2

Correlation and distribution comparisons between conventional metrics and cumulative drug exposure metrics. This table shows the results of Spearman correlation assessing the relationship between conventional metrics at admission and discharge and cumulative drug exposure (CDE) and cumulative drug exposure density (CDED) for each prescribing determinant: polypharmacy (PP), hyperpolypharmacy (HPP), drug-drug interactions (DDIs), and potentially inappropriate medications (PIMs). Table A corresponds to CDE, quantifying total exposure days. Table B corresponds to CDED, which normalizes exposure to hospital stay length. Spearman rank correlation (r) assessed the relationship between the number of drug administrations at admission or discharge and CDE or CDED. A strong correlation (r close to 1 or −1) suggests that the number of administrations at a single time point is predictive of cumulative exposure. The difference in median CDE or CDED values between exposed (1) and nonexposed (0) patients provides a complementary perspective on the extent to which conventional, binary time point metrics reflect overall cumulative drug burden.
[DOCX File, 6925 KB - medinform_v14i1e76961_app2.docx]

Multimedia Appendix 3

Representation and contribution of variables in the factorial analysis of mixed data (FAMD) performed on cumulative drug exposure (CDE) and conventional metrics. This table shows the results of the FAMD performed to compare CDE with conventional metrics, including the number of drug administrations, the number of drug interactions, and the presence of prescribing determinants—polypharmacy (PP), hyperpolypharmacy (HPP), and drug-drug interactions (DDIs)—at admission and discharge. For each variable, two parameters are provided across the first 3 principal dimensions. (1) Squared cosine (cos^2): It indicates the quality of representation of a variable on a given dimension. Higher values (closer to 1) suggest strong alignment with that dimension, while lower values indicate dispersion across multiple dimensions. (2) Contribution (%): It represents the extent to which a variable influences the formation of a given dimension. Variables with higher contributions play a more prominent role in defining the corresponding factorial axis.
[DOCX File, 5174 KB - medinform_v14i1e76961_app3.docx]

Multimedia Appendix 4

Representation and contribution of variables in the factorial analysis of mixed data (FAMD) performed on cumulative drug exposure density (CDED) and conventional metrics. This table shows the results of the FAMD performed to compare CDED with conventional drug exposure metrics, including the number of drug administrations, the number of drug interactions, and the presence of prescribing determinants—polypharmacy (PP), hyperpolypharmacy (HPP), and drug-drug interactions (DDIs)—at admission and discharge. For each variable, two parameters are provided across the first 3 principal dimensions. (1) Squared cosine (cos^2): It indicates the quality of representation of a variable on a given dimension. Higher values (closer to 1) suggest strong alignment with that dimension, while lower values indicate dispersion across multiple dimensions. (2) Contribution (%): It represents the extent to which a variable influences the formation of a given dimension. Variables with higher contributions play a more prominent role in defining the corresponding factorial axis.
[DOCX File, 5173 KB - medinform_v14i1e76961_app4.docx]

XSL•FO
RenderX

References

1. Bates DW, Cullen DJ, Laird N. Incidence of adverse drug events and potential adverse drug events. JAMA 1995 Jul 5;274(1):29. [doi: 10.1001/jama.1995.03530010043033] [Medline: 7791255]
2. Pirmohamed M, James S, Meakin S, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 2004 Jul 3;329(7456):15-19. [doi: 10.1136/bmj.329.7456.15] [Medline: 15231615]
3. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 2003 Jun 23;163(12):1409-1416. [doi: 10.1001/archinte.163.12.1409] [Medline: 12824090]
4. Reeve E, Gnjidic D, Long J, Hilmer S. A systematic review of the emerging definition of “deprescribing” with network analysis: implications for future research and clinical practice. Br J Clin Pharmacol 2015 Dec;80(6):1254-1268. [doi: 10.1111/bcp.12732] [Medline: 27006985]
5. Park H, Chae MK, Jeong W, et al. Appropriateness of alerts and physicians’ responses with a medication-related clinical decision support system: retrospective observational study. JMIR Med Inform 2022 Oct 4;10(10):e40511. [doi: 10.2196/40511] [Medline: 36194461]
6. Lee TA, Pickard AS. Exposure definition and measurement. In: Developing a Protocol for Observational Comparative Effectiveness Research: A User’s Guide Agency for Healthcare Research and Quality (US) 2013. URL: https://www.ncbi.nlm.nih.gov/books/NBK126191 [accessed 2025-04-08]
7. Exposure-response relationships—study design, data analysis, and regulatory applications. US Food and Drug Administration. 2025. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/exposure-response-relationships-study-design-data-analysis-and-regulatory-applications [accessed 2025-04-08]
8. Ouchi D, Giner-Soriano M, Gómez-Lumbreras A, Vedia Urgell C, Torres F, Morros R. Automatic estimation of the most likely drug combination in electronic health records using the smooth algorithm: development and validation study. JMIR Med Inform 2022 Nov 15;10(11):e37976. [doi: 10.2196/37976] [Medline: 36378514]
9. Wang Z, Syed M, Syed S, et al. Clinical data warehousing: a scoping review. J Soc Clin Data Manag 2024;4(2). [doi: 10.47912/jscdm.320] [Medline: 41602775]
10. Dimitsaki S, Natsiavas P, Jaulent MC. Applying AI to structured real-world data for pharmacovigilance purposes: scoping review. J Med Internet Res 2024 Dec 30;26:e57824. [doi: 10.2196/57824] [Medline: 39753222]
11. Patadia VK, Coloma P, Schuemie MJ, et al. Using real-world healthcare data for pharmacovigilance signal detection - the experience of the EU-ADR project. Expert Rev Clin Pharmacol 2015 Jan;8(1):95-102. [doi: 10.1586/17512433.2015.992878] [Medline: 25487079]
12. Ledieu T, Bouzillé G, Polard E, Plaisant C, Thiessard F, Cuggia M. Clinical data analytics with time-related graphical user interfaces: application to pharmacovigilance. Front Pharmacol 2018;9:717. [doi: 10.3389/fphar.2018.00717] [Medline: 30233354]
13. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights 2016;8(1–10):1-10. [doi: 10.4137/BII.S31559] [Medline: 26843812]
14. Madec J, Bouzillé G, Riou C, et al. eHOP clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019 Aug 21;264:1536-1537. [doi: 10.3233/SHTI190522] [Medline: 31438219]
15. Husson MC. Theriaque: independent-drug database for good use of drugs by health practitioners. Ann Pharm Fr 2008;66(5-6):268-277. [doi: 10.1016/j.pharma.2008.07.009] [Medline: 19061726]
16. Bories M, Bouzillé G, Cuggia M, Le Corre P. Drug-drug interactions in elderly patients with potentially inappropriate medications in primary care, nursing home and hospital settings: a systematic review and a preliminary study. Pharmaceutics 2021 Feb 16;13(2):266. [doi: 10.3390/pharmaceutics13020266] [Medline: 33669162]
17. Hshieh TT, DuMontier C, Jaung T, et al. Association of Polypharmacy and potentially inappropriate medications with frailty among older adults with blood cancers. J Natl Compr Canc Netw 2022 Aug;20(8):915-923. [doi: 10.6004/jnccn.2022.7033] [Medline: 35948031]
18. Bhagavathula AS, Tesfaye W, Vidyasagar K, Fialova D. Polypharmacy and hyperpolypharmacy in older individuals with Parkinson’s disease: a systematic review and meta-analysis. Gerontology 2022;68(10):1081-1090. [doi: 10.1159/000521214] [Medline: 35026767]
19. Hines LE, Murphy JE. Potentially harmful drug-drug interactions in the elderly: a review. Am J Geriatr Pharmacother 2011 Dec;9(6):364-377. [doi: 10.1016/j.amjopharm.2011.10.004] [Medline: 22078863]
20. Laroche ML, Bouthier F, Merle L, Charmes JP. Potentially inappropriate medications in the elderly: a list adapted to French medical practice. Rev Med Interne 2009 Jul;30(7):592-601. [doi: 10.1016/j.revmed.2008.08.010] [Medline: 18814939]
21. Sugidono M, Lo M, Young R, et al. Impact of polypharmacy prior to allogeneic hematopoietic stem cell transplantation in older adults. Transplant Cell Ther 2021 Apr;27(4):344. [doi: 10.1016/j.jtct.2021.01.001] [Medline: 33836888]
22. Bhargava D, Arora M, DeFor TE, et al. Use of potentially inappropriate medications in older allogeneic hematopoietic cell transplantation recipients. Biol Blood Marrow Transplant 2020 Dec;26(12):2329-2334. [doi: 10.1016/j.bbmt.2020.08.031] [Medline: 32919078]

XSL•FO
RenderX

23. Visintini C, Mansutti I, Palese A. Medication adherence among allogeneic haematopoietic stem cell transplant recipients: a systematic review. Cancers (Basel) 2023 Apr 25;15(9):2452. [doi: 10.3390/cancers15092452] [Medline: 37173924]
24. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018 Sep;18(3):91-93. [doi: 10.1016/j.tjem.2018.08.001] [Medline: 30191186]
25. Majhail NS, Rizzo JD, Lee SJ, et al. Recommended screening and preventive practices for long-term survivors after hematopoietic cell transplantation. Bone Marrow Transplant 2012 Mar;47(3):337-341. [doi: 10.1038/bmt.2012.5] [Medline: 22395764]
26. Rotz SJ, Bhatt NS, Hamilton BK, et al. International recommendations for screening and preventative practices for long-term survivors of transplantation and cellular therapy: a 2023 update. Transplant Cell Ther 2024 Apr;30(4):349-385. [doi: 10.1016/j.jtct.2023.12.001] [Medline: 38413247]
27. Hammami MB, Talkin R, Al-Taee AM, Schoen MW, Goyal SD, Lai JP. Autologous graft-versus-host disease of the gastrointestinal tract in patients with multiple myeloma and hematopoietic stem cell transplantation. Gastroenterology Res 2018 Feb;11(1):52-57. [doi: 10.14740/gr925w] [Medline: 29511407]
28. Tomblyn M, Chiller T, Einsele H, et al. Guidelines for preventing infectious complications among hematopoietic cell transplantation recipients: a global perspective. Biol Blood Marrow Transplant 2009 Oct;15(10):1143-1238. [doi: 10.1016/j.bbmt.2009.06.019]
29. Lin RJ, Ma H, Guo R, Troxel AB, Diefenbach CS. Potentially inappropriate medication use in elderly non-Hodgkin lymphoma patients is associated with reduced survival and increased toxicities. Br J Haematol 2018 Jan;180(2):267-270. [doi: 10.1111/bjh.15027] [Medline: 29143301]
30. Elliot K, Tooze JA, Geller R, et al. The prognostic importance of polypharmacy in older adults treated for acute myelogenous leukemia (AML). Leuk Res 2014 Oct;38(10):1184-1190. [doi: 10.1016/j.leukres.2014.06.018] [Medline: 25127690]
31. Malnoe D, Lamande T, Jouvance-Le Bail A, Marchand T, Le Corre P. Therapeutic pathways of allogeneic and autologous hematopoietic stem cell transplantation recipients: a hospital pharmacist's perspective. Ars Pharm 2024;65(3):240-257. [doi: 10.30827/ars.v65i3.30246]
32. Alshammari H, Al-Saeed E, Ahmed Z, Aslanpour Z. Reviewing potentially inappropriate medication in hospitalized patients over 65 using explicit criteria: a systematic literature review. Drug Healthc Patient Saf 2021;13:183-210. [doi: 10.2147/DHPS.S303101] [Medline: 34764701]
33. Smith MT, de la Rosa R, Daniels SI. Using exposomics to assess cumulative risks and promote health. Environ Mol Mutagen 2015 Dec;56(9):715-723. [doi: 10.1002/em.21985] [Medline: 26475350]
34. Wu Q, Bagdad Y, Taboureau O, Audouze K. Capturing a comprehensive picture of biological events from adverse outcome pathways in the drug exposome. Front Public Health 2021;9:763962. [doi: 10.3389/fpubh.2021.763962] [Medline: 34976924]
35. Kelly TL, Salter A, Pratt NL. The weighted cumulative exposure method and its application to pharmacoepidemiology: a narrative review. Pharmacoepidemiol Drug Saf 2024 Jan;33(1):e5701. [doi: 10.1002/pds.5701] [Medline: 37749615]
36. Wabe N, Urwin R, Seaman K, Timothy A, Raban MZ, Westbrook J. Analysis of longitudinal patterns and predictors of medicine use in residential aged care using group-based trajectory modelling: the MEDTRAC-polypharmacy longitudinal cohort study. Br J Clin Pharmacol 2024 Dec;90(12):3308-3319. [doi: 10.1111/bcp.16220] [Medline: 39183449]
37. Baysari MT, Duong MH, Hooper P, et al. Supporting deprescribing in hospitalised patients: formative usability testing of a computerised decision support tool. BMC Med Inform Decis Mak 2021 Apr 5;21(1):116. [doi: 10.1186/s12911-021-01484-z] [Medline: 33820536]
38. Mouazer A, Tsopra R, Sedki K, Letord C, Lamy JB. Decision-support systems for managing polypharmacy in the elderly: a scoping review. J Biomed Inform 2022 Jun;130:104074. [doi: 10.1016/j.jbi.2022.104074] [Medline: 35470079]
39. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
40. Bakker T, Klopotowska JE, Dongelmans DA, et al. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial. Lancet 2024 Feb 3;403(10425):439-449. [doi: 10.1016/S0140-6736(23)02465-0] [Medline: 38262430]

Abbreviations

ADR: adverse drug reaction
ATC: Anatomical Therapeutic Chemical
CDE: cumulative drug exposure
CDED: cumulative drug exposure density
CDSS: clinical decision support system
CDW: clinical data warehouse
DDI: drug-drug interaction
EHR: electronic health record
FAMD: factorial analysis of mixed data

XSL•FO
RenderX

HPP: hyperpolypharmacy
HSCT: hematopoietic stem cell transplant
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
LHS: learning health system
PIM: potentially inappropriate medication
PP: polypharmacy

Edited by A Benis; submitted 05.May.2025; peer-reviewed by A Patel, M Cesare; accepted 30.Nov.2025; published 16.Apr.2026.

Please cite as:
Bories M, Bouzillé G, Cuggia M, Le Corre P, Bannay A
Methodological Development and Assessing Prescribing Determinants Through Cumulative Drug Exposure in Hospitalized Patients: Proof-of-Concept Retrospective Study
JMIR Med Inform 2026;14:e76961
URL: https://medinform.jmir.org/2026/1/e76961
doi:10.2196/76961

© Mathilde Bories, Guillaume Bouzillé, Marc Cuggia, Pascal Le Corre, Aurélie Bannay. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research on Risk Transfer Pathways for Lung Cancer Among Middle-Aged and Older Individuals Using Deep Reinforcement Learning: Retrospective Cohort Study

Songjing Chen, PhD; Sizhu Wu, PhD

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 3 Yabao Road, Chaoyang District, Beijing, China

Corresponding Author:
Sizhu Wu, PhD
Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 3 Yabao Road, Chaoyang District, Beijing, China

Abstract

Background: Nowadays, lung cancer has a significantly high incidence rate worldwide. The mortality rate of lung cancer continues to rise; it is more common in middle-aged and older individuals and poses a great threat to human health.

Objective: This study aimed to assess the lung cancer risk among middle-aged and older individuals in a timely manner and to establish an efficient pathway for the risk transfer.

Methods: We proposed a deep reinforcement learning model based on deep Q-network (DQN) to explore the risk transfer pathway for lung cancer among middle-aged and older individuals. Risk stratification of lung cancer occurrence was deduced through deep neural network. The DQN model was developed using the Health and Retirement Study cohort for model training and internal validation. We also used the China Health and Retirement Longitudinal Study cohort for model external validation. Transfer simulation of multiple pathways in different cycles was calculated in a stratified risk groups–leveraged DQN model.

Results: We developed and evaluated the DQN method for optimizing the risk transfer pathway among middle-aged and older individuals, with accuracy ranging from 0.917 (95% CI 0.896 - 0.928) to 0.949 (95% CI 0.909 - 0.961) and area under curve ranging from 0.906 (95% CI 0.887 - 0.933) to 0.927 (95% CI 0.893 - 0.938). External validation was conducted to assess the model's effectiveness and availability. A total of 8780 and 3763 samples from the Health and Retirement Study were used for model training and internal testing, respectively, and 16,442 samples from the China Health and Retirement Longitudinal Study were used for external validation. The results showed that DQN models illuminated the optimal risk transfer pathways for stratified risk groups. Lung cancer incidence in the high risk group had declined by 68.2% through risk transfer of model-based simulation, which had declined by 56.9% in the medium risk group. Through simulative intervention and risk transition deduced from the DQN model, lung cancer incidences of the high risk, medium risk, and low risk groups were obviously decreased.

Conclusions: A DQN-based deep reinforcement learning model was proposed and validated to develop and simulate a risk transfer pathway of lung cancer among middle-aged and older individuals. Risk stratification supplied an effective foundation for lung cancer risk transition.

(JMIR Med Inform 2026;14:e74990) doi:10.2196/74990

KEYWORDS

lung cancer; risk transfer pathway; deep reinforcement learning; deep Q-learning; middle-aged and older individuals

Introduction

Nowadays, lung cancer has a significantly high incidence rate worldwide. The mortality rate of lung cancer continues to rise; it is more common in middle-aged and older individuals and poses a great threat to human health. Moreover, 50% of patients with nonsmall cell lung cancer are aged 65 years and older at the time of initial diagnosis [1,2]. The pathogenesis of lung cancer is insidious, which is prone to misdiagnosis with nonspecific clinical manifestations. Therefore, it is important that we timely assess the risk for middle-aged and older individuals and establish an efficient pathway of the risk transfer for lung cancer.

Actively carrying out the risk assessment of lung cancer will help reduce the incidence rate of lung cancer. In the United Kingdom, a research team from the University of Oxford used the Cox proportional-hazards method in a population-based cohort to develop a lung cancer risk assessment model [3]. Henriksen et al [4] adopted a Bayesian network for lung cancer risk prediction based on a Danish population according to risk factors, such as age, sex, smoking, and laboratory results, and so on. Most previous studies conducted lung cancer risk

XSL•FO
RenderX

assessment through traditional statistical methods or other machine learning methods [5-7] and lacked detailed stratification of lung cancer risk. A deep learning–based lung cancer risk assessment model with high performance was developed in this research, which provided a sufficiently detailed division for the risk process of lung cancer occurrence. Risk transfer in this research referred to lung cancer risk changes among high risk, medium risk, low risk, and non-risk states. High risk, medium risk, and low risk reflected that the risk of lung cancer decreased from high to low, and non-risk meant there was no risk of developing lung cancer. Deep Q-network (DQN) of deep reinforcement learning was adopted to explore risk transfer pathway to avoid lung cancer occurrence, which could effectively help reduce the incidence rate of lung cancer.

Research on the risk transfer pathways of lung cancer mainly focused on risk transitions among different risk states before lung cancer occurrence, as well as risk transitions between these risk states and lung cancer occurrence. However, several studies [8,9] were focused on the risk transition between high risk and lung cancer incidence. Huang et al [10] used lung cancer–associated risk factors, temporal trends, global incidence, and the mortality of lung cancer based on multiple open-access datasets to conduct research on high risk detection and analyze the trend of lung cancer incidence. Kowada [11] developed the state transition model to assess the cost-effectiveness of lung cancer screening in never smokers of the United States and Japan. Wang and Lukito [12] proposed the semi-Markov model to perform lung cancer risk and economic evaluation, which provided a health care decision support tool for the survivability and expenditure of lung cancer. Related research rarely explored the optimal risk transfer pathway through refined lung cancer risk to prevent disease occurrence [13-15]. Therefore, research on the risk transfer pathway for lung cancer among middle-aged and older individuals was conducive to reducing lung cancer incidence and helping early diagnosis and treatment of lung cancer.

The purpose of this research was to assess the lung cancer risk and develop a risk transfer pathway using deep reinforcement learning for middle-aged and older individuals. External validation was simulatively conducted to evaluate the DQN-based model through the large longitudinal cohort. We also explored risk transfer pathway evaluation to demonstrate effectiveness and availability of the proposed approach.

Methods

Data Sources

We used retrospective datasets from the Health and Retirement Study (HRS) [16] of the United States for model training and internal testing, which were selected between the year of 1992 and 2020 and followed up every 2 years, composing a population of more than 20,000. The cohort of the China Health and Retirement Longitudinal Study (CHARLS) [17] was adopted for model external validation, which involved a population of over 17,000 aged 45 years and older from 2010 to 2020. Lung cancer incidence, lung cancer survival, and overall survival were involved in this research. The inclusion criteria were aged 50 years and older and having respiratory disease–related questionnaire responses in follow-up surveys. Those without complete follow-up data were excluded. Risk factors such as person number, birth year, gender, marital status, BMI, years of education, smoking status, physical activity, drinking status, preventive behaviors, cancer history, and lung disease condition were involved in risk prediction and transfer modeling. We used R console to extract and map variables between the HRS and CHARLS datasets. For example, in HRS, the variable “PN” was used to represent a person number, while in CHARLS, the variable “ID” was used to denote a person number, as shown in Table 1. For years of education, specific education years were listed in HRS, while education types were supplied in CHARLS. We matched “no formal education illiterate” in CHARLS and “0 year of education” in HRS, and other education types were also matched similarly. Smoking cessation duration was calculated through “respondent smoking ever” and “respondent smoking now”; for example, if a respondent reported smoking ever in survey wave 1 and reported not smoking now in survey wave 2, we then deduced the respondent’s smoking cessation duration. Preventive behavior for cholesterol was involved in HRS and missed in CHARLS, which had no significant correlation with this research. Therefore, we did not include such variable.

XSL•FO
RenderX

Table . Variable comparison between HRS[a] and CHARLS[b].

Variables	Code in HRS	Code in CHARLS
Person number	PN	ID
Respondent's reported birth year	RABYEAR	RABYEAR
Respondent's reported gender	RAGENDER	RAGENDER
BMI in survey wave 1	R1BMI	R1MBMI
Marital status in survey wave 1	R1MSTAT	R1MSTAT
Years of education	RAEDYRS	RAEDUC_C
Respondent smoking ever in survey wave 1	R1SMOKEV	R1SMOKEV
Respondent smoking now in survey wave 2	R2SMOKEN	R2SMOKEN
Physical activity in survey wave 1	R1VGACTF	R1VGACT_C
Preventive behavior for cholesterol in survey wave 1	R1CHOLST	__[c]

[a]HRS: Health and Retirement Study.

[b]CHARLS: China Health and Retirement Longitudinal Study.

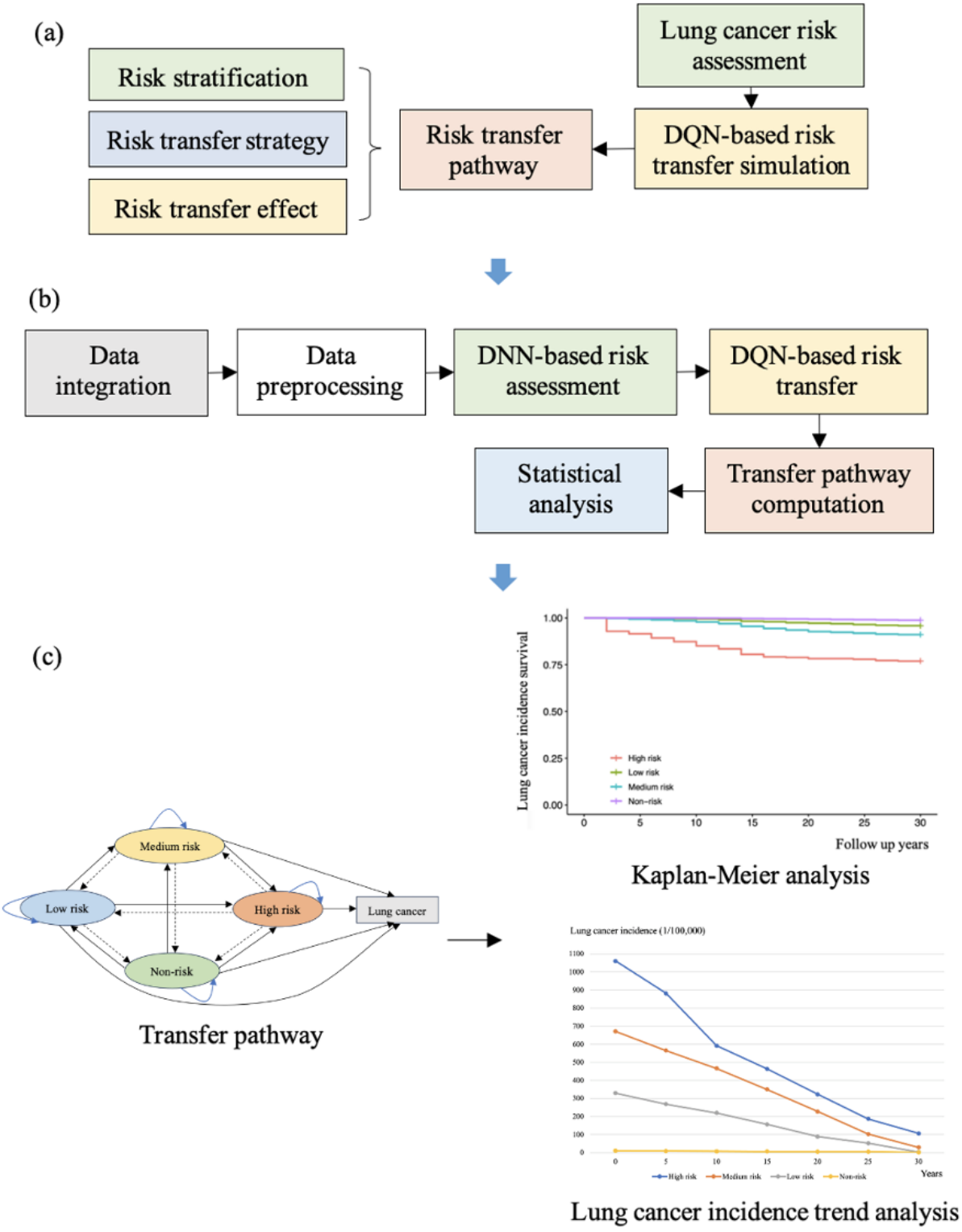
[c]Not applicable: the variable of "R1CHOLST" was involved in HRS, which was not applicable in CHARLS.

Experimental Design

Lung cancer risk stratification was conducted based on the deep neural network (DNN) model and the risk transfer pathway was simulated using the DQN model. Lung cancer incidence, lung cancer survival, and overall survival were calculated to evaluate lung cancer risk stratification. Lung cancer incidence survival referred to the time from the initial follow-up to lung cancer diagnosis. Lung cancer survival indicated the interval between the initial follow-up and death from lung cancer or the last follow-up. Overall survival represented the time from the initial follow-up to death from any cause or the last follow-up. Four stratified risk degrees of lung cancer were generated using the weight thresholds of the DNN model, which included high risk, medium risk, low risk, and non-risk. Further, 3 risk transfer directions were deduced and analyzed through the DQN model, including direction of risk improvement, direction of risk deterioration, and risk maintenance (maintaining the original risk state). Risk improvement represented the risk decrease of lung cancer occurrence; risk deterioration reflected the risk increase of lung cancer occurrence; and risk maintenance indicated that the risk of lung cancer remained unchanged. The accuracy and the area under the receiver operating characteristic curve (AUROC) were computed to evaluate the performance of the DQN model. Risk transfer pathway for lung cancer was developed based on DQN models, as shown in Figure 1.

XSL•FO
RenderX

Figure 1. The main research technology route: (a) the outline of this research; (b) schematic diagram of the deep q-network model's construction; and (c) survival analysis and lung cancer incidence trend analysis based on the risk transfer pathway. DNN: deep neural network; DQN: deep Q-network.



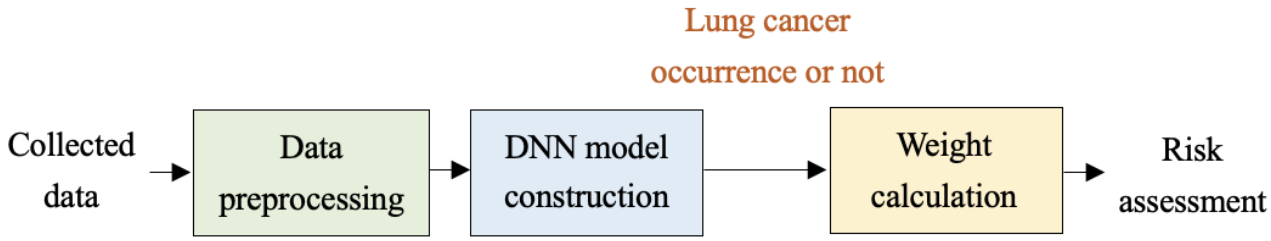
Lung Cancer Risk Assessment

The risk assessment model of lung cancer occurrence was constructed based on the DNN method as shown in Figure 2. Because of the long follow-up time in HRS and CHARLS, it was inevitable that some indicators might have missing values. We adopted the multiple imputation method [18] to fill in the missing values. We used the mean value to perform initial imputation of missing values and used linear regression for iterative imputation. Then we updated the imputation value until realizing model convergence and reaching the maximum iteration. Therefore, missing values were predicted through multiple iterations. We also used singularly valuable decomposition [19] to reduce the white noises. We matched follow-up data according to each unique person number. Data imbalance problems were mitigated through the synthetic

XSL•FO
RenderX

minority oversampling technique [20]. Datasets of HRS were randomly divided into the training set and the internal test set for the DNN-based risk assessment model, which included 8780 samples in the training set and 3763 samples in the internal test set. Further, 16,442 samples from CHARLS were adopted for external testing. The synthetic minority oversampling technique was used with the HRS training dataset to construct models. The k-nearest neighbor algorithm was applied to calculate the minority sample. The simulated samples were added to the HRS training dataset.

Figure 2. Schematic diagram of the lung cancer risk assessment model. DNN: deep neural network.



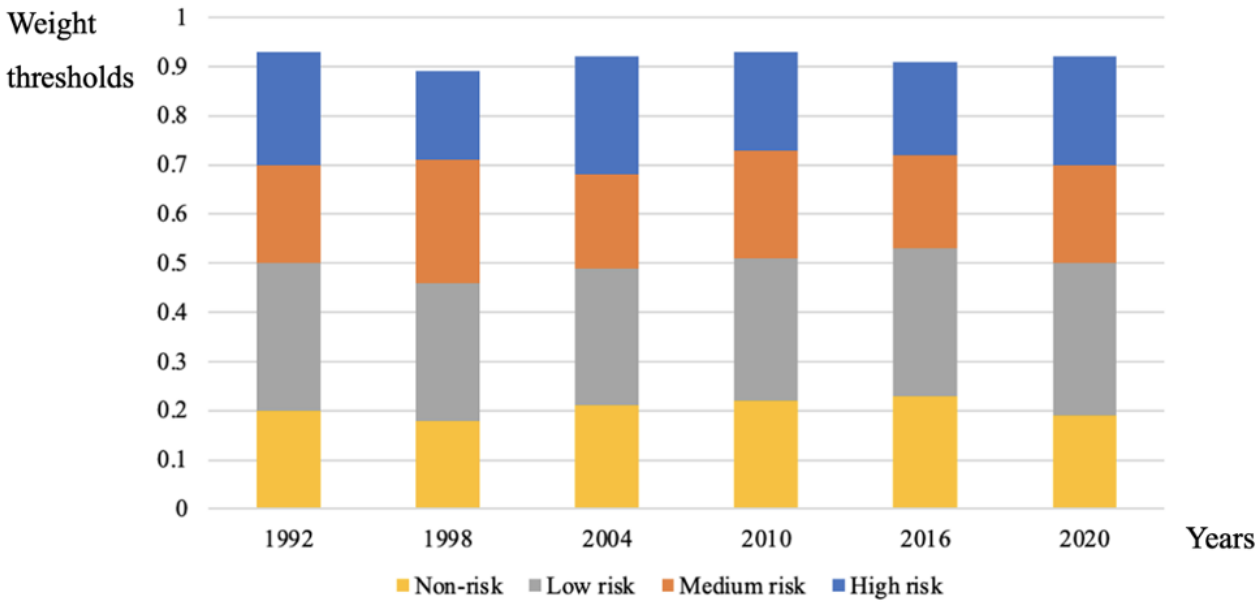
The preprocessed data were input to the DNN model's input layer. We iteratively calculate the weight matrix of risk factors using Equation (1). Specifically, W_k and b_k were the weight matrix and the bias vector of the *k*th layer of the DNN model. We used Equation (1) to compute the output of the *k*th layer. The linear rectification function, as shown in Equation (2), was used as the activation function to improve representation capability of the DNN model. Through multiple iterations and calculations, we computed weight thresholds using Equation (3), which are shown in Figure 3. When the average weight of the sample was in the interval 0 - 0.2 in the year of 1992, we defined it as non-risk. When the average weight was in the interval 0.2 - 0.5, 0.5 - 0.71, or 0.71 - 0.93 in the year of 1992, we defined it as low risk, medium risk, or high risk, respectively. Similarly, risk stratifications of other follow-up years were obtained according to weight thresholds as shown in Figure 3. The DNN-based lung cancer risk assessment model's outputs were risk stratification of lung cancer occurrence, including high risk, medium risk, low risk, and non-risk. The calibration plots of the DNN model with Brier score and log loss are shown in Multimedia Appendix 1, and the confusion matrix of the DNN model is listed in Multimedia Appendix 2. The random seed of the DNN model was 42. Person number, birth year, gender, marital status, BMI, years of education, smoking status, physical activity, drinking status, prevention behaviors, cancer history, and lung disease condition, and so on, were involved features. The initial learning rate was 0.01, the batch size was 32, and the initial episode was 30. Through multiple iterations, 3 layers and 22 nodes were derived from the hidden layer, 56 nodes of the input layer, and 1 node of the output layer. The activation function was the linear rectification function. The *deepnet* package and h2o framework were leveraged for DNN model construction.

$$(1) lk = Wklk-1 + bk$$

$$(2) fT(x) = \{0, x \leq 0x, x > 0$$

$$(3) y = f(x,\theta) = fL-1(WL-1fL-2(\ f1(W1X+b1)) + bL-1)$$

Figure 3. Weight thresholds of risk assessment in follow-up years.



XSL•FO
RenderX

DQN Model Development and Validation

DQN models were constructed through different risk groups, and internal and external testing were conducted for model validation. Datasets including 8780 samples from HRS were used for DQN model training. A total of 3763 samples from HRS were used as an internal test set and 16,442 samples from CHARLS were used as an external test set. DQN models were built through 4 risk groups. First, Q-learning and convolutional neural networks (CNNs) were combined to construct the DQN model. The stratified risk group was the model's input. We used CNN to approximate the target value function (target Q function) of the transfer strategy as closely as possible. Target Q functions for different risk groups were obtained. Second, the DQN model was trained using Q-learning. We calculated the loss function through Equation (4). *a* was the transfer strategy corresponding to different cycles. *s* was the different risk states including high risk, medium risk, low risk, and non-risk. *Q* was the output value function, which represented the maximum cumulative transfer effect of transfer strategy *a*. $Q(s,a;\theta_i)$ was the output of the current network. Q_i was the output of the target network. Third, we adopted the stochastic gradient descent function to optimize the loss function. The target Q_i function of the transfer strategy was obtained using Equation (5), and the transfer strategies of different risk states were derived accordingly.

The state space of the DQN setup was the vector (high risk state, medium risk state, low risk state, and non-risk state), and the action space was the set of transfer strategies, such as extending the duration of smoking cessation, reducing the frequency of smoking, and so on. We constructed DQN models for the 4 risk groups separately. As shown in Figure 4, we took the low risk group as an example to illustrate state representation and transition processes. We simulated risk transition of the low risk group in multiple cycles through corresponding transfer strategies. Similarly, DQN models for the high risk group, medium risk group, and non-risk group were developed to explore risk transfer pathways of lung cancer. The reward function is described in Equation (6), which was the transfer effect of current network. We used the stochastic gradient descent function to optimize the loss function. $Q(s', a'; \theta_{i-1})$ was the output of the target network, $Q(s,a;\theta_i)$ was the output of the current network, *r* was the transfer effect of the current network, ε was the transfer environment, and γ was the discount factor between 0.9 and 1. The random seed of the DQN model was 50. Person number, birth year, gender, marital status, BMI, years of education, smoking status, physical activity, drinking status, prevention behaviors, cancer history, and lung disease condition, and so on, were involved features. The initial learning rate was 0.01, the batch size was 256, and the episode number was 500. One input layer with 32×32 neurons; 3 convolutional layers with 5×5, 4×4, 3×3 convolution kernels; and 1 output layer were leveraged for the DQN model. OpenAI Gym and PyTorch framework were used.

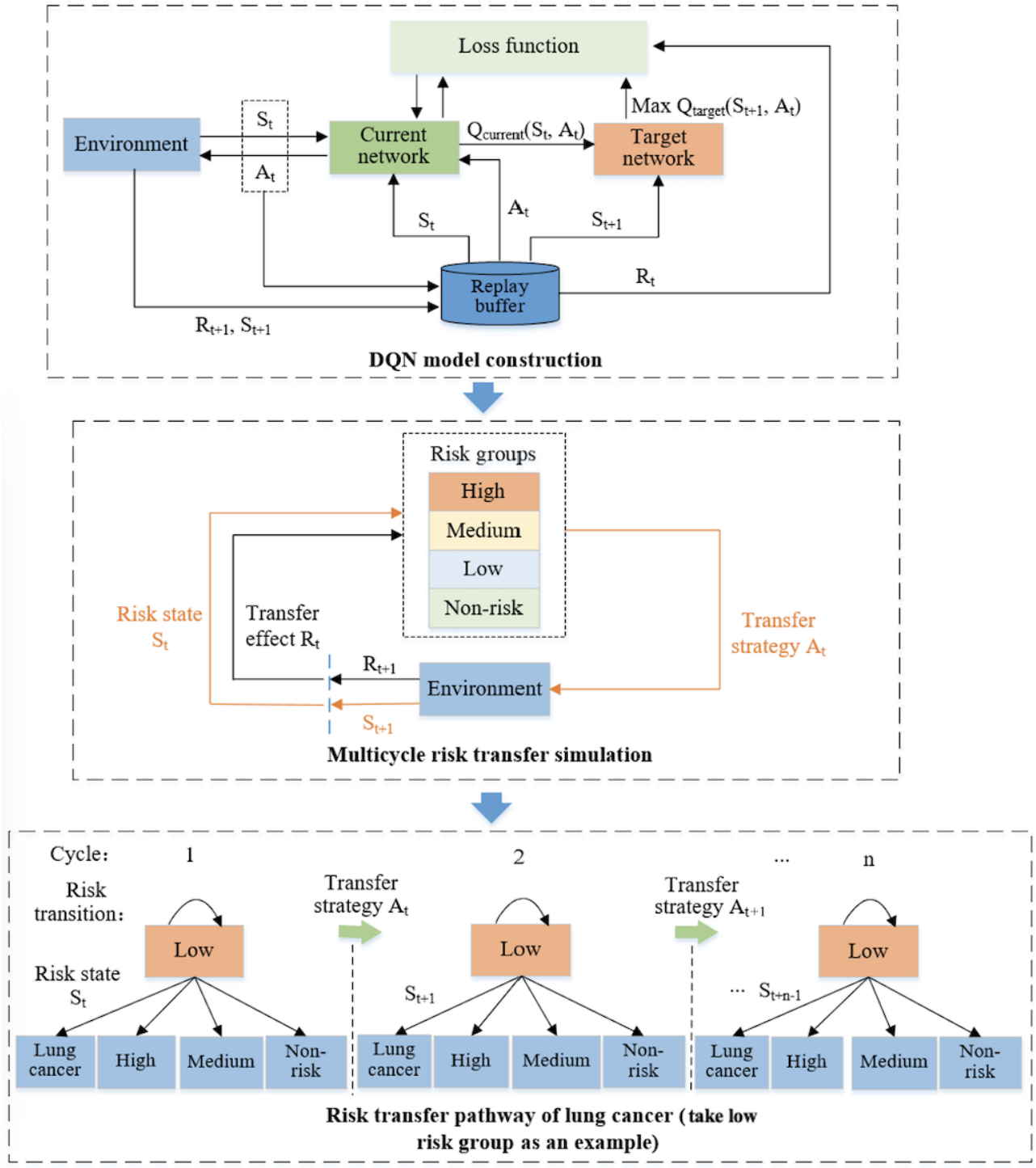
$$(4)\ Li(\theta i)=Es,a\sim\rho(\cdot)[(Qi-Q(s,a;\theta i))2]$$

$$(5)\ Qi=Es'\sim\varepsilon[r+\gamma maxa'Q(s',a';\theta i-1)|s,a]$$

$$(6)\ \nabla\theta iLi(\theta i)=Es,a\sim\rho(\cdot),s'\sim\varepsilon[(r+\gamma maxa'Q(s',a';\theta i-1)-Q(s,a;\theta i))\nabla\theta iQ(s,a;\theta i)]$$

XSL•FO
RenderX

Figure 4. The workflow of deep Q-network model development and validation processes. DQN: deep Q-network.



Risk Transfer Pathways Computation

Transfer simulation of multiple pathways in different cycles for lung cancer occurrence was conducted in stratified risk groups through the DQN model. To explore optimal transfer strategy and transfer effect of different pathways, we calculated and deduced the optimum transfer pathway. The probabilities of lung cancer risk state transition for different risk groups were calculated for each transition cycle. Transfer pathways corresponding to various transfer directions, such as risk improvement, risk deterioration, and risk maintenance, were deduced. Then, the optimal transfer pathway and transfer effect were obtained through the DQN model.

Statistical Analysis

Kaplan-Meier analysis was leveraged to demonstrate survival differences among the 4 stratified risk groups. We used a Cox proportional hazards model for the relative risk analysis between lung cancer incidence and smoking. R version 4.3.3 (R Core Team) was used for statistical analysis. The *deepnet* package and the h2o framework were used for DNN model training. OpenAI Gym and PyTorch framework were used for the DQN

XSL•FO
RenderX

model. The *survival* and *survminer* packages were adopted for Kaplan-Meier analysis.

Ethical Considerations

This study used secondary anonymized data from HRS and CHARLS, which were approved by the Institutional Review Board Committees of Michigan State University (HUM00061128) and the Biomedical Ethics Review Committee of Peking University (IRB00001052–11015). In this study, the individual data were anonymized; patients and the public were not involved in the design, conduct, and reporting stages of the research. All participants provided informed consent, and the study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

Results

Baseline Characteristics of Samples

Baseline characteristics of the training set, internal test set, and external test set are listed in Table 2. A total of 12,543 samples from HRS were used for DQN model development, and 16,442 samples from CHARLS were used for external testing of the DQN model.

XSL•FO
RenderX

Table . Baseline characteristics of the HRS[a] training set, internal test set, and external test set.

Variables	HRS training set (n=8780)	HRS test set (n=3763)	CHARLS[b] test set (n=16,442)
Age group (years), n (%)			
≤45	369 (4.20)	158 (4.19)	1090 (6.63)
45 - 65	8141 (92.72)	3489 (92.72)	11,431 (69.52)
>65	270 (3.08)	115 (3.06)	3766 (22.9)
Missing	1 (0.01)	—[c]	155 (0.94)
Sex, n (%)			
Male	4069 (46.34)	1744 (46.35)	7884 (47.95)
Female	4711 (53.66)	2019 (53.65)	8556 (52.04)
Missing	—	—	2 (0.01)
Ethnicity, n (%)			
Non-Hispanic White	6976 (82.12)	2990 (83.46)	—[d]
Black	1456 (14.37)	624 (13.27)	—
Other	348 (3.51)	149 (3.27)	—
Region, n (%)			
Northeast	1497 (17.05)	641 (17.03)	—
Midwest	2194 (24.99)	940 (24.98)	—
South	3662 (41.71)	1570 (41.72)	—
West	1427 (16.25)	612 (16.27)	—
Missing	—	—	—
Habitation, n (%)			
Urban	—	—	7171 (43.61)
Rural	—	—	9271 (56.39)
Marital status, n (%)			
Married	6815 (77.62)	2921 (77.62)	12,966 (78.86)
Married, spouse absent	40 (0.46)	17 (0.45)	—
Partnered	264 (3.01)	113 (3)	1243 (7.56)
Separated	217 (2.47)	93 (2.47)	71 (0.43)
Divorced	756 (8.61)	324 (8.61)	143 (0.87)
Widows	435 (4.95)	186 (4.94)	1856 (11.29)
Never married	253 (2.88)	109 (2.90)	153 (0.93)
Missing	—	—	10 (0.06)
Years of education, n (%)			
0 - 3	169 (1.92)	73 (1.94)	6578 (40)
4 - 6	385 (4.38)	165 (4.38)	4657 (28.32)
7 - 9	894 (10.18)	383 (10.18)	3119 (18.97)
10 - 12	3637 (41.42)	1559 (41.43)	1266 (7.7)
13 - 16	2848 (32.44)	1220 (32.42)	794 (4.83)
≥17	820 (9.34)	351 (9.33)	12 (0.07)
Missing	28 (0.32)	12 (0.32)	16 (0.10)

[a]HRS: Health and Retirement Study.

XSL•FO
RenderX

[b]CHARLS: China Health and Retirement Longitudinal Study.

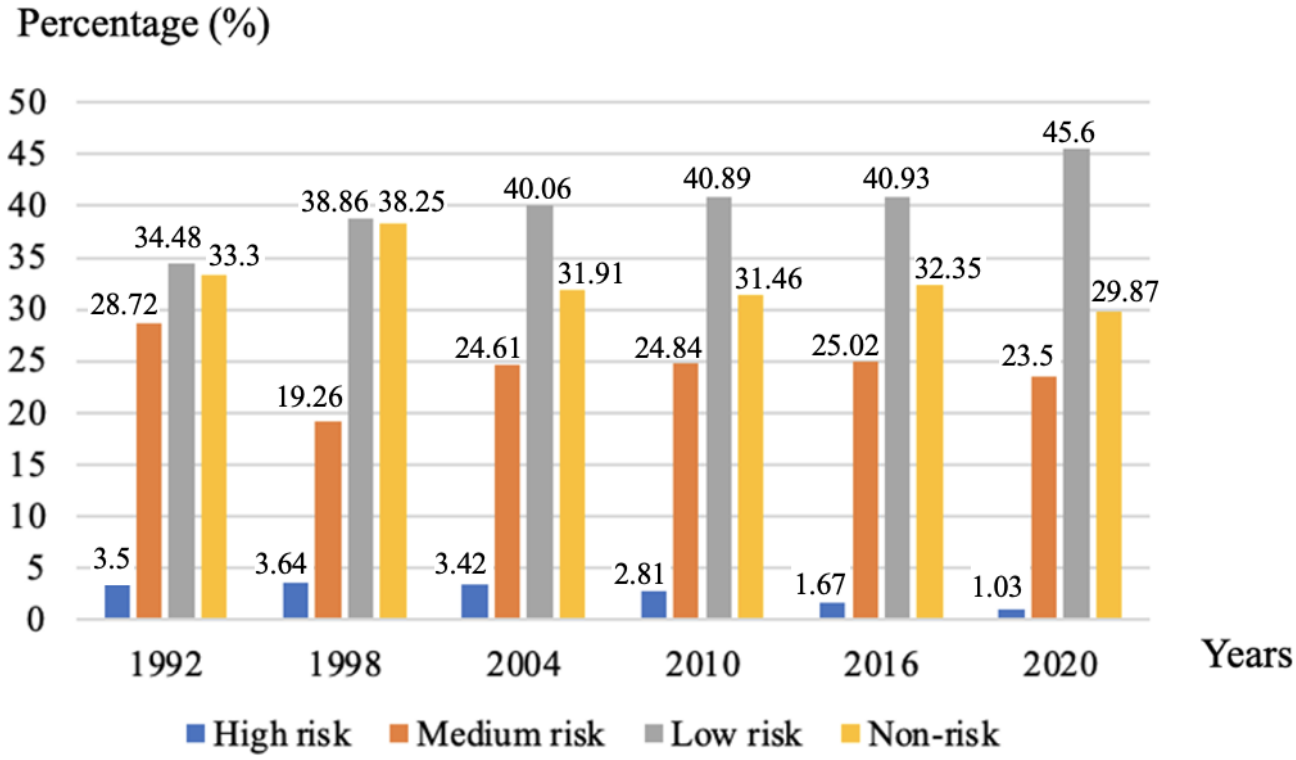
[c]Not available.

[d]Not applicable: the "ethnicity" for HRS in Table 2 consisted of "White," "Black," and "other," which were not applicable for CHARLS.

Lung Cancer Risk Stratification

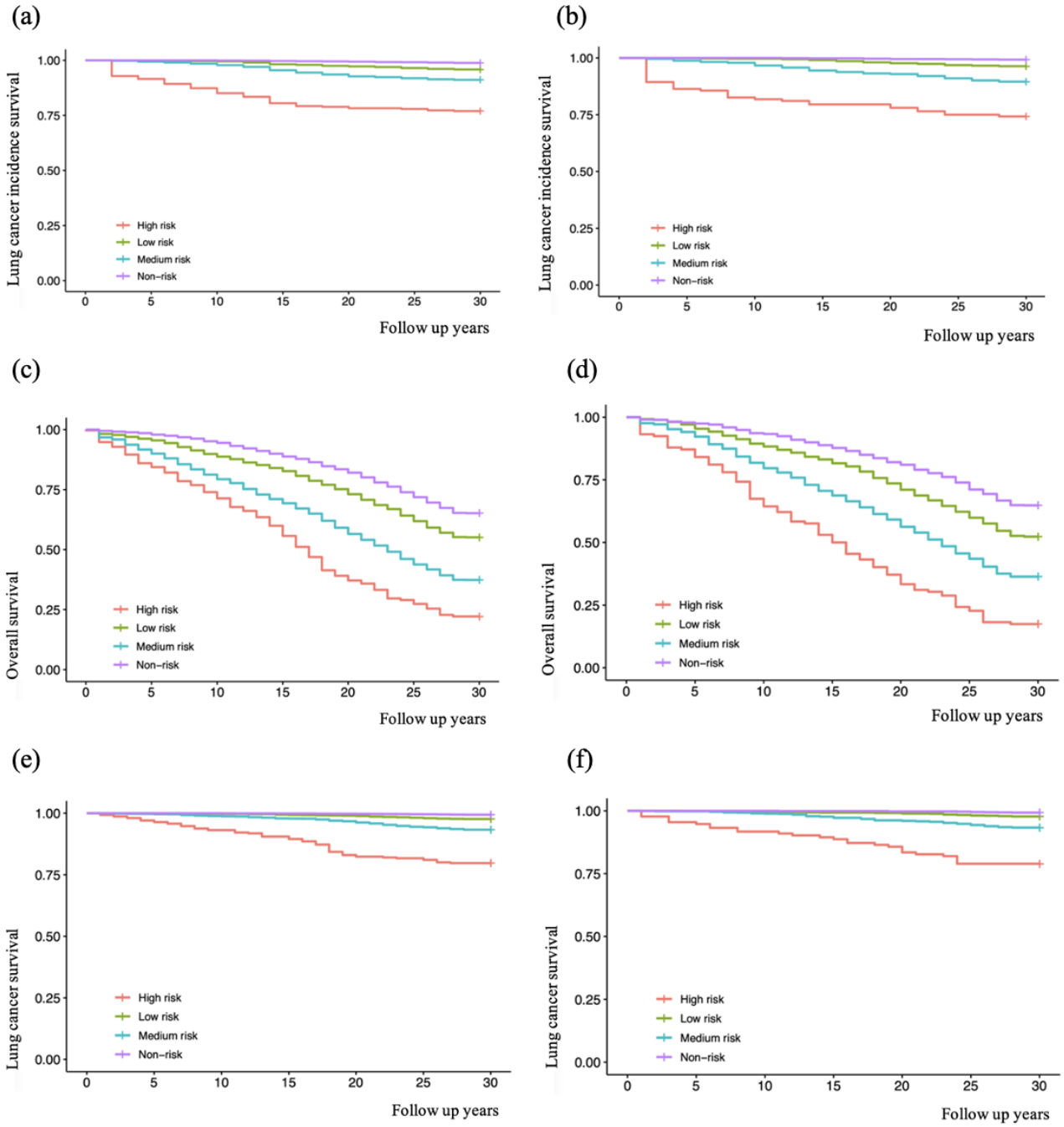
The risk stratification and the trend of risk transfer of lung cancer incidence in different follow-up years are shown in Figure 5, which displays the percentages of the high risk, medium risk, low risk, and non-risk groups. People in the high risk group had higher lung cancer incidence than the medium risk, low risk, and non-risk groups, as shown in Figure 6A and B. The horizontal axis represents the follow-up time, and the value 1.00 on the vertical axis indicates the absence of lung cancer. The value 0.75 represents that the incidence survival rate of lung cancer was 75%, and the corresponding incidence rate of lung cancer was 25%. Lung cancer incidence of the medium risk group was also higher than low risk and non-risk groups. At the last follow-up, lung cancer incidence of the high risk group was about 25%, which was much higher than the incidence of the low risk and non-risk groups. As shown in Figure 6C and D, the high risk group had lower overall survival than other risk groups. Moreover, lung cancer survival of the high risk group was also lower than medium risk, low risk, and non-risk groups, as shown in Figure 6E and F.

Figure 5. The percentage of risk transfer for lung cancer incidence in different risk groups.



XSL•FO
RenderX

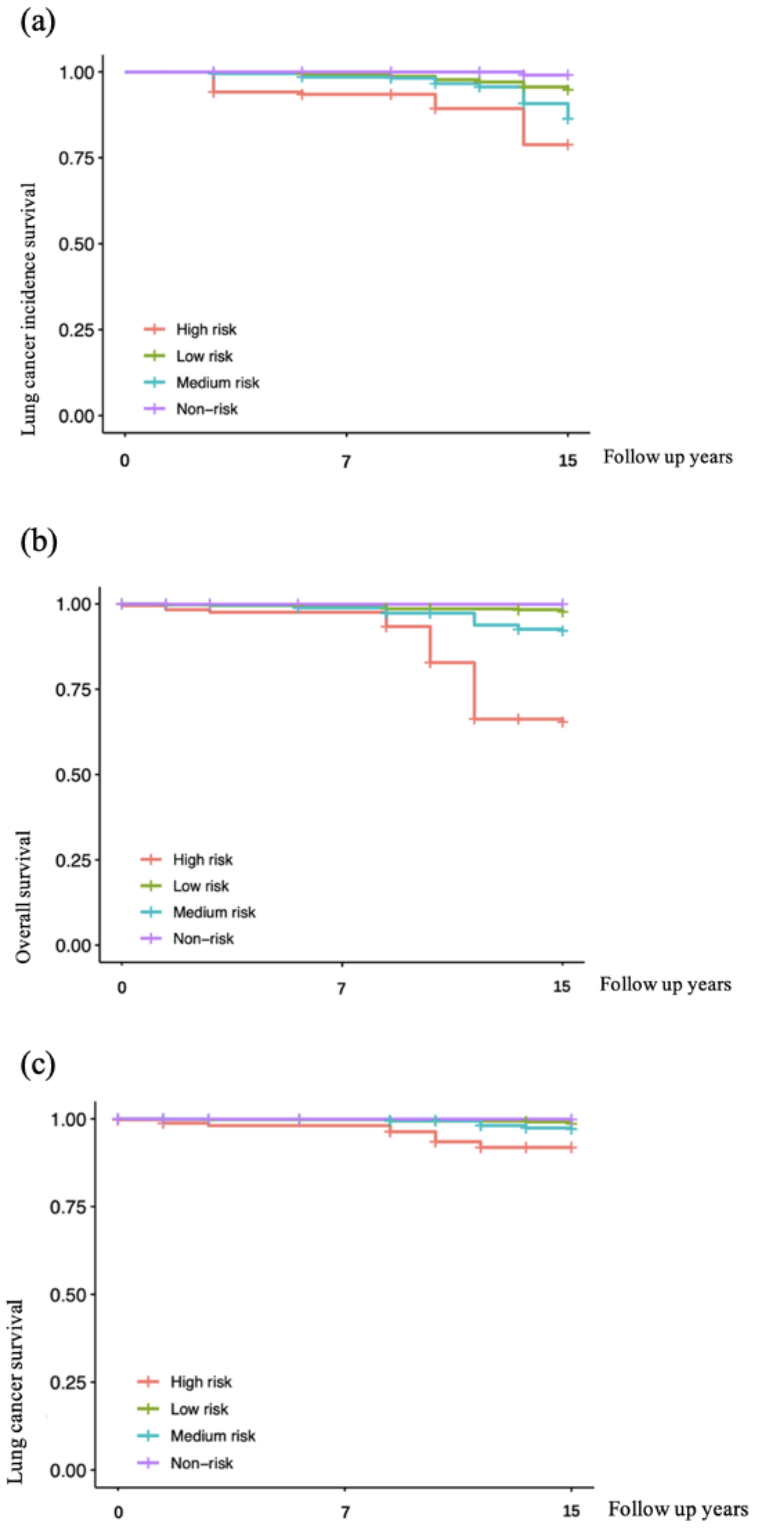
Figure 6. K-M analysis of the high risk, medium risk, low risk, and non-risk subgroups of the HRS cohort: (a) K-M curve of lung cancer incidence survival in HRS training set; (b) K-M curve of lung cancer incidence survival in the HRS test set; (c) K-M curve of overall survival in the HRS training set; (d) K-M curve of overall survival in the HRS test set; (e) K-M curve of lung cancer survival in the HRS training set; and (f) K-M curve of lung cancer survival in the HRS test set. HRS: Health and Retirement Study; K-M: Kaplan-Meier.



Furthermore, external testing of the DNN model was conducted using the CHARLS cohort, as shown in Figure 7. External test results and internal test results of lung cancer risk assessment based on DNN models were mainly consistent.

XSL•FO
RenderX

Figure 7. K-M analysis of the high risk, medium risk, low risk, and non-risk subgroups in the external validation CHARLS cohort: (a) K-M curve of lung cancer incidence survival in the external validation CHARLS cohort; (b) K-M curve of overall survival in the external validation CHARLS cohort; and (c) K-M curve of lung cancer survival in the external validation CHARLS cohort. CHARLS: China Health and Retirement Longitudinal Study; K-M: Kaplan-Meier.



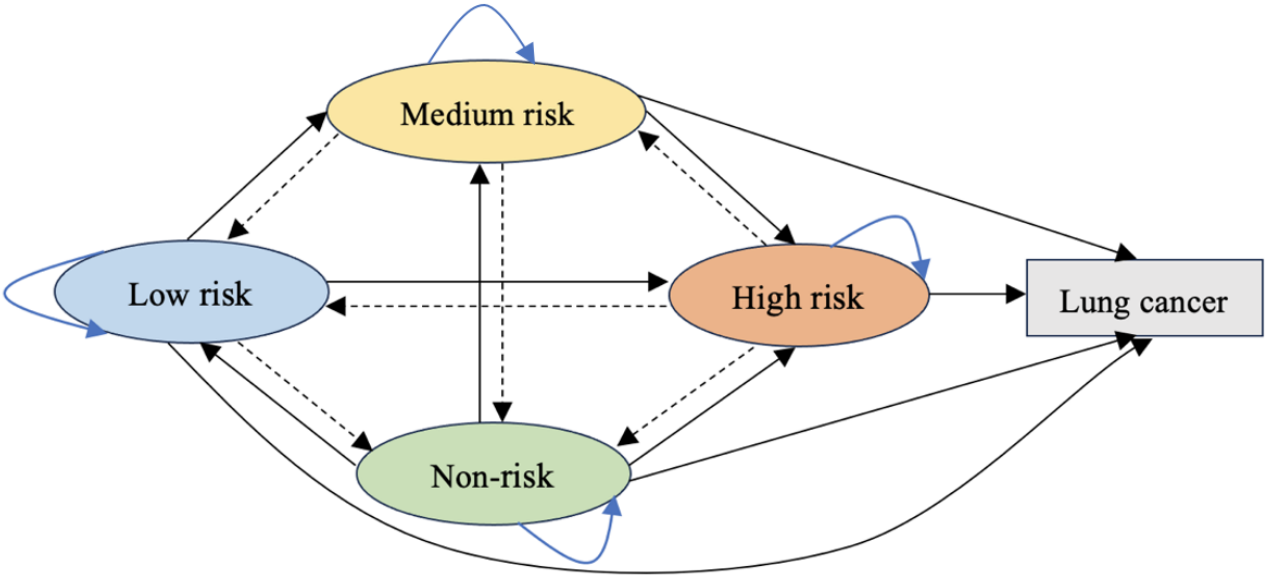
Risk Transfer Strategy and Pathway

Lung cancer risk transfer strategy and pathway were explored through DQN models. We deduced different transfer pathways of lung cancer incidence for stratified risk groups, as shown in Figure 8. Risk improvement represented the reduced risk of lung cancer occurrence, including high risk to medium risk, then to low risk and non-risk; high risk to low risk, then to non-risk; high risk to non-risk; medium risk to low risk, then to non-risk; medium risk to non-risk; and low risk to non-risk. Risk deterioration reflected the increased risk of lung cancer occurrence, including non-risk to low risk, then to medium risk and high risk; non-risk to medium risk, then to high risk; non-risk to high risk; low risk to medium risk, then to high risk;

XSL•FO
RenderX

low risk to high risk; medium risk to high risk; high risk to lung cancer occurrence; medium risk to lung cancer occurrence; low risk to lung cancer occurrence; and non-risk to lung cancer occurrence. Risk maintenance represented that the risk of lung cancer occurrence had not changed, including high risk, medium risk, low risk, and non-risk remaining unchanged separately. In the risk improvement pathways, extending the duration of smoking cessation—such as quitting smoking from less than 1 month to 1 year, quitting smoking from less than 1 month to 3 years, or quitting smoking from less than 1 month to more than 5 years—and reducing the frequency of smoking—such as from everyday smoking to someday smoking, from someday smoking to quitting smoking, or from everyday smoking to quitting smoking—were the effective intervention strategies.

Figure 8. Risk transfer pathways of lung cancer incidence. Black dashed arrows represent risk improvement, black solid arrows represent risk deterioration, and blue solid arrows represent risk maintenance.



The accuracy and AUROC of DQN models are shown in Multimedia Appendix 3. We used the decrease of lung cancer incidence to evaluate the effectiveness of the transfer strategy. The transfer strategy brought the largest decline of lung cancer incidence compared to others, which could be considered an optimal transfer strategy. Otherwise, the transfer strategy should be adjusted through a feedback mechanism. The whole workflow would be reworked as shown in Figure 4; the transfer pathway was comprehensively evaluated until an optimal transfer strategy was derived. For off-policy evaluation, DQN used experience replay data to update the strategy, and existing historical data were reused, as shown in Figure 4. Extending the duration of smoking cessation and reducing the frequency of smoking were the optimal transfer strategies for risk improvement pathway. The tangible intervention strategies were extending the duration of smoking cessation from less than 1 month to more than 5 years and reducing smoking frequency from everyday smoking to quitting. A forest plot of lung cancer incidence and smoking is shown in Figure 9. Lung cancer incidence of the high risk group had the closest relationship with smoking than the medium risk, low risk, and non-risk groups. The optimal transfer pathways and the decline of lung cancer incidence of stratified risk groups are displayed in Table 3. Lung cancer incidence of the high risk group had declined from 1060.38 per 100,000 to 336.78 per 100,000, which had decreased by 68.2% through risk transfer. The lung cancer incidence of the medium risk group had reduced from 671.20 per 100,000 to 289.13 per 100,000, which had declined by 56.9%.

XSL•FO
RenderX

Figure 9. Forest plot of lung cancer incidence and smoking.

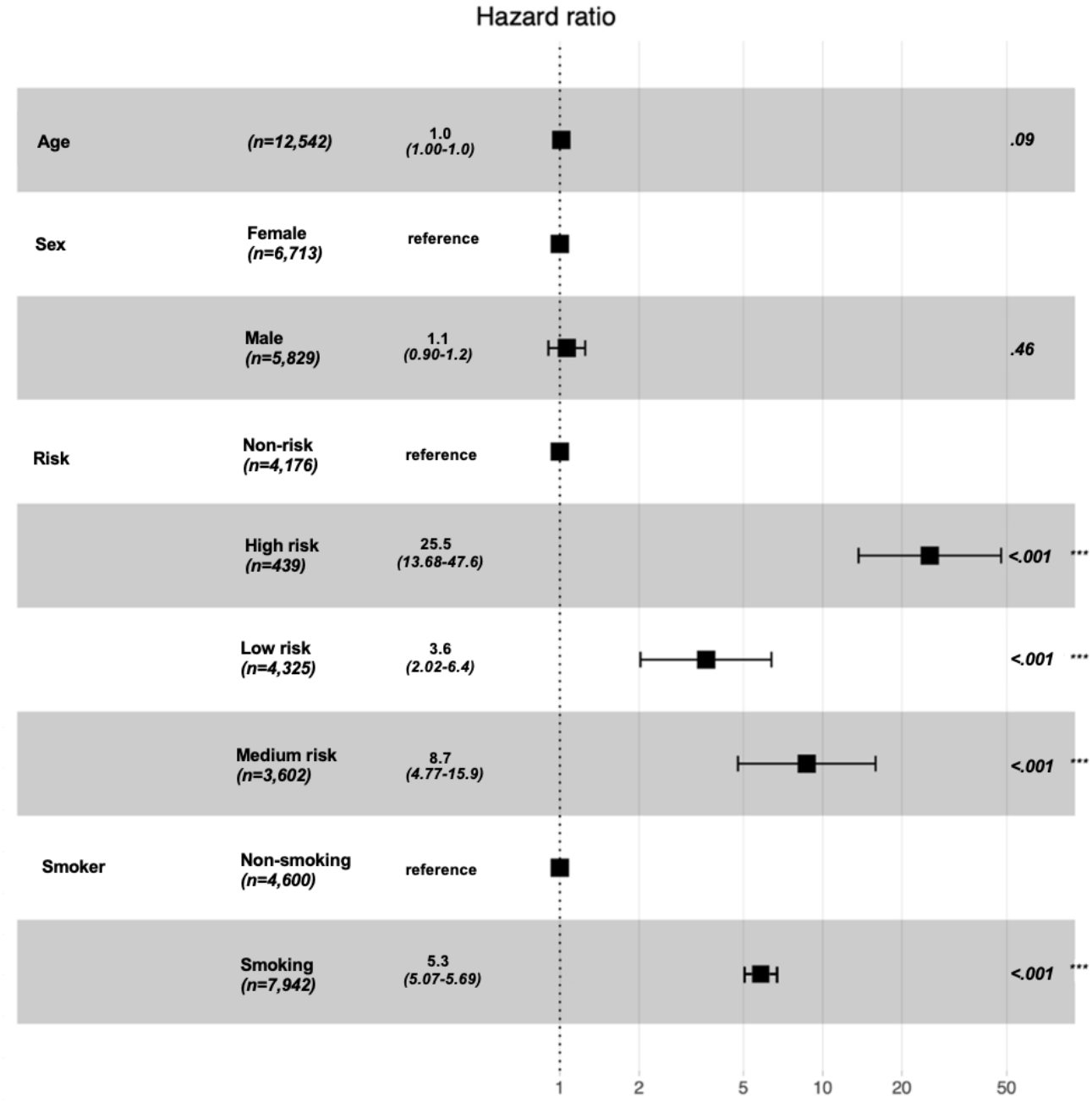


Table . Optimal transfer pathway and transfer effect of stratified risk groups.

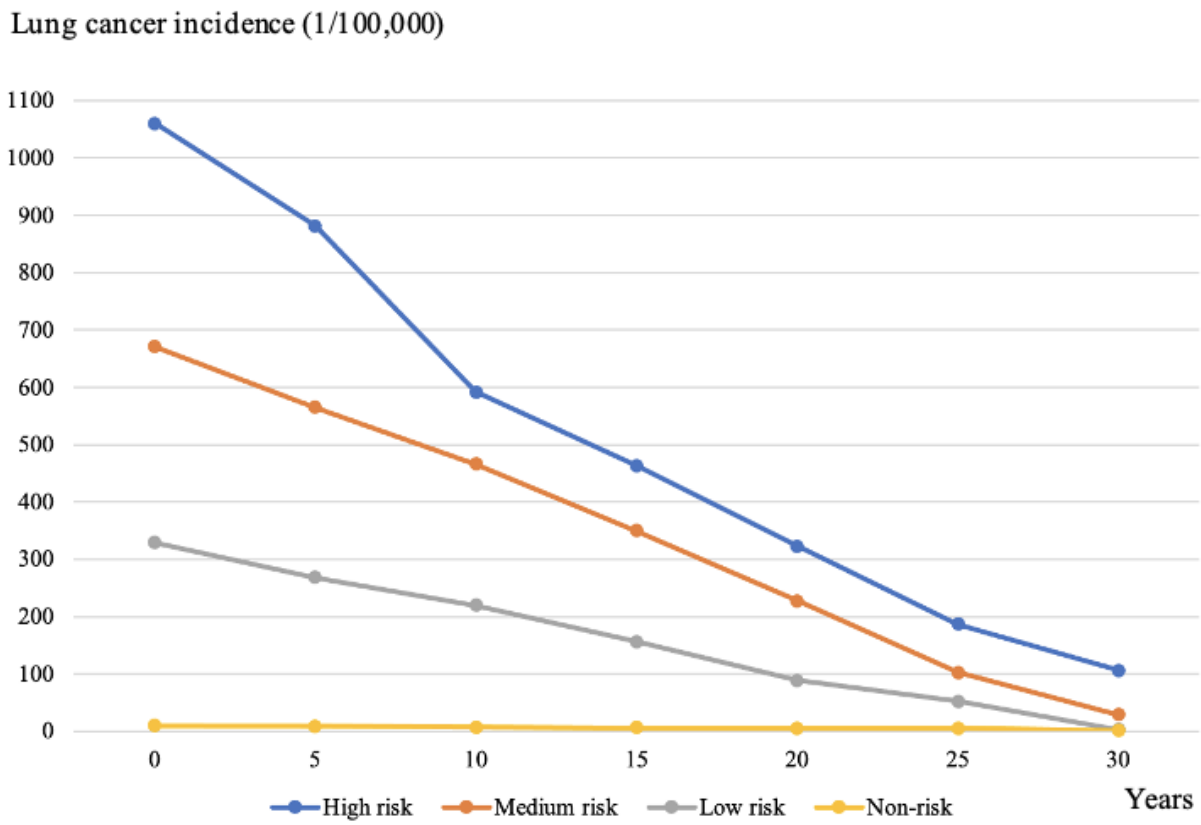
Transfer pathway	Lung cancer incidence after transfer (per 100,000)	Incidence decline (%)
High risk→medium risk→low risk→non-risk	336.78	68.24
Medium risk→low risk→non-risk	289.13	56.92
Low risk→non-risk	3.02	99.08

Transfer Pathway Evaluation

The incidence trend of lung cancer for stratified risk groups was simulated and analyzed to evaluate the optimal transfer pathway and transfer strategy, as shown in Figure 10. According to the optimal transfer strategy and the optimal transfer pathway, lung cancer incidence of the high risk group reduced to 106.78 per 100,000 after simulative intervention and risk transition, lung cancer incidence of the medium risk group declined to 29.13 per 100,000, and the incidence of the low risk group also decreased obviously. The results validated the effectiveness and availability of the transfer pathway.

XSL•FO
RenderX

Figure 10. The incidence trend of lung cancer for stratified risk groups.



Comparison With Other Methods

We used random forest to conduct comparison work of the DQN-based risk transfer pathway simulation. The training dataset and internal test dataset of HRS were leveraged to construct a random forest model for lung cancer incidence, lung cancer survival, and overall survival. The CHARLS dataset was used for external validation. Risk transfer pathways of lung cancer incidence were derived through random forest models, and optimal transfer pathways were also deduced, which were mainly the same as those of the DQN models. For example, in the risk improvement pathway, extending the duration of smoking cessation and reducing the frequency of smoking were the optimal transfer strategies through random forest models, which were consistent with the DQN models. The accuracies of random forest external validation models for stratified risk groups were from 0.823 to 0.851, and the AUROCs were from 0.778 to 0.836, which were almost lower by 10% than the DQN models, as shown in Table 4. Therefore, DQN models had better performance for transfer pathways computation than random forest models.

Table . Comparison of performance between random forest models and deep Q network models of external validation.

Risk groups	Random forest		DQN[b]	
	Accuracy (95% CI)	AUROC[a] (95% CI)	Accuracy (95% CI)	AUROC (95% CI)
High risk	0.837 (0.791 - 0.852)	0.793 (0.776 - 0.826)	0.949 (0.909 - 0.961)	0.927 (0.893 - 0.938)
Medium risk	0.823 (0.803 - 0.865)	0.812 (0.785 - 0.845)	0.936 (0.913 - 0.952)	0.919 (0.886 - 0.941)
Low risk	0.826 (0.796 - 0.848)	0.778 (0.761 - 0.827)	0.917 (0.896 - 0.928)	0.913 (0.902 - 0.937)
Non-risk	0.851 (0.825 - 0.876)	0.836 (0.793 - 0.859)	0.921 (0.902 - 0.946)	0.906 (0.887 - 0.933)

[a]AUROC: area under the receiver operating characteristic curve.

[b]DQN: deep Q network.

Discussion

Lung cancer risk transfer pathways were simulated through proposed DQN models for stratified risk groups among middle-aged and older individuals. The DQN-based risk transfer pathway calculation model performed well, not only for internal validation of the HRS dataset but also for external validation of the CHARLS dataset, with accuracy ranging from 0.917 (95% CI 0.896 - 0.928) to 0.949 (95% CI 0.909 - 0.961) and AUROC ranging from 0.906 (95% CI 0.887 - 0.933) to 0.927 (95% CI 0.893 - 0.938). Moreover, optimal risk transfer pathways through model-based simulations were effective

XSL•FO
RenderX

solutions to reducing lung cancer incidence for the high risk, medium risk, low risk, and non-risk groups. The optimal intervention strategies for decreasing lung cancer incidence were extending the duration of smoking cessation from less than 1 month to more than 5 years and reducing smoking frequency from everyday smoking to quitting, which could have significant effects in the high risk and medium risk groups.

We developed DNN models for lung cancer risk stratification, which refined different lung cancer risk factors and supplied research foundation for lung cancer risk transition. The optimal risk transfer strategy and the optimal transfer effect were deduced using DQN models, which improve the effectiveness and availability of the proposed risk transfer pathway. In recent years, deep learning has been adopted by more research [21-24] for lung cancer risk prediction. Landy et al [25] used the recalibrated CNNs to identify risk individuals in the National Lung Screening Trial and predict lung cancer risk. Bhatia et al [26] developed and validated lung cancer risk screening and prediction model using transfer learning. Jiang et al [27] applied CNNs to predict lung cancer risk, and the National Lung Screening Trial cohort was used to evaluate the model's performance. The DQN model, which we proposed to explore risk transfer pathway, could also demonstrate the stratified risk classification and the optimal transfer strategy for lung cancer prevention.

We developed and evaluated the DQN model for optimizing the risk transfer pathway among middle-aged and older individuals. External validation was conducted to assess the model's effectiveness and availability. The results illuminated that DQN models could deduce the optimal risk transfer pathways for stratified risk groups. Lung cancer incidence of the high risk group had declined by 68.2% through risk transfer, which had declined by 56.9% in the medium risk group. Lung cancer incidences of high risk, medium risk, and low risk were obviously decreased through simulated risk intervention and transition, which were consistent with previous studies [14,28,29] and indicated that actively quitting smoking before being diagnosed with lung cancer could effectively reduce the incidence and mortality of lung cancer.

Previous studies adopted statistical models or other machine learning models to predict lung cancer risk, such as multiple logistic regression [30-32], artificial neural network [33], random forest [34,35], extreme gradient boosting [36,37], decision tree [38], and risk prediction models [39]. Compared to these models, the proposed DQN model explored risk transfer pathways of lung cancer among middle-aged and older individuals with high accuracy and AUROC. External validation was conducted to verify the proposed model based on stratified risk states, which enhanced the credibility and effectiveness of the DQN model. Moreover, proposed risk transfer pathways might supply clues for revealing the pathogenesis of lung cancer.

Several limitations existed in this research. First, this study explored lung cancer risk pathways of middle-aged and older individuals based on a large longitudinal retrospective cohort, which did not involve a causal framework. A causal framework, such as directed acyclic graphs, could be constructed in the further research. Second, variables available in the HRS and CHARLS survey datasets were mostly demographic and self-reported health questionnaire data in this study. Genomic mutation data, chest computed tomography scan image, outpatient attendance records, laboratory tests, and so on, could be involved in future work to improve the biological plausibility and real-world use of the derived risk transfer pathways. Third, conducting follow-up surveys over the years for the large longitudinal cohort was a long-term process. Transfer pathway evaluation was conducted through simulative prediction at present. In the future, this evaluation could be done according to a real-world longitudinal cohort. Fourth, lung cancer incidence had a significant decrease through the proposed optimal risk transfer pathway, which needs clinical validation in our future work.

In conclusion, a DQN-based deep reinforcement learning model was proposed and validated to develop and simulate the risk transfer pathway of lung cancer among middle-aged and older individuals. Risk stratification supplied an effective foundation for lung cancer risk transition. This modeling approach of the risk transfer pathway might supply a prevention clue for other chronic diseases.

Funding

This study was supported by the National Natural Science Foundation of China (NSFC; grant 62106286); National Science and Technology Major Project for the Prevention and Treatment of Cancer, Cardiovascular, Respiratory, and Metabolic Diseases (grant 2023ZD0509702); CAMS Innovation Fund for Medical Sciences (CIFMS; grant 2021-I2M-1-057); and the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (grant 2024-ZHCH630-01).

Data Availability

Analyzed data supporting the funding of this research are available from the corresponding author on reasonable request.

Authors' Contributions

SC contributed to research design, methodology, data analysis, visualization, validation, and writing–review and editing. SW revised the manuscript. Both authors read and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Calibration plots of the deep neural network model.
[DOCX File, 90 KB - medinform_v14i1e74990_app1.docx]

Multimedia Appendix 2
Confusion matrix of the deep neural network model.
[DOCX File, 13 KB - medinform_v14i1e74990_app2.docx]

Multimedia Appendix 3
Deep Q-network models' performance for internal and external validation.
[DOCX File, 14 KB - medinform_v14i1e74990_app3.docx]

References

1. Filho AM, Laversanne M, Ferlay J, et al. The GLOBOCAN 2022 cancer estimates: data sources, methods, and a snapshot of the cancer burden worldwide. Int J Cancer 2025 Apr 1;156(7):1336-1346. [doi: 10.1002/ijc.35278] [Medline: 39688499]
2. Data visualization tools for exploring the global cancer burden in 2022. International Agency for Research on Cancer, World Health Organization. 2022. URL: https://gco.iarc.fr/today/home [accessed 2025-02-26]
3. Liao W, Coupland CAC, Burchardt J, et al. Predicting the future risk of lung cancer: development, and internal and external validation of the CanPredict (lung) model in 19·67 million people and evaluation of model performance against seven other risk prediction models. Lancet Respir Med 2023 Aug;11(8):685-697. [doi: 10.1016/S2213-2600(23)00050-4] [Medline: 37030308]
4. Henriksen MB, Van Daalen F, Wee L, et al. Lung cancer detection using Bayesian networks: a retrospective development and validation study on a Danish population of high-risk individuals. Cancer Med 2025 Feb;14(3):e70458. [doi: 10.1002/cam4.70458] [Medline: 39887592]
5. Feng X, Goodley P, Alcala K, et al. Evaluation of risk prediction models to select lung cancer screening participants in Europe: a prospective cohort consortium analysis. Lancet Digit Health 2024 Sep;6(9):e614-e624. [doi: 10.1016/S2589-7500(24)00123-7] [Medline: 39179310]
6. Zhong D, Sidorenkov G, Jacobs C, et al. Lung nodule management in low-dose CT screening for lung cancer: lessons from the NELSON trial. Radiology 2024 Oct;313(1):e240535. [doi: 10.1148/radiol.240535] [Medline: 39436294]
7. Fatima S, Kumar V, Kumar D. Molecular mechanism of genetic, epigenetic, and metabolic alteration in lung cancer. Med Oncol 2025 Feb 2;42(3):61. [doi: 10.1007/s12032-025-02608-5] [Medline: 39893601]
8. Zhao Z, Gu S, Yang Y, et al. A cost-effectiveness analysis of lung cancer screening with low-dose computed tomography and a polygenic risk score. BMC Cancer 2024 Jan 13;24(1):73. [doi: 10.1186/s12885-023-11800-7] [Medline: 38218803]
9. Kim HJ, Lee MK. Effectiveness of nursing interventions based on lung cancer trajectory: a systematic review and meta-analysis. Int Nurs Rev 2025 Sep;72(3):e13074. [doi: 10.1111/inr.13074] [Medline: 39604008]
10. Huang J, Deng Y, Tin MS, et al. Distribution, risk factors, and temporal trends for lung cancer incidence and mortality: a global analysis. Chest 2022 Apr;161(4):1101-1111. [doi: 10.1016/j.chest.2021.12.655] [Medline: 35026300]
11. Kowada A. Cost-effectiveness and health impact of lung cancer screening with low-dose computed tomography for never smokers in Japan and the United States: a modelling study. BMC Pulm Med 2022 Jan 8;22(1):19. [doi: 10.1186/s12890-021-01805-y] [Medline: 34996423]
12. Wang KJ, Lukito H. Lifespan and medical expenditure prognosis for cancer metastasis - a simulation modeling using semi-Markov process. Comput Methods Programs Biomed 2023 Jun;234:107509. [doi: 10.1016/j.cmpb.2023.107509] [Medline: 37003040]
13. Yang F, Sun D, Ding C, et al. Global patterns of cancer transitions: a modelling study. Int J Cancer 2023 Nov 1;153(9):1612-1622. [doi: 10.1002/ijc.34650] [Medline: 37548247]
14. Adams SJ, Stone E, Baldwin DR, Vliegenthart R, Lee P, Fintelmann FJ. Lung cancer screening. Lancet 2023 Feb 4;401(10374):390-408. [doi: 10.1016/S0140-6736(22)01694-4] [Medline: 36563698]
15. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol 2023 Sep;20(9):624-639. [doi: 10.1038/s41571-023-00798-3] [Medline: 37479810]
16. The Health and Retirement Study. Institute for Social Research, University of Michigan. 2024. URL: https://hrs.isr.umich.edu [accessed 2025-02-26]
17. China Health and Retirement Longitudinal Study. National School of Development, Peking University. 2024. URL: http://charls.pku.edu.cn/en/ [accessed 2025-02-26]
18. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999 Mar;8(1):3-15. [doi: 10.1177/096228029900800102] [Medline: 10347857]
19. Kalman D. A singularly valuable decomposition: the SVD of a matrix. Coll Math J 1996 Jan;27(1):2-23. [doi: 10.1080/07468342.1996.11973744]

XSL•FO
RenderX

20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR 2002;16(1):321-357. [doi: 10.1613/jair.953]
21. Cellina M, Cè M, Irmici G, et al. Artificial intelligence in lung cancer imaging: unfolding the future. Diagnostics (Basel) 2022 Oct 31;12(11):2644. [doi: 10.3390/diagnostics12112644] [Medline: 36359485]
22. Lee HA, Chao LR, Hsu CY. A 10-year probability deep neural network prediction model for lung cancer. Cancers (Basel) 2021 Feb 23;13(4):928. [doi: 10.3390/cancers13040928] [Medline: 33672275]
23. Senthil P, Kuhan S, Potter AL, Jeffrey Yang CF. Update on lung cancer screening guideline. Thorac Surg Clin 2023 Nov;33(4):323-331. [doi: 10.1016/j.thorsurg.2023.04.002] [Medline: 37806735]
24. Chen SJ, Wu SZ. Ensemble machine learning models for lung cancer incidence risk prediction in the elderly: a retrospective longitudinal study. BMC Cancer 2025 Jan;25(1):126. [doi: 10.1186/s12885-025-13562-w] [Medline: 39844064]
25. Landy R, Wang VL, Baldwin DR, et al. Recalibration of a deep learning model for low-dose computed tomographic images to inform lung cancer screening intervals. JAMA Netw Open 2023 Mar 1;6(3):e233273. [doi: 10.1001/jamanetworkopen.2023.3273] [Medline: 36929398]
26. Bhatia I, Ansarullah SI, Amin F, Alabrah A. An advanced lung carcinoma prediction and risk screening model using transfer learning. Diagnostics (Basel) 2024 Jun 28;14(13):1378. [doi: 10.3390/diagnostics14131378] [Medline: 39001268]
27. Jiang Y, Ebrahimpour L, Després P, Manem VSK. A benchmark of deep learning approaches to predict lung cancer risk using National Lung Screening Trial cohort. Sci Rep 2025 Jan 11;15(1):1736. [doi: 10.1038/s41598-024-84193-7] [Medline: 39799226]
28. Li J, Chung S, Martinez MC, Luft HS. Smoking-cessation interventions after lung cancer screening guideline change. Am J Prev Med 2020 Jul;59(1):88-97. [doi: 10.1016/j.amepre.2020.01.031] [Medline: 32417022]
29. González Maldonado S, Motsch E, Trotter A, et al. Overdiagnosis in lung cancer screening: estimates from the German Lung Cancer Screening Intervention Trial. Int J Cancer 2021 Mar 1;148(5):1097-1105. [doi: 10.1002/ijc.33295] [Medline: 32930386]
30. Tang W, Peng Q, Lyu Y, et al. Risk prediction models for lung cancer: perspectives and dissemination. Chin J Cancer Res 2019 Apr;31(2):316-328. [doi: 10.21147/j.issn.1000-9604.2019.02.06] [Medline: 31156302]
31. Yang JJ, Wen W, Zahed H, et al. Lung cancer risk prediction models for Asian ever-smokers. J Thorac Oncol 2024 Mar;19(3):451-464. [doi: 10.1016/j.jtho.2023.11.002] [Medline: 37944700]
32. Rubin KH, Haastrup PF, Nicolaisen A, et al. Developing and validating a lung cancer risk prediction model: a nationwide population-based study. Cancers (Basel) 2023 Jan 12;15(2):487. [doi: 10.3390/cancers15020487] [Medline: 36672436]
33. Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS ONE 2018;13(10):e0205264. [doi: 10.1371/journal.pone.0205264] [Medline: 30356283]
34. Pathan RK, Shorna IJ, Hossain MS, Khandaker MU, Almohammed HI, Hamd ZY. The efficacy of machine learning models in lung cancer risk prediction with explainability. PLoS ONE 2024;19(6):e0305035. [doi: 10.1371/journal.pone.0305035] [Medline: 38870229]
35. Zhang W, Guan X, Jiao S, Wang G, Wang X. Development and validation of an artificial intelligence prediction model and a survival risk stratification for lung metastasis in colorectal cancer from highly imbalanced data: a multicenter retrospective study. Eur J Surg Oncol 2023 Dec;49(12):107107. [doi: 10.1016/j.ejso.2023.107107] [Medline: 37883884]
36. Li Y, Zou Z, Gao Z, et al. Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. Cancer Med 2022 Dec;11(23):4469-4478. [doi: 10.1002/cam4.4800] [Medline: 35499292]
37. Lee AR, Park H, Yoo A, Kim S, Sunwoo L, Yoo S. Risk prediction of emergency department visits in patients with lung cancer using machine learning: retrospective observational study. JMIR Med Inform 2023 Dec 6;11:e53058. [doi: 10.2196/53058] [Medline: 38055320]
38. Trujillo JC, Soriano JB, Marzo M, et al. Cost-effectiveness of a machine learning risk prediction model (LungFlag) in the selection of high-risk individuals for non-small cell lung cancer screening in Spain. J Med Econ 2025 Dec;28(1):147-156. [doi: 10.1080/13696998.2024.2444781] [Medline: 39697091]
39. Juang YR, Ang L, Seow WJ. Predictive performance of risk prediction models for lung cancer incidence in Western and Asian countries: a systematic review and meta-analysis. Sci Rep 2025 Mar 4;15(1):4259. [doi: 10.1038/s41598-024-83875-6] [Medline: 40038330]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CHARLS: China Health and Retirement Longitudinal Study
CNN: convolutional neural network
DNN: deep neural network
DQN: deep Q-network
HRS: Health and Retirement Study

XSL•FO
RenderX

Edited by A Benis; submitted 26.Mar.2025; peer-reviewed by M Al-Agil, S Matsuda; accepted 17.Feb.2026; published 15.Apr.2026.

Please cite as:
Chen S, Wu S
Research on Risk Transfer Pathways for Lung Cancer Among Middle-Aged and Older Individuals Using Deep Reinforcement Learning: Retrospective Cohort Study
JMIR Med Inform 2026;14:e74990
URL: https://medinform.jmir.org/2026/1/e74990
doi:10.2196/74990

© Songjing Chen, Sizhu Wu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Real-Time Symptom Ratings Using Ecological Momentary Assessment Versus Traditional Questionnaires in Patients With Chronic Obstructive Pulmonary Disease: Observational Study

Banchia Palmen[1,2], MSc; Maarten van Herck[1], PhD; Yvonne M J Goërtz[3,4], PhD; Zjala Ebadi[5,6], MSc; Qichen Deng[1], PhD; Melissa S Y Thong[7,8], PhD; Chris Burtin[9], Prof Dr; Jeannette B Peters[6], PhD; Roy T M Sprooten[10], PhD; Erik W M A Bischoff[11,12], Prof Dr MD; Emiel F M Wouters[1,10,13], Prof Dr MD; Jan H Vercoulen[5,6], PhD; Sarah Houben-Wilke[1], PhD; Daisy J A Janssen[3,4,14], Prof Dr; Martijn A Spruit[1,15], Prof Dr; Anouk W Vaes[1], PhD

[1]Department of Research and Development, Centre of Expertise for Chronic Organ Failure (CIRO), Hornerheide 1, Horn, The Netherlands

[10]Department of Respiratory Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands

[11]Department of General Practice, Erasmus University Medical Centre, Rotterdam, The Netherlands

[12]Department of Primary and Community Care, Radboud University Medical Center, Nijmegen, The Netherlands

[13]Department Internal Medicine, Sigmund Freud University Vienna, Vienna, Austria

[14]Department of Family Medicine, Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[15]Department of Respiratory Medicine, Institute of Nutrition and Translational Research in Metabolism, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[2]Institute of Nutrition and Translational Research in Metabolism, Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[3]Living Lab in Ageing and Long-Term Care, Maastricht University, Maastricht, The Netherlands

[4]Department of Health Services Research, Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[5]Department of Medical Psychology, Radboud University Nijmegen, Nijmegen, The Netherlands

[6]Department of Pulmonary Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands

[7]Medical Psychology, Amsterdam UMC Location University of Amsterdam, Amsterdam, The Netherlands

[8]Cancer Survivorship Outcomes and Epidemiology, German Cancer Research Center, Heidelberg, Germany

[9]REVAL—Rehabilitation Research Center, BIOMED—Biomedical Research Institute, Faculty of Rehabilitation Sciences, Hasselt University, Hasselt University, Diepenbeek, Belgium

Corresponding Author:
Banchia Palmen, MSc
Department of Research and Development, Centre of Expertise for Chronic Organ Failure (CIRO), Hornerheide 1, Horn, The Netherlands

Abstract

Background: Questionnaires assessing symptoms in chronic obstructive pulmonary disease (COPD) rely on retrospective reporting, introducing recall bias and missing symptom variability. Ecological momentary assessment (EMA) addresses these limitations by capturing real-time symptom data.

Objective: The aim of this study was to compare EMA symptom ratings with traditional questionnaires in patients with COPD.

Methods: A subsample from the FAntasTIGUE study rated symptoms using questionnaires (visual analog scale, modified Medical Research Council Dyspnea Scale, Physical Activity Rating Scale-Dyspnea Questionnaire, Checklist Individual Strength-subscale Subjective Fatigue, Hospital Anxiety and Depression Scale, and COPD Assessment Test) and EMA. Participants rated breathlessness, fatigue, anxiety, and energy level on a 7-point Likert scale during 8 random EMA prompts daily for 5 days.

Results: Among 54 participants (mean age 67, SD 7 y; forced expiratory volume in 1 s 53%, SD 20% predicted; n=35, 65% men), the EMA response rate was 76.1%. EMA scores were moderately to strongly correlated with questionnaire scores (r_s=0.49-0.78, all P<.05) and varied across severity groups (all P<.05). EMA captured significant intraday and interday symptom variability.

Conclusions: EMA is a valid tool for COPD symptom assessment, capturing real-time fluctuations to improve symptom management.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e79001) doi:10.2196/79001

KEYWORDS

chronic obstructive pulmonary disease; COPD; ecological momentary assessment; questionnaires; symptom assessment

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous chronic lung condition characterized by persistent airflow limitation and respiratory symptoms such as breathlessness and fatigue [1,2]. Additionally, patients with COPD often experience psychological symptoms like anxious and depressed mood and lack of energy, which can substantially affect their health-related quality of life and functional status [3,4]. Despite this, breathlessness, fatigue, anxiety, and lack of energy are frequently underreported [5]. In both daily clinical practice and research, these symptoms are assessed using questionnaires, such as the modified Medical Research Council (mMRC) Dyspnea Scale, Checklist Individual Strength-subscale Subjective Fatigue (CIS-Fatigue), and Hospital Anxiety and Depression Scale (HADS) [6,7]. However, these traditional tools rely on retrospective self-reports, which may be limited by recall bias and fail to fully capture the expected dynamic nature of symptoms both intraday and interday [1,8-13].

Ecological momentary assessment (EMA) is a real-time data collection approach that seeks to overcome these limitations by facilitating data capture in patients' natural environment [1,8,9,14]. By prompting patients to report breathlessness, fatigue, anxiety, and lack of energy at repeated intervals throughout the day, EMA minimizes memory-related biases and provides a more detailed depiction of symptom variability, offering a comprehensive view of daily symptom burden in patients with COPD. Despite the growing use of EMA in chronic respiratory disease research, there remains a limited understanding of how well EMA-based symptom ratings align with traditional questionnaires [9,15,16]. Establishing this relationship is essential to evaluate the measurement properties of EMA, specifically its convergent validity with questionnaires. By exploring to which extent EMA-derived symptom ratings correlate with well-established validated questionnaire scores, it can be determined whether EMA can serve as a valid complementary tool for symptom assessment in COPD.

This study aims to assess the convergent validity of EMA symptom ratings by exploring the associations between EMA-based and questionnaire-based measurements of breathlessness, fatigue, anxiety, and energy in patients with COPD. We seek to determine whether EMA provides unique insights into daily symptom fluctuations in COPD by examining moment-to-moment variability of symptoms both intraday and interday, which cannot be captured by traditional retrospective questionnaires. Our findings may have implications for improving symptom monitoring and management in daily clinical practice and research. By identifying moment-to-moment fluctuations in symptoms such as breathlessness, fatigue, anxiety, and lack of energy, EMA may enable timely and personalized adjustments in treatment or self-management strategies, potentially leading to more responsive and individualized care for patients with COPD. It is hypothesized that EMA symptom ratings would show significant positive correlations with traditional questionnaire scores, supporting convergent validity. Furthermore, we expect EMA to reveal substantial intraday and interday variability in breathlessness, fatigue, anxiety, and lack of energy that is not captured by retrospective questionnaires.

Methods

Design

This research was conducted as part of a multicenter, longitudinal, observational study in patients with COPD investigating physical, systemic, psychological, and behavioral factors associated with fatigue (FAntasTIGUE study) [17]. The prevalence of fatigue in patients with COPD was previously reported, as well as the association between physical activity, sedentarism, and EMA-derived symptom scores [3,18,19].

Ethical Considerations

The study received approval from the Medical Research Ethics Committee United, Nieuwegein, The Netherlands (R17.036/NL60484.100.17), and written informed consent was obtained from all participants. Participation was entirely voluntary, and participants could withdraw at any time without consequence. Participants did not receive any financial compensation. Data were deidentified before analysis.

Recruitment

Patients were recruited between 2018 and 2021 at outpatient clinics of the Department of Respiratory Medicine in Maastricht University Medical Center and the Department of Pulmonary Diseases of the Radboud University Medical Center in Nijmegen, as well as via general practitioners (Research Network Family Medicine Maastricht [20] and the Academic General Practitioner Network in the Nijmegen region). Additionally, patients with COPD who attended a peer support group meeting in Maastricht or Nijmegen, as well as patients recruited from primary and secondary care who previously participated in the Chance Study (NTR3416) and had consented to follow-up research, were invited to take part in the FAntasTIGUE study. To participate in the EMA substudy, patients had to meet the following inclusion criteria:

1. A diagnosis of COPD according to the Global Strategy for the Diagnosis, Management and Prevention of COPD (Global Initiative for Chronic Obstructive Lung Disease, grade 1A–4D) [2].
2. No exacerbation-related hospitalization and/or use of oral corticosteroids and/or antibiotics less than 4 weeks preceding enrollment.
3. Provided written informed consent.
4. Able to use an iPod.

XSL•FO
RenderX

Patients lacking sufficient understanding of the Dutch language and/or participating in concurrent intervention studies were excluded.

Measurements

During baseline assessment, sociodemographic, physical, psychological, and behavioral characteristics were assessed as described elsewhere [17].

Questionnaires

Breathlessness was assessed using 3 methods: mMRC Dyspnea Scale, item 1 of the Physical Activity Rating Scale-Dyspnea Questionnaire (PARS-D), and visual analog scale (VAS). The mMRC Dyspnea Scale assesses the current severity of activity-related dyspnea on a scale ranging from 0 ("only gets breathless with strenuous exercise") to 4 ("too breathless to leave the house or breathless when dressing") [21]. An mMRC score of 2 or higher is used as a threshold for separating "less breathlessness" from "more breathlessness" [2]. On item 1 of the PARS-D, patients had to rate the level of breathlessness on most days of the previous month from 1 ("not at all breathless") to 10 ("very breathless") [22]. Patients rated their level of breathlessness over the past 2 weeks using a VAS, which ranged from 0 ("not") to 100 mm ("worst possible") [23,24].

Fatigue was assessed using 2 methods: CIS-Fatigue and VAS. The CIS-Fatigue contains questions about fatigue-related complaints in the previous 2 weeks. For these analyses, we used both the score from item 1 and the CIS-Fatigue total score. Item 1 asks the patient if they are tired, ranging from 1 ("yes, that's true") to 7 ("no, that's not true"). A total score of 26 points or lower on the CIS-Fatigue indicates normal fatigue, between 27 and 35 reflects mild fatigue, and 36 points or higher signifies severe fatigue [6]. Patients rated their level of fatigue over the past 2 weeks using the VAS, which ranged from 0 ("not") to 100 mm ("worst possible") [23,24].

Symptoms of anxiety and depression were measured using 2 methods: HADS and VAS. The HADS is divided into an anxiety and a depression subscale. Each subscale consists of 7 items and ranges from 0 to 21 points. A higher score on these subscales indicates a higher level of symptoms of anxiety or depression, respectively. A subscale score of 8 points or higher indicates possible symptoms of anxiety or depression, while a score of 11 points or higher indicates probable symptoms of anxiety or depression [7]. Patients rated their level of anxiety over the past 2 weeks using the VAS, which ranged from 0 ("not") to 100 mm ("worst possible") [23,24].

Energy was assessed with item 8 of the COPD Assessment Test (CAT), which assesses the current energy level on a scale from 0 to 5, with a higher score indicating less energy [25]. All questionnaires used in this study are well-validated and widely applied instruments for assessing symptoms in patients with COPD [2,26,27]. However, for some analyses, single items from these questionnaires were used to allow direct comparison with specific EMA items reflecting corresponding symptom dimensions.

Ecological Momentary Assessment

Electronic EMA questionnaires were administered over 5 consecutive days using an Apple iPod Touch with a custom-installed app PsyMate (version 1.1b). If possible, the EMA assessments began the day immediately following the baseline assessment. Patients were instructed to respond to each questionnaire as soon as possible upon hearing the beep and to keep their usual day and night routine. If patients did not respond within 15 minutes after the beep, the questionnaire was skipped. PsyMate generated a beep at 8 random times each day between 7:30 AM and 10:30 PM. Patients had to rate 10 symptoms ("I feel: relaxed; short of breath; energetic; cheerful; insecure; irritated; satisfied; anxious; tired; and mentally fit") with a 7-point Likert scale ranging from 1 ("Not") to 7 ("Very"). For these analyses, 4 symptoms were included: breathlessness, fatigue, anxiety, and energy. Patients who completed less than one-third of the total beeps (ie, <13 out of 40 beeps) were excluded [28]. An EMA questionnaire completion time of 10 minutes or longer was considered invalid, and this EMA data point was excluded.

Data Analyses

All analyses were performed using SPSS (version 28.0.1.1) and R statistical software (version 4.4.2). The distribution of data was tested using the Shapiro-Wilk test with a 2-tailed significance of less than .05. Data were reported as mean (SD), median (IQR), percentages, and numbers, where appropriate. In the analyses, EMA item "I feel short of breath" was compared with mMRC Dyspnea Scale score, PARS-D score, and VAS breathlessness score. EMA item "I feel tired" was compared with CIS-Fatigue item 1 score, CIS-Fatigue total score, and VAS Fatigue score. EMA item "I feel anxious" was compared with HADS anxiety score and VAS anxiety score, while EMA item "I feel energetic" was compared with CAT item 8 score. To evaluate the convergent validity of EMA measures, associations between EMA symptom ratings and corresponding questionnaire scores were examined using Spearman correlation coefficients and multilevel linear mixed models (MLMs).

Median EMA scores for all patients were calculated. Spearman correlation analyses were performed between median EMA scores and corresponding questionnaire scores [29]. Subsequently, MLMs were applied to analyze all EMA data points. In these models, EMA scores were treated as dependent variables, while questionnaire scores were included as fixed effects. Patient ID was specified as a random intercept to account for the correlation of intraindividual repeated observations. MLMs were performed using the "mixed" command, restricted maximum likelihood estimations, and a significance of less than .05. Mann-Whitney *U* tests were used to assess differences in median EMA scores across symptom severity groups based on established cutoff values of the mMRC Dyspnea Scale (0 - 1/≥2), CIS-Fatigue (8 - 35/≥36), and HADS questionnaire (1 - 7/≥8). Patients were categorized into the symptom severity groups based on their corresponding questionnaire scores.

To assess intraday and interday symptom variability, we used MLMs with EMA symptom scores as the dependent variable. For each symptom (breathlessness, fatigue, anxiety, and energy), time of day (expressed as hours since 8:00 AM) and day number

XSL•FO
RenderX

(days 1 - 5) were included as fixed effects to model temporal trends. Random intercepts for each participant accounted for interindividual differences in baseline symptom levels, while random slopes for time of day allowed the trajectory of symptoms to change throughout the day to vary interindividuals. Variance components were extracted from the MLMs to estimate the proportion of total variance attributable to intraindividual fluctuations and interindividual differences. Intraclass correlation coefficients (ICC) were calculated as the ratio of interindividual variance to total variance after accounting for time of day and day effects. ICC values close to 1 indicate that most variability is explained by differences between individuals, whereas lower ICC values indicate greater intraindividual variability.

Results

Overview

Sixty patients were assessed for eligibility. Six patients were excluded because of no available EMA data (n=2), invalid EMA data (<13 beeps in total; n=3), or incorrect loading of the EMA data (n=1). Consequently, 54 patients were included in the current analysis.

Participants

The majority of patients were men (35/54, 65%) with an average age of 67 years (Table 1). Most patients had moderate-to-severe COPD (49/54, 91%), and almost half had at least one exacerbation in the previous 12 months (25/54, 46%). Across the 9 questionnaires that were used in the analysis, the overall response rate was high (471/486, 97%). Severe dyspnea was reported by 39% (21/52) and 43% (23/53) experienced severe fatigue. In addition, 9% (5/52) and 11% (6/52) of the patients had symptoms of anxiety and depression, respectively. The EMA subsample included in the current analyses is highly comparable to the overall FAntasTIGUE cohort in terms of key demographic and clinical characteristics. Results of the full FAntasTIGUE cohort have recently been published [19].

XSL•FO
RenderX

Table . Baseline characteristics.

Variable[a]	N	Value
Age (y), mean (SD)	54	67 (7)
Male, n (%)	54	35 (65)
Married/living together, n (%)	53	35 (65)
Currently employed, n (%)	52	7 (13)
Current smoker, n (%)	54	9 (17)
CCI[b] (points), median (IQR)	54	1 (1-2)
Exacerbation in previous 12 mo (0), n (%)	54	29 (54)
Exacerbation in previous 12 mo (1), n (%)	54	9 (17)
Exacerbation in previous 12 mo (≥2), n (%)	54	16 (30)
Exacerbation-related hospitalization in previous 12 mo (0), n (%)	54	45 (83)
Exacerbation-related hospitalization in previous 12 mo (1), n (%)	54	6 (13)
Exacerbation-related hospitalization in previous 12 mo (≥2), n (%)	54	3 (6)
BMI (kg/m^2), median (IQR)	54	25 (23 - 30)
FEV_1[c] (% predicted), mean (SD)	54	53 (20)
GOLD[d] I/II/III/IV, n (%)	54	5 (9); 22 (41); 22 (41); 5 (9)
FEV_1/FVC[e] (%), median (IQR)	54	46 (35 - 57)
mMRC[f] Dyspnea (grade), median (IQR)	52	1 (0 - 2)
mMRC[f] ≥2, n (%)	52	21 (39)
PARS-D[g] item 1 (Shortness of breath) (score), median (IQR)	54	5 (2-6)
VAS[h] breathlessness (score), median (IQR)	52	25 (10 - 62)
CIS-Fatigue[i] item 1 (I feel tired) (score), median (IQR)	53	4 (2-7)
CIS-Fatigue[i] (points), median (IQR)	53	33 (21 - 49)
Severe fatigue (≥36), n (%)	53	23 (43)
VAS[h] fatigue (score), median (IQR)	52	40 (13 - 60)
HADS[j]-Anxiety (points), median (IQR)	52	4 (1-8)
Symptoms of anxiety (8 - 10 points), n (%)	52	9 (17)
Symptoms of anxiety (≥11 points), n (%)	52	5 (9)
HADS[j]-Depression (points), median (IQR)	52	4 (1-8)
Symptoms of depression (8 - 10 points), n (%)	52	7 (13)
Symptoms of depression (≥11 points), n (%)	52	6 (11)
VAS[h] anxiety (score), median (IQR)	52	2 (1-15)
CAT[k] item 8 (I have a lot of energy) (score), median (IQR)	52	3 (2-4)

[a]Quantitative variables are presented as mean (SD) or median (IQR) for skewed variables.

[b]CCI: Charlson Comorbidity Index.

[c]FEV1 % predicted: Forced expiratory volume in 1 second as percentage from the predicted volume.

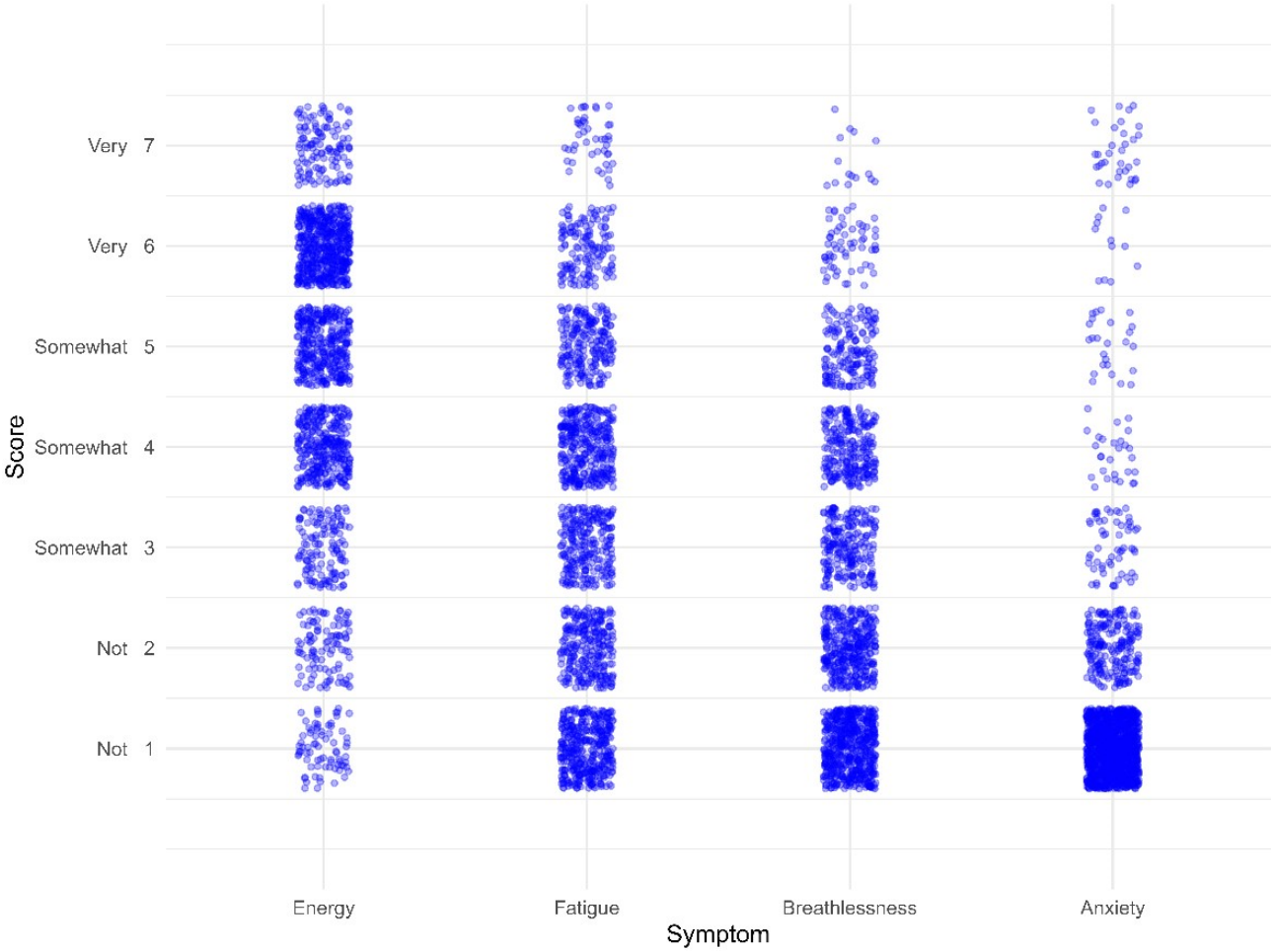
XSL•FO
RenderX

[d]GOLD: Global Initiative for Chronic Obstructive Lung Disease.
[e]FEV1/FVC %: the forced expiratory volume in 1 second to forced vital capacity ratio.
[f]mMRC: modified Medical Research Council Dyspnea Scale.
[g]PARS-D: Physical Activity Rating Scale-Dyspnea questionnaire.
[h]VAS: visual analog scale.
[i]CIS-Fatigue: Checklist Individual Strength-subscale Subjective fatigue.
[j]HADS: Hospital Anxiety and Depression Scale.
[k]CAT: COPD Assessment Test.

Ecological Momentary Assessment

One patient completed 4 days of EMA measurements, and 53 patients completed 5 days. Of the 2160 maximal possible EMA responses (54 participants × 5 days × 8 prompts per day), 1644 EMA responses were valid and could be analyzed (response rate: 76.1%). Median (IQR) time of completing the EMA questionnaire was 1 minute and 2 seconds (00:49–01:21). During the 5 days, patients generally rated feeling somewhat energetic (median 5 points, IQR 2), somewhat tired (median 3 points, IQR 2), not short of breath (median 2 points, IQR 3), and not anxious (median 1 point, IQR 0). Nevertheless, a large variation across the 7-point Likert scale was observed between patients (Figure 1). In Multimedia Appendix 1, the median EMA scores and total number of EMA responses per patient are shown.

Figure 1. Jittered scatterplot of the ecological momentary assessment–based symptoms scored on a 7-point Likert scale for 5 consecutive days.



Association

Median EMA scores and corresponding questionnaires showed moderate-to-strong correlations (Table 2). Patients with an mMRC grade of 0 to 1 reported significantly lower EMA scores for “I feel short of breath” compared to those with an mMRC grade of 2 or higher (P<.001; Figure 2). Similarly, patients with normal-to-mild CIS-Fatigue total scores reported significantly lower EMA scores for “I feel tired” compared to those with severe CIS-Fatigue scores (P<.001). Patients with no symptoms of anxiety had significantly lower EMA scores for “I feel anxious” compared to those with possible or probable anxiety (P=.01).

XSL•FO
RenderX

Table . Correlations between median ecological momentary assessment scores and corresponding questionnaires.

EMA[a] item and questionnaire		Spearman correlation coefficient[b]
I feel short of breath		
	mMRC[c] Dyspnea Scale	0.63
	PARS-D[d] item 1 Shortness of breath	0.76
	VAS[e] breathlessness	0.74
I feel tired		
	CIS-Fatigue[f] item 1 (I feel tired)	0.74
	CIS-Fatigue[f] total score	0.76
	VAS[e] fatigue	0.78
I feel anxious		
	HADS[g]-Anxiety	0.55
	VAS[e] anxiety	0.49
I feel energetic		
	CAT[h] Item 8 (I have a lot of energy)	−0.65

[a]EMA: ecological momentary assessment.

[b]$P<.001$.

[c]mMRC: modified Medical Research Council Dyspnea Scale.

[d]PARS-D: Physical Activity Rating Scale-Dyspnea questionnaire.

[e]VAS: visual analog scale.

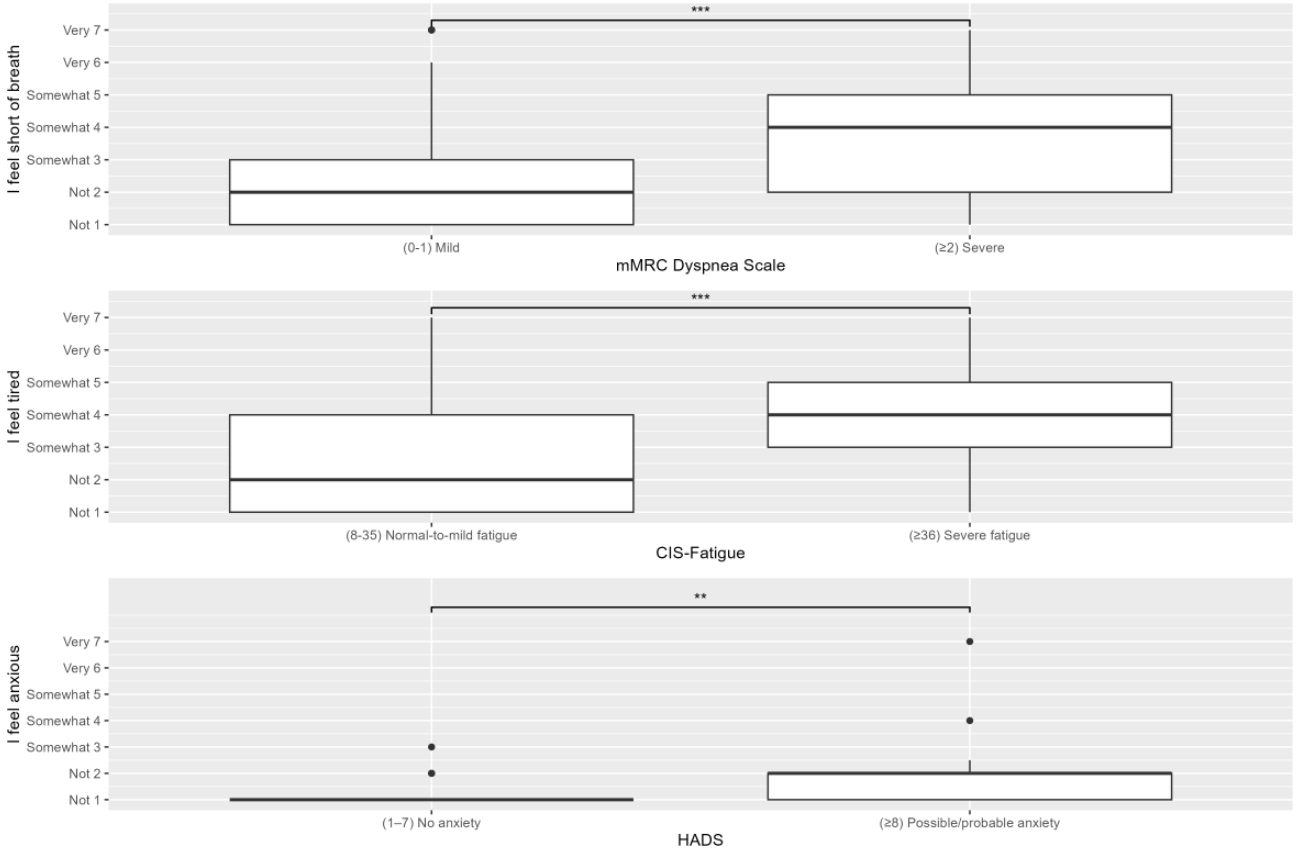
[f]CIS-Fatigue: Checklist Individual Strength-subscale Subjective Fatigue.

[g]HADS: Hospital Anxiety and Depression Scale.

[h]CAT: COPD Assessment Test.

XSL•FO
RenderX

Figure 2. Ecological momentary assessment (EMA) breathlessness scores divided by the modified Medical Research Council (mMRC) Dyspnea cutoff values. EMA tiredness scores divided by Checklist Individual Strength-subscale Subjective Fatigue (CIS-Fatigue) cutoff values. EMA anxious scores divided by the Hospital Anxiety and Depression Scale (HADS) cutoff values.



Multilevel models revealed that higher scores on the mMRC Dyspnea Scale, PARS-D item 1, and VAS breathlessness were associated with higher EMA scores for “I feel short of breath” (Table 3). Specifically, a one-unit increase in mMRC Dyspnea Scale score corresponded to a 0.6-point increase in EMA score. For example, a patient with an mMRC score of 1 (the median in our sample) would be expected to report an average EMA breathlessness score of approximately 3.1 on the 7-point scale. Corresponding expected EMA scores for each questionnaire value are provided in Multimedia Appendix 2. Similarly, higher scores on CIS-Fatigue item 1, CIS-Fatigue total score, and VAS Fatigue were associated with higher EMA scores for “I feel tired.” Elevated HADS-Anxiety and VAS anxiety scores were also associated with higher EMA scores for “I feel anxious.” Conversely, higher scores on CAT item 8 were associated with lower EMA scores for “I feel energetic.”

XSL•FO
RenderX

Table . Associations between symptoms and corresponding questionnaires.

EMA[a] item and questionnaire (range)		Estimate (SE)[b]
I feel short of breath		
	mMRC[c] Dyspnea Scale (0-4)	0.600 (0.126)
	PARS-D[d] Item 1 Shortness of breath (1-10)	0.399 (0.046)
	VAS[e] breathlessness (0 - 100)	0.035 (0.004)
I feel tired		
	CIS-Fatigue[f] item 1 (I feel tired) (1-7)	0.438 (0.053)
	CIS-Fatigue[f] total score (8-56)	0.071 (0.009)
	VAS[e] Fatigue (0 - 100)	0.037 (0.005)
I feel anxious		
	HADS[g]-Anxiety (0-21)	0.166 (0.022)
	VAS[e] anxiety (0 - 100)	0.035 (0.004)
I feel energetic		
	CAT[h] Item 8 (I have a lot of energy) (0-5)	−0.602 (0.099)

[a]EMA: ecological momentary assessment.

[b]$P<.001$.

[c]mMRC: Modified Medical Research Council Dyspnea Scale.

[d]PARS-D: Physical Activity Rating Scale-Dyspnea Questionnaire.

[e]VAS: visual analog scale.

[f]CIS-Fatigue: Checklist Individual Strength-subscale Subjective Fatigue.

[g]HADS: Hospital Anxiety and Depression Scale.

[h]CAT: COPD Assessment Test.

Variability

The ICC for EMA "I feel short of breath" was 0.624, indicating that 62.4% of the variance in EMA scores was explained by interindividual differences, while 37.6% was attributable to intraindividual fluctuations over time. Time of day (expressed as hours since 8:00 AM) had a significant negative association with EMA breathlessness ratings ($\beta=-1.3\times10^{-2}$, $P=.03$; Table 4), with scores being highest in the morning and declining throughout the day. Additionally, scores on day 2 were significantly lower than day 1 ($\beta=-1.6\times10^{-1}$, $P=.03$), whereas differences for subsequent days were not statistically significant.

XSL•FO
RenderX

Table . Intraclass correlation coefficient (ICC) and associations between the ecological momentary assessment (EMA) symptoms, time of day, and day number, with the first day being the reference day.

EMA item[a]	ICC[b]	Estimate (SE)				
		Time of day[c]	Day 2[d]	Day 3[d]	Day 4[d]	Day 5[d]
I feel short of breath	0.624	-1.3×10^{-2} (5.8×10^{-3})[e]	-1.6×10^{-1} (7.8×10^{-2})[e]	-1.4×10^{-1} (7.8×10^{-2})	-9.7×10^{-2} (7.9×10^{-2})	-1.3×10^{-1} (7.9×10^{-2})
I feel tired	0.584	3.7×10^{-2} (1.1×10^{-2})[f]	8.9×10^{-3} (9.0×10^{-2})	-8.1×10^{-3} (9.1×10^{-2})	4.4×10^{-2} (9.2×10^{-2})	-2.1×10^{-2} (9.2×10^{-2})
I feel anxious	0.683	5.2×10^{-4} (3.9×10^{-3})	-1.2×10^{-1} (5.4×10^{-2})[e]	-8.6×10^{-2} (5.4×10^{-2})	-1.1×10^{-1} (5.5×10^{-2})	-1.0×10^{-1} (5.5×10^{-2})
I feel energetic	0.605	-1.6×10^{-2} (9.1×10^{-3})	7.6×10^{-2} (7.7×10^{-2})	6.0×10^{-2} (7.7×10^{-2})	9.4×10^{-3} (7.8×10^{-2})	1.8×10^{-2} (7.8×10^{-2})

[a]EMA: ecological momentary assessment.

[b]The ICC represents the proportion of total variance in EMA scores attributable to interindividual differences, after accounting for temporal effects.

[c]Estimates (β) for time of day indicate the change in symptom rating per hour since 8:00 AM. Positive β values indicate increasing symptom ratings per hour and negative β values indicate decreasing ratings. SEs are provided in parentheses.

[d]Estimates for days 2-5 represent the difference in symptom ratings compared to the first day (reference).

[e]$P<.05$.

[f]$P<.01$.

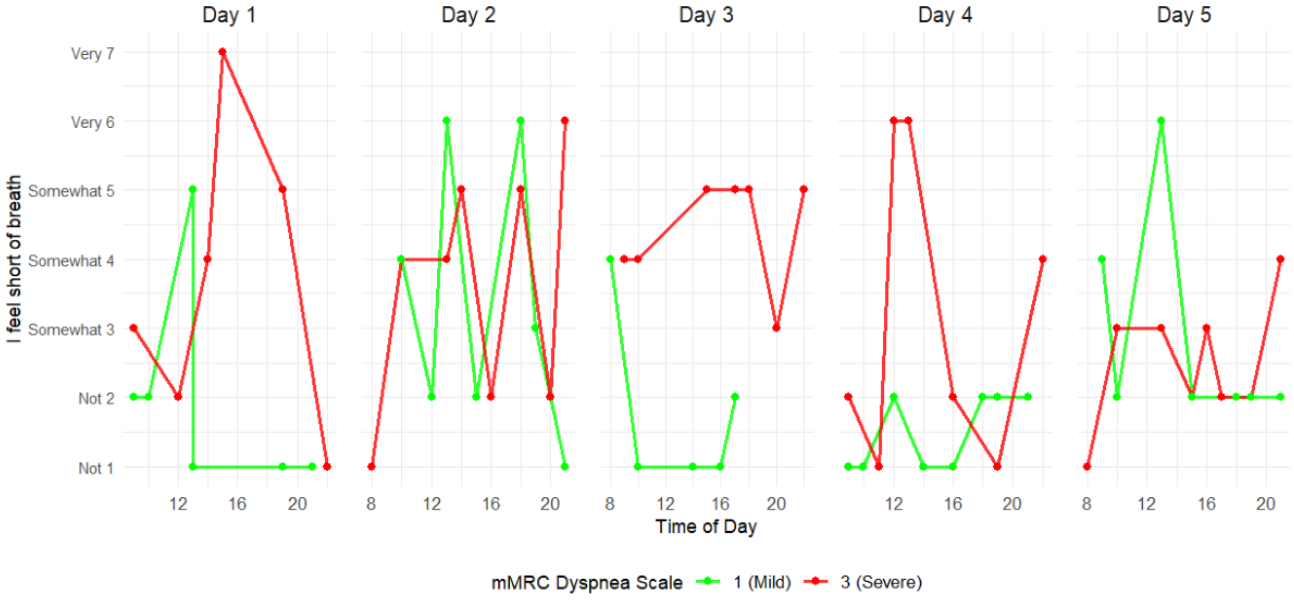
Figure 3 illustrates EMA data for "I feel short of breath" for 2 patients with low and high mMRC grades. The figure demonstrates substantial intraday and interday variability in scores for both patients, showing that low and high EMA scores occurred regardless of the mMRC dyspnea grade.

For EMA "I feel tired," the ICC was 0.584. Time of day had a significant effect on EMA scores ($\beta=3.7\times10^{-2}$, $P=.002$), where scores were lowest in the morning (≈8:00 AM) and increased during the day. Day number had no significant effect. Figure 4 depicts daily intraday variations in EMA scores for patients in both normal, mild, and severe CIS-Fatigue groups.

ICC for EMA "I feel anxious" showed the highest interindividual variability, which was 0.683. Day 2 scores were significantly lower than day 1 ($\beta=-1.2\times10^{-1}$, $P=.02$), but there was no significant effect of time of day. Figure 5 highlights intraday fluctuations in EMA scores across all patient groups, irrespective of their HADS subgroup.

For EMA "I feel energetic," the ICC was 0.605. Neither time of day nor day number showed significant effects in the model.

Figure 3. Five consecutive days of ecological momentary assessment breathlessness scores of 2 patients from different modified Medical Research Council (mMRC) Dyspnea subgroups. Each point represents a momentary breathlessness rating (1="not at all breathless" to 7="very breathless") collected across 8 random prompts per day. The green line is of a patient with low breathlessness severity according to the mMRC Dyspnea Scale, and the red line is of a patient with high breathlessness severity.



XSL•FO
RenderX

Figure 4. Five consecutive days of ecological momentary assessment fatigue scores of three patients from different Checklist Individual Strength-subscale Subjective Fatigue (CIS-Fatigue) subgroups. Each point represents a momentary tiredness rating (1="not at all tired" to 7="very tired") collected across 8 random prompts per day. The green line is of a patient with normal fatigue according to the CIS-Fatigue, the orange line is of a patient with mild fatigue, and the red line is of a patient with severe fatigue.

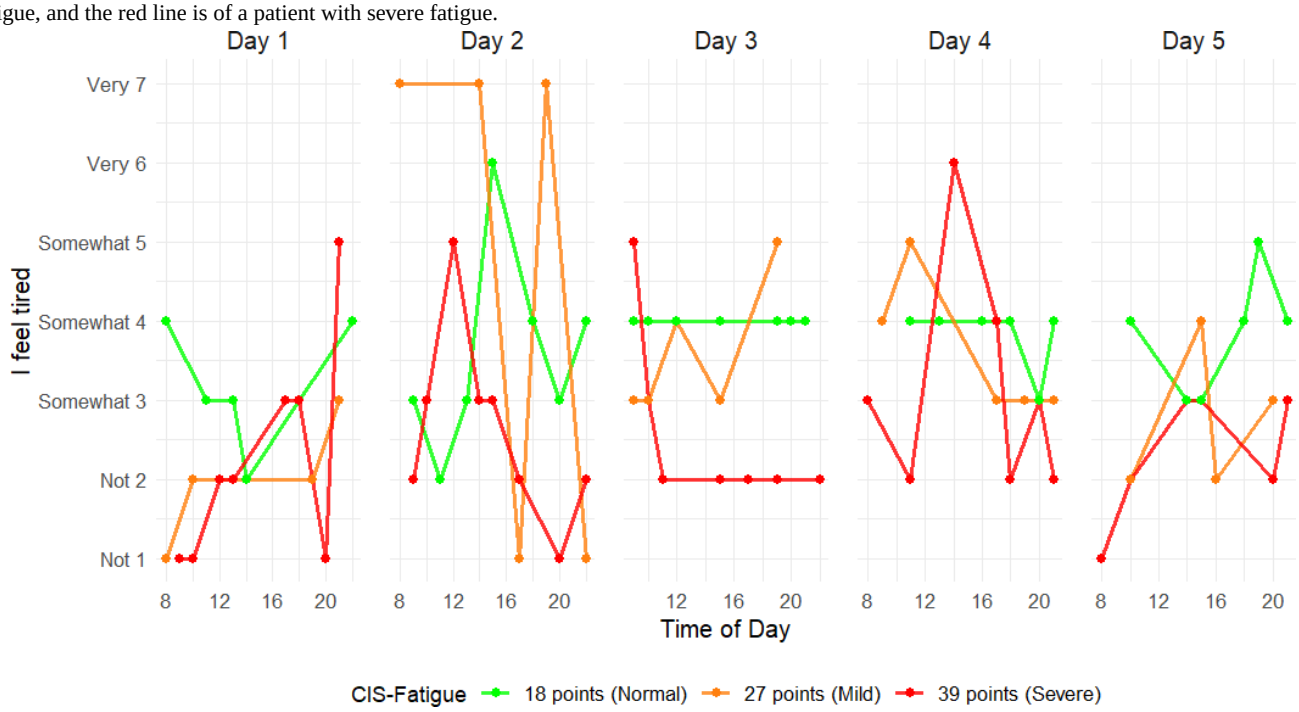
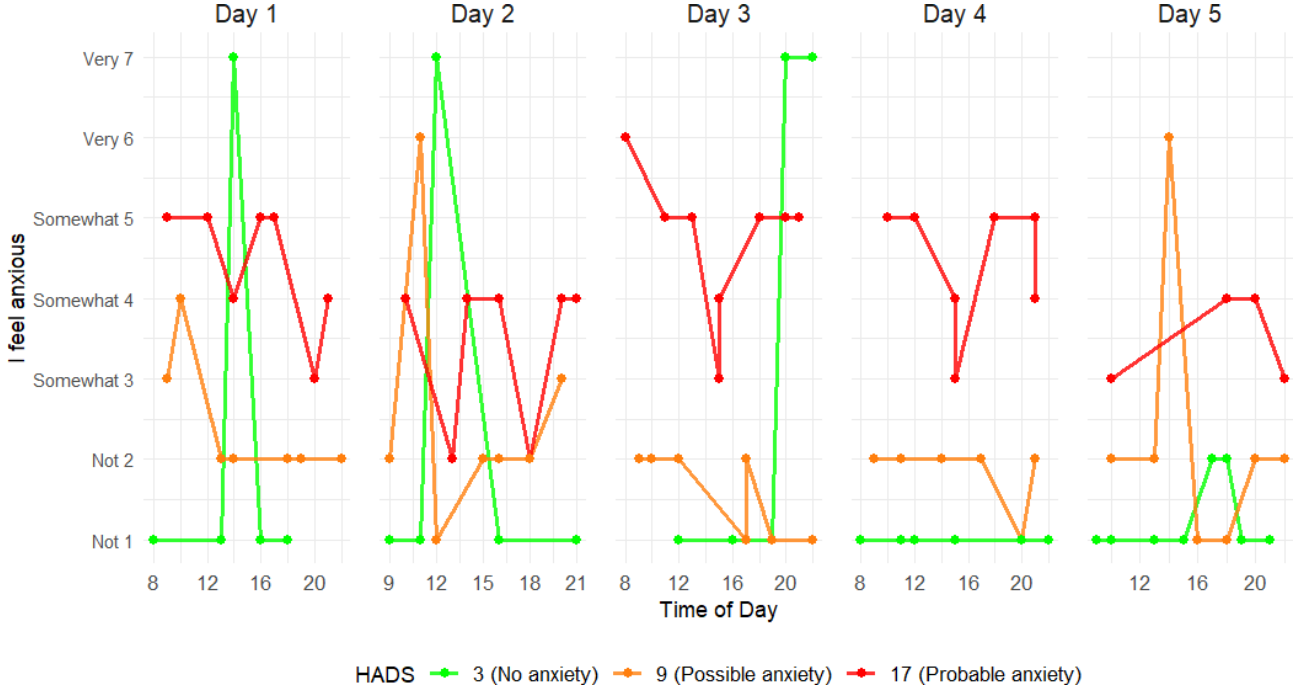


Figure 5. Five consecutive days of ecological momentary assessment anxiety scores of 3 patients from different Hospital Anxiety and Depression Scale (HADS) subgroups. Each point represents a momentary anxious rating (1="not at all anxious" to 7="very anxious") collected across 8 random prompts per day. The green line is a patient with no anxiety symptoms according to the HADS, the orange line is a patient with possible symptoms of anxiety, and the red line is a patient with probable symptoms of anxiety.



Discussion

Principal Findings

This study aimed to assess convergent validity of EMA symptom ratings by exploring associations between measuring symptoms with established questionnaires and EMA in patients with COPD. Moderate-to-strong correlations and significant associations support the convergent validity of EMA symptom ratings, with EMA effectively differentiating symptom severity and capturing intraday and interday variability which is present in the current study population.

Questionnaires provide a retrospective single-point observation that summarizes symptoms experienced over the previous weeks or months, whereas EMA collects real-time data in patients'

XSL•FO
RenderX

natural environment, minimizing recall bias [8]. While questionnaires may rely on patients' general recollection of symptom burden, EMA focuses on the immediate context in which symptoms are experienced, offering a complementary perspective. Our findings indicate that both assessment tools measure symptoms accordingly; however, EMA can measure variability. Interestingly, the moderate-to-strong correlations observed in our study align with findings from previous studies conducted in different populations, which reported similar relationships between EMA and retrospective self-report measures in assessing clinical symptoms and behaviors [30-33]. This consistency across studies suggests that, despite their methodological differences, both approaches approximate the same symptom burden over time. However, momentary assessments using EMA may also capture subtle fluctuations influenced by contextual factors, which retrospective self-reports may overlook [34].

The substantial interday and intraday variability observed in EMA scores for breathlessness, fatigue, and anxiety highlights a key advantage of EMA over traditional assessments [16]. This emphasizes that traditional self-report measures often fail to capture such variability due to recall biases and the aggregation of experiences over extended timeframes. In the study of Solhan et al [16], EMA indices only showed modest agreement with retrospective reports of affective instability, further underscoring the limitations of relying solely on retrospective assessment. Similarly, our findings showed that patients with low mMRC, CIS-Fatigue, and HADS scores sometimes reported high EMA scores of the corresponding symptoms, illustrating the heterogeneity of symptoms intraindividual and across different time points [10].

Clinical Practice and Research

EMA's ability to capture real-time symptom variability provides clinicians with a richer and more nuanced dataset to inform personalized treatment plans. For example, when patients report a consistent pattern of increasing fatigue through EMA, this could lead to targeted behavioral strategies such as activity pacing (eg, scheduling lower-demand activities during that period) and timing of short rest breaks. Similarly, if EMA reveals increased feelings of anxiety around specific situations (eg, during exertion, social activity, or waking up), then this could trigger timely psychological-support interventions such as brief monitored relaxation techniques, guided breathing exercises, or preemptive alert via a mobile app to initiate mindfulness at forecasted high-risk times. In pulmonary rehabilitation, incorporating EMA can help patients gain insight into how daily lifestyle behaviors trigger certain symptoms and identify windows of vulnerability when symptoms tend to worsen. By visualizing the intraday and interday symptom fluctuations, clinicians and patients can co-develop personalized coping strategies such as adjusting nonpharmacological tactics (eg, paced walking, energy-conserving breathing techniques) during windows of higher symptom burden to remind patients to apply these tactics when indicated. EMA has been shown to enhance the detection of treatment effects, as evidenced in studies where EMA demonstrated superior sensitivity to change and convergence with structured clinical assessments [9,35]. By incorporating EMA alongside traditional measures, researchers can achieve a more comprehensive understanding of symptom variability and treatment efficacy. A potential drawback is that patient burden for completing EMA could be high, particularly for patients who are not very technically proficient or not stable [16].

Methodological Considerations and Limitations

Both assessment tools demonstrated high response rates. While the EMA response rate in this study was comparable to previous research, adherence remains a critical factor for its successful implementation. Strategies to improve compliance, such as financial incentives or tailored EMA schedules, could enhance data quality [9]. Additionally, exclusion of patients with insufficient EMA data underscores the importance of establishing minimum response thresholds to ensure data reliability. Another limitation is that the self-reported symptoms may still be influenced by momentary factors such as mood, context, or tasks, even with EMA. The integration of EMA with objective measures, such as actigraphy or physiological monitoring, could provide a better understanding of how symptoms relate to specific activities and daily behaviors. This approach gives a more nuanced and objective assessment of symptom burden, capturing the dynamic interplay between symptoms and activity levels and potentially enhancing the validity of findings [15]. The median EMA score was used to summarize symptom burden across the 5-day period, which reduces the influence of extreme values and aligns with questionnaire measures. However, this approach does not capture fluctuations at specific times of day, such as morning or evening peaks. Future studies could explore time-specific EMA scores to provide more detailed insight into diurnal symptom variability. Additionally, symptom experiences often differ between sexes, and the predominance of male participants in this study may have influenced the results [36,37]. Furthermore, selection bias should also be considered, as participants who are more technologically proficient or motivated may be more likely to adhere to EMA protocols [38]. This may limit generalizability, particularly among older adults or those with severe COPD who could find frequent digital reporting burdensome [39]. Future studies should explore adaptive EMA designs or simplified interfaces to minimize this barrier and broaden participation. Despite these limitations, convergent validity remained high, even though EMA assessments were conducted after patients completed the questionnaires. Patients may have experienced a better or worse week in terms of symptoms before or during the EMA assessment period. Notably, some questionnaires assess symptoms over longer time frames, such as the past month or two weeks, which may capture a different temporal perspective on symptom burden. Another important methodological aspect concerns potential recall bias inherent in retrospective questionnaires. Traditional self-report instruments require participants to summarize symptoms experienced over extended periods, which may lead to distortions due to memory limitations or recency effects [12,13]. EMA mitigates this limitation by collecting data in real time, reducing recall-related error and providing a more accurate temporal representation of symptom burden. Furthermore, this study used single items from validated questionnaires (eg, PARS-D item 1, CIS-Fatigue item 1, CAT

XSL•FO
RenderX

item 8). Although these items originate from instruments with well-established psychometric properties, individual items may not fully represent the multidimensional constructs assessed by the complete questionnaires. Together, these findings underscore the complementary strengths of EMA to provide a more nuanced understanding of symptom variability while still aligning with traditional measures of symptom severity.

Conclusions

This study demonstrates that EMA is a validated tool for symptom assessment in patients with COPD. EMA showed moderate-to-strong correlations and significant associations with established questionnaires while capturing both interday and intraday symptom variability. The ability to reflect real-time fluctuations provides a more detailed and dynamic understanding of daily symptom burden, underscoring EMA's added value for enhancing symptom monitoring and informing personalized management in COPD.

Funding

This project is supported by grant 4.1.16.085 from Lung Foundation Netherlands, Leusden, the Netherlands; AstraZeneca Netherlands; Boehringer Ingelheim Netherlands; and Stichting Astma Bestrijding, Amsterdam, the Netherlands.

Data Availability

Participants in this study did not provide written consent for their data to be shared publicly. Therefore, due to the sensitive nature of the research, supporting data are not publicly available.

Authors' Contributions

Conceptualization: MVH, YMJG, ZE, MSYT, CB, JBP, EFMW, MAGS, JHV, DJAJ, MAS
Data curation: MVH, YMJG, ZE, JBP, RTMS, EWMAB, JHV, AWV
Formal analysis: BP, QD
Methodology: MVH, YMJG, ZE, MSYT, CB, JBP, EFMW, MAGS, JHV, DJAJ, MAS
Writing – original draft: BP, QD, SH-W, DJAJ, MAS
Writing – review & editing: MVH, YMJG, ZE, MSYT, CB, JBP, RTMS, EWMAB, EFMW, MAGS, JHV, AWV, SH-W, DJAJ, MAS

Conflicts of Interest

MVH and YMJG report research grant 4.1.16.085 from Lung Foundation Netherlands. MSYT reports research grant from Lung Foundation Netherlands. RTMS reports payment from GSK for lectures and from PENTAX Medical. EWMAB reports research grants from The Dutch Research Council (NWO) and The Netherlands Organization for Health Research and Development (ZonMw). DJAJ reports research grants from Dutch Foundation for Asthma Prevention and the Netherlands Respiratory Society (award), payment from Chiesi, Abbott, and AstraZeneca (all paid to the institution). MAS reports grants from Lung Foundation Netherlands, Stichting Astma Bestrijding, Boehringer Ingelheim, AstraZeneca, GSK, TEVA, Chiesi, Sanofi—all falling outside the scope of the submitted work—were paid to Ciro. MAS reports consulting fees from AMES B.V., O2-matic, and GSK, all outside the scope of the submitted work, were paid to Care2Know B.V. MAS is the owner of Care2Know B.V. BP, ZE, QD, CB, JBP, EW, MAGS, JHV, SH-W, and AWV declare that they have no conflicts of interest.

Multimedia Appendix 1
Median ecological momentary assessment (EMA) symptom scores and the total number of EMA responses per patient over 5 consecutive days.
[DOCX File, 29 KB - medinform_v14i1e79001_app1.docx]

Multimedia Appendix 2
Minimum, median, and maximum EMA symptom scores derived from questionnaire scores using the specified formulas.
[DOCX File, 26 KB - medinform_v14i1e79001_app2.docx]

References

1. Miller SN, Nichols M, Teufel RJ, Silverman EP, Walentynowicz M. Use of ecological momentary assessment to measure dyspnea in COPD. Int J Chron Obstruct Pulmon Dis 2024;19:841-849. [doi: 10.2147/COPD.S447660] [Medline: 38566847]
2. Venkatesan P. GOLD COPD report: 2025 update. Lancet Respir Med 2025 Jan;13(1):e7-e8. [doi: 10.1016/S2213-2600(24)00413-2] [Medline: 39647487]
3. Goërtz YMJ, Spruit MA, Van 't Hul AJ, et al. Fatigue is highly prevalent in patients with COPD and correlates poorly with the degree of airflow limitation. Ther Adv Respir Dis 2019;13:1753466619878128. [doi: 10.1177/1753466619878128] [Medline: 31558115]

XSL•FO
RenderX

4. Janssen DJA, Spruit MA, Leue C, et al. Symptoms of anxiety and depression in COPD patients entering pulmonary rehabilitation. Chron Respir Dis 2010 Aug;7(3):147-157. [doi: 10.1177/1479972310369285] [Medline: 20688892]
5. Houben-Wilke S, Deng Q, Janssen DJA, Franssen FME, Spruit MA. Symptom burden and its associations with clinical characteristics in patients with COPD: a clustering approach. ERJ Open Res 2024 Jul;10(4):01052-2023. [doi: 10.1183/23120541.01052-2023] [Medline: 39104954]
6. Vercoulen JH, Swanink CM, Fennis JF, Galama JM, van der Meer JW, Bleijenberg G. Dimensional assessment of chronic fatigue syndrome. J Psychosom Res 1994 Jul;38(5):383-392. [doi: 10.1016/0022-3999(94)90099-x] [Medline: 7965927]
7. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
8. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
9. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016 Apr;75:116-123. [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
10. Wu M, Wang Z, Li M, Li K. Daily symptom variability in patients with stable COPD: a narrative review. West J Nurs Res 2018 Oct;40(10):1543-1561. [doi: 10.1177/0193945917705132] [Medline: 28459177]
11. Mesquita R, Spina G, Pitta F, et al. Physical activity patterns and clusters in 1001 patients with COPD. Chron Respir Dis 2017 Aug;14(3):256-269. [doi: 10.1177/1479972316687207] [Medline: 28774199]
12. Sandberg J, Ekström M. Differences between experienced and recalled breathlessness: a review. Curr Opin Support Palliat Care 2019 Sep;13(3):161-166. [doi: 10.1097/SPC.0000000000000448] [Medline: 31365460]
13. Bennett AV, Amtmann D, Diehr P, Patrick DL. Comparison of 7-day recall and daily diary reports of COPD symptoms and impacts. Value Health 2012 May;15(3):466-474. [doi: 10.1016/j.jval.2011.12.005] [Medline: 22583457]
14. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994 Sep 1;16(3):199-202. [doi: 10.1093/abm/16.3.199]
15. Tamm J, Takano K, Just L, Ehring T, Rosenkranz T, Kopf-Beck J. Ecological momentary assessment versus weekly questionnaire assessment of change in depression. Depress Anxiety 2024;2024(1):9191823. [doi: 10.1155/2024/9191823] [Medline: 40226649]
16. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol Assess 2009 Sep;21(3):425-436. [doi: 10.1037/a0016869] [Medline: 19719353]
17. Goërtz YMJ, Looijmans M, Prins JB, et al. Fatigue in patients with chronic obstructive pulmonary disease: protocol of the Dutch multicentre, longitudinal, observational FAntasTIGUE study. BMJ Open 2018 Apr 10;8(4):e021745. [doi: 10.1136/bmjopen-2018-021745] [Medline: 29643168]
18. Palmen B, Ebadi Z, van Herck M, et al. Bidirectional associations between physical activity, sedentary behavior, and daily symptoms in patients with chronic obstructive pulmonary disease: longitudinal observational study. JMIR Mhealth Uhealth 2025 Oct 22;13:e65653. [doi: 10.2196/65653] [Medline: 41124569]
19. Van Herck M, Goërtz YMJ, Ebadi Z, et al. Associated factors of fatigue in patients with COPD: results from the FAntasTIGUE study. ERJ Open Res 2025 Sep;11(5):00857-02024. [doi: 10.1183/23120541.00857-2024] [Medline: 41089566]
20. Metsemakers JF, Höppener P, Knottnerus JA, Kocken RJ, Limonard CB. Computerized health information in The Netherlands: a registration network of family practices. Br J Gen Pract 1992 Mar;42(356):102-106. [Medline: 1493025]
21. Bestall JC, Paul EA, Garrod R, Garnham R, Jones PW, Wedzicha JA. Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. Thorax 1999 Jul;54(7):581-586. [doi: 10.1136/thx.54.7.581] [Medline: 10377201]
22. Peters JB, Daudey L, Heijdra YF, Molema J, Dekhuijzen PNR, Vercoulen JH. Development of a battery of instruments for detailed measurement of health status in patients with COPD in routine care: the Nijmegen Clinical Screening Instrument. Qual Life Res 2009 Sep;18(7):901-912. [doi: 10.1007/s11136-009-9502-2] [Medline: 19543807]
23. Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. Pain 1983 May;16(1):87-101. [doi: 10.1016/0304-3959(83)90088-X] [Medline: 6602967]
24. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain 2003 Sep;4(7):407-414. [doi: 10.1016/s1526-5900(03)00716-8] [Medline: 14622683]
25. Jones PW, Harding G, Berry P, Wiklund I, Chen WH, Kline Leidy N. Development and first validation of the COPD assessment test. Eur Respir J 2009 Sep;34(3):648-654. [doi: 10.1183/09031936.00102509] [Medline: 19720809]
26. Vercoulen JH, Daudey L, Molema J, et al. An integral assessment framework of health status in chronic obstructive pulmonary disease (COPD). Int J Behav Med 2008;15(4):263-279. [doi: 10.1080/10705500802365474] [Medline: 19005926]
27. Peters JB, Heijdra YF, Daudey L, et al. Course of normal and abnormal fatigue in patients with chronic obstructive pulmonary disease, and its relationship with domains of health status. Patient Educ Couns 2011 Nov;85(2):281-285. [doi: 10.1016/j.pec.2010.08.021] [Medline: 20884162]

XSL•FO
RenderX

28. Palmier-Claus JE, Myin-Germeys I, Barkus E, et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 2011 Jan;123(1):12-20. [doi: 10.1111/j.1600-0447.2010.01596.x] [Medline: 20712828]
29. Dodge Y. Spearman rank correlation coefficient. In: The Concise Encyclopedia of Statistics: Springer; 2008:502-505. [doi: 10.1007/978-0-387-32833-1_379]
30. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: a prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord 2023 Mar 1;324:170-174. [doi: 10.1016/j.jad.2022.12.127] [Medline: 36586594]
31. McIntyre RS, Lee Y, Rong C, et al. Ecological momentary assessment of depressive symptoms using the mind.me application: convergence with the Patient Health Questionnaire-9 (PHQ-9). J Psychiatr Res 2021 Mar;135:311-317. [doi: 10.1016/j.jpsychires.2021.01.012] [Medline: 33540296]
32. Bucci R, Manfredini D, Lenci F, Simeon V, Bracci A, Michelotti A. Comparison between ecological momentary assessment and questionnaire for assessing the frequency of waking-time non-functional oral behaviours. J Clin Med 2022 Oct 5;11(19):5880. [doi: 10.3390/jcm11195880] [Medline: 36233747]
33. Harvey PD, Miller ML, Moore RC, Depp CA, Parrish EM, Pinkham AE. Capturing clinical symptoms with ecological momentary assessment: convergence of momentary reports of psychotic and mood symptoms with diagnoses and standard clinical assessments. Innov Clin Neurosci 2021;18(1-3):24-30. [Medline: 34150360]
34. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess 2009 Dec;21(4):463-475. [doi: 10.1037/a0017075] [Medline: 19947781]
35. Oreel TH, Delespaul P, Hartog ID, et al. Ecological momentary assessment versus retrospective assessment for measuring change in health-related quality of life following cardiac intervention. J Patient Rep Outcomes 2020 Nov 16;4(1):98. [doi: 10.1186/s41687-020-00261-2] [Medline: 33196959]
36. Han MK, Postma D, Mannino DM, et al. Gender and chronic obstructive pulmonary disease: why it matters. Am J Respir Crit Care Med 2007 Dec 15;176(12):1179-1184. [doi: 10.1164/rccm.200704-553CC] [Medline: 17673696]
37. Zhang H, Wu F, Yi H, et al. Gender differences in chronic obstructive pulmonary disease symptom clusters. Int J Chron Obstruct Pulmon Dis 2021;16:1101-1107. [doi: 10.2147/COPD.S302877] [Medline: 33907396]
38. Stone AA, Schneider S, Smyth JM, et al. Shedding light on participant selection bias in ecological momentary assessment (EMA) studies: findings from an internet panel study. PLoS One 2023;18(3):e0282591. [doi: 10.1371/journal.pone.0282591] [Medline: 36893179]
39. Wang S, Yang CH, Brown D, Cheng A, Kwan MYW. Participant compliance with ecological momentary assessment in movement behavior research among adolescents and emerging adults: systematic review. JMIR Mhealth Uhealth 2025 Feb 11;13:e52887. [doi: 10.2196/52887] [Medline: 39933165]

Abbreviations:

CAT: Chronic Obstructive Pulmonary Disease Assessment Test
CIS-Fatigue: Checklist Individual Strength-subscale Subjective Fatigue
COPD: chronic obstructive pulmonary disease
EMA: ecological momentary assessment
HADS: Hospital Anxiety and Depression Scale
ICC: intraclass correlation coefficient
MLM: multilevel linear mixed model
mMRC: modified Medical Research Council
PARS-D: Physical Activity Rating Scale-Dyspnea questionnaire
VAS: visual analog scale

Edited by A Benis; submitted 14.Jun.2025; peer-reviewed by E Felekoglu, H Demeyer; revised version received 27.Jan.2026; accepted 05.Feb.2026; published 07.Apr.2026.

Please cite as:
Palmen B, van Herck M, Goërtz YMJ, Ebadi Z, Deng Q, Thong MSY, Burtin C, Peters JB, Sprooten RTM, Bischoff EWMA, Wouters EFM, Vercoulen JH, Houben-Wilke S, Janssen DJA, Spruit MA, Vaes AW
Real-Time Symptom Ratings Using Ecological Momentary Assessment Versus Traditional Questionnaires in Patients With Chronic Obstructive Pulmonary Disease: Observational Study
JMIR Med Inform 2026;14:e79001
URL: https://medinform.jmir.org/2026/1/e79001
doi:10.2196/79001

XSL•FO
RenderX

© Banchia Palmen, Maarten van Herck, Yvonne M J Goërtz, Zjala Ebadi, Qichen Deng, Melissa S Y Thong, Chris Burtin, Jeannette B Peters, Roy T M Sprooten, Erik W M A Bischoff, Emiel F M Wouters, Jan H Vercoulen, Sarah Houben-Wilke, Daisy J A Janssen, Martijn A Spruit, Anouk W Vaes. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

CT Radiomics–Based Machine Learning Model for Predicting Capsular and Neural Invasion in Thyroid Carcinoma: Diagnostic Accuracy Study

Fang-fang Cong[1*], MM; Ke Tian[2*], MM; Qian Gao[2*], MB; Fulin Wang[3], MM; Peng Sun[2], MM; Nan Xu[2], MD

[1]Department of MRI, Zhoukou Medical Science Research Center, Zhoukou Central Hospital, Zhoukou, China

[2]Department of Radiology, Air Force Medical Center, Air Force Medical University, Number 30 Fucheng Road, Haidian District, Beijing, China

[3]Department of CT, Zhoukou Medical Science Research Center, Zhoukou Central Hospital, Zhoukou, China

[*]these authors contributed equally

Corresponding Author:
Nan Xu, MD
Department of Radiology, Air Force Medical Center, Air Force Medical University, Number 30 Fucheng Road, Haidian District, Beijing, China

Abstract

Background: Thyroid carcinoma is the most prevalent endocrine malignancy, with a worldwide increasing incidence. Capsular invasion and neural invasion (NI) are pivotal prognostic factors for recurrence and survival; however, their preoperative noninvasive assessment remains challenging.

Objective: We aimed to identify computed tomography (CT) radiomic biomarkers associated with capsular invasion in thyroid carcinoma, construct machine learning models integrating radiomic and clinical data, and evaluate their utility for NI risk stratification.

Methods: In this retrospective cohort, 111 patients with thyroid carcinoma were divided into capsular invasion–positive (n=63) and capsular invasion–negative (n=48) groups, with 37 (33.3%) cases presenting concurrent NI. Radiomic features were extracted from arterial and venous phase CT images at original resolution, including 111 gray-level co-occurrence matrix features. Nine key radiomic features (A1-A9) were selected via least absolute shrinkage and selection operator regression (λ=0.017). To preserve the physical meaning of texture features (eg, spatial correlation and contrast reflecting tumor microstructural heterogeneity), no resampling or scaling was performed on the regions of interest during radiomic feature extraction. Nomogram models and random forest (RF) models were constructed based on clinical indicators (galectin-3, etc) and radiomic features, respectively. Additionally, a neural network (NN) model integrating multimodal data was developed. Model stability was verified using 5-fold cross-validation and 1000-time bootstrap resampling, while performance was evaluated via receiver operating characteristic curves, calibration curves, and decision curve analysis.

Results: Model performance analysis revealed that among the nomogram models, the clinical indicator-based nomogram achieved an internally estimated area under the curve (AUC) of 0.9418 (95% CI 0.892 - 0.976) in the capsular invasion prediction task. The radiomic-based nomogram had an internally estimated AUC of 0.9334 (95% CI 0.881 - 0.968) in the capsular invasion prediction task and 0.8001 (95% CI 0.663 - 0.898) in the cross-label association analysis task. In RF models, clinical indicator-based and radiomic-based RFs exhibited an AUC of 0.7646 (95% CI 0.651 - 0.857) and 0.8102 (95% CI 0.703 - 0.892) in the cross-label association analysis task, respectively. The NN model performed promisingly, with an AUC of 0.775 (95% CI 0.621 - 0.903) in the cross-label association analysis task and a mean absolute error of <0.05 on the calibration curve.

Conclusions: Capsular invasion is a strong predictor of NI risk in thyroid carcinoma. Radiomic models based solely on preoperative CT images show potential for preoperative NI risk stratification. Models incorporating clinical parameters (obtained from postoperative tissue), including the integrated multimodal model, are more accurately characterized as postoperative risk stratification tools. The NN model, which integrated raw CT images with clinical data, achieved an AUC of 0.775 (95% CI 0.621 - 0.903), underscoring the potential of such multimodal analysis to capture complex relationships between imaging phenotypes and tissue-level biomarkers for enhanced postoperative assessment. This framework's radiomic component points toward purely image-based, preoperative evaluation tools' development.

(JMIR Med Inform 2026;14:e77349) doi:10.2196/77349

XSL•FO
RenderX

KEYWORDS

thyroid carcinoma; capsular invasion; neural invasion; tadiomics; computed tomography; CT; artificial intelligence; machine learning

Introduction

Thyroid carcinoma represents the most prevalent malignancy of the endocrine system, with a worldwide increasing incidence in recent years [1]. The biological behavior of thyroid carcinoma varies considerably. Some patients experience favorable outcomes, while others face higher risks of recurrence and metastasis [2]. Accurate assessment of tumor aggressiveness and prognostic factors is critical for developing personalized therapeutic strategies and improving patients' quality of life. Capsular invasion and neural invasion (NI) are key prognostic determinants in thyroid carcinoma. Research has demonstrated that thyroid carcinoma patients with capsular invasion or NI exhibit significantly higher postoperative recurrence and mortality rates compared to those without such invasive features [3]. According to previous studies, thyroid carcinoma biomarkers galectin-3 (Gal-3), Hector Battifora Mesothelial Epitope-1 (HBME-1), and cytokeratin 19 (CK19) are often overexpressed in malignant tumors compared to benign lesions. However, whether they are related to capsular invasion and NI is unclear [4]. Radiomics, an emerging computational technology, enables high-throughput extraction of quantitative features from medical imaging data, which reflect multidimensional tumor characteristics, including morphology (eg, ranges of tumor boundaries), texture, and functional heterogeneity [5,6]. Computed tomography (CT)–based radiomic analysis holds promise for identifying imaging biomarkers closely associated with tumor biology and may provide novel insights for diagnosis and prognosis [7,8]. The development of CT-derived radiomic biomarkers linked to capsular invasion in thyroid carcinoma may facilitate early and precise prediction of tumor aggressiveness. Furthermore, clarifying the stratification value of these biomarkers for NI risk could empower clinicians to more accurately assess risk preoperatively and provide a reliable basis for targeted neuroprotective strategies [9]. Nevertheless, current research on molecular imaging biomarkers for capsular invasion in thyroid carcinoma remains in its exploratory phase, and their potential for stratifying NI risks needs further investigation. This study aims to systematically identify CT-based radiomic biomarkers associated with capsular invasion in thyroid carcinoma and comprehensively evaluate their risk stratification utility for NI, thereby advancing novel methodologies for precision oncology in thyroid carcinoma management.

Methods

Study Population

A total of 111 patients diagnosed with thyroid carcinoma and hospitalized were retrospectively enrolled. Based on the presence of capsular invasion, patients were categorized into a control group (capsular invasion–negative, n=48) and a study group (capsular invasion–positive, n=63). No significant differences in baseline characteristics were observed between the 2 groups (Table 1), ensuring comparability for subsequent analyses.

XSL•FO
RenderX

Table . Comparative analysis of baseline characteristics between patients with and without membrane invasion.[a]

		Control group (n=48)	Study group (n=63)	Statistics	*P* value
Gal-3[b], n (%)				9.3 (1)[c]	.002
	High expression	18 (37.5)	42 (66.67)		
	Low or no expression	30 (62.5)	21 (33.33)		
HBME-1[d], n (%)				1.7 (1)[c]	.19
	High expression	45 (93.75)	62 (98.41)		
	Low or no expression	3 (6.25)	1 (1.59)		
CK19[e], n (%)				29.3 (1)[c]	.001
	High expression	26 (54.17)	61 (96.83)		
	Low or no expression	22 (45.83)	2 (3.17)		
NI[f], n (%)				27.9 (1)[c]	<.001
	Yes	3 (6.25)	34 (53.97)		
	No	45 (93.75)	29 (46.03)		
CEA[g] (ng/mL), mean (SD)		31.66 (5.12)	34.91 (4.07)	−3.73（109）[h]	<.001
CCSA-2[i] (ng/mL), mean (SD)		19.26 (4.98)	19.86 (4.65)	−0.66（109）[h]	.51
CA199[j] (U/mL), mean (SD)		59.95 (11.03)	74.17 (10.07)	−7.07（109）[h]	<.001
CA125[k] (U/mL), mean (SD)		77.96 (13.75)	87.12 (14.87)	−3.32（109）[h]	.001
Age (y), mean (SD)		45.29 (12.37)	44.97 (10.83)	0.99（109）[h]	.64
BMI (kg/m^2), mean (SD)		21.83 (2.16)	22.51 (2.17)	−1.49（109）[h]	.14

[a]High expression was defined as ≥2-fold expression relative to peripheral blood mononuclear cell controls.
[b]Gal-3: galectin-3.
[c]Chi-square test.
[d]HBME-1: Hector Battifora Mesothelial Epitope-1.
[e]CK19: cytokeratin 19.
[f]NI: neural invasion.
[g]CEA: carcinoembryonic antigen.
[h]*t* test.
[i]CCSA-2: circulating cancer-specific antigen.
[j]CA199: carbohydrate antigen 199.
[k]CA125: carbohydrate antigen 125.

Inclusion criteria were as follows: (1) diagnosis of thyroid carcinoma confirmed by postoperative histopathological examination and compliance with the SEOM-GETNE-TTCC （ Spanish Society of Medical Oncology, Spanish Group for Neuroendocrine and Endocrine Tumors, and Spanish Group for Head and Neck Tumors） Clinical guideline for thyroid cancer (2023) [10]; (2) preoperative neck CT scan; (3) availability of complete clinical records, including postoperative histopathological reports explicitly documenting capsular invasion and NI status; and (4) high-quality imaging data with standardized acquisition protocols, clearly delineated tumor boundaries, and absence of motion artifacts. Exclusion criteria included having (1) concurrent diagnosis of other malignancies or (2) severe hepatic or renal dysfunction, active autoimmune diseases, or pregnancy or lactation.

Ethical Considerations

This study was approved by the Ethics Committee of Air Force Medical Center, Air Force Medical University (No. 2025‐47-YJ01) and adhered to the principles outlined in the Declaration of Helsinki. The ethics committee specifically evaluated and approved a waiver of informed consent for this retrospective study, considering the use of anonymized imaging and clinical data and the impracticability of obtaining individual consent. All data were anonymized and deidentified before analysis to ensure patient privacy and confidentiality. Participants did not receive any compensation.

Imaging Feature Extraction

CT scans were performed on 2 scanner models: a GE Discovery energy-spectrum CT (GE Healthcare) and a Siemens SOMATOM Definition Flash dual source CT (IncSiemens Healthineers). Scanner parameters were as follows: GE

XSL•FO
RenderX

parameters: 120 kV, automatic tube current (250 - 465 mA), 512 × 512 matrix, pitch 0.9. Siemens parameters: 120 kV, reference 210 mAs (CARE Dose4D), 512 × 512 matrix, pitch 0.8. All scans used iohexol contrast (2 mL/kg, 3 - 4 mL/s) with arterial/venous phases at 30/50 seconds. Contiguous 5 mm slices were acquired, and 1.25 mm thin-section reconstructions were generated for analysis.

For data harmonization, to mitigate interscanner variability, all images were reconstructed to a 1.25 mm slice thickness. For radiomic features (A1-A9), *z* score normalization was applied across the entire dataset. For deep learning input, minimum-maximum scaling to [0,1] was applied per scan.

Before feature extraction, raw CT images underwent preprocessing to enhance image quality and ensure accurate subsequent analyses. Thyroid cancer lesion boundaries were independently delineated by 2 associate chief physicians with 6 and 8 years of experience in head and neck imaging diagnosis, using ITK-SNAP (developed by the ITK-SNAP Development Team, University of Pennsylvania) software to manually delineate regions of interest (ROIs) on arterial phase CT images. For venous-phase features (A3, A5, A7, and A9), arterial-phase ROIs were propagated to venous-phase images using rigid image registration, followed by manual verification and adjustment. To verify consistency, CT images of 30 (27% of the cohort) randomly selected patients were analyzed. Intraclass correlation coefficients (ICC) were calculated to assess both ROI volume consistency and numerical consistency of core gray level cooccurrence matrix (GLCM) features (contrast, correlation, and energy; distance d=1, angle θ=0°, 45°, 90°, 135°). The ICC for ROI volume was 0.92 (95% CI 0.85 - 0.96, P<.001), and the ICC value range of GLCM core features was 0.88 - 0.94 (both P<.001), indicating excellent reproducibility. Discrepancies in 3 (10%) cases were resolved through consensus review with postoperative pathology as reference.

Dimensionality reduction was performed via principal component analysis (PCA) and linear discriminant analysis (LDA). Intensity-based features were concurrently extracted and subjected to least absolute shrinkage and selection operator (LASSO) regression to identify robust predictors. To mitigate interscan variability in gray-level intensity values caused by heterogeneous acquisition protocols, for the extraction of hand-crafted radiomic features (A1-A9), *z* score normalization was applied across the entire dataset to standardize the gray-scale distribution and eliminate interscan variability caused by heterogeneous CT acquisition protocols. It should be noted that applying normalization to the full dataset before cross-validation could lead to potential information leakage, as the mean and SD used for normalization may inadvertently include information from samples designated for validation. Although our primary aim was to compare the relative performance of different models rather than to provide absolute performance estimates, we acknowledge this methodological limitation and have further addressed its potential impact in the Discussion section. For the deep learning model (DenseNet121) input only: the original-resolution ROIs were first normalized to a [0, 1] range using minimum-maximum scaling, then resized to 224×224 pixels (to meet the fixed-size input requirement of the pretrained DenseNet121 backbone) and converted to 3-channel format by duplicating the grayscale image across RGB (red, green, and blue) channels. This 224×224 resizing step was exclusive to deep learning model training and did not involve radiomic feature extraction.

Notably on ROI resolution for feature extraction, a critical distinction was made between radiomic feature extraction and deep learning model input preparation in this study. All radiomic texture features (111 GLCM features and 9 LASSO-selected A1-A9 features) were extracted from original-resolution CT ROIs to avoid systematic scaling bias and retain the physical meaning of texture metrics. The resizing of ROIs to 224×224 pixels was a dedicated preprocessing step for the DenseNet121-based deep learning model, as pretrained convolutional neural networks (NN) typically require fixed-size inputs. This separation of technical processes ensures the reliability of both radiomic feature analysis and deep learning model training.

Model Development

Given the limited total sample size (111 patients), to maximize the use of data, a full-sample modeling approach was adopted.

Model construction: all 111 cases were used as the analysis cohort, with capsular invasion status (1=capsular invasion–positive, n=63; 0=capsular invasion–negative, n=48) as the labeling variable. Input features were tailored to each model type: a common set of clinical biomarkers (for the complete list, see Table 1) was used across all models, while pixel-level CT features A1-A9 (selected via LASSO regression from GLCM texture features of CT ROIs) were used specifically in the traditional radiomic models (nomogram and random forest [RF]).

The NN model used a multimodal input, comprising both raw arterial-phase CT ROI images and the suite of clinical biomarkers (for the complete list, see Table 1), to autonomously learn high-level spatial and clinical patterns associated with capsular invasion (Multimedia Appendix 1). It is crucial to note that the hand-crafted radiomic features (A1-A9) were not direct inputs to this NN model. For each patient, a single 2D axial slice with the largest tumor cross-section was selected, standardized to 224×224 pixels, and converted to 3-channel format. Rationale for arterial-phase only input: this single-phase design was implemented for two primary reasons: (1) to maintain model simplicity and mitigate overfitting risk given the limited sample size, and (2) because the arterial phase provides the most direct imaging correlate of tumor vascularity, a key biological process underlying invasion. We hypothesized that the convolutional NN could learn high-level features from the arterial phase that are informative for predicting outcomes potentially associated with venous-phase patterns. Rationale despite venous-phase radiomic feature importance: it is noteworthy that in our traditional radiomics analysis, a feature derived from the venous phase (A9) showed high importance. The decision to use only arterial-phase images for the deep learning model was based on a complementary hypothesis: that the convolutional NN could learn high-level, abstract features from the arterial phase that are biologically or spatially correlated with the tumor properties captured by the venous-phase radiomic signature (A9). This approach allowed

XSL•FO
RenderX

us to test whether a single, routinely acquired phase could provide sufficient information for the NN to make accurate predictions, thereby simplifying the model and its potential clinical translation.

Concerning model validation, NI status (1=NI-positive, n=37; 0=NI-negative, n=74) was used for a cross-label association analysis. The core goal was to evaluate whether the features associated with capsular invasion (as captured by our models) were also rank-informative for NI status within the same cohort. The full 111-case dataset was used for both (1) model training (with capsular invasion status as the label) and (2) cross-label association evaluation (with NI status as the label). It is crucial to note that this analysis evaluates the association between capsular invasion–related features and NI status on the same samples, primarily reflecting their strong pathological correlation rather than constituting an independent predictive validation on unseen data. The full 111-case cohort, labeled based on the presence or absence of capsular invasion, was used to construct 3 predictive models (nomogram, RF, and NN) to capture capsular invasion–associated features. The same full cohort, labeled based on the presence or absence of NI, was then used to evaluate the models' cross-predictive performance for NI risk stratification. To address potential concerns regarding the unconventional data usage design and to empirically validate our model's generalizability, we performed a supplementary traditional 7:3 hold-out validation.

For the nomogram development, clinically significant variables associated with capsular invasion were incorporated into a clinical nomogram, while gray-level features selected via LASSO regression were integrated into an imaging-based nomogram. Calibration curves (predicted probability vs observed probability) were generated to assess agreement between model predictions and observed outcomes, with internal validation performed using a bootstrap resampling method (B=1000 iterations).

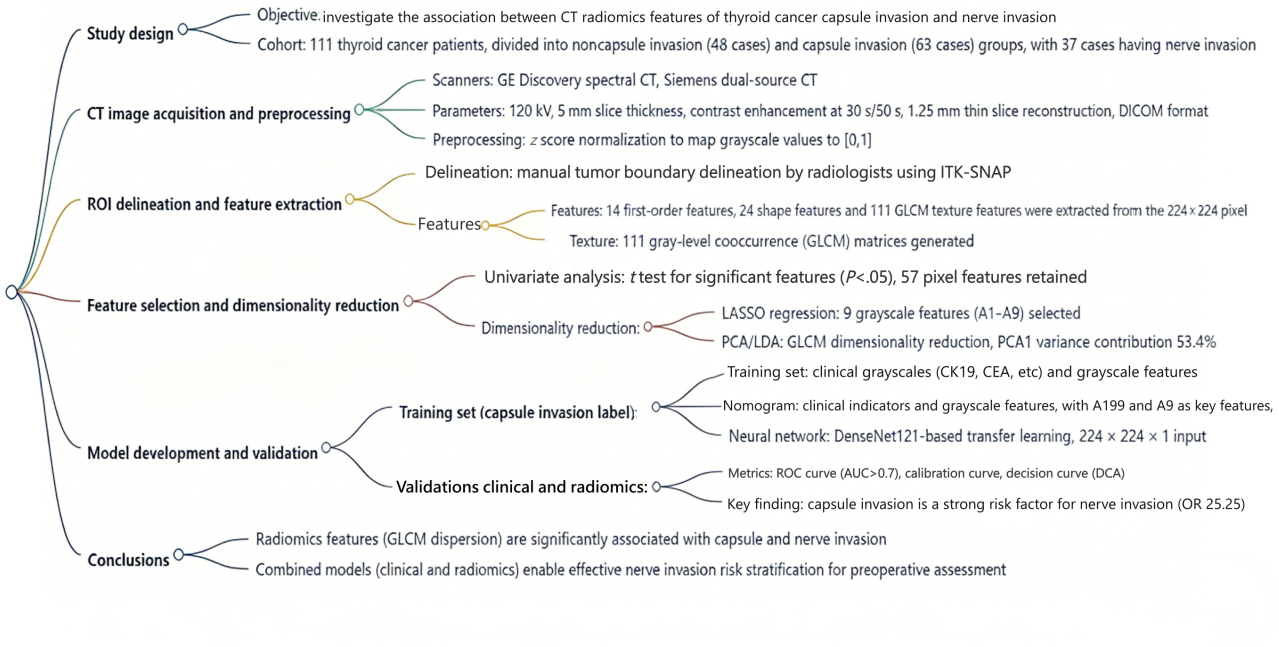
The RF model architecture comprised 2 components: a clinical RF model incorporating capsular invasion–associated variables and an imaging RF model integrating LASSO-selected radiomic features. Hyperparameter optimization was conducted via 5-fold cross-validation using the caret package, with the number of trees (ntree) tuned from 50 to 500 (step size=5) and the minimum number of features per node split (mtry) ranging from 2 to 10. Model interpretability was enhanced through feature importance plots, decision tree visualization, receiver operating characteristic (ROC) curves, and out-of-bag error analysis.

For the input of the NN model (based on DenseNet121): to comply with the fixed-size input requirement of the pretrained DenseNet121 backbone, the original-resolution ROIs were first normalized to a [0, 1] range using minimum-maximum scaling, then resized to 224×224 pixels. The single-channel grayscale images were further converted to a 3-channel format by duplication across RGB channels. The first 10 convolution blocks of DenseNet121 were reserved for extracting high-level image features, and the feature map was compressed into fixed-length vectors through global average pooling. Image feature vectors (extracted by DenseNet121) and clinical feature vectors were normalized, respectively, and directly spliced into multimodal feature vectors. The spliced features were input to the fully connected layer and were finally classified (capsular invasion: positive or negative) through the Sigmoid output layer. The 9 radiomic features (A1-A9) were selected from GLCM features of the ROI via LASSO regression. The model was optimized using the Adam optimizer (learning rate=1×10^{4}) with binary cross-entropy loss over 50 training epochs (batch size=16). Early stopping was triggered if validation loss plateaued for 10 consecutive epochs (Multimedia Appendices 1 and 2).

The technical workflow is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Workflow of CT imaging feature extraction and predictive model development. A9: venous-phase coefficient of variation; AUC: area under the curve; CA199: carbohydrate antigen 199; CEA: carcinoembryonic antigen; CK19: cytokeratin 19; CT: computed tomography; DCA: decision curve analysis; DICOM: Digital Imaging and Communications in Medicine; GE: General Electric; GLCM: gray-level co-occurrence matrix; LASSO: least absolute shrinkage and selection operator; LDA: linear discriminant analysis; OR: odds ratio; PCA: principal component analysis; ROC: receiver operating characteristic; ROI: regions of interest.

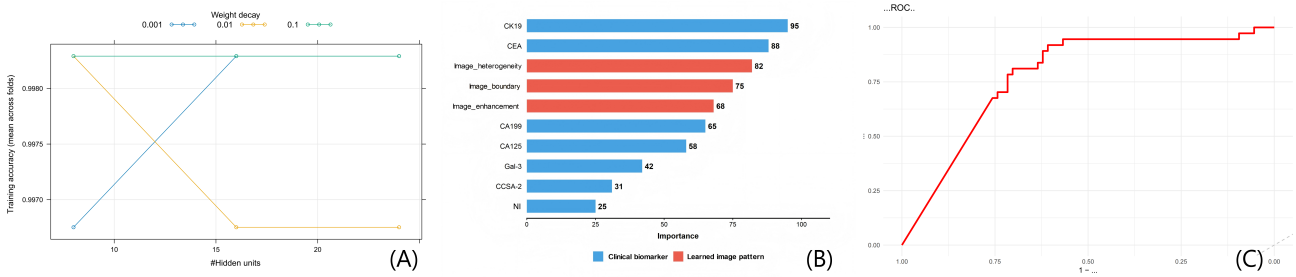


Model Interpretability Analysis

To interpret the NN's predictions and derive the importance scores in Figure 2B, we used Shapley Additive Explanations (SHAP). The analysis pipeline was as in the following paragraphs.

Concerning calculation for the validation set, SHAP values were computed for each input feature (clinical variables and the high-dimensional features extracted from raw CT images by DenseNet121). These values quantify a feature's contribution to a specific prediction.

Figure 2. Neural network model performance. (A) Mean training accuracy curves across the 5-fold cross-validation splits, plotted for different numbers of hidden units. (B) SHAP-based feature importance (normalized 0-100). Blue: clinical biomarkers (direct inputs); red: image patterns (learned from raw CT). (C) ROC curve for neural invasion prediction. Note for (B): normalized importance scores derived from SHAP analysis (see subsection in Methods: Model interpretability analysis). Imaging patterns were semantically annotated based on activation map visualizations. CA125: carbohydrate antigen 125; CA199: carbohydrate antigen 199; CCSA-2: circulating cancer-specific antigen; CEA: carcinoembryonic antigen; CK19: cytokeratin 19; CT: computed tomography; Gal-3: galectin-3; NI: neural invasion; NN: neural network; ROC: receiver operating characteristic; SHAP: Shapley Additive Explanations.



For aggregation and semantic mapping, global importance for each clinical variable was defined as its mean absolute SHAP value. For image features, activation maps corresponding to features with high mean absolute SHAP values were analyzed. Recurring visual patterns were inductively categorized into 3 semantic concepts: "Image_Heterogeneity," "Image_Boundary," and "Enhancement_Dynamics." The importance score for each concept was calculated as the mean absolute SHAP value of its constituent image feature dimensions.

In normalization, all resulting scores (for clinical variables and image concepts) were linearly normalized to a 0 - 100 scale for comparative visualization in Figure 2B.

Data Collection

All thyroid carcinoma tissue specimens were collected immediately after surgical resection, and 10% neutral buffered formalin was used for fixation within 30 minutes for morphological preservation. After paraffin embedding, 5-µm-thick sections were prepared, and tumor tissue areas (with

XSL•FO
RenderX

tumor cell content ≥80%, confirmed by 2 pathologists) were microdissected for quantitative polymerase chain reaction (qPCR) analysis. This ensured that the detected expression of Gal-3, HBME-1, and CK19 accurately reflected the biological characteristics of thyroid carcinoma cells.

Real-time fluorescence qPCR was performed to measure the levels of Gal-3, HBME-1, and CK19 in thyroid carcinoma tissue specimens from the 111 enrolled patients. Glyceraldehyde-3-phosphate dehydrogenase served as the endogenous reference gene for normalization. Peripheral blood mononuclear cells (PBMCs) from age- and sex-matched healthy individuals undergoing routine health screenings were used as the control group for qPCR analysis. Relative gene expression levels ≥2-fold compared to PBMC controls were pragmatically defined as high expression for subsequent categorical analysis. It is important to note that the use of PBMCs as a control group is a methodological consideration due to the practical and ethical challenges in obtaining matched healthy thyroid tissue samples. As PBMCs (of nonepithelial origin) and thyroid epithelial cells have fundamentally different baseline gene expression profiles, this comparison is intended to establish a highly sensitive detection threshold for identifying genes that are actively expressed at levels substantially above this technical background in tumor tissue, rather than to provide precise biological fold-change values across tissue types. The ≥2-fold threshold therefore serves as a pragmatic cutoff for distinguishing tumors with detectably high expression of the target genes. Therefore, it is crucial to emphasize that “high expression” as defined in this study indicates detectability above a nonepithelial (PBMC) baseline, and should not be interpreted as pathological overexpression relative to normal thyroid tissue.

Concurrently, baseline characteristics and serum levels of tumor biomarkers, including carcinoembryonic antigen (CEA), carbohydrate antigen 199 (CA199), and carbohydrate antigen 125 (CA125), and the second-generation assay for circulating cancer-specific antigen, were systematically collected.

Statistical Methods

All statistical analyses were performed using SPSS Statistics (version 27.0; IBM Corp) and R software (version 4.4.3; R Foundation). Continuous variables with normal distribution were expressed as mean ± SD (x−± s), while categorical data were presented as n (%). Intergroup comparisons were conducted using independent samples *t* tests for continuous variables and chi-square tests for categorical variables. Multivariable binary logistic regression analysis was used to identify independent risk factors. The RF model was developed using the randomForestSRC, *ggRandomForests*, *pdp*, and *GGally* packages in R. Nomogram construction and calibration were implemented with the *rms*, *dcurves*, and *pROC* packages in R. NN architecture was built using the keras interface in R. Model accuracy was evaluated via the confusionMatrix() function from the *caret* package. All visualizations were generated using the *ggplot2* package in R. A *P* value <.05 was considered statistically significant.

Results

Overview

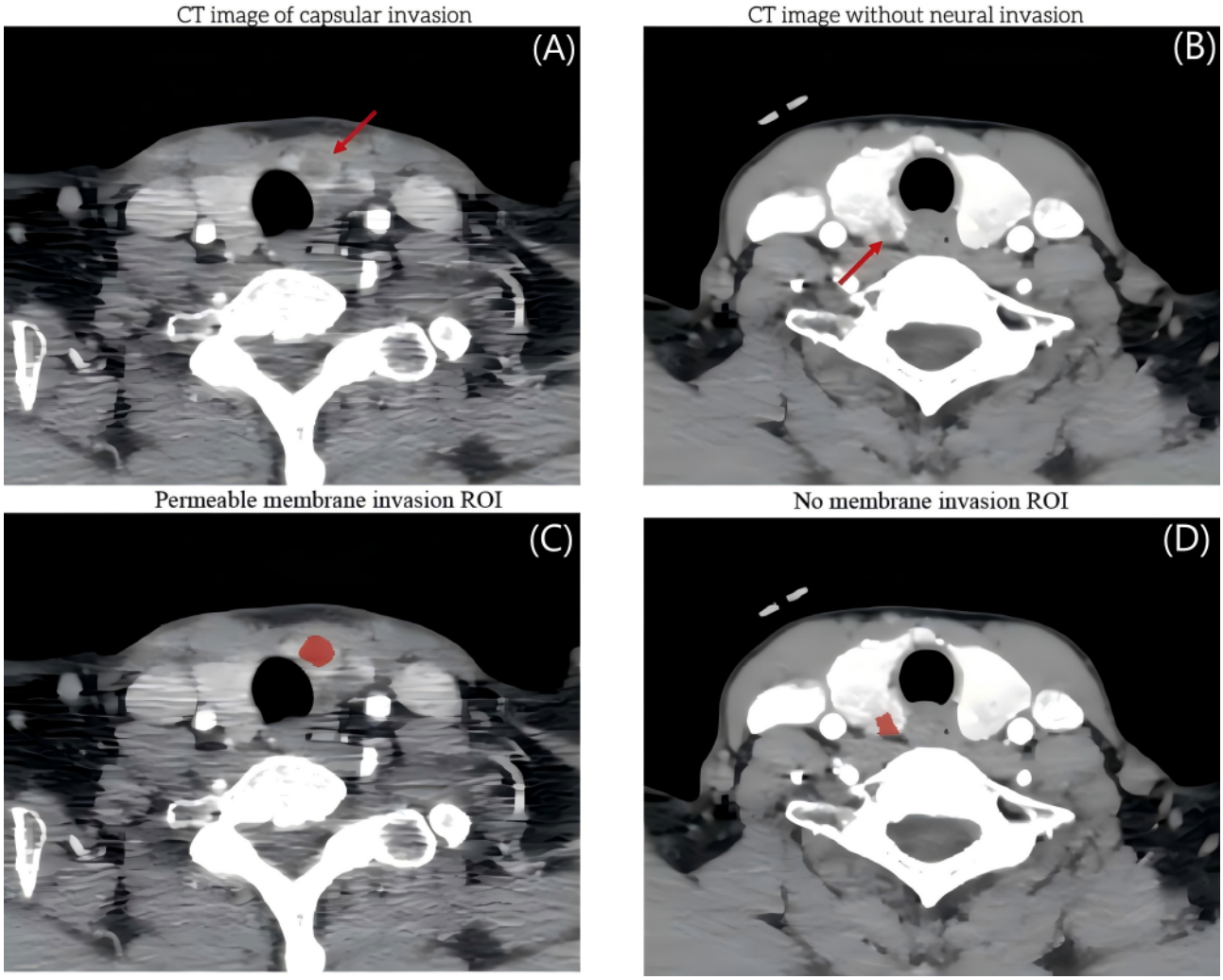
In this section, Table 1 presents the baseline characteristics of patients. The detailed multivariate logistic regression results (originally Tables 2 and 3) have been moved to the supplementary material; see Tables S1 and S2 in Multimedia Appendix 3.

Baseline Characteristics Comparison

The comparative analysis of baseline characteristics between patients with and without capsular invasion revealed statistically significant differences in the categorized expression levels (high vs low or no expression) of Gal-3 and CK19, as well as in the serum concentrations of tumor markers including CEA, CA199, and CA125 (*P*<.05, Table 1). Representative arterial-phase CT images and corresponding ROI delineations from patients with thyroid carcinoma are shown in Figure 3A-D.

XSL•FO
RenderX

Figure 3. Representative arterial-phase CT images of thyroid carcinoma. (A-C) A case with pathologically confirmed capsular invasion showing irregular tumor margins (arrow). (B-D) A case without NI demonstrating well-defined tumor boundaries. CT: computed tomography; NI: neural invasion; ROI: regions of interest.

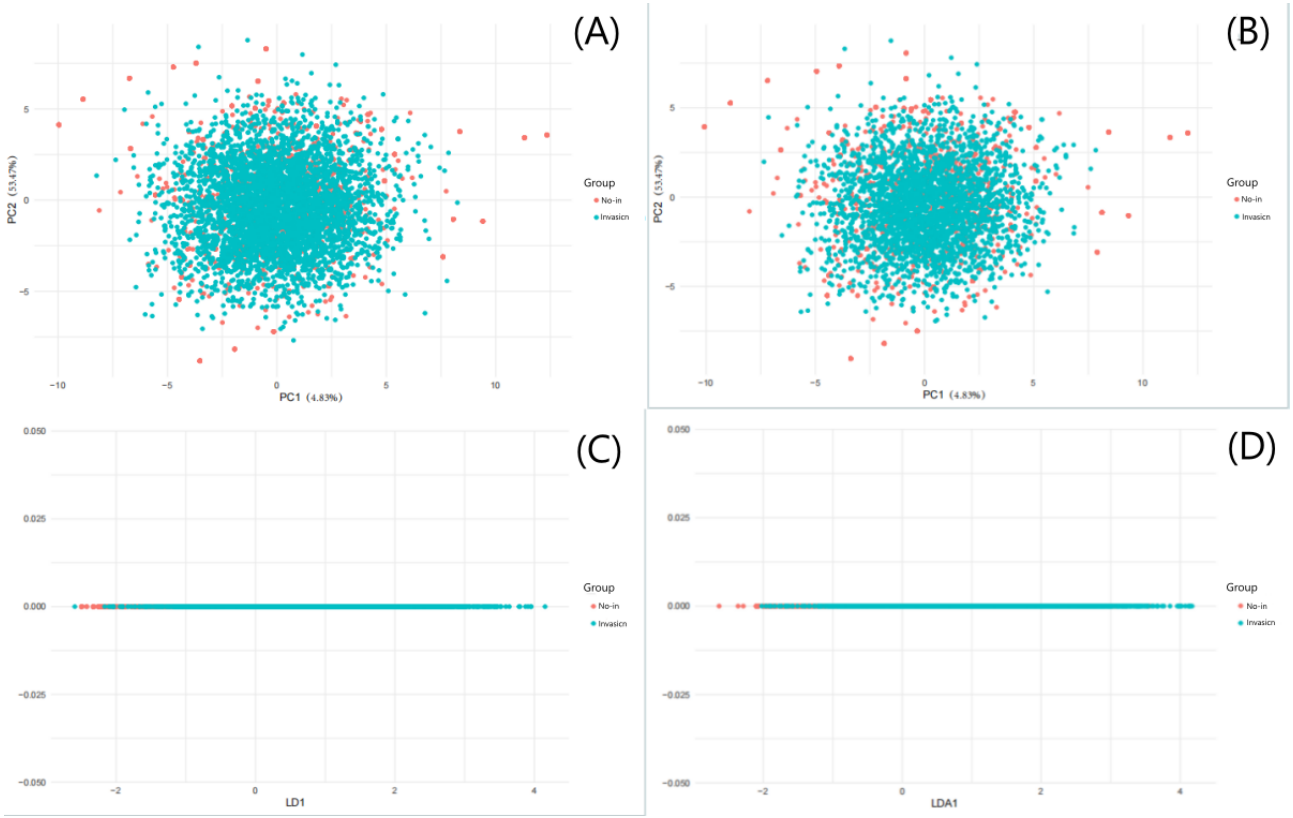


GLCM Analysis

The GLCM analysis yielded 111 texture features, which were subsequently subjected to dimensionality reduction via PCA and LDA. A total of 14 principal components were extracted for capsular invasion classification, with PCA1 and PCA2 accounting for 53.47% and 4.83% of the total variance, respectively. Significant intergroup differences were observed in PCA1 (t_{109}=11.671, P=.001, 2-tailed), where patients without capsular invasion exhibited greater dispersion in GLCM patterns, as evidenced by their larger Euclidean distance (6.98 vs 4.12 in the capsular invasion group). Similarly, NI analysis demonstrated significant differences in both PCA1 (t_{109}=10.339, P=.001) and PCA2 (t_{109}=−2.0187, P=.04). Non-NI cases again displayed greater GLCM dispersion, with Euclidean distances measuring 7.02 vs 4.05 in the NI group, a pattern consistent with the capsular invasion findings. LDA projections further corroborated this dispersion pattern, showing broader distribution clusters in noninvasion groups compared to their invasion counterparts. Visual representations of these analytical outcomes are provided in Figure 4.

XSL•FO
RenderX

Figure 4. Dimensionality reduction analysis. (A) PCA of capsular invasion status; (B) PCA of NI status; (C) LDA of capsular invasion status; (D) LDA of NI status. LD1: linear discriminant 1; LDA: linear discriminant analysis; LDA1: linear discriminant 1; NI: neural invasion; no-in: noninvasion; PC1: principal component 1; PC2: principal component 2; PCA: principal component analysis.



Multivariate Binary Logistic Regression Analysis Based on Clinical Indicators

It is important to emphasize that the following multivariate logistic regression analyses are intended to explore factors associated with the co-occurrence of pathological features, rather than to build preoperative predictive models. As both capsular invasion and NI are postoperative pathological diagnoses, their inclusion as variables in models predicting each other primarily reflects their strong pathological correlation, not a causal predictive relationship usable for preoperative assessment.

A multivariate binary logistic regression model was constructed with capsular invasion status as the dependent variable (1=presence, 0=absence). The analysis identified CK19, CEA, CA199, CA125, and NI as significant independent risk factors for capsular invasion. Notably, CK19 demonstrated an exceptionally strong association with capsular invasion, yielding an odds ratio (OR) of 60.491 (P<.05, see Table S1 in Multimedia Appendix 3).

A parallel analysis was performed with NI status as the dependent variable (1=presence, 0=absence). Among all clinical indicators, only capsular invasion emerged as a significant predictor of NI, with an OR of 25.25 (P<.05, Table S2 in Multimedia Appendix 3), underscoring the pronounced influence of capsular invasion on NI risk.

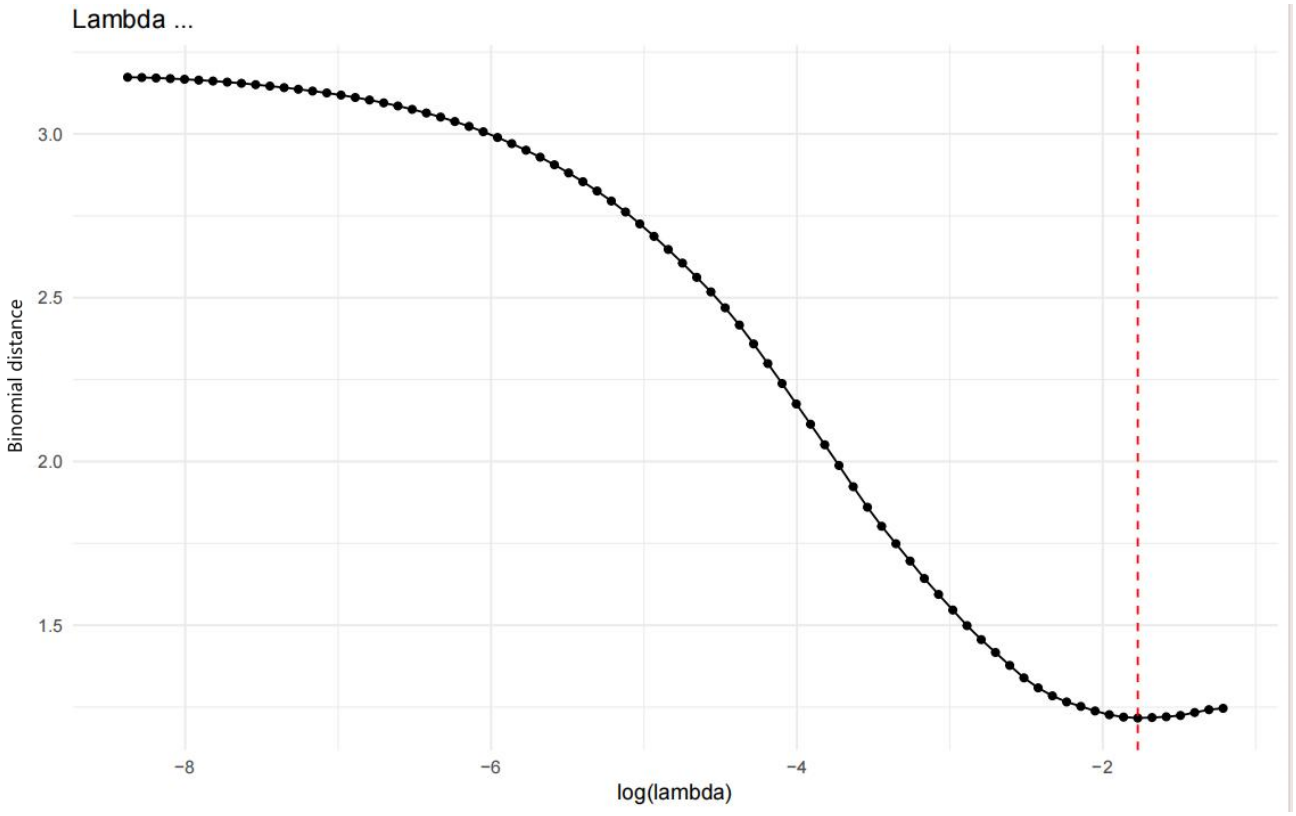
Gray-Level Feature Selection via LASSO Regression Analysis

A set of radiomic features was extracted from the pixels within the ROI. Following exclusion of features without significant differential expression, 57 discriminative features were retained for LASSO regression analysis. During variable selection, the penalty coefficient (λ) was iteratively adjusted to compress the initial 57 predictors. Optimal λ value (λ=0.017) was determined through cross-validation, achieving minimal model deviance while balancing parsimony and predictive accuracy. This process yielded a refined model incorporating 9 pivotal features (designated A1-A9), as presented in Figure 5. The 9 pixel-level features (A1-A9) selected in this study directly reflect the heterogeneity of contrast enhancement and tissue composition of thyroid carcinoma on CT images. Specifically: (1) differences in arterial-phase pixel intensity mean (A1) reflect variations in tumor vascular density: tumors with higher A1 values (stronger arterial enhancement) are more likely to have capsular invasion/NI, as increased angiogenesis provides conditions for tumor invasion; (2) variations in pixel intensity SD (A2) correspond to intratumoral necrosis: higher A2 values (uneven enhancement) often indicate necrotic foci, which are associated with poor differentiation (CK19 positivity) and higher NI risk; and (3) changes in venous-phase pixel percentile (A3 and A8) reflect capsule integrity: lower A3 or A8 values suggest capsule rupture, a direct marker of capsular invasion. These imaging phenotypes serve as quantifiable surrogates for pathological characteristics, such as vascular density, necrosis, and capsular status, enabling noninvasive preoperative assessment of tumor aggressiveness. Detailed feature definitions are provided in

XSL•FO
RenderX

Multimedia Appendix 2, with additional descriptions in Multimedia Appendix 1.

Figure 5. Cross-validated deviance profile versus log(λ) values.



Development of Nomogram Models for Predicting NI Based on Capsular Invasion

A tumor biomarker-derived nomogram was constructed to assess the predictive value of capsular invasion for NI. ROC curve analysis demonstrated robust discriminative performance, with an internally estimated area under the curve (AUC) of 0.9418 for the capsular invasion-prediction task (modeled on the full 111-case dataset) and 0.7253 for the cross-label association analysis task (evaluated on the full 111-case dataset). Calibration curves indicated high consistency between predicted probabilities and observed outcomes across both tasks. Decision curve analysis further validated the clinical utility of the model, as its net benefit across all threshold probabilities outperformed the “treat-all” and “treat-none” strategies (Figure 6).

XSL•FO
RenderX

Figure 6. Clinically parameterized nomogram model for neural invasion prediction. (A) Nomogram integrating clinical biomarkers (Gal-3, CK19, CEA, CA199, and CA125) to predict NI risk, with key variables (CK19 and CA199) marked by "*" (high OR values from Tables S1 and S2 in Multimedia Appendix 3); red dashed lines guide point assignment for each variable. (B and C) Calibration curves for the capsular invasion–prediction task ((B) modeled on full 111-case dataset) and the cross-label association analysis task ((C) cross-prediction on full 111-case dataset), showing agreement between predicted NI probability (x-axis) and observed NI status (y-axis); 95% CI bands (light blue or red) and MAE are labeled. (D and E) DCA curves for both tasks (using the full 111-case dataset, with no traditional train or test split), comparing the net benefit of the nomogram (dark blue line) with "treat all" (red line) and "treat none" (gray line) strategies across different NI risk thresholds. (F and G) ROC curves for the capsular invasion-prediction task ((F) modeled on full 111-case dataset) and the cross-label association analysis task ((G) evaluated on full 111-case dataset), with exact AUC values and 95% CIs labeled; light blue or red bands represent 95% CIs. All analyses were based on the full 111-case cohort (capsular invasion–labeled modeling and NI-labeled evaluation). AUC: area under the curve; CA125: carbohydrate antigen 125; CA199: carbohydrate antigen 199; CCSA-2: circulating cancer-specific antigen; CEA: carcinoembryonic antigen; CK19: cytokeratin 19; DCA: decision curve analysis; Gal-3: galectin-3; MAE: mean absolute error; NI: neural invasion; OR: odds ratio; Prob: probability; RDATA: research data; ROC: receiver operating characteristic; .

XSL•FO
RenderX

A parallel nomogram model incorporating gray-level radiomic features achieved internally estimated AUC values of 0.9334 for the capsular invasion-prediction task (full 111-case dataset) and 0.8001 for the cross-label association analysis task (full 111-case dataset). Calibration curves similarly confirmed strong predictive accuracy for both tasks, and decision curve analysis revealed that the model achieved superior net clinical benefit across the entire spectrum of risk thresholds compared to reference strategies (Figure 7).

XSL•FO
RenderX

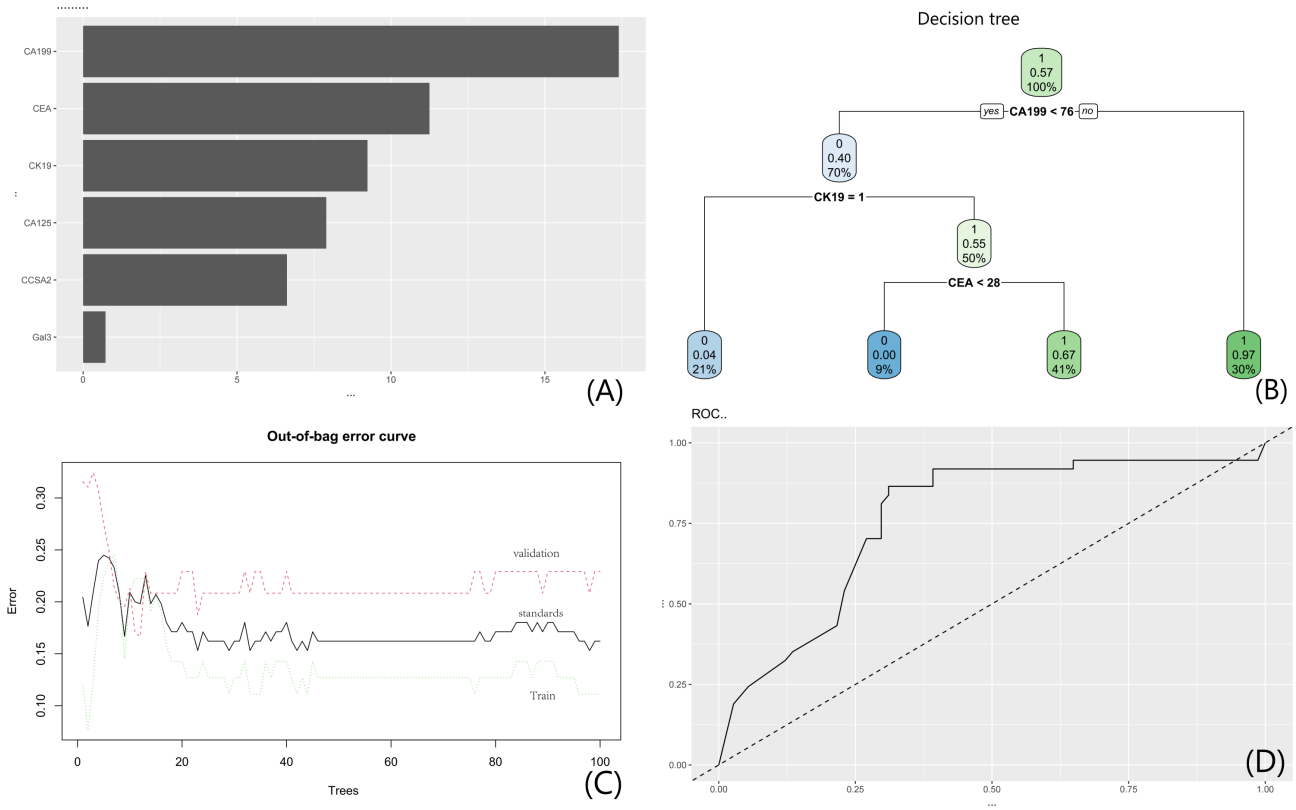
Figure 7. Gray-level feature-integrated nomogram model for neural invasion prediction. (A) Nomogram integrating LASSO-selected pixel-level features (A1-A9) to predict NI risk; footnotes define key features (A1 and A9) per the Gray-Level Feature Selection via LASSO Regression Analysis section. (B and C) Calibration curves for capsular invasion-prediction task ((B) modeled on full 111-case dataset) and cross-label association analysis task ((C) evaluated on full 111-case dataset); hollow circles (dark blue) and solid squares (coral red) represent data points, with 95% CI bands and MAE ((B) 0.05; (C) 0.049) labeled. (D and E) DCA curves, with the nomogram (dark blue line) showing higher net benefit than reference strategies across NI risk thresholds. (F and G) ROC curves for the capsular invasion-prediction task ((F) modeled on full 111-case dataset) and the cross-label association analysis task ((G) evaluated on full 111-case dataset); exact internally estimated AUC values (capsular invasion-prediction task: 0.9334; cross-label association analysis task: 0.8001) and 95% CIs are labeled, with light blue or red bands representing 95% CIs. All analyses used the full 111-case cohort (capsular invasion–labeled modeling and NI-labeled evaluation). AUC: area under the curve; DCA: decision curve analysis; LASSO: least absolute shrinkage and selection operator; MAE: mean absolute error; NI: neural invasion; Prob: probability; RDATA: research data.

XSL•FO
RenderX

RF Model Development

An RF model was constructed using tumor biomarkers, with CA199 demonstrating the highest feature importance (visualized in Figure 8A). A representative surrogate tree from the RF ensemble (visualized in Figure 8B) initiated classification through sequential partitioning: the root node split on "CA199<76," followed by a "yes" branch subdivided by "CK19=1," and further partitioned via "CEA<28." Out-of-bag error curves (visualized in Figure 8C) revealed stabilized misclassification rates for both training and validation sets as tree numbers increased, with training set errors consistently exceeding validation set values. The 5-fold cross-validation AUC (for capsular invasion prediction) was 0.782 (95% CI 0.671 - 0.874), and the out-of-bag error rate stabilized at 0.213 when ntree ≥300 (visualized in Figure 8D). When this model was used to predict NI (full-sample NI label), the model exhibited discriminative efficacy with an AUC of 0.7646, only 2.5% lower than the cross-validation AUC for capsular invasion, confirming no significant overfitting (Figure 8).

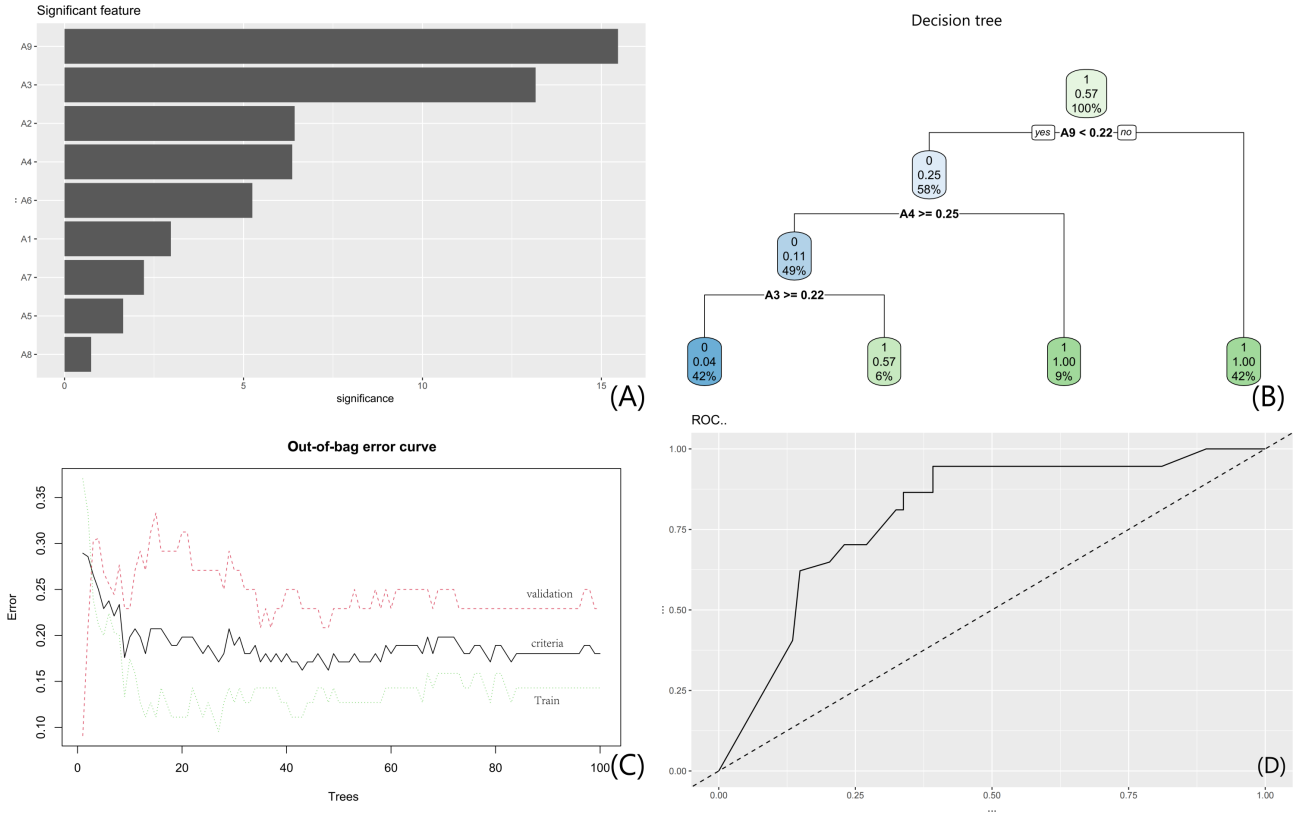
Figure 8. RF model based on clinical biomarkers. (A) Feature importance plot; (B) a representative surrogate tree from the RF ensemble; (C) out-of-bag error trajectories; and (D) ROC curve. CA125: carbohydrate antigen 125; CA199: carbohydrate antigen 199; CCSA2: colon cancer-specific antigen-2; CEA: carcinoembryonic antigen; CK19: cytokeratin 19; Gal3: galectin-3; RF: random forest; ROC: receiver operating characteristic.



A parallel RF model incorporating gray-level features identified feature A9 as the primary contributor (corresponding to Figure 9A). A representative surrogate tree from the RF ensemble (corresponding to Figure 9B) partitioned data at "A9 <0.22," with subsequent splits at "A4 ≥0.25" and "A3 ≥0.22." Out-of-bag error curves mirrored the tumor biomarker-based model, showing convergence to stability with increasing tree numbers (corresponding to Figure 9C). The 5-fold cross-validation AUC (for capsular invasion prediction) was 0.825 (95% CI 0.723 - 0.906), and the out-of-bag error rate stabilized at 0.187 when ntree ≥350. ROC analysis confirmed enhanced predictive performance (corresponding to Figure 9D); its AUC for NI prediction was 0.8102% - 1.8% lower than the capsular invasion cross-validation AUC, further confirming low overfitting risk (Figure 9). These data collectively validate the clinical utility of both RF models.

XSL•FO
RenderX

Figure 9. RF model based on gray-level features. (A) Feature importance plot (features A1-A9 represent the 9 key radiomic grayscale features selected via LASSO regression from GLCM texture features of CT regions of interest. They quantify tumor heterogeneity in arterial and venous phases (eg, A1, A9); (B) a representative surrogate tree from the RF ensemble; (C) out-of-bag error trajectories; (D) ROC curve. A1: arterial-phase mean intensity; A9: venous-phase coefficient of variation; CT: computed tomography



NN Model Development

The NN model used exclusively two input modalities: (1) raw arterial-phase CT ROI images and (2) clinical biomarkers. No hand-crafted radiomic features (A1-A9) were directly input to the NN.

An NN model integrating multimodal data demonstrated high learning efficiency, with training accuracy values exceeding 0.99 across learning curves (Figure 2A). It is important to note that Figure 2A displays the mean training accuracy across the 5-fold cross-validation splits, not a cross-validated test accuracy. The significant gap between the near-perfect training accuracy (>0.99) and the more moderate cross-validation AUC for the primary capsular invasion prediction task (0.798) is indicative of overfitting, a common challenge in training complex models on limited datasets. While early stopping was used to mitigate overtraining by halting training when validation loss plateaued, this gap suggests that the model still learned patterns highly specific to the training samples. The architecture used hidden units ranging from 10 to 30, where variations in hidden unit counts under distinct weight decay coefficients differentially influenced cross-validated accuracy values. Feature importance analysis based on SHAP values (as detailed in the Model Interpretability Analysis section) revealed that clinical biomarkers were the most influential predictors for NI risk, particularly CK19 (normalized importance score: 95) and CEA (normalized importance score: 88; Figure 2B). Concurrently, the semantically annotated image patterns—"Image_Heterogeneity" (score: 82), "Image_Boundary" (score: 75), and "Enhancement_Dynamics" (score: 68), which were autonomously learned and inductively summarized by the NN from the raw CT inputs, also contributed substantially to the predictions. Importantly, these image patterns were learned directly from raw CT images and are conceptually distinct from the hand-crafted radiomic features (A1-A9) used in traditional radiomic models. The model achieved an AUC of 0.775 in the validation set, confirming its satisfactory generalizability beyond the training cohort. The 5-fold cross-validation AUC (for capsular invasion prediction) of the integrated NN model was 0.798 (95% CI 0.695 - 0.883), with a coefficient of variation of 7.2% across folds—indicating stable performance. Early stopping was triggered at an average of 32 (SD 2.8, range: 28 - 35 epochs) across the 5 folds, effectively preventing overtraining. When predicting NI (full-sample NI label), the model achieved an AUC of 0.775 (95% CI 0.621 - 0.903)—only 2.3% lower than the capsular invasion cross-validation AUC, indicating that the model's predictive ability for NI remains robust despite some degree of overfitting on the primary capsular invasion task. Detailed architectural and performance metrics are illustrated in Figure 2. Feature importance analysis within the NN identified that its predictions were driven by several key learned representations: an image pattern capturing intratumoral heterogeneity, a feature related to venous-phase enhancement dynamics, and another reflecting textural uniformity, in conjunction with the clinical biomarkers CK19 and CEA. Feature importance analysis indicated that intratumoral heterogeneity was the most influential learned representation (importance score 0.82, full score=1.0).

XSL•FO
RenderX

About validation on an internal hold-out set, to further assess generalizability, a traditional 7:3 hold-out validation was performed. The RF model based on radiomic features (A1-A9) was trained on a subset of 78 cases and was subsequently evaluated on the separate, internal hold-out set of 33 cases. The observed AUC of 0.827 was promising, but the extremely wide 95% CI (0.652 - 0.981) indicates substantial uncertainty due to the small test sample size (n=33), necessitating cautious interpretation of the model's generalizability. This additional validation yielded a promising but variable performance (AUC=0.827, 95% CI 0.652 - 0.981) for NI prediction (see Multimedia Appendices 1 and 4 for details).

Discussion

Research Background and Basis

Thyroid carcinoma, the most prevalent malignancy of the head and neck region, has exhibited a rapid surge in incidence over the past decade, emerging as a critical threat to global human health [11]. Beyond its capacity to disrupt thyroid function and induce hormonal dysregulation, thyroid carcinoma frequently metastasizes to adjacent cervical tissues, lymph nodes, and even distant organs such as the lungs, liver, and bones, severely compromising physiological integrity [12]. Equally concerning are its profound psychosocial impacts, as patients endure debilitating fear, anxiety, and uncertainty regarding prognosis and treatment outcomes [13]. Current advances in thyroid carcinoma research have identified molecular markers such as *BRAF* (OMIM 164757) and *NRAS* (OMIM 164790) gene mutations, which have emerged as critical tools for diagnosis, prognostic stratification, and therapeutic decision-making [14]. Nevertheless, significant gaps persist in risk stratification for capsular invasion and NI, 2 pivotal prognostic determinants. Capsular invasion serves as a key predictor of lymph node metastasis, whereas NI directly correlates with postoperative survival quality and recurrence risk [3]. Elucidating the relationship between capsular invasion and NI is therefore essential for precise clinical assessment and optimized treatment planning. Therefore, the identification of radiogenomic biomarkers for capsular invasion using CT radiomics and the evaluation of their stratification value for NI are both scientifically imperative and clinically urgent.

Identification of Radiomic Signatures for Invasion Risk

In this study, GLCM analysis of CT images from thyroid carcinoma patients revealed distinct radiomic signatures between tumors with and without capsular invasion or NI. A total of 111 GLCM features were extracted and subjected to dimensionality reduction via PCA and LDA, which uncovered critical associations between key imaging biomarkers and invasion risk. The heterogeneous distribution of GLCM features in noninvasive cohorts, characterized by greater dispersion, reflects heightened spatial heterogeneity in gray-level pixel relationships within tumor microenvironments. This phenomenon may correlate with histopathological characteristics such as tumor boundary demarcation and cellular alignment integrity, a concept supported by the work of Wei et al [11], who demonstrated that radiomic features from multiparametric magnetic resonance imaging could effectively predict extrathyroidal extension in papillary thyroid carcinoma by capturing similar underlying tumor heterogeneity. In contrast, tumors with capsular invasion demonstrate enhanced propensity for boundary breaching [15], resulting in homogenization of imaging features. Noninvasive lesions, by comparison, maintained relatively intact tissue architecture [16,17], which manifests as increased textural complexity on imaging. This inverse relationship is further corroborated by a recent meta-analysis [18], which confirmed that textural homogenization in medical imaging often reflects the loss of normal tissue architecture due to invasive growth patterns. Thus, these findings suggest that GLCM feature dispersion may serve as a universal imaging biomarker of tumor invasiveness [19].

Mechanistically, tumors with NI exhibit the proliferative spread of malignant cells along nerve bundles [20], resulting in textural homogeneity on imaging. In contrast, noninvasive lesions present greater histological heterogeneity in interstitial architecture and cellular distribution [21]. Although direct radiomic studies on NI in thyroid carcinoma are scarce, the work of Mu et al [22] reinforces this biological rationale by showing that a radiomic signature on CT could effectively detect metastatic lymph nodes in thyroid carcinoma, demonstrating that imaging features can capture the tumor's propensity for local invasion and spread. LDA further corroborated these findings, demonstrating broader feature dispersion in noninvasive groups and supporting PCA findings, which suggest the utility of radiomic heterogeneity as a quantifiable biomarker for invasion risk in thyroid carcinoma [18].

Key Molecular Markers and Risk Pathways for Invasion

Notably, 3 (6.25%) patients in the capsular invasion–negative group were pathologically diagnosed with NI. This observation indicates that while capsular invasion is a dominant driver, NI can occur through alternative pathways (eg, direct perineural infiltration). Moreover, multivariate binary logistic regression identified CK19, CEA, CA199, CA125, and NI as independent risk factors for capsular invasion. Notably, CK19, a member of the cytokeratin family, exhibited an OR of 60.491, indicating that patients with significant CK19 expression had approximately 60.5 times the odds of having capsular invasion compared to those with nonsignificant expression. This finding is consistent with prior studies reporting CK19 overexpression in papillary thyroid carcinoma [23], where its upregulation enhances tumor cell proliferation and invasive potential via disruption of intercellular adhesion molecules, thereby facilitating tumor penetration through capsular barriers [24]. Laishram et al [25] further identified CK19 as a highly sensitive and specific immunohistochemical marker for thyroid carcinoma in a cross-sectional study using histopathology as the reference standard. Elevated serum levels of CEA, CA199, and CA125 further signify increased tumor burden and epithelial-mesenchymal transition, synergistically promoting invasive phenotypes [26,27]. Intriguingly, NI itself emerged as a covariate for capsular invasion, implying shared molecular pathways or microenvironmental crosstalk. Conversely, when NI was modeled as the dependent variable, only capsular invasion remained significant, underscoring its role as a

XSL•FO
RenderX

dominant predictor for NI. This finding is in line with the anatomical invasion pathway of thyroid carcinoma. Following capsular penetration, tumors demonstrate increased propensity for either direct NI (eg, recurrent laryngeal nerve involvement) or chemokine-mediated perineural infiltration through paracrine signaling modulation [28]. The resultant loss of tumor boundary integrity and disruption of the local microenvironment may jointly provide a structural basis and inflammatory background for NI. Importantly, the exclusion of variables such as age and tumor size from the model highlights the preeminence of capsular status in neural risk stratification, offering clinicians an actionable biomarker for preoperative assessment.

Predictive Model Construction and Clinical Translation Value

For preoperative assessment, the radiomic model prioritizes spatial heterogeneity derived from preoperative CT, enabling noninvasive risk prediction via texture features [6]. Such insights may guide surgical planning, such as the extent of resection or the consideration of intraoperative nerve monitoring. For postoperative stratification, the clinical biomarker-based nomogram and the integrated multimodal model use data available after surgery. They emphasize the biological aggressiveness associated with capsular invasion, aiding in postoperative pathological substaging and recurrence risk prediction. For instance, concurrent positivity for CA199 and CK19 in patients with capsular invasion could signal high-risk subgroups for NI, indicating the need for intensified adjuvant therapy. Specifically, by integrating quantitative molecular data with imaging phenotypes, these postoperative models could provide a supplementary risk assessment that might identify patients with occult invasive features or a higher propensity for recurrence, thereby adding prognostic granularity beyond standard histopathological examination alone. NN architectures further synergized clinical and imaging data, capturing nonlinear interactions between capsular invasion and neural risk [29]. Biological markers such as CEA may reflect intrinsic invasive potential, whereas radiomic features (eg, A9) represent its spatial manifestation. Notably, tumors with capsular invasion characterized by elevated CEA levels and disorganized textures (low A9 values) exhibited markedly increased NI rates, illustrating the model's ability to decode multifactorial invasion pathways. The high performance of all models in the cross-label association analysis (eg, AUCs ≈0.80) is largely concordant with the strong pathological correlation between capsular invasion and NI identified by logistic regression (OR=25.25). This indicates that the models effectively captured the shared features underlying both invasion types, rather than demonstrating independent predictive power for NI on entirely unseen data.

Concerning clinical translation and distinction between model types, it is crucial to distinguish the intended clinical context of the different models developed in this study. The radiomic models (nomogram and RF based on features A1-A9), derived solely from preoperative CT, hold promise as preoperative decision-support tools. They could inform surgical planning, such as the extent of resection or the need for intraoperative nerve monitoring, based on noninvasive imaging. In contrast, the clinical biomarker-based models and the multimodal NN model integrate data (Gal-3 and CK19) that are only available postoperatively from resected tissue. Therefore, their primary utility lies in postoperative risk stratification. They may help pathologists and clinicians identify high-risk subsets among patients already diagnosed with capsular invasion, potentially predicting the likelihood of concurrent or subsequent NI, and thus informing adjuvant therapy strategies or the intensity of follow-up. This postoperative application remains valuable, as it could refine prognostic assessment beyond standard histology.

Limitations and Future Directions

However, this study has several limitations. First, the multivariate logistic regression models (Tables S1 and S2 in Multimedia Appendix 3) are methodologically limited by circular reasoning. Using 1 postoperative pathological finding (eg, NI) to predict another (eg, capsular invasion) inflates the observed associations. The resultant astronomically high ORs with extremely wide CIs are classic signs of overfitting and multicollinearity within our dataset, rather than reliable measures of preoperative risk. These results should be interpreted as strong indicators of pathological co-occurrence, not as valid predictive effects.

Second, the single-center design, limited sample size (n=111), absence of external validation, and reliance on CT scanners from a single manufacturer may restrict generalizability and introduce selection bias.

Third, an additional limitation is the unconventional data usage design: the full dataset was used for both capsular invasion-modeling and cross-label association analysis without a separate held-out test set. Therefore, the reported AUCs for NI prediction are not directly comparable to studies using independent test sets. However, this design was necessary to maximize data use given the small sample size and to specifically test the cross-predictive value of capsular invasion–related features for NI, an exploratory goal that prioritizes verifying feature relevance across related clinical outcomes. Supplementary traditional 7:3 split validation confirmed consistent performance, mitigating concerns about overfitting to some extent (Multimedia Appendix 4).

Fourth is the sampling bias from single-slice analysis. Our NN model used only the single 2D slice with the largest cross-sectional tumor area. As capsular invasion predominantly occurs at the tumor boundaries or poles, this central slice may systematically exclude the imaging features most directly associated with capsular invasion, potentially limiting the model's sensitivity to capture these specific patterns. Consequently, the "Image_Boundary" pattern identified by the network (Figure 2B) likely reflects the general sharpness or regularity of the tumor-parenchyma interface on the central slice, rather than specific imaging signs of focal capsular penetration at the tumor periphery.

Fifth, there is a limitation regarding the control group for qPCR analysis. The use of PBMCs from healthy individuals as a control does not yield biologically precise fold-change values for epithelial markers such as CK19, due to their near absence in PBMCs. In this study, the ≥2-fold threshold was used primarily as a sensitive, technical cutoff to categorize tumors

XSL•FO
RenderX

into those with detectably high versus low or no expression of the target genes, facilitating subsequent clinical correlation analysis. This pragmatic approach should be distinguished from a biologically rigorous quantification of overexpression. Future studies should use paired adjacent normal thyroid tissue as the control to obtain accurate biological quantifications.

Sixth is the normalization-induced information leakage. The *z* score normalization of radiomic features was performed on the entire dataset before model development and cross-validation. This approach could theoretically allow information from validation samples (mean and SD) to leak into the training process, potentially inflating the model's performance estimates. While we used cross-validation and an independent hold-out set to mitigate overfitting, this methodological choice may still contribute to an optimistic bias in the reported metrics. Future studies should consider performing normalization within each fold of the cross-validation loop to provide a more rigorous and unbiased evaluation of model generalizability.

Seventh is feature selection bias. In this study, the LASSO regression was applied to the entire dataset to select the 9 key radiomic features (A1-A9) before model development and cross-validation. Performing feature selection outside the cross-validation loop can introduce selection bias, as the selection process is informed by data that would later serve for validation. This may lead to overfitting and an optimistic estimation of the model's generalizable performance. Consequently, the independence of the hold-out validation for the RF model is partially compromised. Future work should use a nested cross-validation framework, where feature selection is performed independently within each training fold, to obtain unbiased feature sets and more rigorous performance estimates.

Eighth is the potential misalignment from rigid registration. This study used rigid image registration to propagate arterial-phase ROIs to the venous-phase images for feature extraction (eg, feature A9, the venous-phase coefficient of variation). While efficient and commonly used, rigid registration assumes no tissue deformation between phases. Given that the thyroid is a soft organ susceptible to movement from swallowing or breathing, perfect alignment cannot be guaranteed. Any misalignment could cause venous-phase features to partially capture the texture of the tumor's immediate periphery or surrounding tissue rather than the tumor itself. This may introduce noise or bias, particularly for texture-based features such as A9. Therefore, the high importance score of A9 in our model may partially reflect registration artifacts (such as edge effects at the ROI boundaries) rather than solely genuine intratumoral heterogeneity. Although we performed manual verification and adjustment of the propagated ROIs, minor residual misalignments may persist. Future studies could consider using deformable (nonrigid) registration techniques or independent manual delineation on venous-phase images to ensure more accurate feature localization and interpretation.

Ninth is the resolution loss due to image resizing for deep learning. In this study, to comply with the input requirements of the pretrained DenseNet121 network, the entire ROI was resized as a whole from its original resolution to 224×224 pixels. This means that no prior cropping to a tumor-specific bounding box was performed. For smaller thyroid tumors, this global down-sampling can substantially reduce the number of pixels representing the tumor tissue, leading to a loss of fine spatial detail and high-frequency texture information. This process may diminish the model's ability to learn subtle textural patterns that are crucial for predicting capsular invasion, particularly in small lesions. Future studies could improve upon this by implementing a crop-to-bound-box step before resizing or by using NN architectures that support variable input sizes. Furthermore, the "learned image patterns" (eg, "Image_Heterogeneity" and "Image_Boundary") identified through our interpretability analysis (SHAP) may therefore reflect macrolevel structures or artifacts introduced by the global down-sampling process, rather than solely the fine, tumor-specific textural details that are theoretically of interest. For instance, patterns interpreted as heterogeneity could be influenced by the interpolation artifacts during resizing or by the inclusion of nontumoral perinodular tissues within the ROI. While these patterns still contributed to the model's predictive performance, their biological interpretability in relation to capsular invasion should be interpreted with this technical limitation in mind.

Tenth is the uncertainty in hold-out set performance. The performance of the NN model on the small independent hold-out set (n=33; AUC=0.827, 95% CI 0.652 - 0.981) showed a point estimate higher than that from the full-sample cross-validation (AUC=0.775). The wide 95% CI reflects the substantial uncertainty inherent in evaluating a model with such a limited test sample. While the hold-out set was constructed using stratified sampling to maintain a representative distribution of capsular invasion status, the observed difference in point estimates may be attributable to the small sample size and its associated variability. Therefore, the hold-out performance should be interpreted with caution, and the model's generalizability needs to be confirmed in larger, prospective cohorts.

Eleventh is the potential confounding in model comparison due to input differences. The radiomic RF model identified a venous-phase feature (A9) as highly predictive, while the NN was designed with arterial-phase-only inputs for simplicity and to mitigate overfitting. This design difference means that the lower AUC of the NN model (0.775) compared to the RF model (0.8102) may be partially attributable to the exclusion of venous-phase information rather than an inherent limitation of the deep learning architecture itself. Our study aimed to test whether a single phase (arterial) could provide a sufficient predictive signal for a deep learning model; however, this choice introduces a confounding factor when directly comparing performance across methodologies. Future studies could develop multimodal NN architectures incorporating both arterial and venous phases to enable a more equitable comparison and potentially enhance performance.

Future work will expand multicenter collaboration to include over 300 cases across diverse CT devices, incorporate data augmentation and explainable artificial intelligence methods, and establish standardized protocols to improve robustness. Most importantly, future models will focus on using exclusively

XSL•FO
RenderX

preoperative variables to achieve genuine preoperative risk stratification.

Conclusions

In conclusion, our exploratory analysis confirms a strong association between capsular invasion and NI in thyroid carcinoma. This study delineates complementary approaches for invasion risk assessment: (1) radiomic signatures from preoperative CT alone demonstrate potential for noninvasive, preoperative risk stratification; and (2) models incorporating postoperative tissue-derived biomarkers (eg, Gal-3 and CK19), including the multimodal model, provide a valuable framework for postoperative risk stratification. These postoperative tools could aid in identifying patients with high-risk pathological features, potentially guiding decisions on adjuvant therapy or surveillance intensity following surgery. The performance of the radiomic models lays the groundwork for future research toward fully preoperative predictive tools.

Acknowledgments

This study used traditional radiomics and machine learning methodologies for CT (computed tomography)-based capsular or neural invasion risk stratification in thyroid carcinoma. No generative artificial intelligence (GenAI) tools were used in data collection, preprocessing, feature extraction, model development, or interpretation. All analyses followed institutional ethics guidelines, with authors retaining full responsibility for methodological rigor and result validity.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used are available from the corresponding author upon reasonable request.

Authors' Contributions

FfC and KT designed this study. FW and QG collected and analyzed the data. PS and NX wrote and reviewed this paper. All authors have read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed architecture of the neural network model for capsular invasion prediction.
[DOCX File, 15 KB - medinform_v14i1e77349_app1.docx]

Multimedia Appendix 2
Definitions and pathophysiological correlations of the 9 key radiomic features (A1-A9).
[DOCX File, 16 KB - medinform_v14i1e77349_app2.docx]

Multimedia Appendix 3
Detailed multivariate logistic regression analyses for capsular invasion (Table S1) and neural invasion (Table S2).
[DOCX File, 16 KB - medinform_v14i1e77349_app3.docx]

Multimedia Appendix 4
Hold-out validation results of the random forest model for neural invasion prediction.
[DOC File, 102 KB - medinform_v14i1e77349_app4.doc]

References

1. Mahkamova K, Latar NM, Aspinall S, Meeson A. Side population cells in anaplastic thyroid cancer and normal thyroid. Exp Cell Res 2019 Jan 1;374(1):104-113. [doi: 10.1016/j.yexcr.2018.11.012] [Medline: 30465733]
2. Latar NM, Mahkamova K, Elson J, et al. Impact of transforming growth factor beta 1 on normal and thyroid cancer side population cells. Endocrine 2022 May;76(2):359-368. [doi: 10.1007/s12020-022-02990-4] [Medline: 35118633]
3. Coca-Pelaz A, Shah JP, Hernandez-Prera JC, et al. Papillary thyroid cancer-aggressive variants and impact on management: a narrative review. Adv Ther 2020 Jul;37(7):3112-3128. [doi: 10.1007/s12325-020-01391-1] [Medline: 32488657]
4. Baloch Z, Mete O, Asa SL. Immunohistochemical biomarkers in thyroid pathology. Endocr Pathol 2018 Jun;29(2):91-112. [doi: 10.1007/s12022-018-9532-9] [Medline: 29744727]

XSL•FO
RenderX

5. Zhou Y, Zhou G, Zhang J, Xu C, Zhu F, Xu P. DCE-MRI based radiomics nomogram for preoperatively differentiating combined hepatocellular-cholangiocarcinoma from mass-forming intrahepatic cholangiocarcinoma. Eur Radiol 2022 Jul;32(7):5004-5015. [doi: 10.1007/s00330-022-08548-2] [Medline: 35128572]
6. Chen J, Xi J, Chen T, Yang L, Liu K, Ding X. Diagnostic performance of computed tomography-based artificial intelligence for early recurrence of cholangiocarcinoma: systematic review and meta-analysis. J Med Internet Res 2025 Sep 18;27:e78306. [doi: 10.2196/78306] [Medline: 40905766]
7. Stoyanova R, Takhar M, Tschudi Y, et al. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res 2016 Aug;5(4):432-447. [doi: 10.21037/tcr.2016.06.20] [Medline: 29188191]
8. Tao J, Liu D, Hu FB, et al. Development and validation of a computed tomography–based model for noninvasive prediction of the T stage in gastric cancer: multicenter retrospective study. J Med Internet Res 2024;26:e56851. [doi: 10.2196/56851]
9. Sollini M, Cozzi L, Chiti A, Kirienko M. Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: where do we stand? Eur J Radiol 2018 Feb;99:1-8. [doi: 10.1016/j.ejrad.2017.12.004] [Medline: 29362138]
10. Alonso-Gordoa T, Jimenez-Fonseca P, Martinez-Trufero J, et al. SEOM-GETNE-TTCC clinical guideline thyroid cancer (2023). Clin Transl Oncol 2024 Nov;26(11):2902-2916. [doi: 10.1007/s12094-024-03736-6] [Medline: 39325263]
11. Wei R, Wang H, Wang L, et al. Radiomics based on multiparametric MRI for extrathyroidal extension feature prediction in papillary thyroid cancer. BMC Med Imaging 2021 Feb 9;21(1):20. [doi: 10.1186/s12880-021-00553-z] [Medline: 33563233]
12. Boucai L, Zafereo M, Cabanillas ME. Thyroid cancer: a review. JAMA 2024 Feb 6;331(5):425-435. [doi: 10.1001/jama.2023.26348] [Medline: 38319329]
13. Pitt SC, Zanocco K, Sturgeon C. The patient experience of thyroid cancer. Endocrinol Metab Clin North Am 2022 Dec;51(4):761-780. [doi: 10.1016/j.ecl.2022.04.002] [Medline: 36244692]
14. Nabhan F, Dedhia PH, Ringel MD. Thyroid cancer, recent advances in diagnosis and therapy. Int J Cancer 2021 Sep 1;149(5):984-992. [doi: 10.1002/ijc.33690] [Medline: 34013533]
15. Gritsenko PG, Ilina O, Friedl P. Interstitial guidance of cancer invasion. J Pathol 2012 Jan;226(2):185-199. [doi: 10.1002/path.3031] [Medline: 22006671]
16. Chapman JD. Measurement of tumor hypoxia by invasive and non-invasive procedures: a review of recent clinical studies. Radiother Oncol 1991;20 Suppl 1:13-19. [doi: 10.1016/0167-8140(91)90181-f] [Medline: 2020762]
17. Khan A, Hasana U, Nadeem IA, et al. Advances in colorectal cancer screening and detection: a narrative review on biomarkers, imaging and preventive strategies. J Egypt Natl Canc Inst 2025 Apr 11;37(1):20. [doi: 10.1186/s43046-025-00277-z] [Medline: 40214837]
18. Zhang L, Li D, Su T, Xiao T, Zhao S. Effectiveness of radiomics-based machine learning models in differentiating pancreatitis and pancreatic ductal adenocarcinoma: systematic review and meta-analysis. J Med Internet Res 2025 Jul 31;27:e72420. [doi: 10.2196/72420] [Medline: 40744488]
19. Liang C, Bian Z, Lv W, Chen S, Zeng D, Ma J. A computer-aided diagnosis scheme of breast lesion classification using GLGLM and shape features: combined-view and multi-classifiers. Phys Med 2018 Nov;55:61-72. [doi: 10.1016/j.ejmp.2018.10.016] [Medline: 30471821]
20. Pundavela J, Roselli S, Faulkner S, et al. Nerve fibers infiltrate the tumor microenvironment and are associated with nerve growth factor production and lymph node invasion in breast cancer. Mol Oncol 2015 Oct;9(8):1626-1635. [doi: 10.1016/j.molonc.2015.05.001] [Medline: 26009480]
21. Bilello M, Akbari H, Da X, et al. Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. Neuroimage Clin 2016;12:34-40. [doi: 10.1016/j.nicl.2016.03.007] [Medline: 27358767]
22. Mu J, Cao Y, Zhong X, Diao W, Jia Z. Prediction of cervical lymph node metastasis in differentiated thyroid cancer based on radiomics models. Br J Radiol 2024 Feb 28;97(1155):526-534. [doi: 10.1093/bjr/tqae010] [Medline: 38366237]
23. Dencic TI, Cvejic D, Paunovic I, Tatic S, Havelka M, Savin S. Cytokeratin19 expression discriminates papillary thyroid carcinoma from other thyroid lesions and predicts its aggressive behavior. Med Oncol 2013 Mar;30(1):362. [doi: 10.1007/s12032-012-0362-1] [Medline: 23269585]
24. Yan Q, Chen P, Chen X, et al. Expression of CK7, CK19 and p16 in HPV-mediated oropharyngeal squamous cell carcinoma. PeerJ 2024;12:e18286. [doi: 10.7717/peerj.18286] [Medline: 39619189]
25. Laishram H, Jagadale K, Shelke P, Bharadwaj R. Cross-sectional study to determine the diagnostic accuracy of cytokeratin-19 immunomarker in the diagnosis of thyroid malignancy with histopathology as gold standard. Niger Postgrad Med J 2025 Jan 1;32(1):61-67. [doi: 10.4103/npmj.npmj_3_25] [Medline: 40091473]
26. Chen J, Wang H, Zhou L, Liu Z, Tan X. A combination of circulating tumor cells and CA199 improves the diagnosis of pancreatic cancer. J Clin Lab Anal 2022 May;36(5):e24341. [doi: 10.1002/jcla.24341] [Medline: 35334495]
27. Ashi A, Al-Hajeili M, Almaghrabi S, et al. Prevalence of CEA, CA 125, and CA 15-3 serum tumour markers in different regions of Saudi Arabia. Saudi Med J 2024 Jun;45(6):565-571. [doi: 10.15537/smj.2024.45.6.20230878] [Medline: 38830664]
28. Chen SH, Zhang BY, Zhou B, Zhu CZ, Sun LQ, Feng YJ. Perineural invasion of cancer: a complex crosstalk between cells and molecules in the perineural niche. Am J Cancer Res 2019;9(1):1-21. [Medline: 30755808]

XSL•FO
RenderX

29. Sun P, Zhao Y, Men J, et al. Application of virtual and augmented reality technology in hip surgery: systematic review. J Med Internet Res 2023 Mar 10;25:e37599. [doi: 10.2196/37599] [Medline: 36651587]

Abbreviations

AUC: area under the curve
CA125: carbohydrate antigen 125
CA199: carbohydrate antigen 199
CCSA-2: circulating cancer-specific antigen
CEA: carcinoembryonic antigen
CK19: cytokeratin 19
CT: computed tomography
Gal-3: galectin-3
GLCM: gray-level co-occurrence matrix
HBME-1: Hector Battifora Mesothelial Epitope-1
ICC: intraclass correlation coefficient
LASSO: least absolute shrinkage and selection operator
LDA: linear discriminant analysis
MAE: mean absolute error
NI: neural invasion
NN: neural network
OR: odds ratio
PBMC: peripheral blood mononuclear cell
PCA: principal component analysis
qPCR: quantitative polymerase chain reaction
RF: random forest
RGB: red, green, and blue
ROC: receiver operating characteristic
ROI: regions of interest
SHAP: Shapley Additive Explanations
XAI: explainable artificial intelligence

Edited by A Coristine; submitted 12.May.2025; peer-reviewed by J Zhang, T Yusuff; revised version received 25.Jan.2026; accepted 26.Jan.2026; published 12.Mar.2026.

Please cite as:
Cong FF, Tian K, Gao Q, Wang F, Sun P, Xu N
CT Radiomics–Based Machine Learning Model for Predicting Capsular and Neural Invasion in Thyroid Carcinoma: Diagnostic Accuracy Study
JMIR Med Inform 2026;14:e77349
URL: https://medinform.jmir.org/2026/1/e77349
doi:10.2196/77349

© Fang-fang Cong, Ke Tian, Qian Gao, Fulin Wang, Peng Sun, Nan Xu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Image Degradation on Deep Neural Network Classification of Scaphoid Fracture Radiographs: Comparison Study of Different Noise Types

Chihung Lin[1,2*], PhD; Alfred P Yoon[3*], MD; Chien-Wei Wang[4], MD; Tung Chao[1], MS; Kevin C Chung[4*], MS, MD; Chang-Fu Kuo[1,5,6,7*], MD, PhD

[1]Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

[2]Department of Artificial Intelligence, Chang Gung University, Taoyuan, Taiwan

[3]Division of Plastic and Reconstructive Surgery, Section of Hand and Microvascular Surgery, University of California Davis Medical Center, Sacramento, CA, United States

[4]Comprehensive Hand Center, Michigan Medicine, Ann Arbor, MI, United States

[5]Division of Rheumatology, Allergy and Immunology, Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, No.5, Fuxing Street, Guishan District, Taoyuan City 333, Taoyuan, Taiwan

[6]College of Medicine, Chang Gung University, Taoyuan, Taiwan

[7]Division of Rheumatology, Orthopaedics and Dermatology, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Chang-Fu Kuo, MD, PhD
Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Abstract

Background: Deep learning models have shown strong potential for automated fracture detection in medical images. However, their robustness under varying image quality remains uncertain, particularly for small and subtle fractures, such as scaphoid fractures. Understanding how different types of image perturbations affect model performance is crucial for ensuring reliable deployment in clinical practice.

Objective: This study aimed to evaluate the robustness of a deep learning model trained to detect scaphoid fractures in radiographs when exposed to various image perturbations. We sought to identify which perturbations most strongly impact performance and to explore strategies to mitigate performance degradation.

Methods: Radiographic datasets were systematically modified by applying Gaussian noise, blurring, JPEG compression, contrast-limited adaptive histogram equalization, resizing, and geometric offsets. Model accuracy was evaluated across different perturbation types and levels. Image quality was quantified using peak signal-to-noise ratio and structural similarity index measure to assess correlations between degradation and model performance.

Results: Model accuracy declined with increasing perturbation severity, but the extent varied across perturbation types. Gaussian blur caused the most substantial performance drop, whereas contrast-limited adaptive histogram equalization increased the false-negative rate. The model demonstrated higher resilience to color perturbations than to grayscale degradations. A strong linear correlation was found between peak signal-to-noise ratio–structural similarity index measure and accuracy, suggesting that better image quality led to improved detection. Geometric offsets and pixel value rescaling had minimal influence, whereas resolution was the dominant factor affecting performance.

Conclusions: The findings indicate that image quality, especially resolution and blurring, substantially influences the robustness of deep learning–based fracture detection models. Ensuring adequate image resolution and quality control can enhance diagnostic reliability. These results provide valuable insights for designing more accurate and resilient medical imaging models under real-world variability.

(JMIR Med Inform 2026;14:e65596) doi:10.2196/65596

KEYWORDS

scaphoid fractures; contrast-limited adaptive histogram equalization; CLAHE; image quality; peak signal-to-noise ratio; PSNR; structural similarity index measure; SSIM; neural network.

XSL•FO
RenderX

Introduction

Significant advances in computer vision have led to remarkable improvements in deep neural network (DNN) performance on tasks such as medical image classification [1]. Despite these achievements, DNN-based systems exhibit limited robustness compared with human perception, especially in the context of image perturbations and corruption [2]. Adversarial attacks can cause substantial misclassifications, and even minor noise can adversely affect performance [3-5]. Nevertheless, state-of-the-art classifiers appear to exhibit some ability to overcome random noise [6].

In real-world medical imaging, DNN model robustness is critical because low-quality images and noise are not uncommon; these issues include radiographic underexposures and motion artifacts as well as downstream processing or transfer distortions that can alter image fidelity and affect artificial intelligence (AI) performance [7,8]. In addition, real-world clinical workflows may involve low-resolution screen recapturing or smartphone-captured radiographs in urgent or resource-limited settings, which introduce compounded degradations, such as resizing, compression, and display-related artifacts [9]. Our previous work on a DNN model that detected both visible and occult scaphoid fractures demonstrated that it was possible to reliably detect fractures of small bones and to assist in the radiographic detection of occult fractures that are not visible to human observers [10]. This experience indicated that image preprocessing, file formatting, and data storage could negatively impact model performance. We discovered that image processing techniques introduced noise into input files, potentially misleading the DNN model. Such noise, often imperceptible to the human eye, can significantly affect model performance and potentially lead to incorrect diagnoses [11]. Consequently, evaluation of performance is vital in convolutional neural network (CNN)–based models that process noisy images; this evaluation poses frequent challenges in clinical settings. Prior studies have revealed that neural networks are most accurate when the data to be classified exhibit quality similar to that of the model training data; it was recommended that noise be injected into the training data to increase model robustness [12]. However, the addition of many possible noises to a training dataset is both computationally expensive and impractical. To the best of our knowledge, no study has investigated the impacts of different types of noise on neural networks designed for fracture classification when radiographs serve as inputs.

This study investigated the robustness of our CNN-based scaphoid fracture classification model when various types of image noise were present. To comprehensively evaluate model performance, we simulated real-world conditions that yield low-quality clinical images. Using fine-tuning techniques to create noisy samples, we sought to enhance model performance under noisy or degraded conditions, thereby mitigating the adverse effects of image corruption [7]. Additionally, we examined the impacts of specific forms of image degradation on model performance; we explored particular model vulnerabilities that might warrant further refinement. Our goals were to ascertain model accuracy when processing low-quality clinical images and to identify the types of noise that most confounded the model, thereby providing valuable insights to aid the development of more resilient DNN-based medical image classifiers appropriate for real-world applications.

Methods

Ethical Considerations

The study protocol was approved by the institutional review board of Chang Gung Memorial Hospital (202202256B0).

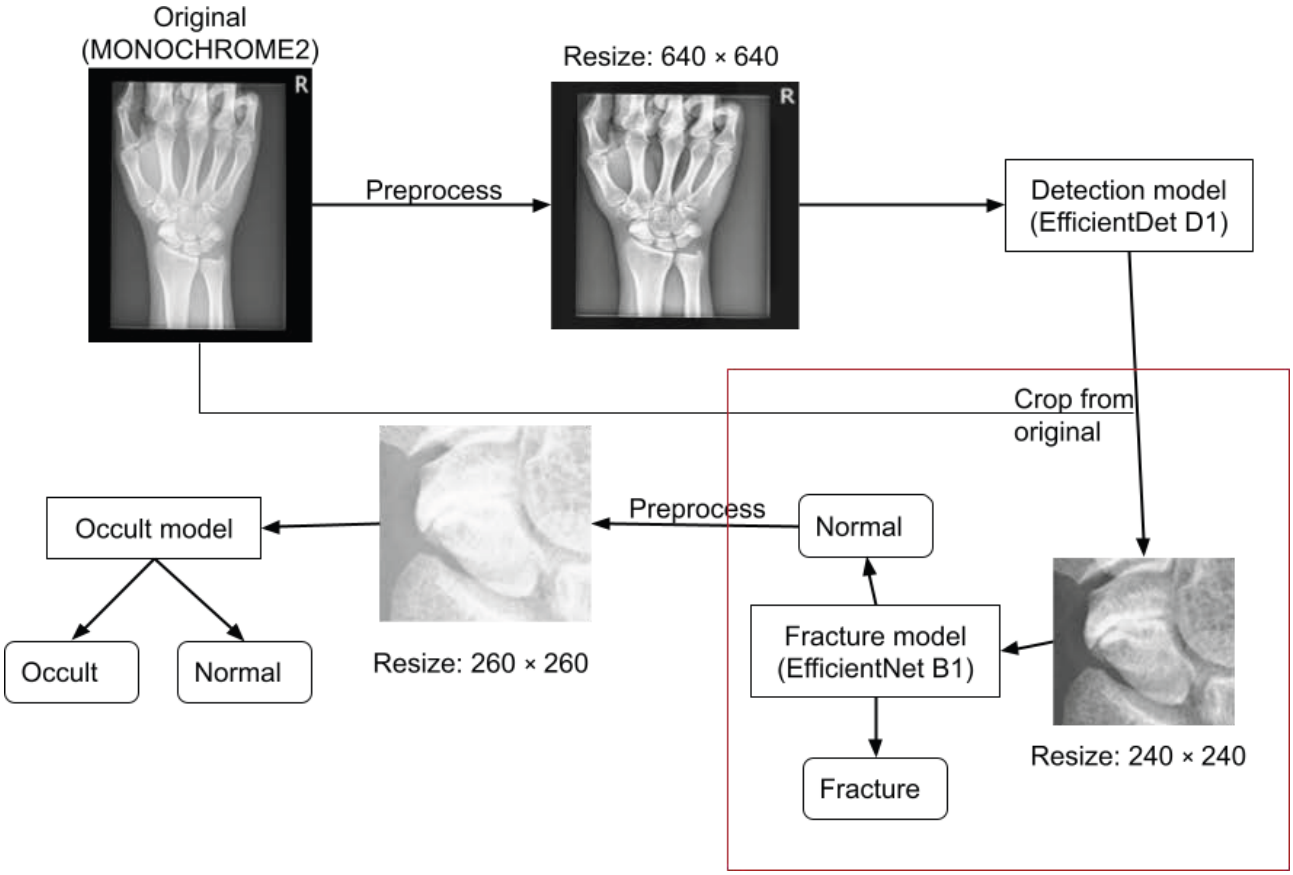
Scaphoid Fracture Classification Model

As previously described in our earlier work [10], the scaphoid fracture classification model was built using an EfficientNetB1 [13] backbone and 240×240-pixel red, green, and blue (RGB) images with a classification threshold of 0.5. In a study by Yoon et al [10], the model was originally trained and validated using 3991 scaphoid fracture radiographs and 5542 normal scaphoid radiographs.

In this study, we adopted the same model architecture as in the study by Yoon et al [10] and initialized the network with the finalized pretrained weights from that study, as shown in Figure 1. We then further fine-tuned this model using the 5286 training radiographs from the dataset described in the “Dataset and Preprocessing” section and evaluated it on an independent test set. Fine-tuning was performed using the AdamW optimizer, with appropriate adjustments to the learning rate, weight decay, and batch size. The learning rate was reduced if the validation loss failed to improve over 6 epochs, and training was stopped early when model performance did not increase further after 15 epochs.

XSL•FO
RenderX

Figure 1. Complete pipeline for inference of a scaphoid fracture, including fracture (red box) and occult fracture. This report solely focuses on the performance of the fracture model.



Dataset and Preprocessing

The scaphoid fracture classification model used in this study was fine-tuned and evaluated using 5954 radiographs stored in Chang Gung Memorial Hospital in Taiwan. Of these radiographs, 1483 were of fractured scaphoids, and the remaining 4471 were of normal scaphoids. Images were captured by commercially available x-ray machines from multiple manufacturers. All radiographs were reviewed by 3 experienced radiologists and labeled as *fractured* or *normal*. This dataset is distinct from the dataset used in our previous work [10], in which the model was originally developed using 3991 fracture and 5542 normal radiographs. In this study, we adopted the same model architecture as in the study by Yoon et al [10], initialized the network with the finalized pretrained weights from that study, and further fine-tuned the model using the 5286 training images described below before evaluating it on an independent test set.

The training dataset consisted of 5286 images, and the test dataset comprised 668 images. A total of 28 images were excluded for reasons including anatomical anomalies secondary to arthritis (12 images), 3 due to unclear laterality, 2 due to wrong-sided imaging, image artifacts introduced by the hardware (1 image), and imaging findings of likely previous fractures or chronic nonunion (10 images). Thus, the final test dataset contained 640 images. The images varied in size and position, but all were 12-bit grayscale images of posteroanterior views of the wrist. Most images were rectangular, with widths ranging from 1000 to 1600 pixels, and heights ranging from 1600 to 2200 pixels. If the photometric interpretation of an image was MONOCHROME1, the image was converted to MONOCHROME2. The scaphoid was isolated from each hand radiograph using a bounding box generated by a separate scaphoid detection model. This detector (not to be confused with the scaphoid fracture classifier) was not the focus of this study because its performance is robust against common perturbations.

Rescaling Images to 8-Bit Depth

All radiographs were 12-bit grayscale posteroanterior wrist views that varied in terms of size and position, with widths between 1000 and 1600 pixels, and heights between 1600 and 2200 pixels. The images were rescaled to 8-bit grayscale and compiled into RGB images using the value of interest (VOI) lookup table and windowing operations that were also used during model training.

The DICOM images were stored in 12-bit grayscale, but the model accepted only 24-bit RGB (color) images. Thus, the images were first rescaled to 8-bit grayscale (with pixel values ranging from 0 to 255) and then compiled into RGB images for interpretation. This rescaling process was based on the maximum and minimum values of the image, rather than the actual bit depth. A VOI lookup table and windowing operations were applied to all images to adjust the pixel values based on the DICOM VOI LUT and windowing tags (pydicom. pixel_data_handlers.apply_voi_lut, version 1.4, 2022).

Differences emerge when images do not contain the lowest and highest pixel values. For 12-bit grayscale images, these pixel

XSL•FO
RenderX

values are 0 and 4095, respectively. The resulting difference can brighten or darken the entire image. Furthermore, the use of a VOI lookup process may change the brightness and contrast. These differences are minor and usually imperceptible to the human eye.

Generating Datasets for Noise Testing

Overview

To assess model performance in the presence of distortions, we generated several low-quality datasets using different noise perturbations. First, we read the test images using method 1. Images were cropped to the scaphoid regions demarcated by the bounding boxes, resized to 240×240 pixels, converted to 8-bit grayscale, and saved in PNG format. We refer to these baseline images (without noise) as “clean” hereafter.

To systematically evaluate the robustness of our scaphoid fracture classification model, we created multiple “noisy” or “degraded” datasets by applying a variety of image perturbations to our *clean baseline test set* of 640 scaphoid radiographs. Each perturbation type and severity level corresponds to a row in Table 1, which lists the name of the treatment, the number of images, and the specific operation performed. Next, we describe the main categories of perturbations and their implementations.

XSL•FO
RenderX

Table . Information on all datasets, including the datasets from which they were modified, sample counts, and methods of modification.

Treatment name	Modified from	N	Operation
Clean dataset	Original DICOM file	640	Read original DICOM file with (1)
Gaussian blur_0.5	Scaphoid	640	Apply Gaussian blur (σ=0.5) to the cropped scaphoid
Gaussian blur_1.0	Scaphoid	640	Apply Gaussian blur (σ=1.0) to the cropped scaphoid
Gaussian blur_1.5	Scaphoid	640	Apply Gaussian blur (σ=1.5) to the cropped scaphoid
Gaussian blur_2.0	Scaphoid	640	Apply Gaussian blur (σ=2.0) to the cropped scaphoid
Gaussian blur_2.5	Scaphoid	640	Apply Gaussian blur (σ=2.5) to the cropped scaphoid
Gaussian blur_3.0	Scaphoid	640	Apply Gaussian blur (σ=3.0) to the cropped scaphoid
Gaussian noise (RGB)_1	Scaphoid	640	Add RGB Gaussian noise with SD 1.00 to the cropped scaphoid
Gaussian noise (RGB)_2	Scaphoid	640	Add RGB Gaussian noise with SD 2.00 to the cropped scaphoid
Gaussian noise (RGB)_3	Scaphoid	640	Add RGB Gaussian noise with SD 3.00 to the cropped scaphoid
Gaussian noise (RGB)_4	Scaphoid	640	Add RGB Gaussian noise with SD 4.00 to the cropped scaphoid
Gaussian noise (RGB)_5	Scaphoid	640	Add RGB Gaussian noise with SD 5.00 to the cropped scaphoid
Gaussian noise (RGB)_6	Scaphoid	640	Add RGB Gaussian noise with SD 6.00 to the cropped scaphoid
Gaussian noise (grayscale)_1	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 1.00 to the cropped scaphoid
Gaussian noise (grayscale)_2	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 2.00 to the cropped scaphoid
Gaussian noise (grayscale)_3	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 3.00 to the cropped scaphoid
Gaussian noise (grayscale)_4	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 4.00 to the cropped scaphoid
Gaussian noise (grayscale)_5	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 5.00 to the cropped scaphoid
Gaussian noise (grayscale)_6.0	Scaphoid	640	Convert image to grayscale, add Gaussian noise with SD 6.00 to the cropped scaphoid
JPEG compression_10	Scaphoid	640	Degrade scaphoid images with compression strength=10 (slightest)
JPEG compression_30	Scaphoid	640	Degrade scaphoid images with compression strength=30
JPEG compression_50	Scaphoid	640	Degrade scaphoid images with compression strength=50
JPEG compression_70	Scaphoid	640	Degrade scaphoid images with compression strength=70
JPEG compression_90	Scaphoid	640	Degrade scaphoid images with compression strength=90 (strongest)

XSL•FO
RenderX

Treatment name	Modified from	N	Operation
CLAHE[a]_1	Scaphoid	640	Apply CLAHE (clip limit=1.0, tile grid size=[8, 8]) to the cropped scaphoid
CLAHE_2	Scaphoid	640	Apply CLAHE (clip limit=2.0, tile grid size=[8, 8]) to the cropped scaphoid
CLAHE_3	Scaphoid	640	Apply CLAHE (clip limit=3.0, tile grid size=[8, 8]) to the cropped scaphoid
CLAHE_4	Scaphoid	640	Apply CLAHE (clip limit=4.0, tile grid size=[8, 8]) to the cropped scaphoid
CLAHE_5	Scaphoid	640	Apply CLAHE (clip limit=5.0, tile grid size=[8, 8]) to the cropped scaphoid
Resize_400	Whole image	640	Resize whole images by width=400 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Resize_600	Whole image	640	Resize whole images by width=600 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Resize_800	Whole image	640	Resize whole images by width=800 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Resize_1000	Whole image	640	Resize whole images by width=1000 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Resize_1200	Whole image	640	Resize the whole images by width=1200 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Resize_1400	Whole image	640	Resize whole images by width=1400 while keeping the aspect ratio and crop out the scaphoid by projecting the recorded bounding box coordinates
Geometrics_1	Whole image	5120	Modify the center point of bounding box (bbox) x,y coordinates by −20% and 0% and 20% of bbox length
Geometrics_2	Whole image	5120	Modify the center point of bbox x,y coordinates by −10% and 0% and 10% of bbox length
Geometrics_3	Whole image	10,240	Modify the 4 bbox x,y coordinates (xmin, xmax, ymin, ymax) by −10% and 10% of bbox length
Geometrics_4	Whole image	10,240	Modify the 4 bbox x,y coordinates (xmin, xmax, ymin, ymax) by −5% and 5% of bbox length

XSL•FO
RenderX

Treatment name	Modified from	N	Operation
12-bit-rescale_1	Original DICOM file	640	Read original DICOM file using method (2) and crop the scaphoid with the recorded bbox coordinates
12-bit-rescale_2	Original DICOM file	640	Read original DICOM file using method (2) and crop the scaphoid with the detector
Screenshot_MicroDicom_1	Original DICOM file	640	Read DICOM files using MicroDicom with the default settings, screenshoot the whole image at a resolution of 550×780, and crop the scaphoid with the detector
Screenshot_MicroDicom_2	Original DICOM file	640	Read DICOM files using MicroDicom with the default settings, enter fullscreen mode, screenshoot the whole image at a resolution of 900×1050 and crop the scaphoid with the detector
Screenshot_ImageJ	Original DICOM file	640	Read DICOM files using ImageJ, adjust contrast to 12-bit, screenshoot the whole image at a resolution width of 600 while retaining the height-width ratio, and crop the scaphoid with the detector

[a]CLAHE: contrast-limited adaptive histogram equalization.

Clean Dataset

We refer to the original 640 cropped scaphoid images (taken directly from DICOM files and converted to 8-bit depth) as the "clean" dataset (row 1 in Table 1). This set serves as our baseline for comparison.

Gaussian Blur

We simulated *blur* by convolving each scaphoid region with a Gaussian kernel using SDs (σ) ranging from 0.5 to 3.0 (rows 2-7 in Table 1). Specifically, we used a Python image-augmentation library (eg, *imgaug*; Python Software Foundation) to apply GaussianBlur(σ=x). Each σ setting generated a separate dataset of 640 images.

Gaussian Noise

For grayscale noise (rows 8-13 in Table 1), we converted each cropped scaphoid image to 8-bit grayscale, then added random Gaussian noise with SD values in the set {1, 2, 3, 4, 5, 6}. The resulting noised images were reconverted to RGB (by replicating the grayscale channel 3 times) to match the model's 3-channel input requirement.

For color (RGB) noise (rows 14-19 in Table 1), we similarly added Gaussian noise to the 3 RGB channels, resulting in "colored" noise. SD values were identical (1-6). Each level produced 640 modified images.

JPEG Compression

We degraded the quality of the cropped scaphoid images by JPEG compression (rows 20-24 in Table 1), with a compression strength of 10, 30, 50, 70, and 90 (higher values indicating more severe compression in our chosen library). Each compression level formed a dataset of 640 images.

This process simulates the impact of lossy image storage on fracture detection performance.

Contrast-Limited Adaptive Histogram Equalization

Contrast-limited adaptive histogram equalization (CLAHE; rows 25-29 in Table 1) enhances local contrast in radiographs, potentially exaggerating edges and intensifying noise. We generated 5 datasets by applying OpenCV's createCLAHE() with *cliplimit* in {1.0, 2.0, 3.0, 4.0, 5.0} and tileGridSize=(8,8). Each CLAHE setting yielded 640 images.

Resizing Whole Images

Before cropping the scaphoid region, we resized the *entire* original wrist radiograph (rows 30-35 in Table 1) to widths of 400, 600, 800, 1000, 1200, or 1400 pixels (preserving the aspect ratio), then reapplied our scaphoid bounding box coordinates to crop out the scaphoid. Finally, the cropped regions were resized to 240×240 pixels for model input. This procedure mimics variations in image resolution and scaling during clinical acquisition or display.

Geometric Offsets

To examine the robustness of our model to bounding box inaccuracies, we systematically shifted or distorted the bounding box coordinates by ±5%, 10%, or ±20% of the bounding box size. This sometimes resulted in only partial scaphoid capture. Four separate datasets (Geometrics_1 to Geometrics_4) covered different offset ranges (rows 36-39 in Table 1). Each row includes multiple transformations, so the total number of images can exceed 640.

Twelve-bit Rescaling

Instead of converting the DICOM images with our typical method (method 1), we used an alternative approach (method

XSL•FO
RenderX

2) that directly scales 12-bit raw pixel values (0-4095) into 8-bit (0-255). We generated 2 variations: (1) 12-bit-rescale_1: recorded bounding box coordinates were applied to these rescaled images; (2) 12-bit-rescale_2: we reran the scaphoid detection model on these rescaled images to obtain new bounding boxes.

Screenshot Datasets

We opened the original DICOM files in different DICOM viewing software (eg, MicroDicom; ImageJ developed by National Institutes of Health and the Laboratory for Optical and Computational Instrumentation), adjusted the default display or resolution, and took screenshots of the entire wrist X-ray. Screen resolutions varied (eg, 550×780 or 900×1050; rows 42-44 in Table 1). We then cropped out the scaphoid region using our detection model, resizing the final images to 240×240. Each screenshot setting introduced different display parameters, simulating suboptimal clinical scenarios where images may be shared or interpreted via screenshots instead of original DICOM files.

Each of these modifications produced a new dataset of 640 images (except when multiple bounding box transformations were applied, resulting in a large number of images). Collectively, these datasets allowed us to evaluate the effects of image degradations on the classification model's performance. Table 1 provides a concise summary of all transformations, whereas Multimedia Appendix 1 contains more detailed code snippets and pseudocode for each operation.

Additionally, 5 datasets were produced using the rescale method or by capturing screenshots from the original DICOM files on image viewers. Figure 2 and Table 1 detail all 44 perturbations. The details of noise introduction into the dataset are described in Multimedia Appendix 1.

Figure 2. Flowchart of the pipeline used to generate clean and noisy datasets. CLAHE: contrast-limited adaptive histogram equalization;

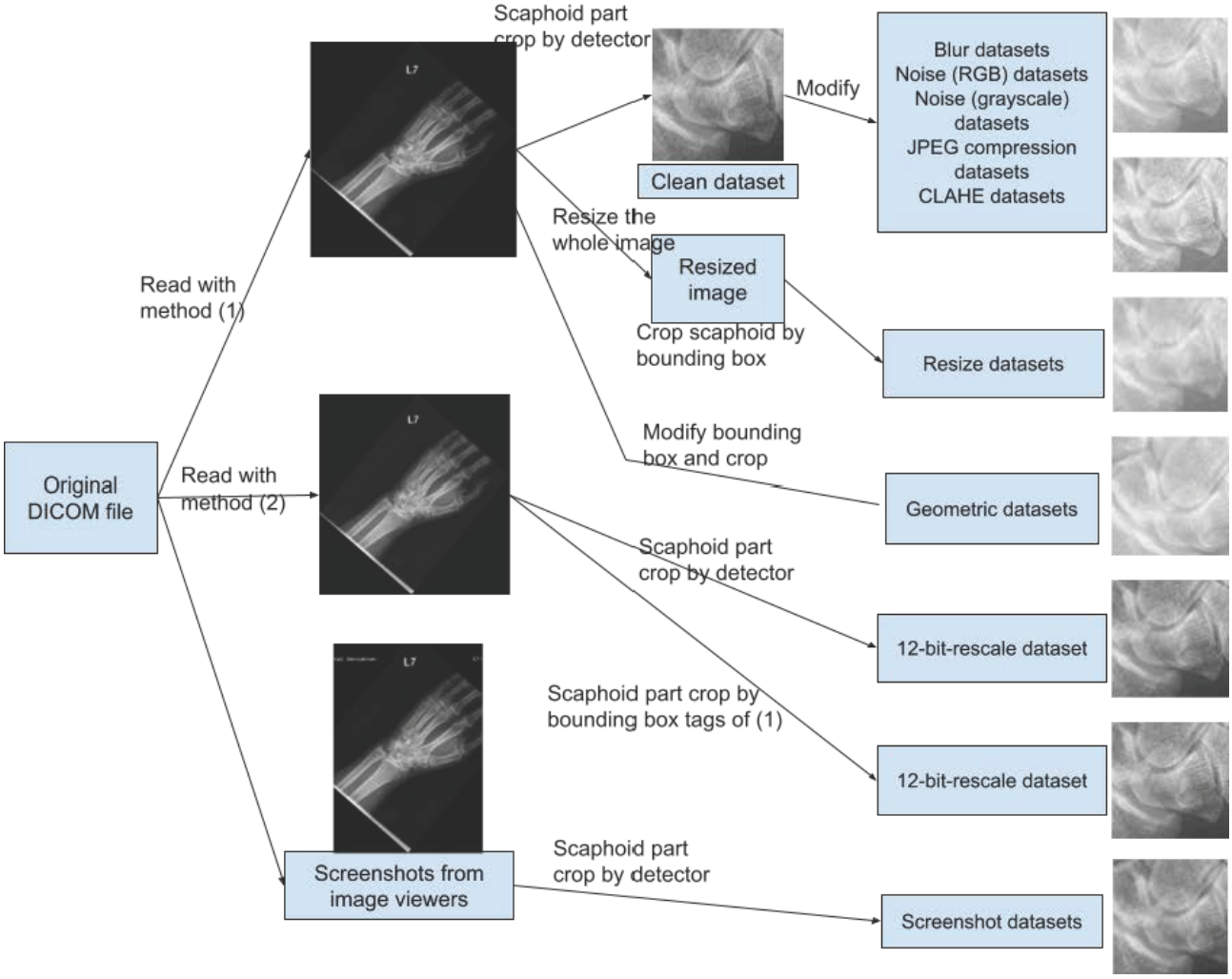


Image Quality Assessment

The peak signal-to-noise ratio (PSNR) and the structural similarity indexing method (SSIM) [14,15] were used as image quality assessment methods when gauging alterations in images. PSNR is a widely used metric that assesses the fidelity of an image compared with its original or uncompressed version. PSNR quantifies the difference between 2 images by calculating the ratio of the maximum possible power of the signal (the image) to the power of the noise (the error that is introduced). A higher PSNR value indicates better image quality because less noise is introduced during changes. In contrast, SSIM is a more advanced metric for measuring image quality that considers structural information, luminance, and contrast when comparing 2 images. SSIM calculates local similarities between the 2 images and combines them into a single score, ranging

XSL•FO
RenderX

from −1 to 1, where a higher value indicates greater similarity between the images. Both PSNR and SSIM values were calculated using TensorFlow image module (TensorFlow Module: tf.image), which allows efficient and accurate computations of both metrics. However, neither PSNR nor SSIM can be used for geometric transformations, such as affine and rotational adjustments, because both rely on pixel-by-pixel comparisons. Such comparisons become less meaningful when the spatial arrangement of image content is altered via geometric transformations. Consequently, datasets with modifications to the labeled scaphoid regions were excluded from comparison to maintain assessment integrity.

We calculated and compared PSNR and SSIM values for the clean and noisy images. Importantly, neither assessment is amenable to geometric transformations, such as affine and rotational adjustments. Therefore, the following datasets with modifications to the labeled scaphoid regions were excluded from the PSNR and SSIM comparisons: the 4 geometric datasets, the 12-bit rescale_2 datasets, and the 3 screenshot datasets. Both PSNR and SSIM were calculated using TensorFlow image functions (PSNR and SSIM).

Model Evaluation and Performance Metrics

The scaphoid fracture classification model was evaluated by analyzing its performance on various test datasets that included clean images and distorted images with diverse noise levels. Performance metrics, including model accuracy, sensitivity, specificity, and F_1-score, were calculated by comparing the model predictions with ground truth labels provided by experienced orthopedic and hand surgeons.

Robustness Against Image Quality Degradation

To assess model robustness against image quality degradation, model performances on distorted images were compared with model performances on clean images. This comparison sought to estimate fracture classification accuracy when perturbations were present, offering insights into potential real-world applications of the model and its robustness when image quality varies in clinical settings.

The Environment

Model inferences were executed on Linux Ubuntu 18.04 LTS (GPU: NVIDIA GeForce RTX 3080 (10 GB); Python version 3.7.13). Implementation was conducted with a TensorFlow backend and TensorFlow version 2.9.1.

Results

Performance of the Deep Learning Model With Varied Perturbations

We investigated the efficacy of a deep learning model designed to detect scaphoid fractures in radiographs, specifically when image quality had been deliberately compromised by adding perturbations. The primary datasets were altered using various methods: addition of Gaussian noise, blurring, JPEG compression, CLAHE, resizing, and geometric adjustments. Table 2 presents the quality assessment and model performance results across the evaluated datasets. Specifically, datasets that underwent blurring, Gaussian noise addition, JPEG corruption, CLAHE, or resizing are reported in rows 1–35; geometric datasets in rows 36–39; bit-rescaled datasets in rows 40–41; and screenshot datasets in rows 42–44. The model achieved an accuracy of 92.03% on the original unaltered dataset, which thus served as the performance benchmark for other evaluations.

XSL•FO
RenderX

Table . The quality assessment and model performance results across the evaluated datasets.

Treatment	PSNR[a], mean (SD)	Average SSIM[b], mean (SD)	Accuracy	Precision	Recall
Clean dataset	Inf	1.0000	0.9203	0.915	0.9119
Gaussian blur _0.5	52.5622 (2.6778)	0.9975 (0.0010)	0.9031	0.8896	0.9017
Gaussian blur _1.0	42.7419 (2.5447)	0.9758 (0.0096)	0.8422	0.7836	0.9085
Gaussian blur _1.5	39.9809 (2.5143)	0.9545 (0.0172)	0.8078	0.7299	0.9254
Gaussian blur _2.0	37.2251 (2.3945)	0.9183 (0.0272)	0.7625	0.6819	0.9085
Gaussian blur _2.5	35.4838 (2.3231)	0.8842 (0.0356)	0.7250	0.6530	0.8610
Gaussian blur _3.0	34.8328 (2.3077)	0.8674 (0.0399)	0.7203	0.6526	0.8407
Gaussian noise (grayscale)_1	47.7838 (0.1157)	0.9917 (0.0018)	0.9109	0.9190	0.8847
Gaussian noise (grayscale)_2	42.0190 (0.0432)	0.9697 (0.0066)	0.8906	0.8840	0.8780
Gaussian noise (grayscale)_3	38.5493 (0.0308)	0.9357 (0.0135)	0.8688	0.8436	0.8780
Gaussian noise (grayscale)_4	36.0689 (0.0271)	0.8929 (0.0216)	0.8422	0.7994	0.8780
Gaussian noise (grayscale)_5	34.1382 (0.0267)	0.8444 (0.0298)	0.8172	0.7602	0.8814
Gaussian noise (grayscale)_6	32.5600 (0.0276)	0.7932 (0.0376)	0.7734	0.7027	0.8814
Gaussian noise (RGB)_1	47.7851 (0.1134)	0.9917 (0.0018)	0.9156	0.9288	0.8847
Gaussian noise (RGB)_2	42.0211 (0.0368)	0.9697 (0.0065)	0.9094	0.9100	0.8915
Gaussian noise (RGB)_3	38.5487 (0.0225)	0.9357 (0.0135)	0.8969	0.8962	0.8780
Gaussian noise (RGB)_4	36.0686 (0.0180)	0.8929 (0.0215)	0.8891	0.8784	0.8814
Gaussian noise (RGB)_5	34.1376 (0.0173)	0.8444 (0.0299)	0.8547	0.8279	0.8644
Gaussian noise (RGB)_6	32.5599 (0.0168)	0.7931 (0.0376)	0.8219	0.7701	0.8746
JPEG compression_10 (slightest)	44.5076 (1.7566)	0.9829 (0.0046)	0.9141	0.9255	0.8847
JPEG compression_30	40.8137 (1.6190)	0.9619 (0.0091)	0.8922	0.9154	0.8441
JPEG compression_50	38.7102 (1.4519)	0.9411 (0.0119)	0.8688	0.8893	0.8169
JPEG compression_70	36.7313 (1.3293)	0.9125 (0.0150)	0.8609	0.8732	0.8169
JPEG compression_90 (strongest)	32.6132 (1.1419)	0.8146 (0.0233)	0.7656	0.6845	0.9119
CLAHE[c]_1	27.5678 (1.9698)	0.932 (0.0095)	0.8875	0.9176	0.8305
CLAHE_2	21.3350 (1.4853)	0.7836 (0.0182)	0.8250	0.9256	0.6746
CLAHE_3	18.7906 (1.3643)	0.6860 (0.0220)	0.7953	0.9457	0.5898
CLAHE_4	16.6896 (1.2363)	0.5904 (0.0262)	0.7812	0.9641	0.5458
CLAHE_5	15.6361 (1.1716)	0.5395 (0.0306)	0.7703	0.9353	0.5390
Resize_400	27.4436 (2.8561)	0.6578 (0.1013)	0.7188	0.6353	0.9153
Resize_600	29.2871 (3.0279)	0.7108 (0.1042)	0.8172	0.7514	0.9017
Resize_800	30.8330 (3.07574)	0.7651 (0.1007)	0.8516	0.8106	0.8847

XSL•FO
RenderX

Treatment	PSNR[a], mean (SD)	Average SSIM[b], mean (SD)	Accuracy	Precision	Recall
Resize_1000	32.3258 (3.6219)	0.8130 (0.0932)	0.8672	0.8454	0.8712
Resize_1200	33.3882 (3.8631)	0.8432 (0.08724)	0.8797	0.8682	0.8712
Resize_1400	34.5316 (3.6730)	0.8754 (0.0748)	0.8938	0.8874	0.8814
Geometrics_1	N/A[d]	N/A	0.8635	0.8524	0.8513
Geometrics_2	N/A	N/A	0.8949	0.9003	0.8682
Geometrics_3	N/A	N/A	0.8827	0.8765	0.8678
Geometrics_4	N/A	N/A	0.9021	0.9041	0.8809
12-bit-rescale_1	48.0188 (8.0564)	0.9945 (0.0181)	0.9188	0.9119	0.9119
12-bit-rescale_2	N/A	N/A	0.9141	0.9027	0.9119
Screenshot_MicroDicom_1	N/A	N/A	0.8281	0.7666	0.9017
Screenshot_MicroDicom_2	N/A	N/A	0.8625	0.8439	0.8610
Screenshot_ImageJ	N/A	N/A	0.7812	0.7123	0.8814

[a]PSNR: peak signal-to-noise ratio.

[b]SSIM: structural similarity index measure.

[c]CLAHE: contrast-limited adaptive histogram equalization.

[d]N/A: not applicable.

The Effects of Noise and Perturbations

This section discusses model performance on datasets directly modified from the cropped scaphoid radiographs and resized datasets without geometric modifications. Table 2 presents the quality assessment and model performance results across the evaluated datasets. Specifically, datasets that underwent blurring, Gaussian noise addition, JPEG corruption, CLAHE, or resizing are reported in rows 1–35; geometric datasets in rows 36-39; bit-rescaled datasets in rows 40-41; and screenshot datasets in rows 42-44. Further details can be found in Multimedia Appendix 1.

For datasets that included cropped scaphoid radiographs and resized images, if the severity of image degradation was minimal, the effect on model accuracy was negligible. An important observation was that the model exhibited varying degrees of resilience against different image perturbations. Even in some datasets with similar PSNR and SSIM values and comparable image quality, some discrepancies in model performances were observed. These findings underscore the nuanced robustness of the deep learning model against different image distortions.

On some noisy datasets, such as Gaussian blur_0.5, Gaussian noise (RGB)_1, Gaussian noise (grayscale)_1, and JPEG compression_10, the model performances were similar to that on the clean dataset. These datasets had the least severe distortions; the pixel values changed minimally, as indicated by the high PSNR and SSIM values. However, as the perturbations increased in severity, image features deteriorated further, and model performance declined. This trend was observed across all treatments, although the extent of performance decline varied according to the type of perturbation. We conclude that if the severity of a perturbation can be maintained below a specific level in degraded images, the model can maintain good performance.

Robustness Against Different Types of Perturbations

Next, we compared model robustness across different image perturbations. Although some datasets yielded similar average PSNR and SSIM assessments, model performances differed. This finding suggested that the model is more robust against certain types of distortions but more vulnerable to others.

Color and Grayscale Gaussian Noise

We expected that the PSNR and SSIM assessments would be similar after the introduction of Gaussian noise, regardless of whether the noise was in color or grayscale. This expectation was confirmed by the similar average PSNR and SSIM values of the corresponding RGB and grayscale Gaussian noise datasets. Nevertheless, model accuracy was considerably lower when grayscale noise was present, suggesting that the model solely enhanced resilience to color perturbations. This may be explained by the fact that the training set was exclusively composed of grayscale samples, although the model input layer accepted 3-channel color images.

Gaussian Blurring

Gaussian blurring substantially degraded model performance, comparatively more than other noises with similar image quality metrics. For images with PSNR values between 35 and 50, linear regression analysis revealed that Gaussian blurring was associated with a 5% to 12% decrease in accuracy compared with other nongeometric transformation techniques. Although blurring an image does not significantly decrease image quality, blurring likely affects fracture features and thus negatively

XSL•FO
RenderX

affects detection. This finding is consistent with the results of a previous study, which concluded that neural networks are very sensitive to blurring, probably because textures and edges are removed [16-18].

JPEG Compression

The effect of JPEG compression on model performance was similar to the effect of grayscale Gaussian noise. The neural network was surprisingly resilient to JPEG compression. Even after a 90% file size reduction via JPEG compression, the model achieved an accuracy of greater than 81%. This finding has important clinical implications: DICOM images, which are often larger than 20 MB, can be compressed by up to 70% via JPEG, but the model will maintain 90% accuracy in terms of detecting scaphoid fractures. This capability will render model implementation more computationally efficient.

Resizing

The resizing perturbations were designed to simulate a realistic low-resolution workflow in which radiographs may be downsampled before region extraction and later rescaled for AI inference; therefore, this setting reflects the combined effects of downscaling information loss and upscaling or interpolation artifacts from small cropped regions of interest (ROIs), rather than a purely isolated resolution test. The downscaling treatments required the new image to store information using fewer pixels than the original, which forced the image to compress its content. As a result, a substantial amount of information was lost when the number of available storage units was reduced, leading to blurry images and lowered image quality, as demonstrated by the PSNR and SSIM values. However, fracture detection accuracy did not significantly decline until the PSNR values fell below 30. Compared with other noise treatments, resizing an image may adequately preserve the features required for fracture detection.

Contrast-Limited Adaptive Histogram Equalization

CLAHE treatment enhances image contrast and thus dramatically alters the image [19,20]. A CLAHE-enhanced image can inadvertently mislead the model, especially if the model has not been trained with CLAHE-augmented data. Model accuracy decreased as the parameter “clip limit” (maximum limit of the adaptive histogram equalization) increased, considerably lowering the PSNR and SSIM values. However, the poor metric values do not necessarily imply that CLAHE transformation worsens the model performance more severely than that of other perturbations. Even at a cliplimit of 2, the PSNR declined to 21, lower than the PSNR after any Gaussian blur or JPEG treatment, and the model accuracy decreased to 83%. As the cliplimit was subsequently increased, the accuracy precipitously fell to 77% (Table 2). This finding differs from the result of a previous study, in which neural networks were resilient against changes in image contrast [17]. This discrepancy may be because Dodge and Karam [5] investigated contrast reduction only via grayscale image superimposition and assessed correct image classification (eg, a dog and a cat) using network models. In contrast, a scaphoid DNN must detect subtle linear features when identifying fractures, and CLAHE likely obscures contrast along the fracture line. CLAHE is useful when enhancing x-ray images before human interpretation, but it should only be used in neural networks that are trained via CLAHE augmentation. Additionally, if a chosen medical imaging software exhibits built-in CLAHE-enhancing features, the neural network must be trained with CLAHE-augmented data.

Differences in the Reductions of Precision and Recall Rates

The prevalence of scaphoid fractures in the test dataset was 46%. As all noisy datasets were derived from this dataset, the prevalences were identical. Accordingly, a decline in model accuracy can be attributed to either reduced precision or recall rates, with the former creating more false positives and the latter creating more false negatives.

In the last step of training, the model was fine-tuned to achieve precision and recall rates of 91%. We expected that these performance metrics would decrease when interpreting noisy datasets. Intriguingly, the severities of performance decline differed for precision and recall. Although most noisy datasets triggered declines in both precision and recall, the performance deteriorations exhibited by the precision rates were more pronounced. However, the recall rate was usually acceptable, even in heavily altered datasets. As the model seeks to identify all possibly useful scaphoid fractures, it remains clinically robust in terms of detecting fractures despite image perturbation, but at the cost of increased false-positive rates.

Only a few datasets showed the opposite, with a precision rate much greater than the recall rate; these were the JPEG compression_50, JPEG compression_70, and all CLAHE datasets. Conversely, the remaining JPEG compression datasets demonstrated a higher recall than precision rate, similar to other changes. Therefore, JPEG compression is likely not the principal explanation for such findings.

However, we found that CLAHE image treatment must be performed with caution. CLAHE primarily reduces recall (increasing false negatives) while precision tends to increase, indicating a heightened risk of missed fractures at higher clip limits (Table 2). If the images to be interpreted are of low quality and CLAHE enhancements are applied, machine learning scientists should be wary of inadvertently increasing the false-negative rates.

Relationship Between Image Quality Assessments and Model Performance

We combined all noisy datasets (excluding the CLAHE datasets, given their heterogeneity) into a single dataset with 18,560 samples exhibiting various perturbations. We used this combined dataset to investigate the relationship between image quality and model performance.

We calculated 25 quantiles of SSIM value distributions and grouped the images according to quantile; this approach yielded 25 groups with 742 or 743 samples each. Our grouping method effectively randomized the images and eliminated the effect of any particular degradation treatment when grouping images from different datasets by image quality. We regrouped images by SSIM quantiles, rather than PSNR quantiles, because the

XSL•FO
RenderX

distributions of SSIM values within each noisy dataset were wider and enabled easier stratification.

For each group, the average PSNR, average SSIM, and model accuracy were calculated (Figure 3). We found a strong linear relationship between average image quality and model accuracy. The model performed better on images with higher quality assessments.

XSL•FO
RenderX

Figure 3. The images were divided into 25 groups according to the structural similarity index measure (SSIM) value quantiles. PSNR: peak signal-to-noise ratio.

XSL•FO
RenderX

The average PSNR, average SSIM, and group accuracy for each were plotted; the accuracy exhibited strong linear relationships with both image quality assessments.

Both PSNR and SSIM regressions yielded high adjusted R-squared values (P<.001). As the PSNR and SSIM values are not independent, it is reasonable to expect that the 2 assessments would yield similar results for most image degradations [21]. However, the SSIM was normalized, whereas the PSNR was not. Robust linear associations between image quality evaluations and model performances were evident. Specifically, PSNR and SSIM, indicators of image quality, served as reliable independent tests that predicted AI performance under various perturbations. Higher image quality evaluations were invariably correlated with superior model performances. Although certain perturbations, such as geometric offsets or pixel value rescaling, did not appear to influence model performance, resolution was identified as a key factor, particularly in screenshot images. An enhanced focus on the ROI, such as zooming, effectively averted accuracy reduction. Recent research introduced a framework for the creation of Robust Medical Imaging AI models, which focuses on developing robust AI models for chest radiographs by addressing real-world sources of image degradation such as device heterogeneity, screen-captured inputs, and compression artifacts [22].

Effects of Geometric Offsets

CNN models extract hierarchical features from local regions of an image through convolution and pooling operations. In principle, a moderate change in object location within the image should therefore not drastically confuse the model. Our results generally reflected this expectation: the 4 geometrically modified datasets retained good overall accuracy when the scaphoid remained fully within the field of view. However, accuracy declined as the strength of the affine transformations increased. A likely explanation is that parts of the scaphoid were shifted outside the image boundaries, resulting in cropped or incomplete scaphoid regions being presented to the model. Such extreme geometric distortions are unlikely to occur under real-world clinical conditions.

To further assess whether the detection model contributed to performance degradation under realistic settings, we evaluated its robustness on the 3 real-world screenshot datasets and the 12-bit-rescale_2 dataset and compared these results with those from the 4 geometric offset datasets. As summarized in Table 3, the mean intersection over union (IoU) values for the screenshot and 12-bit-rescale_2 datasets (0.8728, SD 0.0679 to 0.9836, SD 0.0268) were consistently higher than those of the geometric offset datasets (0.5752, SD 0.0952 to 0.8530, SD 0.0222). The geometric offset datasets (Geometrics_1 - 4) were intentionally designed as synthetic baselines to simulate pure localization errors of the detector. The lower IoU values in Geometrics_1 - 3 (0.5752 - 0.7546) correspond to increasingly large artificial offsets, whereas Geometrics_4 (mean 0.8530, SD 0.0222) represents the least-modified geometric condition in which the scaphoid remains fully visible after offset treatment. Notably, Table 3 presents that the detector achieved IoU values exceeding Geometrics_4 on all screenshot datasets, suggesting that the detection model remains stable under realistic screenshot-related degradation.

Table . IoU values of the datasets which the detection model was involved with and the geometric datasets.

Datasets involve detection model	IoU[a], mean (SD)
Screenshot_MicroDicom_1	0.8768 (0.0625)
Screenshot_MicroDicom_2	0.8728 (0.0679)
Screenshot_ImageJ	0.8907 (0.0554)
12-bit-rescale_2	0.9836 (0.0268)
Geometric datasets[b]	
Geometrics_1	0.5752 (0.0952)
Geometrics_2	0.7546 (0.0662)
Geometrics_3	0.7063 (0.0529)
Geometrics_4	0.8530 (0.0222)

[a]IoU: intersection over union.

[b]Considering geometric datasets IoU as baseline, the detection model performed better on all the screenshot datasets, showing its robustness toward perturbations, and that the performance of the detection model would not be a major factor leading to failure of the classification model when image degradation occurs.

As the classification model still maintained reasonable accuracy across the geometric baselines, IoU values comparable to or higher than Geometrics_4 indicate that under real-world screenshot- and rescaling-related perturbations, the detector's localization accuracy is at least similar to its best geometric baseline performance and therefore is less likely to be a dominant contributor to the observed classification performance changes. Overall, these findings suggest that detector-related bounding box offsets likely play a limited role in the performance degradation observed in these realistic conditions.

Effects of Pixel Value Rescaling Methods

The 2 methods for rescaling images from 12-bit to 8-bit did not materially affect model performance. The 2 datasets derived from the alternative rescaling method, 12-bit-rescale_1 and 12-bit-rescale_2, showed only a minor decrease in accuracy

XSL•FO
RenderX

(<1%), although the rescaled images were nearly identical. Examination of the few inconsistent cases revealed that the classification confidence scores for "fracture" and "nonfracture" were much closer in these instances, with differences typically on the order of 10^2. In contrast, in most other cases, the confidence score gap exceeded 10^3, indicating stronger and more stable model decisions. These relatively uncertain confidence scores suggest that the inconsistent cases were inherently difficult for the model to judge, even without perturbation. As the model's decisions for such cases were already unstable, slight modifications introduced by the alternative rescaling algorithms—despite causing only minimal pixel value changes—could flip the predicted label. This likely explains the minor performance differences observed among the clean dataset and the 2 rescaled datasets (12-bit-rescale_1 and 12-bit-rescale_2).

Screenshot Datasets

Taking screenshots was the most complex perturbation of all the studied treatments. Resolution may be the factor that most strongly affects model performance. The only difference between the datasets Screenshot_MicroDicom_1 and Screenshot_MicroDicom_2 was the resolution; changing the resolution from 900×1050 to 550×780 pixels resulted in an accuracy reduction from 86.3% to 82.8%. The dataset Screenshot_ImageJ, the dataset with images resized to a width of 600 pixels while keeping the aspect ratio collected using ImageJ, exhibited the worst resolution and worst read accuracy (78.1%). Such declines in accuracy, precision rate, and recall rate were similar to declines observed in the corresponding resized datasets (Screenshot_MicroDicom_1 with Resize_600, Screenshot_MicroDicom_2 with Resize_1000, and Screenshot_ImageJ with Resize_600). Although the detection model had been involved in the preparation of these datasets, Table 3 implies that in these cases, scaphoids were correctly detected and cropped. In conjunction with the fact that the classification performance of screenshot datasets was worse than that of any of the geometric datasets, the degradation in model performance indeed stems from resizing effects. We also observed that although resolution deterioration was the major factor, the screenshot process did not impact model performance identically to pure resizing. Specifically, Screenshot_ImageJ showed a modest but consistent performance drop compared with its corresponding resized dataset (Resize_600) and also performed worse than another screenshot dataset (Screenshot_MicroDicom_1) collected at a similar resolution. This suggests that viewer-specific display or resampling algorithms during screenshot capture may also influence model performance.

On the basis of these results, one strategy that may prevent accuracy reduction after taking screenshots would be to zoom into the image and then take a screenshot at the 100% level. Although the radiographic image subsequently may not fit within the screen display, it is unnecessary to capture the entire image because the only ROI is the scaphoid.

Computational Cost and Inference Time

We evaluated the average inference time by running through the 640 whole-hand x-ray images from the clean testing dataset using the complete preprocessing, bounding box detection, and classification inference pipeline. We obtained a computation time of approximately 6 to 8 ms per image on an NVIDIA RTX 3080 (10 GB RAM), while warmup runs were not included. The end-to-end pipeline used less than 200 MB GPU memory per 240×240×3 input (including framework overhead), enabling near real-time inference on a single RTX 3080. Consequently, our approach is compatible with near-real-time applications, assuming a well-optimized implementation. Although resizing or blurring images may incur a minor preprocessing overhead, these operations add only approximately 1 ms to the total inference time, indicating that real-time usage in clinical workflow is feasible on a modern GPU.

Discussion

Principal Findings

Taken together, our findings highlight that the deep learning model's performance is inversely related to the severity of image degradation, with Gaussian blur, grayscale noise, and CLAHE standing out as the most disruptive factors. These types of distortions can destroy or obscure the crucial edge information that the network relies upon for detecting subtle fracture lines. In contrast, moderate JPEG compression, resizing, or color noise had a less pronounced effect. Notably, the resizing experiments in this study reflect a realistic low-resolution workflow involving downscaling, then cropping, and then upscaling, which is consistent with real-world recaptured or low-resolution imaging scenarios described in the literature [9]. While accuracy decreases were comparable between Screenshot_ImageJ and the corresponding Resize datasets, Screenshot_ImageJ showed a modest additional reduction in recall, suggesting that screenshot recapture can affect sensitivity beyond pure resizing.

From a clinical perspective, these results underscore the importance of maintaining sufficient image fidelity and avoiding overly aggressive postprocessing steps. If CLAHE-based enhancement or heavy contrast adjustments are used, training the model with matching augmentations may be necessary to preserve performance. Similarly, our analysis indicates that modest compression does not necessarily compromise diagnostic accuracy, suggesting an avenue for reducing file sizes without compromising model output.

Although the network proved fairly robust across various perturbations, investigators should remain cautious when applying transformations, such as extreme blurring, under or overexposure, or very low-resolution screenshots. In these scenarios, crucial details may be irretrievably lost, leading to significantly higher misclassification rates. Our results also emphasize that the precision rate is especially sensitive to noise, often dropping faster than recall. This behavior means that some perturbations can inflate false positives, which may still be acceptable in screening contexts that favor high recall but could impact workflows that depend on precise diagnoses.

XSL•FO
RenderX

In certain clinical scenarios, obtaining the original radiographic file may not be immediately feasible—such as in emergency settings, remote consultations, or when imaging systems have limited data access privileges. In such situations—particularly in emergency department workflows where the original radiographic DICOM file may not be readily accessible because of time pressure, remote consultation needs, or limited system privileges—clinicians or technicians may rely on on-screen screenshots or smartphone photographs for rapid review, sharing, or AI-assisted interpretation. Although practical, these recaptured images can introduce unintended degradation due to variations in capture resolution, scaling distortions, and secondary compression, which may obscure subtle fracture cues. This real-world practice motivated our inclusion of screenshot-based perturbations as a clinically relevant proxy for urgent or resource-limited conditions. To replicate such real-world conditions, our study incorporated a "screenshot" perturbation by capturing radiographs at different image sizes and resolutions. The resulting artifacts mimic the geometric and pixel-level distortions that may arise in urgent or resource-limited workflows, providing insight into how these factors influence model performance.

Although this study was conducted using data from a single institution, the dataset represents one of the largest and most diverse radiographic collections in Taiwan, drawn from multiple campuses within the Chang Gung Memorial Hospital system, with different scanners and protocols, introducing natural variability while maintaining consistent image quality and annotation standards. The main objective of this study was to test a methodological approach for evaluating model robustness to image degradation, rather than to claim broad generalization to all clinical environments. The single-institution design, therefore, provided a stable yet sufficiently varied dataset for controlled experimentation.

Comparison With Prior Work

Our findings are consistent with prior studies showing that CNNs are sensitive to image degradation and domain shifts in medical imaging. Recent research introduced the RoMIA framework, which aims to develop robust AI models for chest radiographs by addressing real-world sources of image degradation, such as device heterogeneity, screen-captured inputs, and compression artifacts. In parallel, previous studies have shown that AI can enhance image quality and improve diagnostic reliability across multiple imaging modalities, including computed tomography and radiography [1,7,23]. Moreover, real-world deployments have demonstrated that recaptured or smartphone-captured radiographs introduce compounded degradations (eg, scaling, compression, and display artifacts) that can alter AI outputs, underscoring the clinical relevance of evaluating robustness under such workflows [9]. Taken together, these studies highlight the growing emphasis on image fidelity and consistent acquisition parameters as key factors for ensuring robust and generalizable AI performance. Previous research [18] has demonstrated that blurring and contrast alteration primarily affect the high-frequency components essential for delineating structural boundaries—similar to the degradation patterns observed in this study. The alignment between our results and those of earlier works reinforces that these degradation effects are likely intrinsic to CNN-based architectures rather than model-specific artifacts.

Limitations

This study has several limitations. First, the robustness results reported here are relative to our standard clinical preprocessing pipeline. Specifically, the "clean" baseline was generated by detector-based cropping followed by resizing each scaphoid ROI to 240×240 pixels to match the fixed input size of EfficientNetB1. As this crop-and-resize step is itself a lossy operation, our robustness conclusions should be interpreted within this trained-and-deployed pipeline context, rather than as absolute performance relative to the raw, unprocessed DICOM images. Second, our analysis was confined to a single deep learning architecture, EfficientNetB1. Different backbones (eg, ResNet and DenseNet) may exhibit varying levels of robustness to specific noise types, such as Gaussian blur or CLAHE. Nevertheless, we expect that many of the general trends—such as vulnerability to edge-destroying blur—would hold true across CNN-based architectures. Third, all experiments were performed on retrospective datasets from a single institution and focused on controlled pixel-level perturbations. As a result, the data may not fully capture the diversity of real-world clinical acquisition conditions or additional artifacts, such as motion blur, partial occlusions, and multiview variability. Future prospective, multi-institutional studies incorporating these realistic factors would provide a more comprehensive assessment of model resilience and generalizability.

Future Directions

Future work could extend our methodology by evaluating multiple architectures under identical perturbations to quantify differences in susceptibility and to determine whether certain network designs are inherently more robust. Incorporating multiview data and realistic clinical artifacts (eg, motion blur and underexposure) may also help develop more generalizable models. Additionally, advanced data augmentation strategies that simulate image degradation during training could enhance robustness. Exploring hybrid or ensemble models that integrate texture- and shape-based cues, as well as integrating domain-specific priors, could further mitigate the impact of low-quality inputs.

Conclusions

Neural network models designed to complement radiographic interpretation in clinical practice will inevitably encounter image quality distortions due to variations in acquisition, processing, and storage. In this study, we systematically evaluated the effects of image degradation on the performance of a DNN for scaphoid fracture classification. We found a strong negative correlation between image quality and model accuracy, with Gaussian blur, grayscale Gaussian noise, and CLAHE exerting the greatest influence on performance.

Performance decline was primarily driven by decreases in precision, whereas recall remained relatively stable. When developing neural networks for fracture detection in clinical radiography, training with targeted perturbations—particularly

XSL•FO
RenderX

Gaussian blur, grayscale noise, and CLAHE—may improve the model robustness and ensure more reliable performance in diverse clinical environments.

Acknowledgments

No generative artificial intelligence was used in the preparation of this manuscript.

Funding

This work was supported by a grant from the Center for Artificial Intelligence in Medicine (grant CORPG3L0181, CORPG3M0231, CORPG3K0201, and CLRPG3H0015) at Chang Gung Memorial Hospital and National Science and Technology Council (112-2410-H-182A-001).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional review board regulations and patient privacy considerations but are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: CFK, KCC, CL
Data curation: CFK, CWW
Formal analysis: CL, TC
Methodology: CL, APY
Project administration: CFK, KCC
Resources: CFK, CWW
Software: CL, TC
Supervision: CFK, KCC
Validation: CL, TC
Writing – original draft: CL, APY, TC
Writing – review & editing: APY, KCC, CWW
Co-first authors: CL, APY (corresponding author: CFK, KCC)

Conflicts of Interest

KC receives funding from the National Institutes of Health and book royalties from Wolters Kluwer and Elsevier. All other authors declare that they have no conflicts of interest related to this work.

Multimedia Appendix 1
Representative examples of all image perturbation types evaluated in this study.
[DOCX File, 16385 KB - medinform_v14i1e65596_app1.docx]

References

1. Gupta V, Erdal B, Ramirez C, et al. Current state of community-driven radiological AI deployment in medical imaging. JMIR AI 2024 Dec 9;3:e55833. [doi: 10.2196/55833] [Medline: 39653370]
2. Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. arXiv. Preprint posted online on Mar 28, 2019. [doi: 10.48550/arXiv.1903.12261]
3. Habibi Aghdam H, Jahani Heravi E, Puig D. Analyzing the stability of convolutional neural networks against image degradation. Presented at: International Conference on Computer Vision Theory and Applications; Feb 27-29, 2016; Rome, Italy p. 370-382. [doi: 10.5220/0005720703700382]
4. Hosseini H, Xiao B, Poovendran R. Google's cloud vision API is not robust to noise. 2017 Presented at: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA); Dec 18-21, 2017. [doi: 10.1109/ICMLA.2017.0-172]
5. Dodge S, Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. 2017 Presented at: 2017 26th International Conference on Computer Communication and Networks (ICCCN); Jul 31 to Aug 3, 2017; Vancouver, BC, Canada p. 1-7. [doi: 10.1109/ICCCN.2017.8038465]
6. Fawzi A, Moosavi-Dezfooli SM, Frossard P. Robustness of classifiers: from adversarial to random noise. Presented at: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems; Dec 5-10, 2016; Barcelona, Spain p. 1632-1640. [doi: 10.5555/3157096.3157279]

XSL•FO
RenderX

7. Zhang S, Zhu Z, Yu Z, et al. Effectiveness of AI for enhancing computed tomography image quality and radiation protection in radiology: systematic review and meta-analysis. J Med Internet Res 2025 Feb 27;27:e66622. [doi: 10.2196/66622] [Medline: 40053787]
8. Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif Intell Rev 2025;58(1):12. [doi: 10.1007/s10462-024-11005-9]
9. Kuo PC, Tsai CC, López DM, et al. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. NPJ Digit Med 2021 Feb 15;4(1):25. [doi: 10.1038/s41746-021-00393-9] [Medline: 33589700]
10. Yoon AP, Lee YL, Kane RL, Kuo CF, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open 2021 May 3;4(5):e216096. [doi: 10.1001/jamanetworkopen.2021.6096] [Medline: 33956133]
11. Hena B, Wei Z, Castanedo CI, Maldague X. Deep learning neural network performance on NDT digital x-ray radiography images: analyzing the impact of image quality parameters—an experimental study. Sensors (Basel) 2023 Apr 27;23(9):4324. [doi: 10.3390/s23094324] [Medline: 37177528]
12. Nazaré TS, da Costa GBP, Contato W. A, Ponti M. Deep convolutional neural networks and noisy images. : Springer International Publishing; 2018 Presented at: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications; Nov 7-10, 2017; Valparaíso, Chile. [doi: 10.1007/978-3-319-75193-1_50]
13. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. 2019 Presented at: Proceedings of the 36th International Conference on Machine Learning; Jun 9-15, 2019; Long Beach, California, USA p. 6105-6114. [doi: 10.48550/arXiv.1905.11946]
14. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004 Apr;13(4):600-612. [doi: 10.1109/tip.2003.819861] [Medline: 15376593]
15. Wang Z. The SSIM index for image quality assessment. Laboratory for Computational Vision, New York University. URL: http://www.cns.nyu.edu/~lcv/ssim [accessed 2025-12-23]
16. Vernon D. Machine Vision: Automated Visual Inspection and Robot Vision: Prentice-Hall; 1991. URL: https://archive.org/details/machinevisionaut0000vern [accessed 2025-12-23]
17. Dodge S, Karam L. Understanding how image quality affects deep neural networks. Presented at: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX); Jun 6-8, 2016; Lisbon, Portugal p. 1-6. [doi: 10.1109/QoMEX.2016.7498955]
18. Gonzalez RC, Woods RE. Digital Image Processing: Addison-Wesley Publishing Company; 1992.
19. Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert PS, editor. Graphics Gems IV: Academic Press Professional Inc; 1994:474-485. [doi: 10.1016/B978-0-12-336156-1.50061-6]
20. Pizer SM, Amburn EP, Austin JD, et al. Adaptive histogram equalization and its variations. Comput Vis Graph Image Process 1987 Sep;39(3):355-368. [doi: 10.1016/S0734-189X(87)80186-X]
21. Horé A, Ziou D. Image quality metrics: PSNR vs. SSIM. Presented at: 2010 20th International Conference on Pattern Recognition; Aug 23-26, 2010; Istanbul, Turkey p. 2366-2369. [doi: 10.1109/ICPR.2010.579]
22. Anand A, Krithivasan S, Roy K. RoMIA: A framework for creating robust medical imaging AI models for chest radiographs. Front Radiol 2023;3:1274273. [doi: 10.3389/fradi.2023.1274273] [Medline: 38260820]
23. Thirunavukarasu AJ, Elangovan K, Gutierrez L, et al. Democratizing artificial intelligence imaging analysis with automated machine learning: tutorial. J Med Internet Res 2023 Oct 12;25:e49949. [doi: 10.2196/49949] [Medline: 37824185]

Abbreviations

AI: artificial intelligence
CLAHE: contrast-limited adaptive histogram equalization
CNN: convolutional neural network
DNN: deep neural network
IoU: intersection over union
PSNR: peak signal-to-noise ratio
RGB: red, green, and blue
ROI: region of interest
SSIM: structural similarity index measure
VOI: value of interest

XSL•FO
RenderX

Edited by A Coristine; submitted 20.Aug.2024; peer-reviewed by AN Marginean, T Panboonyuen; revised version received 07.Dec.2025; accepted 08.Dec.2025; published 22.Jan.2026.

Please cite as:
Lin C, Yoon AP, Wang CW, Chao T, Chung KC, Kuo CF
Effects of Image Degradation on Deep Neural Network Classification of Scaphoid Fracture Radiographs: Comparison Study of Different Noise Types
JMIR Med Inform 2026;14:e65596
URL: https://medinform.jmir.org/2026/1/e65596
doi:10.2196/65596

© Chihung Lin, Alfred P Yoon, Chien-Wei Wang, Tung Chao, Kevin C Chung, Chang-Fu Kuo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 22.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Self-Management–Based Mobile Health User Typologies and Associations Between User Types and Satisfaction With Key Mobile Health Functions: Comparative Study of Various Fitness and Weight Management App User Types

Tong Wang[1,2], PhD; Yiwen Fan[3], MD; Zheng Li[4], MSc; Xiaoyi Jiao[5,6], MSc; Qichuan Fang[7], MSc; Junhao Ma[8], MSc; Jianbo Lei[9,10,11], MD, PhD

[1]Department of Cardiology and Institute of Vascular Medicine, Peking University Third Hospital; State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China

[2]University of Health and Rehabilitation Sciences Qingdao Hospital (Qingdao Municipal Hospital), School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao, Shandong, China

[3]Department of Oncology, The Affiliated Hospital of Southwest Medical University, Luzhou, China

[4]Department of Rehabilitation Medicine & Physical Therapy, The Affiliated Hospital of Southwest Medical University, Luzhou, China

[5]Center for Health Policy Studies, School of Public Health, Zhejiang University, Hangzhou, China

[6]Department of AI and IT, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China

[7]School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China

[8]School of Public Health, Hangzhou Medical College, Hangzhou, China

[9]Institute of Advanced Clinical Medicine, Center for Medical Informatics, Peking University, Beijing, China

[10]Clinical Research Center, Affiliated Hospital of Southwest Medical University, Luzhou, China

[11]The First Affiliated Hospital, Hainan Medical University, Haikou, China

Corresponding Author:
Jianbo Lei, MD, PhD
Institute of Advanced Clinical Medicine
Center for Medical Informatics
Peking University
38 Xueyuan Rd, Haidian District
Beijing, 100191
China
Phone: 86 8280 5901
Email: jblei@hsc.pku.edu.cn

Abstract

Background: Exploring user satisfaction is crucial for enhancing and ensuring the sustainable development of mobile health (mHealth) apps, particularly in the fitness and weight management sectors. Analyzing user types and developing user profiles are valuable for understanding differences in satisfaction. However, prior research lacks a classification of user types based on self-management characteristics and an analysis of satisfaction disparities among these types.

Objective: This study analyzes user heterogeneity from a self-management perspective among fitness and weight management app users by identifying user types and constructing profiles. It further explores differences in satisfaction with the functional design of these mHealth apps across user types.

Methods: First, 8 feature indicators were selected based on the Health Belief Model and the Behavior Change Wheel to evaluate users' levels of health knowledge and beliefs, as well as self-regulation related to self-management. Existing research was integrated to categorize mHealth app functional design into 5 categories: health guidance, health education, health monitoring, social features, and gamification. Second, a questionnaire survey was used to collect data on users' 8 health management characteristics and their satisfaction with the 5 functional design categories. A total of 2518 responses were collected, of which 1025 were included in the analysis. Cluster analysis was conducted to classify users into distinct types based on the 8 health management characteristics, and user profiles were constructed according to the distribution of these characteristics within each

XSL•FO
RenderX

type. Finally, the Kruskal-Wallis test was used to analyze differences in satisfaction across user types with respect to the 5 functional design categories of mHealth apps.

Results: Cluster analysis revealed that users could be categorized into 6 types based on the 8 self-management characteristics: positively proactive energizers, proactive intenders, negatively proactive energizers, low health management demanders, potential health management demanders, and passive attitude holders. Significant differences were observed across all 8 health management characteristics among the 6 user types (all P<.001). The Kruskal-Wallis test indicated significant variations in user satisfaction with the 5 functional designs of mHealth apps: H(4)=445.388, (P<.001). Overall, users reported the highest satisfaction with health guidance and health monitoring (median 4.00, IQR 1.00) and the lowest satisfaction with gamification (median 3.00, IQR 1.00). Positively proactive energizers, proactive intenders, and negatively proactive energizers demonstrated the highest satisfaction with health education and health guidance (median 4.00). Potential health management demanders, proactive intenders, positively proactive energizers, and negatively proactive energizers reported the highest satisfaction with health monitoring (median 4.00). Proactive intenders reported the highest satisfaction with social features and gamification (median 4.00).

Conclusions: Users of mHealth apps exhibit diverse types, with significant differences in health management characteristics and satisfaction with the 5 functional designs of fitness and weight management apps. This study clarifies individual-level differences in user satisfaction with mHealth apps.

(JMIR Med Inform 2026;14:e64860) doi:10.2196/64860

KEYWORDS

mHealth app; user satisfaction; self-management characteristics; user profile; unsupervised learning

Introduction

The poor continued use of mobile health (mHealth) apps poses a significant threat to the sustainability of the mHealth industry, while also weakening the benefits of health management for users. Approximately 40% of smartphone users utilize mHealth apps [1]. Research indicates that health management based on mHealth apps often requires continued use for over 3 weeks to demonstrate effective outcomes [2]. However, a real-world observational study revealed that only 2.58% of the 189,770 users who downloaded mHealth apps actively used them for at least one week [3], indicating that continued use of mHealth faces serious challenges. Furthermore, app uninstallations result in average monthly losses of US $57,000 for app companies [4], highlighting the negative impact of poor continued use on the industry. Therefore, it is necessary to research the usability of mHealth. Previous studies have shown that user satisfaction is a key determinant of acceptance and continued use [5].

User satisfaction is an important indicator for measuring the quality of mHealth services and has a significant impact on the continued usage and intention to use mHealth [6]. Exploring the factors related to user satisfaction has important reference value for promoting the use of mHealth. A review of existing research on mHealth satisfaction shows that studies mainly follow 3 perspectives: technology acceptance, functional design, and user characteristics. Research from the perspective of technology acceptance is primarily based on the technology acceptance model and its derivative theories, focusing on how users' perceived characteristics of mHealth technology influence satisfaction [7-11]. Studies on the influencing factors of satisfaction have gradually expanded from a practical dimension to multidimensional value [12], including perceived entertainment and other nonpractical characteristics [13]. Research from the perspective of functional design mainly focuses on how specific functional characteristics of mHealth meet user needs and improve satisfaction, such as data analysis, monitoring, and information sharing [14,15]. This line of research shows a trend of transition from overall evaluation to modular analysis, with the research focus shifting from overall satisfaction with mHealth to the exploration of multidimensional satisfaction [16-18].

Current research has conducted relatively in-depth explorations of mHealth satisfaction from the perspectives of technology acceptance and functional design; however, evidence regarding the relationship between heterogeneity in user characteristics and satisfaction remains limited [19]. A survey on the use of digital health tools among the US population revealed that there is no "one-size-fits-all" model in digital health, as consumers are not a homogeneous group [20]. However, digital health users cannot be classified solely based on sociodemographic characteristics, as each generation interacts with technology differently due to its unique experiences, needs, and levels of trust in the health care system [21]. Previous studies have confirmed that eHealth literacy [22], strong self-reporting capabilities [19], health needs [23], and poor health status [24] significantly influence mHealth satisfaction. Intrinsic characteristics, such as health knowledge, beliefs, and self-regulation skills and abilities, are central to self-management [25]. These findings suggest that self-management attributes provide important explanations for user heterogeneity and differences in satisfaction among users. Compared with traditional health care models, mHealth relies more heavily on self-management service models [26]. Systematic exploration of individual self-management–related characteristics can help construct user classifications that more deeply capture internal features, thereby promoting personalized and refined mHealth design [27]. However, there is a lack of research systematically integrating self-management characteristics and analyzing user-group heterogeneity based on combinations of these characteristics. In addition, empirical evidence on the differential distribution of satisfaction across different user types with respect to mHealth function design remains scarce, posing challenges for developers seeking to deliver mHealth services that meet the needs of diverse users.

XSL•FO
RenderX

To address the above limitations, this study, guided by self-management theory, analyzes relevant self-management characteristics based on the Health Belief Model (HBM) and Behavior Change Wheel (BCW) among users of fitness and weight management apps. A refined classification of these mHealth users is conducted based on characteristic attributes, and differences in satisfaction with mHealth functional design across user types are examined. Through this investigation, the study seeks to address the following research questions:

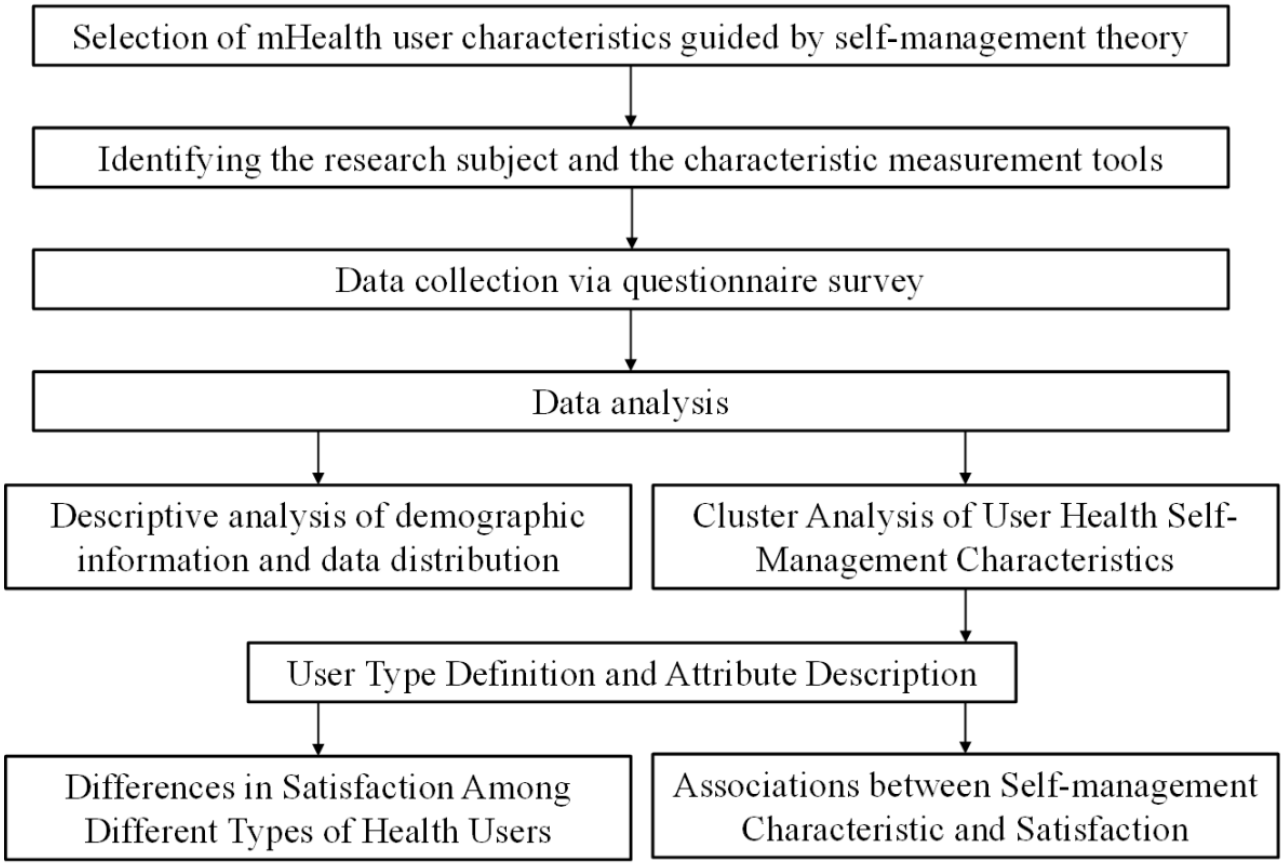
- Research question 1: User typology: How can users of fitness and weight management apps be systematically categorized using quantitative methods based on self-management characteristics? What are the defining attributes of each user type?
- Research question 2: Satisfaction differences. Do user groups with distinct self-management characteristics demonstrate statistically significant differences in satisfaction with various mHealth functional designs? What are the specific manifestations of these differences, and which factors are relevant?

Methods

Design

Guided by self-management theory, this study utilizes the HBM and BCW to select variables representing the self-management characteristics of mHealth users. Data were collected through a questionnaire survey, and unsupervised clustering analysis was used to identify user clusters. Subsequently, the characteristic distributions of different user clusters were analyzed to construct user profiles. Through intergroup difference analysis, this study compared satisfaction levels across user types with respect to various mHealth functional designs and examined the correlations between user characteristics and satisfaction with functional designs. The overall workflow of the study is illustrated in Figure 1.

Figure 1. Study flowchart. mHealth: mobile health.



Selection of Self-Management–Related Characteristics

mHealth mainly improves health outcomes by empowering self-management abilities [28]. Therefore, this study selects variables representing user characteristics guided by self-management theory. In individual and family self-management theory [29], self-management refers to the process by which individuals and families use knowledge and beliefs, self-regulation skills and abilities, and social facilitation strategies to achieve health-related outcomes in their real-life environments. Among these components, knowledge and beliefs refer to information and perceptions related to a health condition or health behavior. The HBM is one of the most widely used models for explaining beliefs and perceptions underlying health behaviors [30]. HBM posits that individuals' perceptions of their susceptibility to a health issue and the potential severity of its consequences, together with perceived benefits of adhering to health recommendations and their intention and ability to engage in health management, are the most direct drivers of health behaviors [31]. HBM operationalizes these factors through perceived susceptibility, perceived severity, perceived benefits, perceived barriers, and self-efficacy [32]. Accordingly,

XSL•FO
RenderX

this study adopts these 5 variables from HBM to characterize the knowledge and beliefs component of self-management theory.

Self-regulation within self-management theory is an iterative process in which individuals utilize a variety of skills and abilities to achieve healthy behavior change [29]. The BCW, a classic theoretical model for explaining behavior change—particularly in the contexts of health promotion and disease prevention [33]—provides an important reference for selecting self-regulation–related characteristic variables in this study. The BCW posits that healthy behavior change can be achieved only when individuals possess capability, motivation, and opportunity [34]. Capability refers to the knowledge and skills required to engage in a particular behavior [35]. In the context of using mHealth for health management, essential knowledge and skills are reflected in 2 aspects: health management and the use of digital technology. eHealth literacy is among the most commonly used indicators for assessing an individual's ability to acquire health management information and utilize electronic health information to address health issues [36,37]. Accordingly, this study selects eHealth literacy to characterize self-regulation skills and abilities. In addition, motivation in the BCW refers to the beliefs that trigger behavioral change, encompassing reflective motivation and intrinsic motivation [34]. Motivation for health management can be divided into proactive health management intention and passive health management intention triggered by perceived health status, corresponding to reflective motivation and intrinsic motivation, respectively. Therefore, this study selects health management intention [38] and perceived health status [39] to further characterize the motivational dimension of self-management behavior. In summary, this study selects a total of 8 variables to analyze self-management characteristics among mHealth users. The definitions of each variable are presented in Table 1.

Table 1. mHealth[a] user self-management characteristics: a framework of 8 variables for user categorization.

User self-management characteristics	Definition
Perceived susceptibility	Beliefs about the likelihood of getting a disease or condition [31].
Perceived severity	Feelings about the seriousness of contracting an illness or of leaving it untreated, including evaluations of both medical and clinical consequences and possible social consequences [31].
Perceived benefits	Beliefs that engaging in health behaviors can help decrease the risk of diseases and health problems [40].
Perceived barriers	An individual's assessment of the difficulties and cost of adopting behaviors [32].
Health self-efficacy	The confidence individuals have in their ability to use their skills to complete health management tasks [41].
eHealth literacy	The ability to seek, find, evaluate, and use electronic health information from electronic resources to address health issues [36].
Perceived health status	The perception of one's own health status [39].
Health management intention	The degree of intention to engage in health protection [38].

[a]mHealth: mobile health.

Identification of Study Targets

The main purpose of this study is to delineate individuals' health management characteristics and classify users based on their levels of health management characteristics, and to further analyze differences in satisfaction with various mHealth app functional designs across user categories. Therefore, it is necessary to clarify the main functional designs of mHealth apps. Previous studies have summarized common mHealth app functional designs through systematic reviews and content analyses [42-44], categorizing functions according to their primary implementation purposes, such as education, tracking, social interaction, gamification, and motivational features [42]. Building on prior research, this study reorganizes the functional characteristics reported in the literature to propose a classification of mHealth app functional designs for analysis (see Table 2).

XSL•FO
RenderX

Table 2. Five functional designs and features of the mHealth[a] apps analyzed in this study

Functional designs	Definition	Features
Health guidance	Providing advice based on health plans and applications may provide real-time feedback and advice based on health plans and status [45], or offer strategies for behavior change that can contribute to improving health conditions [46].	Intelligent virtual assistants and communication with health care personnel [43]; interaction with medical staff [45]; and development of health plans and goal setting [46].
Health education	Offering necessary knowledge to achieve health behavior change [47].	Personalized education and general knowledge education [42]; image-based education, video education, and audio education [47].
Health monitoring	Recording past and current health statuses [47].	Health records and health diaries [43]; trackers, data entry, and data export [42]; and self-tracking [46].
Social function	Providing a platform for user interaction, communication, and maintaining connections [47].	Community forums and social media [42]; and social sharing [47].
Gamification	Integrating gaming mechanisms or elements into mHealth apps and enhancing the user experience with gamelike features [48].	Personalized avatars, challenges, tasks, health, and rewards [43]; points, badges, and levels [42].

[a]mHealth: mobile health.

To achieve the aforementioned research objectives, this study selected appropriate objects of study for empirical analysis. First, the comparability of different mHealth app functional designs was considered. To minimize the influence of differences in core service content across app types on comparability, this study selected a homogeneous category of apps as the empirical focus, thereby reducing potential confounding effects. Second, the application domains of different mHealth app categories were considered. Exercise, fitness, and weight management align with substantial health needs. These domains represent important modalities for primary prevention, and exercise therapy also serves as a key approach for tertiary prevention (disease management) in various conditions, such as cardiac rehabilitation, pulmonary rehabilitation, tumor rehabilitation, and metabolic disease management [49]. Given that exercise, fitness, and weight management apps account for more than half of all mHealth apps [50], and that these functions represent 48.37% of users' motivations for using mHealth apps [51], we selected exercise, fitness, and weight management apps as the units of analysis to represent a broader and more heterogeneous user population.

This study did not restrict respondents to users of a specific app for the following reasons. First, as the aim of this study is to examine the satisfaction of different user types with various mHealth functional designs in real-world contexts and to provide generalizable recommendations, the focus is on mHealth functional modules rather than on specific app design implementations. Accordingly, the research sample was not limited to users of a single app. Moreover, this questionnaire-based data collection approach captures multiple implementations of the same functional module (eg, health education), thereby enhancing sample diversity and improving the generalizability of the findings [52,53]. Second, this study was conducted in real-world settings, which differ from laboratory-based studies that examine factors influencing mHealth satisfaction. In laboratory environments, variables must be tightly controlled to quantify the idealized effects of specific factors, often requiring the use of a single app as the research object. By contrast, this study emphasizes user satisfaction with mHealth in real-world contexts, where the use of multiple similar apps by the same user is common. Analyzing such data provides real-world evidence that can inform future research and practical mHealth design.

Measurement Instrument

Based on the constructed User Health Management Characteristics Framework, this study measures 8 independent variables and 5 dependent variables reflecting satisfaction with mHealth functionalities. To achieve the aforementioned research objectives, it was necessary to select appropriate and measurable items corresponding to the characteristics defined within the framework. To ensure the rational design of the measurement scale, this study undertook a careful process of adapting previously validated scales, making nuanced adjustments to align them with the specific research objectives.

Specifically, the User Health Management Characteristics Framework includes 4 variables derived from the HBM (perceived severity, perceived susceptibility, perceived barriers, and perceived benefits), with measurement items adapted from HBM-related studies by Ahadzadeh et al [54], Saghafi-Asl et al [32], and McArthur et al [55]. Two variables were derived from the BCW (eHealth literacy and perceived health status), with measurement items drawn from BCW-related studies [39,56,57] and the Brief Health Literacy Screen scale [58]. In addition, 2 variables integrating both models (health self-efficacy and health management intention) were included, with measurement items adapted from health management–related studies, such as Zhou et al [59] and Li et al [38]. For the measurement of satisfaction with the 5 mHealth functional designs, this study first provided definitions and specific design examples of each function in the questionnaire, followed by item-based measurement. The satisfaction items were adapted from relevant studies on mHealth satisfaction [8]. All measurement items used a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

A multistage validation and refinement process was implemented to ensure the conceptual accuracy and cultural

XSL•FO
RenderX

appropriateness of the English-originated scales within the Chinese context. First, the preliminary translation and adaptation of the scale items were collaboratively conducted by research team members proficient in both Chinese and English and with expertise in health informatics. Through group discussions, each item was refined to preserve the core meaning of the original construct while aligning with Chinese linguistic norms and the research context. Second, before the formal survey, 24 participants with backgrounds in medical informatics were invited to complete a pretest of the questionnaire. Participants were asked not only to respond to the questionnaire but also to provide feedback on the clarity and contextual appropriateness of the items. Based on this feedback, the wording of the questionnaire was further refined to eliminate ambiguities and improve readability, resulting in the final version used in the formal survey. Finally, during the data analysis phase, reliability and validity tests were conducted on the measurement model, thereby verifying the effectiveness and reliability of the measurement instrument from a data-driven perspective (see Multimedia Appendix 1).

Data Collection

The survey targeted individuals who had used exercise, fitness, or weight loss apps within the preceding 3 months. According to the N:q hypothesis [60], the ratio of sample size to item parameters in a scale should range from 20:1 to 10:1. In this study, a total of 51 items were included for measurement; therefore, applying the more conservative ratio of 20:1, the required sample size was approximately 1020.

From June 28, 2022, to September 8, 2022, a snowball sampling survey was conducted via the WeChat platform (Tencent Holdings Limited). The questionnaire link, generated using Wenjuanxing, was distributed through Moments (Tencent Holdings Limited) and various group chats. In total, 2518 questionnaires were collected. Of these, 902 respondents reported that they had not used the target apps within the preceding 3 months. A quality control item was included in the questionnaire: “This question is an attention test to ensure careful completion of the questionnaire. For this question, please select ‘strongly disagree’.” The correct response to this item was “1. strongly disagree.” Responses that answered this item incorrectly were considered to have failed the quality control check. In addition, given the length of the questionnaire, responses completed in less than 3 minutes or with identical answers across all scale items were deemed insufficiently attentive and were also excluded. Through this screening process, 591 responses were removed. Consequently, 1025 valid questionnaires remained for subsequent analysis. Before completing the questionnaire, all participants were informed of the study’s overall purpose and detailed procedures.

In this study, item-level missing data were addressed using mode imputation based on each respondent’s answers to other items within the same variable. Specifically, there were 3 missing values in total: 1 in perceived susceptibility, 1 in eHealth literacy, and 1 in gamification. As the questionnaire employed Likert scales, which represent ordinal rather than strictly continuous data, each missing value was replaced with the mode of the remaining items within the same scale for that respondent. The imputed values were 3, 5, and 3, respectively. Given that the dataset comprised a total of 45,100 item responses, the proportion of missing values was 0.0067%, which is negligible and unlikely to affect the results.

Data Analysis

To address the question “What are the common combinations of health management characteristics in users during the process of using mHealth?” we employed cluster analysis. Cluster analysis is an unsupervised machine learning technique that divides a set of observations into multiple clusters composed of similar objects and can identify the structural composition of different clusters. In clustering, similarity between observations within the same cluster is high, whereas similarity between observations from different clusters is low. Among various clustering methods, the k-means algorithm produces results that are easy to interpret, are suitable for clustering high-dimensional data, and demonstrate good scalability and efficiency when dealing with large datasets [61]. The clustering in this study was based on 8 health management characteristics (Table 1), and the sample size was moderate; therefore, k-means was appropriate for this study. The k-means algorithm requires the number of clusters to be predetermined. We determined the optimal number of clusters by integrating the Elbow Method and the Silhouette Coefficient. The Elbow Method identifies the inflection point by analyzing the relationship between the within-cluster sum of squared errors (SSE) and the number of clusters, while the Silhouette Coefficient is used to mathematically assess the compactness and separation of the clustering structure. The results showed that when the number of clusters was 6, an inflection point appeared in the SSE curve, and the Silhouette Coefficient reached a locally optimal value of 0.1514 (see Multimedia Appendix 2), indicating that this clustering solution maintains good within-cluster compactness while achieving favorable between-cluster separation. Therefore, we ultimately set the number of clusters to 6. To further verify the stability of this 6-cluster solution, we performed 10 repeated experiments with random initializations. The results showed a relative fluctuation rate of only 0.002% in the SSE values, demonstrating that the 6-cluster solution exhibits high repeatability and stability. We conducted the cluster analysis to determine the grouping of mHealth app users using sklearn.cluster in Python 3.8 (Python Foundation; see Multimedia Appendix 3).

To answer the question “Is there a difference in satisfaction with the functional design of mHealth apps among different types of users?” we performed rank sum tests to examine satisfaction and preferences regarding functionalities across user groups using SPSS 26 (IBM Corp). To address the question “How do health management characteristics affect users’ satisfaction and preferences for the functional design of mHealth apps?” we conducted a partial least squares analysis using SmartPLS 4 (SmartPLS GmbH). The significance level was set at $P<.05$ for 2-tailed tests.

The data collected for this study were analyzed to assess potential nonresponse bias. To accomplish this and examine nonresponse bias [11], the Mann-Whitney U test was used to compare responses from the first 20% (205/1025) and the last

XSL•FO
RenderX

20% (205/1025) of the sample. The results showed no significant differences between the 2 groups for the health guidance (z=−0.728, P=.47), health education (z=−1.171, P=.24), health monitoring (z=−0.857, P=.39), social function (z=−1.794, P=.07), and gamification (z=−0.260, P=.80) constructs. Therefore, nonresponse bias was not a concern in this study. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting standard (see Multimedia Appendix 4).

Ethical Considerations

This study utilized an online questionnaire survey. On the first page of the questionnaire, the research purpose, content, potential risks, data usage, confidentiality terms, and researcher contact information were clearly explained to participants. By checking a box, participants confirmed that they had read and understood the above information and voluntarily agreed to participate in the study, thereby providing informed consent. Participants who did not agree with the above content were directed to exit the online questionnaire. The questionnaire materials for this study were destroyed after the completion of quality control and transcription, within 6 months after the survey concluded. Any associated information, such as source IP addresses, collection time stamps, and serial numbers, was also destroyed to ensure privacy protection. Following transcription, the data were renumbered sequentially, starting from 1. All images and data presented in the manuscript or multimedia appendices are fully anonymized and cannot be used to identify any individual participant or user. No financial compensation was provided to participants. Approval was obtained from the Biomedical Ethics Committee of the Second Affiliated Hospital, School of Medicine, Zhejiang University, China (approval number IRB-2021-841).

Results

Demographic Information

The demographic characteristics of the included sample are presented in Table 3. The sample exhibited a balanced gender distribution (female: 553/1025, 53.95%; male: 472/1025, 46.05%), with participants predominantly aged 19-30 years (500/1025, 48.78%). In terms of education, most respondents held college degrees (519/1025, 50.63%), followed by master's degrees (269/1025, 26.24%). Income distribution showed that nearly one-quarter of participants (244/1025, 23.80%) earned ≤2500 RMB (1 RMB=US $0.14) monthly, whereas a smaller proportion (49/1025, 4.78%) reported incomes of 50,001 or higher. Regarding usage patterns, the largest group reported using the service 2-3 times per week (370/1025, 36.10%), while a minority were frequent users (>7 times per week: 59/1025, 5.76%). In terms of occupation, students constituted the largest subgroup (283/1025, 27.61%), followed by public servants/clerks (208/1025, 20.29%) and professional technical personnel (196/1025, 19.12%). Overall, 664 of 1025 (64.78%) users reported using 1 app.

XSL•FO
RenderX

Table 3. Descriptive statistics of demographic information for the analyzed sample (n=1025).

Characteristics	Values, n (%)
Gender	
Male	472 (46.05)
Female	553 (53.95)
Age (years)	
≤18	6 (0.59)
19-30	500 (48.78)
31-45	337 (32.88)
46-60	174 (16.98)
≥61	8 (0.78)
Education	
High school	37 (3.61)
Junior college education	111 (10.83)
College	519 (50.63)
Master’s degree	269 (26.24)
Doctor’s degree	89 (8.68)
Income (RMB[a])	
≤2500	244 (23.80)
2501-5000	159 (15.51)
5001-8000	178 (17.37)
8001-15,000	213 (20.78)
15,001-30,000	125 (12.20)
30,001-50,000	57 (5.56)
≥50,001	49 (4.78)
Frequency	
≤1 time per week	265 (25.85)
2-3 times per week	370 (36.10)
4-5 times per week	213 (20.78)
6-7 times per week	118 (11.51)
>7 times per week	59 (5.76)
Occupation	
Students	283 (27.61)
Medicine-related personnel	168 (16.39)
Public servants and clerks	208 (20.29)
Commercial and service personnel	107 (10.44)
Professional technical personnel	196 (19.12)
Manual workers	33 (3.22)
Unemployed and other personnel	30 (2.93)

[a]1 RMB=US $0.14.

Data Characteristics

A descriptive analysis of the 8 self-management characteristics and satisfaction with the 5 functional designs was conducted. First, we calculated the average score of multiple items measuring each characteristic variable to represent the level of that variable. However, because the distributions of the average

XSL•FO
RenderX

values for each variable were not normally distributed, medians and IQRs were used for presentation, as shown in Table 4.

Table 4. Descriptive statistics (median and IQR) of the 8 health management characteristics and satisfaction with the 5 functional designs for the analyzed sample (n=1025).

Characteristics	Median	First quartile (Q1)	Third quartile (Q3)	IQR
Health-specific determinants				
Perceived severity	4.25	3.75	5.00	1.25
Perceived susceptibility	3.50	3.00	4.00	1.00
Perceived barrier	2.50	2.00	3.50	1.50
Perceived benefit	4.75	4.00	5.00	1.00
Health self-efficacy	4.00	3.33	4.33	1.00
Perceived health status	3.67	3.00	4.00	1.00
eHealth literacy	4.00	3.50	4.50	1.00
Health management intention	4.00	3.33	4.33	1.00
Satisfaction				
Satisfaction with health guidance	4.00	3.00	4.00	1.00
Satisfaction with health education	3.67	3.00	4.00	1.00
Satisfaction with health monitoring	4.00	3.33	4.33	1.00
Satisfaction with social function	3.33	3.00	4.00	1.00
Satisfaction with gamification	3.00	3.00	4.00	1.00

Cluster Analysis of User Self-Management Characteristics and Determination of User Types

The determination of the optimal number of clusters in this study, which was set at 6, was based on the relationship between the within-cluster SSE and the number of clusters. Subsequently, cluster analysis was performed on the sample using *z* score–standardized values of the 8 health management indicators. The distribution across the 6 clusters was relatively uniform (140, 135, 189, 218, 180, and 163). Figure 2 shows the distribution of each indicator across the 6 clusters obtained from the analysis. A *z* score close to 0 indicates that the corresponding indicator in a given cluster approximates the overall average level, whereas a positive *z* score indicates that the indicator is above the average level in that cluster.

We further analyzed differences in each indicator across the 6 clusters using the Kruskal-Wallis test to assess the distinctiveness of each cluster. The results of the Kruskal-Wallis test demonstrated statistically significant differences in the *z* scores of all indicators across the 6 clusters (P<.001 for all). Specifically, significant differences were observed for the following: perceived severity, $H(5)=400.947$ (P<.001); perceived susceptibility, $H(5)=522.425$, (P<.001); perceived barrier, $H(5)=673.985$, (P<.001); perceived benefit, $H(5)=492.630$, (P<.001;) health self-efficacy, $H(5)=518.298$, (P<.001); perceived health status, $H(5)=490.411$, (P<.001); eHealth literacy, $H(5)=362.910$, (P<.001); and health management intention, $H(5)=390.573$, (P<.001). These findings indicate that the 6 clusters exhibit distinct patterns and magnitudes of health management characteristics, as reflected by the significant differences in the *z* scores of the respective indicators. The detailed *z* scores and pairwise comparison results for each indicator across the 6 clusters are presented in Table 5, allowing for a detailed examination of how each cluster differs from the others in terms of specific health management characteristics. Detailed test statistics are provided in Multimedia Appendix 5, which presents pairwise comparisons of health management characteristics among user groups.

The distribution of the 8 health management traits varied across individual users. To further understand user health management characteristics and provide references for personalized services in mHealth apps, this study categorized users based on combinations of their 8 current health management factors. Given the distinctiveness of each user group, the 6 types were named according to the indicator profiles within each cluster to reflect the main health management characteristics of users in each type. Cluster 1 was labeled “positively proactive energizers.” Users in this cluster exhibited higher perceived health status, health management self-efficacy, and health management intention than proactive intenders, along with substantially lower perceived barriers and perceived susceptibility.

Cluster 2 was labeled “proactive intenders,” characterized by users whose perceived health status, eHealth literacy, and health management intention are all above average, but who also report higher perceived barriers and perceived susceptibility. Proactive intenders (135/1025, 13.17%), the smallest group, represent users who demonstrate strong proactivity in health management, recognize their potential risk for health problems, and actively manage their health despite the presence of perceived barriers.

Cluster 3 was labeled “negatively proactive energizers.” Compared with positively proactive energizers, users in this cluster have slightly below-average perceived health status and

XSL•FO
RenderX

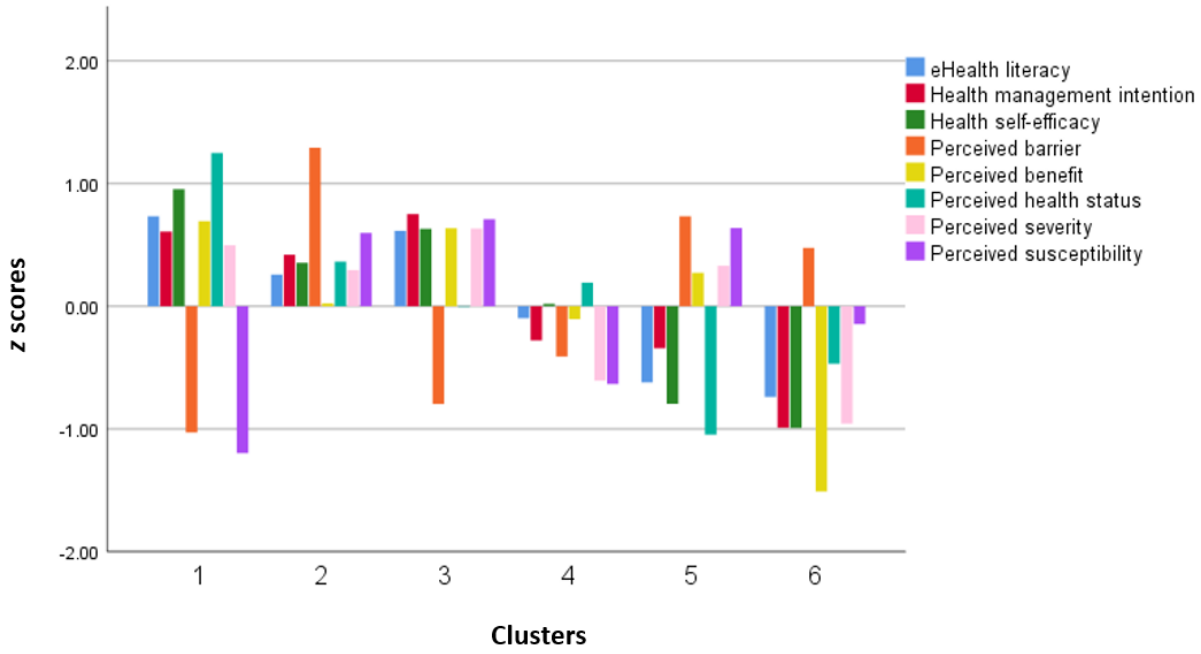
higher perceived susceptibility, yet they exhibit higher health management intention, health management self-efficacy, and eHealth literacy. Both positively proactive energizers (140/1025, 13.66%) and negatively proactive energizers (189/1025, 18.44%) represent groups with high levels of health management cognition, literacy, and capability, along with strong health management intention, making them the primary beneficiaries of current mHealth apps. The key difference between these 2 groups lies in perceived health status, which is higher among positively proactive energizers and below average among negatively proactive energizers. Cluster 4 was labeled "low health management demanders." Users in this cluster tend to have average perceived health status, with perceived susceptibility, perceived severity, and health management intention all below the average level. Low health management demanders constitute the largest proportion of users (218/1025, 21.27%) and represent a group with relatively good current health, limited emphasis on health management, insufficient awareness of health risks and potential harm, and consequently lower intention to engage in health management.

Cluster 5 was labeled "potential health management demanders." Users in this cluster have below-average perceived health status and, although their perceived benefits, perceived susceptibility, and perceived severity are above average, they exhibit lower health management intention and higher perceived barriers. Potential health management demanders (180/1025, 17.56%) represent individuals with poor current health status who are aware of the harms associated with health risks and recognize the benefits of health management for improving their condition, yet face substantial barriers to engaging in health management and therefore have a lower intention to do so. This group has both a clear recognition of the importance of health management and an objective need for it due to their poor health status.

Cluster 6 was labeled "passive attitude holders," comprising users whose perceived health status is below average and who report very low perceived benefits, health management self-efficacy, perceived severity, and health management intention. Passive attitude holders (163/1025, 15.90%) share similarities with low health management demanders, including low perceived benefits, health management self-efficacy, perceived severity, and intention for health management. However, they differ in that passive attitude holders perceive their health status as below average and face higher perceived barriers. This group perceives a decline in health status but lacks the willingness to improve it, placing them at high risk for illness.

From the perspective of age distribution, the 6 user categories exhibited distinct age-structure characteristics. Positively proactive energizers (n=140) comprised 48 (34.3%) aged 19-30 years, 48 (34.3%) aged 31-45 years, 42 (30.0%) aged 46-60 years, and 2 (1.4%) aged ≥61. Proactive intenders (n=135) consisted of 71 (52.6%) aged 19-30 years, 50 (37.0%) aged 31-45 years, 12 (8.9%) aged 46-60 years, and 2 (1.5%) aged ≥61. Negatively proactive energizers (n=189) included 2 (1.1%) aged ≤18, 65 (34.4%) aged 19-30 years, 79 (41.8%) aged 31-45 years, 41 (21.7%) aged 46-60 years, and 2 (1.1%) aged ≥61. Low health management demanders (n=218) comprised 1 (0.5%) aged ≤18, 103 (47.2%) aged 19-30 years, 69 (31.7%) aged 31-45 years, 44 (20.2%) aged 46-60 years, and 1 (0.5%) aged ≥61. Potential health management demanders (n=180) included 2 (1.1%) aged ≤18, 97 (53.9%) aged 19-30 years, 54 (30.0%) aged 31-45 years, 26 (14.4%) aged 46-60 years, and 1 (0.6%) aged ≥61. Passive attitude holders (n=163) were distributed as 1 (0.6%) aged ≤18, 116 (71.2%) aged 19-30 years, 37 (22.7%) aged 31-45 years, and 9 (5.5%) aged 46-60 years. Chi-square tests indicated that the associations between user clusters and both age (Cramer *V*=0.145) and education (Cramer *V*=0.116) groups were weak.

Figure 2. Cluster analysis–derived profiles of 6 user groups across 8 health management characteristics (*z* scores).



XSL•FO
RenderX

Table 5. Mean z scores and Kruskal-Wallis group comparisons of the 8 health management characteristics across the 6 user clusters.

Clustering variables	Cluster 1: positively proactive energizers (n=140)	Cluster 2: proactive intenders (n=135)	Cluster 3: negatively proactive energizers (n=189)	Cluster 4: low health management demanders (n=218)	Cluster 5: potential health management demanders (n=180)	Cluster 6: passive attitude holders (n=163)
Perceived severity	0.496[a,b,c]	0.294[b,e]	0.632[c]	–0.607[d]	0.331[a,e]	–0.956[f]
Perceived susceptibility	–1.197[a]	0.598[b,c,e]	0.709[c,e]	–0.634[d]	0.638[e]	–0.146[f]
Perceived barrier	–1.029[a,c]	1.292[b]	–0.796[c]	–0.410[d]	0.733[e,f]	0.476[f]
Perceived benefit	0.693[a,c]	0.022[b,d]	0.636[c]	–0.106[d]	0.273[e]	–1.510[f]
Health self-efficacy	0.955[a]	0.354[b,c]	0.630[c]	0.020[d]	–0.795[e,f]	–0.993[f]
Perceived health status	1.249[a]	0.363[b,d]	–0.004[c,d]	0.192[d]	–1.047[e]	–0.469[f]
eHealth literacy	0.733[a,c]	0.258[b]	0.615[c]	–0.097[d]	–0.621[e,f]	–0.740[f]
Health management intention	0.608[a,b,c]	0.419[b]	0.751[c]	–0.278[d,e]	–0.342[e]	–0.990[f]

[a-f]In the same row of data, if the same characteristic across the 6 clusters shares the same superscript (eg, "b"), it indicates that the difference between them is not statistically significant. Conversely, if the same characteristic across the 6 clusters does not share the same superscript, it suggests that the difference between them is statistically significant. For instance, cluster 1's perceived barrier is denoted as –1.029[a,c], and cluster 3's perceived barrier is denoted as –0.796[c]; the presence of the superscript "c" in both indicates that the Kruskal-Wallis test for the *z* scores of perceived barrier between cluster 1 and cluster 3 is not statistically significant. Conversely, cluster 2's perceived barrier is denoted as 1.292[b]. The absence of a common superscript implies that the Kruskal-Wallis test for the *z* scores of perceived barrier between cluster 1 and cluster 2 is statistically significant.

Differences in Satisfaction With mHealth App Functional Design Among Different Types of Health Users

First, the Kruskal-Wallis test was used to analyze differences in satisfaction with the 5 functional designs of the mHealth app among all users. The Kruskal-Wallis test indicated significant variations in user satisfaction with the 5 functional designs of mHealth apps—$H(4)=445.388$, $P<.001$. Specifically, the median satisfaction levels for the 5 functional dimensions among all users are presented in Table 6. Among these dimensions, health monitoring and health guidance exhibited the highest levels of satisfaction. Detailed satisfaction scores for the 5 functional designs, as well as pairwise comparisons of satisfaction among the 6 clusters, are also presented in Table 6. These results allow for a detailed examination of how each cluster differs from the others in terms of satisfaction with specific functional designs. Detailed test statistics are provided in Multimedia Appendix 6, which presents pairwise comparisons of satisfaction with the 5 functions across user groups.

Specifically, satisfaction with health guidance was higher in clusters 1, 2, 3, and 5 than in cluster 6. Pairwise comparisons revealed that clusters 1 ($z=6.246$, Bonferroni-adjusted $P<.001$), 2 ($z=8.404$, Bonferroni-adjusted $P<.001$), 3 ($z=6.887$, Bonferroni-adjusted $P<.001$), and 5 ($z=3.674$, Bonferroni-adjusted $P=.004$) differed significantly from cluster 6. Satisfaction with health education and health monitoring was significantly higher in clusters 1–3 than in clusters 4–6, with all 9 cross-group pairwise comparisons for each domain yielding statistically significant differences (health education: $z=2.77$-7.95, Bonferroni-adjusted $P<.001$ to .02 and health monitoring: $z=3.05$-7.88, Bonferroni-adjusted $P<.001$ to .03). With respect to social functions and gamification, users in cluster 2 reported higher satisfaction than those in the other 5 clusters. Pairwise comparisons revealed that cluster 2 differed significantly from cluster 1 ($z=–3.584$, Bonferroni-adjusted $P=.005$), cluster 3 ($z=4.599$, Bonferroni-adjusted $P<.001$), cluster 4 ($z=6.857$, Bonferroni-adjusted $P<.001$), cluster 5 ($z=7.392$, Bonferroni-adjusted $P<.001$), and cluster 6 ($z=5.992$, Bonferroni-adjusted $P<.001$) in social function; and from cluster 1 ($z=–3.999$, Bonferroni-adjusted $P=.001$), cluster 3 ($z=5.015$, Bonferroni-adjusted $P<.001$), cluster 4 ($z=5.880$, Bonferroni-adjusted $P<.001$), cluster 5 ($z=7.200$, Bonferroni-adjusted $P<.001$), and cluster 6 ($z=6.083$, Bonferroni-adjusted $P<.001$) in gamification.

XSL•FO
RenderX

Table 6. Median satisfaction and Kruskal-Wallis group comparisons regarding the 5 functional designs across the 6 user types.

Dependent variables	Median (IQR)						
	All users (n=1025)	Cluster 1: positively proactive energizers (n=140)	Cluster 2: proactive intenders (n=135)	Cluster 3: negatively proactive energizers (n=189)	Cluster 4: low health management demanders (n=218)	Cluster 5: potential health management demanders (n=180)	Cluster 6: passive attitude holders (n=163)
Satisfaction with health guidance	4.00 (1.00)	4.00[a,b,c,e] (1.67)	4.00[b,c] (1.00)	4.00[c] (1.00)	3.67[d,e,f] (1.00)	4.00[e] (1.00)	3.33[f] (1.00)
Satisfaction with health education	3.67 (1.00)	4.00[a,b,c] (1.67)	4.00[b,c] (1.33)	4.00[c] (1.33)	3.67[d,e] (1.00)	3.67[e,f] (1.00)	3.33[f] (1.00)
Satisfaction with health monitoring	4.00 (1.00)	4.00[a,b,c] (0.92)	4.00[b,c] (1.00)	4.00[c] (1.33)	3.67[d,e,f] (1.00)	4.00[e] (0.67)	3.67[f] (1.00)
Satisfaction with social function	3.33 (1.00)	3.33[a,f,c] (1.00)	4.00[b] (1.00)	3.33[c,d,f] (1.00)	3.00[d,e,f] (0.67)	3.00[e,f] (1.00)	3.33[f] (0.67)
Satisfaction with gamification	3.00 (1.00)	3.00[a,c,d,f] (1.00)	4.00[b] (1.33)	3.00[c,d,e,f] (1.33)	3.00[d,e,f] (0.67)	3.00[e,f] (1.33)	3.00[f] (0.33)

[a]In the same row of data, the group with a superscript "a" is statistically distinct from that with "b," "c," "d," and "e". However, the group with a superscript "b" is statistically not distinct from that with "b, c" because they share the superscript "b." Differing grouping labels indicate the heterogeneity across clusters.

Associations Between User Self-Management Characteristics and Satisfaction With mHealth App Functional Design

In this study, confirmatory factor analysis was used to test the reliability and validity of the questionnaire. The results, as shown in Multimedia Appendix 7, indicate that all constructs had Cronbach α and composite reliability values above 0.7; the average variance extracted for each construct was greater than 0.6; and the factor loadings for all items exceeded 0.7, demonstrating good reliability and convergent validity for all constructs. The heterotrait-monotrait ratio (HTMT) method was used to assess discriminant validity. The commonly accepted criterion for establishing discriminant validity is an HTMT value below 0.85. The results indicate that all construct pairs exhibited HTMT values below 0.85, supporting the discriminant validity of the measurement model. Using 8 health management characteristics as independent variables and satisfaction with 5 feature designs as dependent variables, partial least squares analysis was conducted to examine the impact of health management characteristics on satisfaction with the functional design of mHealth apps. The R^2 value represents the proportion of variance in the dependent variable explained by the independent variables and is used to interpret the goodness-of-fit of the regression equation, while β indicates the correlation between variables. Table 7 displays the β values and statistical significance of R^2 for the independent and dependent variables. Specifically, perceived barriers (β=.177, *z*=3.586, *P*<.001), perceived susceptibility (β=.076, *z*=2.056, *P*=.04), health self-efficacy (β=.140, *z*=3.307, *P*=.001), perceived health status (β=.111, *z*=2.967, *P*=.003), eHealth literacy (β=.153, *z*=3.947, *P*<.001), and health management intention (β=.148, *z*=3.966, *P*<.001) had a significant positive impact on satisfaction with health guidance; perceived barriers (β=0.143, *z*=3.130, *P*=.002), health self-efficacy (β=0.147, *z*=3.491, *P*<.001), perceived health status (β= 0.120, *z*=3.072, *P*=.002), eHealth literacy (β= 0.131, *z*=3.362, *P*=.001), and health management intention (β= 0.175, *z*=4.670, *P*<.001) had a significant positive impact on satisfaction with health education; perceived severity (β=.118, *z*=3.235, *P*=.001), perceived barriers (β=.107, *z*=2.194, *P*=.03), perceived benefits (β=.089, *z*=2.428, *P*=.02), perceived health status (β=.088, *z*=2.188, *P*=.03), eHealth literacy (β=.170, *z*=4.561, *P*<.001), and health management intention (β=.083, *z*=2.064, *P*=.04) had a significant positive impact on satisfaction with health monitoring; perceived susceptibility (β=.098, *z*=2.314, *P*=.02), perceived barriers (β=.150, *z*=3.032, *P*=.002), health self-efficacy (β=.195, *z*=4.598, *P*<.001), perceived health status (β=.135, *z*=3.391, *P*=.001), eHealth literacy (β=.118, *z*=2.805, *P*=.005), and health management intention (β=.139, *z*=3.396, *P*=.001) had a significant positive impact on satisfaction with social functions, whereas perceived benefits (β=–.210, *z*=5.906, *P*<.001) had a significant negative impact and perceived susceptibility (β=.099, *z*=2.200, *P*=.03), perceived barriers (β=.139, *z*=3.074, *P*=.002), health self-efficacy (β=.137, *z*=3.076, *P*=.002), perceived health status (β=.122, *z*=2.834, *P*=.005), eHealth literacy (β=.119, *z*=2.771, *P*=.006), and health management intention (β=.100, *z*=2.354, *P*=.02) had a significant positive impact on satisfaction with gamification, while perceived benefits (β=–.165, *z*=5.011, *P*<.001) had a significant negative impact. These results help explain why different types of users exhibit varying levels of satisfaction with different functional designs. The statistical analysis results for each path are presented in Multimedia Appendix 8.

XSL•FO
RenderX

Table 7. Results of partial least squares analyses for 8 health management characteristics on satisfaction of 5 functional designs (β value).

	Satisfaction with health guidance	Satisfaction with health education	Satisfaction with health monitoring	Satisfaction with social function	Satisfaction with gamification
Perceived severity	.058	.050	.118[a]	.034	–.012
Perceived susceptibility	.076[b]	.049	.046	.098[b]	.099[b]
Perceived barrier	.177[c]	.143[a]	.107[b]	.150[a]	.139[a]
Perceived benefit	.048	–.001	.089[b]	–.210[c]	–.165[c]
Health self-efficacy	.140[a]	.147[c]	.070	.195[c]	.137[a]
Perceived health status	.111[a]	.120[a]	.088[b]	.135[a]	.122[a]
eHealth literacy	.153[c]	.131[a]	.170[c]	.118[a]	.119[a]
Health management intention	.148[c]	.175[c]	.083[b]	.139[a]	.100[b]
R^2	.183	.168	.152	.144	.094

[a]$P<.01$.
[b]$P<.05$.
[c]$P<.001$.

Discussion

Principal Findings

Understanding the heterogeneity of mHealth users and analyzing distribution differences in satisfaction with mHealth function design among user groups with distinct characteristics have significant referential value for personalized mHealth design. Therefore, guided by self-management theory and driven by self-management characteristics, this study constructed mHealth user personas encompassing 6 categories, analyzed the main characteristics and attributes of these 6 types of mHealth users, and examined differences in satisfaction with various mHealth function designs across user types. The results show that levels of user self-management characteristics are unevenly distributed and that different combinations of self-management characteristics can explain user heterogeneity. Moreover, a clear correlation exists between user heterogeneity arising from self-management characteristics and satisfaction with specific function designs. The findings of this study mutually confirm and extend existing literature on differences in mHealth intervention effects, user participation motivation, and technology design principles.

User Profiles of Fitness and Weight Management Apps Based on Health Management Characteristics

This study categorizes mHealth users into 6 distinct groups based on their self-management characteristics, revealing a complex classification derived from self-management–related attributes such as knowledge, beliefs, self-regulation skills and abilities, and motivation. This classification system differs from traditional approaches that rely on broad demographic labels (such as age and education level), offering a more refined and interpretable psychographic user profile. Although the sample in this study is predominantly composed of young-to-middle-aged and highly educated users, comparison with the 50th Statistical Report on China's Internet Development (2022) [62] shows that while individuals aged 20-49 account for 56.6% of all Chinese internet users, only 28.5% have used online medical services. This indicates that even within the broadly age-eligible internet population, active adopters of digital health services constitute a specific subgroup. Accordingly, this study employed an online snowball sampling method to reach and recruit active users with genuine interest and engagement in mHealth management. The "young-to-middle-aged, well-educated" characteristics of the sample, to some extent, reflect the profile of individuals who pay close attention to such apps. However, chi-square test results show that the associations between user classification and age (Cramer V=0.145) and education level (Cramer V=0.116) are weak. Furthermore, analysis of the age composition of each user group revealed that every category spans multiple age layers, suggesting that single demographic labels are insufficient to capture the characteristics of the 6 user categories identified in this analysis. The uneven distribution of multiple features across different user categories indicates that there is no complete positive correlation between individuals' perceived health status and their health management abilities and motivations. The differentiation of users' characteristic attributes is consistent with relevant theories such as BCWs. Potential health management demanders and passive attitude holders both exhibit low health management intention, but their underlying mechanisms differ: the former perceive susceptibility and severity but face higher perceived barriers, whereas the latter generally lack recognition of the perceived benefits and health self-efficacy of health management. This is consistent with phenomena observed in studies of smoking cessation apps [63], namely that when users remain at the level of understanding health information but fail to convert it into personalized risk perception and action confidence, behavioral intentions are difficult to activate. Based on the results of this study, targeted interventions in mHealth management are needed according to individuals' motivation status and health perceptions, to facilitate the transformation of different perception characteristics into internal motivation and to address differences

XSL•FO
RenderX

in health management behaviors arising from varying motivation statuses [64]. In addition, although negatively proactive energizers perceive an average health status, they exhibit high health management intention and self-efficacy. This finding indicates that perceived health status alone is not the sole determinant of mHealth service use, which is also supported by existing studies [65], showing that patients' satisfaction with mHealth services may be influenced by individual differences in disease progression, perceived service attitudes, and understanding of diagnostic and treatment outcomes. Correspondingly, passive attitude holders exhibit low levels of perceived health status, perceived benefits, health self-efficacy, and health management intention, which aligns with a previous study [66] identifying lack of motivation and low perceived value as significant barriers to mHealth participation. In general, obtaining a positive mHealth service experience often depends on whether services can be personalized or adaptively adjusted to accommodate different user needs and motivations, and user heterogeneity is strongly correlated with self-management characteristics [67].

Satisfaction of Various User Types With mHealth App Functionality Design

There are significant differences in satisfaction with mHealth function designs among different user types, providing counterevidence to the "one-size-fits-all" function design strategy and underscoring the necessity of personalized design. This study found that proactive intenders and positively proactive energizers, characterized by high self-efficacy and strong health management intention, reported higher satisfaction with goal-oriented and interactive functions such as health guidance, health education, and health monitoring. This finding is consistent with multiple studies [63,68,69], indicating that users who actively focus on their own well-being demonstrate higher levels of engagement with mHealth. Features such as progress tracking, personalized feedback, and access to educational content are often welcomed and perceived as useful by these users.

Low health management demanders and passive attitude holders report lower satisfaction with most mHealth functions. This finding is consistent with a previous study [70], which indicated that for users with low intrinsic motivation or insufficient perceived value, complex functions or interfaces may act as barriers to use; such users require designs that are concise, intuitive, provide immediate and intuitive feedback, and involve low operational barriers. Notably, prior research has demonstrated a positive correlation between eHealth literacy and mHealth usage [71]. However, our findings offer a contrasting perspective, showing that potential health management demanders, despite having lower-than-average eHealth literacy, report high satisfaction with health guidance and health monitoring functions. This may be because this group's higher perceived benefits and perceived susceptibility compensate for their lower capability, thereby increasing their interest in tools that enhance health benefits and reduce health risks. This phenomenon may also be partially explained by the Dunning-Kruger effect, whereby users lacking specialized knowledge may be less able to critically evaluate application limitations, leading to inflated satisfaction ratings for mHealth. Individuals with low health literacy have been shown to report levels of confidence about their knowledge comparable to, or even higher than, those with high literacy [72]. Conversely, users with stronger health knowledge reserves tend to use educational materials and support functions in mHealth apps more frequently, thereby enhancing recognition and evaluation of specific functions [73]. The user-centered design concept, combined with the user types constructed in this study, can therefore provide valuable reference for mHealth designs that address the diverse needs of different user groups.

Recommendations

The findings of this study provide clear implications for the personalized design and practical application of mHealth. mHealth developers should adopt a user-centered design approach and, based on the user types constructed in this study, tailor functional modules accordingly. For positively proactive energizers, further development of data tracking, personalized goal setting, and systematic educational content can be pursued, alongside the appropriate introduction of gamification elements to enhance users' sense of enjoyment [74]. For potential health management demanders, design efforts should focus on lowering usage barriers, providing clear and step-by-step health guidance and monitoring tools, and strengthening health-related beliefs. For passive attitude holders, the primary objective may be to establish initial engagement and trust through simple designs and light-touch interventions, while incorporating social features to help this group build social support [75], increase awareness of health risks, and enhance beliefs in health management. Furthermore, the study's results suggest that when evaluating the effectiveness of mHealth functions, user heterogeneity must be considered, as overall average satisfaction may obscure substantial differences among user subgroups.

Limitations and Future Research

Although this study was carefully designed in terms of methodology and empirical analysis, several limitations should be acknowledged. First, to avoid comparability issues arising from differences in functional design across mHealth apps with different core service content, and to recruit as diverse a user sample as possible, this study selected exercise, fitness, and weight management apps as the objects of study. However, these apps cannot fully represent all types of mHealth apps. Nevertheless, by conducting empirical research on one of the most widely used and numerous categories of mHealth apps, this study verifies that individuals' self-management characteristics have important referential value for mHealth user classification. The results demonstrate significant differences in satisfaction with various functional designs across user types, and that the effects of self-management characteristics on satisfaction differ across functional design dimensions. These results provide direct empirical reference for the personalized design of fitness and weight loss apps, and the research conclusions may also offer inspiration for user classification–related studies in other types of mHealth apps. Second, although questionnaires are an effective research method, they cannot fully represent all users. Access to feedback data from the entire mHealth user population would enable a more refined classification of user types. At present, limited

XSL•FO
RenderX

research on mHealth apps has focused on users' health cognition and self-management abilities, which constrains the analysis of comprehensive user characteristics from a big data perspective. To mitigate sampling bias, data were collected from 34 regions across China, and the sample size was substantially larger than that of comparable studies, supporting the representativeness of the findings. Third, although empirical results indicate that 8 self-management characteristics significantly influence user satisfaction, these characteristics explain only a limited proportion of the variance in satisfaction with mHealth (R^2=0.094-0.183). Although the identified pathways were statistically significant, a substantial proportion of the variance in satisfaction remains unexplained by these characteristics. Given that the primary aim of this study was to examine whether the effects of heterogeneous user types on satisfaction differ—that is, to construct user profiles based on self-management characteristics and analyze differences in satisfaction across user types—we did not conduct an in-depth examination of other potential determinants of satisfaction (eg, usage context, service quality, and user experience). Future research on factors influencing mHealth satisfaction could incorporate additional predictive variables to improve the explanatory power of the model.

Conclusions

This study identified 6 distinct mHealth user typologies (fitness and weight management apps) derived from a comprehensive set of self-management characteristics. The results reveal an uneven distribution of multidimensional characteristics across different user categories. This granular classification provides a deeper understanding of user heterogeneity than traditional demographic-based approaches. Overall, the primary contribution of this work is an empirically validated framework that offers a structured approach to understanding mHealth user diversity. The identified user typologies and their differential functional satisfaction profiles yield actionable insights for the personalized design of fitness and weight management apps. Specifically, for health guidance and education features, proactive intenders, positively proactive energizers, and negatively proactive energizers report higher satisfaction. For health monitoring functions, potential health management demanders, proactive intenders, positively proactive energizers, and negatively proactive energizers demonstrate higher satisfaction. For social features and gamification design, proactive intenders exhibit significantly higher satisfaction than the other 5 user types. From a practical perspective, these findings underscore the importance of adapting mHealth features to the distinct self-management needs and preferences of specific user groups. Such adaptation is essential for enhancing user satisfaction, fostering sustained engagement, and ultimately improving the perceived usefulness of digital health solutions.

Funding

This work was funded by the National Natural Science Foundation of China (grants 72404112 and 72574006), the Municipal Natural Science Foundation of Beijing (grant 7222306), and the National TCM innovation team and talent support projects (grant ZYYCXTD-C-202210).

Data Availability

The data will be made available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: TW
Data curation: JL (lead), TW (equal)
Formal analysis: TW (lead)
Funding acquisition: JL (lead), TW (equal)
Investigation: JL (lead), TW (equal), YF (supporting), ZL (supporting), XJ (supporting), QF (supporting), JM (supporting)
Methodology: TW (lead), YF (supporting), ZL (supporting), XJ (supporting), QF (supporting)
Project administration: JL
Resources: JL
Supervision: JL
Writing – original draft: TW
Writing – review & editing: TW (lead), JL (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measurement instrument.
[PDF File (Adobe PDF File), 298 KB - medinform_v14i1e64860_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Cluster sum of squared errors values and silhouette coefficients.
[PDF File (Adobe PDF File), 121 KB - medinform_v14i1e64860_app2.pdf]

Multimedia Appendix 3
Clustering analysis code.
[PDF File (Adobe PDF File), 111 KB - medinform_v14i1e64860_app3.pdf]

Multimedia Appendix 4
STROBE checklist.
[PDF File (Adobe PDF File), 238 KB - medinform_v14i1e64860_app4.pdf]

Multimedia Appendix 5
Comparison of health management characteristics between pairs of user groups.
[PDF File (Adobe PDF File), 327 KB - medinform_v14i1e64860_app5.pdf]

Multimedia Appendix 6
Comparison of satisfaction across 5 functions between paired user groups.
[PDF File (Adobe PDF File), 230 KB - medinform_v14i1e64860_app6.pdf]

Multimedia Appendix 7
Reliability and validity analysis.
[PDF File (Adobe PDF File), 178 KB - medinform_v14i1e64860_app7.pdf]

Multimedia Appendix 8
Path coefficients.
[PDF File (Adobe PDF File), 131 KB - medinform_v14i1e64860_app8.pdf]

References

1. Khalil AA, Meyliana, Hidayanto AN, Prabowo H. Identification of factor affecting continuance usage intention of mHealth application: a systematic literature review. 2020 Dec 24 Presented at: The 4th International Conference on Informatics and Computational Sciences (ICICoS); November 10-11, 2020; Semarang, Indonesia p. 1-6. [doi: 10.1109/icicos51170.2020.9299038]
2. Wang Y, Xue H, Huang Y, Huang L, Zhang D. A systematic review of application and effectiveness of mHealth interventions for obesity and diabetes treatment and self-management. Adv Nutr 2017 May;8(3):449-462 [FREE Full text] [doi: 10.3945/an.116.014100] [Medline: 28507010]
3. Helander E, Kaipainen K, Korhonen I, Wansink B. Factors related to sustained use of a free mobile app for dietary self-monitoring with photography and peer feedback: retrospective cohort study. J Med Internet Res 2014 Apr 15;16(4):e109. [doi: 10.2196/jmir.3084] [Medline: 24735567]
4. AppsFlyer. The uninstall ThreatApp uninstall benchmarks. AppsFlyer. 2020. URL: https://www.appsflyer.com/infograms/2019-app-uninstall-benchmarks/ [accessed 2024-12-10]
5. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
6. Wang T, Zheng X, Liang J, An K, He Y, Nuo M, et al. Use of machine learning to mine user-generated content from mobile health apps for weight loss to assess factors correlated with user satisfaction. JAMA Netw Open 2022 May 02;5(5):e2215014 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.15014] [Medline: 35639374]
7. Yang M, Al Mamun A, Gao J, Rahman MK, Salameh AA, Alam SS. Predicting m-health acceptance from the perspective of unified theory of acceptance and use of technology. Sci Rep 2024 Jan 03;14(1):339 [FREE Full text] [doi: 10.1038/s41598-023-50436-2] [Medline: 38172184]
8. Wang T, Fan L, Zheng X, Wang W, Liang J, An K, et al. The impact of gamification-induced users' feelings on the continued use of mHealth apps: a structural equation model with the self-determination theory approach. J Med Internet Res 2021 Aug 12;23(8):e24546 [FREE Full text] [doi: 10.2196/24546] [Medline: 34387550]
9. Wang T, Wang W, Feng J, Fan X, Guo J, Lei J. A novel user-generated content-driven and Kano model focused framework to explore the impact mechanism of continuance intention to use mobile apps. Computers in Human Behavior 2024 Aug;157:108252. [doi: 10.1016/j.chb.2024.108252]

XSL•FO
RenderX

10. Hashemi P, Reychav I, Arora A, Sawaed N, Sabherwal R, Azuri J. How does information technology affect individuals' health behavior in a pandemic? Insights from mobile health application use during COVID-19. Computers in Human Behavior 2024 Feb;151:108005. [doi: 10.1016/j.chb.2023.108005]
11. Singh P, Malik G. From trial to triumph! A longitudinal design eliciting social impact perspective for sustained usage of gamified fitness apps. APJML 2024 Aug 20;37(2):382-402. [doi: 10.1108/APJML-03-2024-0315]
12. Han K, Zo H. Understanding the mobile healthcare applications continuance: the regulatory focus perspective. Int J Med Inform 2023 Sep;177:105161. [doi: 10.1016/j.ijmedinf.2023.105161] [Medline: 37544241]
13. Nam M, Song T, Kim D, Jang K, Kim J, Koo B, et al. Effects of content characteristics and improvement in user satisfaction on the reuse of home fitness application. Sustainability 2023 May 22;15(10):8371. [doi: 10.3390/su15108371]
14. He Y, Chen H, Xiang P, Zhao M, Li Y, Liu Y, et al. Establishing an evaluation indicator system for user satisfaction with hypertension management apps: combining user-generated content and analytic hierarchy process. J Med Internet Res 2024 Sep 03;26:e60773 [FREE Full text] [doi: 10.2196/60773] [Medline: 39226103]
15. Nuo M, Zheng S, Wen Q, Fang H, Wang T, Liang J, et al. Mining the influencing factors and their asymmetrical effects of mHealth sleep app user satisfaction from real-world user-generated reviews: content analysis and topic modeling. J Med Internet Res 2023 Jan 31;25:e42856 [FREE Full text] [doi: 10.2196/42856] [Medline: 36719730]
16. Luna-Perejon F, Malwade S, Styliadis C, Civit J, Cascado-Caballero D, Konstantinidis E, et al. Evaluation of user satisfaction and usability of a mobile app for smoking cessation. Comput Methods Programs Biomed 2019 Dec;182:105042 [FREE Full text] [doi: 10.1016/j.cmpb.2019.105042] [Medline: 31473444]
17. Lee H, Uhm KE, Cheong IY, Yoo JS, Chung SH, Park YH, et al. Patient satisfaction with mobile health (mHealth) application for exercise intervention in breast cancer survivors. J Med Syst 2018 Nov 06;42(12):254. [doi: 10.1007/s10916-018-1096-1] [Medline: 30402781]
18. Arensman R, Kloek C, Pisters M, Koppenaal T, Ostelo R, Veenhof C. Patient perspectives on using a smartphone app to support home-based exercise during physical therapy treatment: qualitative study. JMIR Hum Factors 2022 Sep 13;9(3):e35316 [FREE Full text] [doi: 10.2196/35316] [Medline: 36098993]
19. Fu HN, Adam TJ, Konstan JA, Wolfson JA, Clancy TR, Wyman JF. Influence of patient characteristics and psychological needs on diabetes mobile app usability in adults with type 1 or type 2 diabetes: crossover randomized trial. JMIR Diabetes 2019 Apr 30;4(2):e11462 [FREE Full text] [doi: 10.2196/11462] [Medline: 31038468]
20. Screenagers to silver surfers: how each generation clicks with care. Rock Health. 2025. URL: https://rockhealth.com/insights/screenagers-to-silver-surfers-how-each-generation-clicks-with-care/ [accessed 2025-05-07]
21. Conway N, Campbell I, Forbes P, Cunningham S, Wake D. mHealth applications for diabetes: user preference and implications for app development. Health Informatics J 2016 Dec;22(4):1111-1120 [FREE Full text] [doi: 10.1177/1460458215616265] [Medline: 26635324]
22. Nie L, Oldenburg B, Cao Y, Ren W. Continuous usage intention of mobile health services: model construction and validation. BMC Health Serv Res 2023 May 05;23(1):442 [FREE Full text] [doi: 10.1186/s12913-023-09393-9] [Medline: 37143005]
23. Lee J, Kim J. Can menstrual health apps selected based on users' needs change health-related factors? A double-blind randomized controlled trial. J Am Med Inform Assoc 2019 Jul 01;26(7):655-666 [FREE Full text] [doi: 10.1093/jamia/ocz019] [Medline: 30946478]
24. Jamison RN, Mei A, Ross EL. Longitudinal trial of a smartphone pain application for chronic pain patients: predictors of compliance and satisfaction. J Telemed Telecare 2016 Nov 10;24(2):93-100. [doi: 10.1177/1357633x16679049]
25. Ryan P, Sawin KJ. The individual and family self-management theory: background and perspectives on context, process, and outcomes. Nurs Outlook 2009 Jul;57(4):217-225.e6 [FREE Full text] [doi: 10.1016/j.outlook.2008.10.004] [Medline: 19631064]
26. Zhou Y, Li S, Huang R, Ma H, Wang A, Tang X, et al. Behavior change techniques used in self-management interventions based on mHealth apps for adults with hypertension: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2024 Oct 22;26:e54978 [FREE Full text] [doi: 10.2196/54978] [Medline: 39437388]
27. Leziak K, Birch E, Jackson J, Strohbach A, Niznik C, Yee LM. Identifying mobile health technology experiences and preferences of low-income pregnant women with diabetes. J Diabetes Sci Technol 2021 Sep;15(5):1018-1026 [FREE Full text] [doi: 10.1177/1932296821993175] [Medline: 33605158]
28. Irfan Khan A, Gill A, Cott C, Hans PK, Steele Gray C. mHealth tools for the self-management of patients with multimorbidity in primary care settings: pilot study to explore user experience. JMIR Mhealth Uhealth 2018 Aug 28;6(8):e171 [FREE Full text] [doi: 10.2196/mhealth.8593] [Medline: 30154073]
29. Bauer W, Schiffman R, Ellis J, Erickson J, Polfuss M, Taani M, et al. An integrative review of the use of the individual and family self-management theory in research. Advances in Nursing Science 2025 Jun;48(2):E41. [doi: 10.1097/ans.0000000000000512]
30. Suess C, Maddock J, Dogru T, Mody M, Lee S. Using the Health Belief Model to examine travelers' willingness to vaccinate and support for vaccination requirements prior to travel. Tour Manag 2022 Feb;88:104405 [FREE Full text] [doi: 10.1016/j.tourman.2021.104405] [Medline: 34456412]
31. Glanz K, Rimer B, Viswanath K. Health Behavior and Health Education: Theory, Research, and Practice. San Francisco, CA: John Wiley & Sons; 2008:1-470.

XSL•FO
RenderX

32. Saghafi-Asl M, Aliasgharzadeh S, Asghari-Jafarabadi M. Factors influencing weight management behavior among college students: an application of the Health Belief Model. PLoS One 2020;15(2):e0228058 [FREE Full text] [doi: 10.1371/journal.pone.0228058] [Medline: 32032376]
33. Chauhan BF, Jeyaraman MM, Mann AS, Lys J, Skidmore B, Sibley KM, et al. Behavior change interventions and policies influencing primary healthcare professionals' practice-an overview of reviews. Implement Sci 2017 Jan 05;12(1):3 [FREE Full text] [doi: 10.1186/s13012-016-0538-8] [Medline: 28057024]
34. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions. London, UK: Silverback Publishing; 2014:1-184.
35. Michie S, van SMM, West R. The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
36. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
37. World Health Organization (WHO). Health promotion glossary of terms. WHO. Geneva, Switzerland: WHO; 2021. URL: https://apps.who.int/iris/handle/10665/350161 [accessed 2022-03-02]
38. Li W, Yuan K, Yue M, Zhang L, Huang F. Climate change risk perceptions, facilitating conditions and health risk management intentions: evidence from farmers in rural China. Climate Risk Management 2021;32:100283. [doi: 10.1016/j.crm.2021.100283]
39. Jeong H, Lee Y. Sex-based differences in the quality of life of elderly Koreans with chronic musculoskeletal pain. Int J Environ Res Public Health 2020 Jan 23;17(3):743 [FREE Full text] [doi: 10.3390/ijerph17030743] [Medline: 31979306]
40. Du J, Luo C, Shegog R, Bian J, Cunningham RM, Boom JA, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. JAMA Netw Open 2020 Nov 02;3(11):e2022025 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22025] [Medline: 33185676]
41. Williams DM, Rhodes RE. The confounded self-efficacy construct: conceptual analysis and recommendations for future research. Health Psychol Rev 2016 Dec;10(2):113-128 [FREE Full text] [doi: 10.1080/17437199.2014.941998] [Medline: 25117692]
42. Mendiola MF, Kalnicki M, Lindenauer S. Valuable features in mobile health apps for patients and consumers: content analysis of apps and user ratings. JMIR Mhealth Uhealth 2015 May 13;3(2):e40 [FREE Full text] [doi: 10.2196/mhealth.4283] [Medline: 25972309]
43. Gimpel H, Manner-Romberg T, Schmied F, Winkler TJ. Understanding the evaluation of mHealth app features based on a cross-country Kano analysis. Electron Markets 2021 Mar 24;31(4):765-794. [doi: 10.1007/s12525-020-00455-y]
44. Yoon S, Goh H, Nadarajan GD, Sung S, Teo I, Lee J, et al. Perceptions of mobile health apps and features to support psychosocial well-being among frontline health care workers involved in the COVID-19 pandemic response: qualitative study. J Med Internet Res 2021 May 31;23(5):e26282 [FREE Full text] [doi: 10.2196/26282] [Medline: 33979296]
45. Wilhide Iii CC, Peeples MM, Anthony Kouyaté RC. Evidence-based mHealth chronic disease mobile app intervention design: development of a framework. JMIR Res Protoc 2016 Feb 16;5(1):e25 [FREE Full text] [doi: 10.2196/resprot.4838] [Medline: 26883135]
46. Agarwal P, Gordon D, Griffith J, Kithulegoda N, Witteman HO, Sacha Bhatia R, et al. Assessing the quality of mobile applications in chronic disease management: a scoping review. NPJ Digit Med 2021 Mar 10;4(1):46 [FREE Full text] [doi: 10.1038/s41746-021-00410-x] [Medline: 33692488]
47. Wang Y, Collins WB. Systematic evaluation of mobile fitness apps: apps as the tutor, recorder, game companion, and cheerleader. Telematics and Informatics 2021 Jun;59:101552. [doi: 10.1016/j.tele.2020.101552]
48. Deterding S, Sicart M, Nacke L, O'Hara K, Dixon D, editors. Gamification: using game design elements in non-gaming contexts. 2011 Presented at: CHI 2011: The International Conference on Human Factors in Computing Systems; May 7-12, 2011; Vancouver, BC, Canada. [doi: 10.1145/1979742.1979575]
49. Lv H, Zeng N, Li M, Sun J, Wu N, Xu M, et al. Association between body mass index and brain health in adults: a 16-year population-based cohort and mendelian randomization study. Health Data Sci 2024 Mar 15;4:0087 [FREE Full text] [doi: 10.34133/hds.0087] [Medline: 38500551]
50. Serrano KJ, Yu M, Coa KI, Collins LM, Atienza AA. Mining health app data to find more and less successful weight loss subgroups. J Med Internet Res 2016 Jun 14;18(6):e154 [FREE Full text] [doi: 10.2196/jmir.5473] [Medline: 27301853]
51. Albarrak MM, Zakaria N, Meo SA. The role of mobile health applications (mHealth apps) in reshaping the body weight for better healthcare: a cross-sectional study. Pak J Med Sci 2024 Oct 11;40(9):1930-1936. [doi: 10.12669/pjms.40.9.9064] [Medline: 39416646]
52. Yin S, Cai X, Wang Z, Zhang Y, Luo S, Ma J. Impact of gamification elements on user satisfaction in health and fitness applications: a comprehensive approach based on the Kano model. Computers in Human Behavior 2022 Mar;128:107106. [doi: 10.1016/j.chb.2021.107106]
53. Mitchell R, Schuster L, Jin HS. Playing alone: can game design elements satisfy user needs in gamified mHealth services? Health Promot Int 2022 Apr 29;37(2):daab168 [FREE Full text] [doi: 10.1093/heapro/daab168] [Medline: 34651180]

XSL•FO
RenderX

54. Ahadzadeh AS, Pahlevan Sharif S, Ong FS, Khong KW. Integrating health belief model and technology acceptance model: an investigation of health-related internet use. J Med Internet Res 2015 Feb 19;17(2):e45 [FREE Full text] [doi: 10.2196/jmir.3564] [Medline: 25700481]
55. McArthur LH, Riggs A, Uribe F, Spaulding TJ. Health belief model offers opportunities for designing weight management interventions for college students. J Nutr Educ Behav 2018 May;50(5):485-493. [doi: 10.1016/j.jneb.2017.09.010] [Medline: 29097024]
56. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
57. Austin JD, Allicock M, Atem F, Lee SC, Fernandez ME, Balasubramanian BA. A structural equation modeling approach to understanding pathways linking survivorship care plans to survivor-level outcomes. J Cancer Surviv 2020 Dec;14(6):834-846 [FREE Full text] [doi: 10.1007/s11764-020-00896-6] [Medline: 32474862]
58. Wallston KA, Cawthon C, McNaughton CD, Rothman RL, Osborn CY, Kripalani S. Psychometric properties of the brief health literacy screen in clinical practice. J Gen Intern Med 2014 Jan;29(1):119-126 [FREE Full text] [doi: 10.1007/s11606-013-2568-0] [Medline: 23918160]
59. Zhou C, Zheng W, Tan F, Lai S, Yuan Q. Influence of health promoting lifestyle on health management intentions and behaviors among Chinese residents under the integrated healthcare system. PLoS One 2022;17(1):e0263004 [FREE Full text] [doi: 10.1371/journal.pone.0263004] [Medline: 35077472]
60. Jackson DL. Revisiting sample size and number of parameter estimates: some support for the N:q hypothesis. Structural Equation Modeling: A Multidisciplinary Journal 2003 Jan;10(1):128-141. [doi: 10.1207/s15328007sem1001_6]
61. Lou N. Analysis of the intelligent tourism route planning scheme based on the cluster analysis algorithm. Comput Intell Neurosci 2022 Jun 28;2022:3310676-3310610 [FREE Full text] [doi: 10.1155/2022/3310676] [Medline: 35800698]
62. The 50th statistical report on China's internet development. China Internet Network Information Center. 2022. URL: https://www3.cnnic.cn/n4/2022/0914/c88-10226.html [accessed 2026-01-04]
63. Tudor-Sfetea C, Rabee R, Najim M, Amin N, Chadha M, Jain M, et al. Evaluation of two mobile health apps in the context of smoking cessation: qualitative study of cognitive behavioral therapy (CBT) versus non-CBT-based digital solutions. JMIR Mhealth Uhealth 2018 Apr 18;6(4):e98 [FREE Full text] [doi: 10.2196/mhealth.9405] [Medline: 29669708]
64. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. Journal of Consulting and Clinical Psychology 1983;51(3):390-395. [doi: 10.1037/0022-006X.51.3.390]
65. Lu J, Zhao H, Ji Z, Dong Y. Effect of patient satisfaction on the utilization of mHealth services by patients with chronic disease. Digit Health 2025 Apr 15;11:20552076251333983 [FREE Full text] [doi: 10.1177/20552076251333983] [Medline: 40297355]
66. Alzghaibi H. Healthcare practitioners' perceptions of mhealth application barriers: challenges to adoption and strategies for enhancing digital health integration. Healthcare (Basel) 2025 Feb 25;13(5):494 [FREE Full text] [doi: 10.3390/healthcare13050494] [Medline: 40077056]
67. Wittmar S, Frankenstein T, Timm V, Frei P, Kurpiers N, Wölwer S, et al. User experience with a personalized mHealth service for physical activity promotion in university students: mixed methods study. JMIR Form Res 2025 Mar 28;9:e64384 [FREE Full text] [doi: 10.2196/64384] [Medline: 40153787]
68. Kowalski L, Finnes A, Koch S, Bujacz A. User engagement with organizational mHealth stress management intervention - a mixed methods study. Internet Interv 2024 Mar;35:100704 [FREE Full text] [doi: 10.1016/j.invent.2023.100704] [Medline: 38268709]
69. Monachelli R, Davis SW, Barnard A, Longmire M, Docherty JP, Oakley-Girvan I. Designing mHealth apps to incorporate evidence-based techniques for prolonging user engagement. Interact J Med Res 2024 Mar 26;13:e51974 [FREE Full text] [doi: 10.2196/51974] [Medline: 38416858]
70. Santos-Vijande ML, Gómez-Rico M, Molina-Collado A, Davison RM. Building user engagement to mHealth apps from a learning perspective: relationships among functional, emotional and social drivers of user value. Journal of Retailing and Consumer Services 2022 May;66:102956. [doi: 10.1016/j.jretconser.2022.102956]
71. Shaw G, Castro BA, Gunn LH, Norris K, Thorpe RJ. The association of eHealth literacy skills and mHealth application use among US adults with obesity: analysis of Health Information National Trends Survey Data. JMIR Mhealth Uhealth 2024 Jan 10;12:e46656 [FREE Full text] [doi: 10.2196/46656] [Medline: 38198196]
72. Canady BE, Larzo M. Overconfidence in managing health concerns: the Dunning-Kruger effect and health literacy. J Clin Psychol Med Settings 2023 Jun 29;30(2):460-468 [FREE Full text] [doi: 10.1007/s10880-022-09895-4] [Medline: 35768740]
73. Chua JYX, Choolani M, Chee CYI, Yi H, Lalor JG, Chong YS, et al. A mobile app–based intervention (Parentbot–a digital healthcare assistant) for parents: secondary analysis of a randomized controlled trial. J Med Internet Res 2025 Apr 17;27:e64882 [FREE Full text] [doi: 10.2196/64882] [Medline: 40245395]
74. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024 [FREE Full text] [doi: 10.2196/10024] [Medline: 29907557]

XSL•FO
RenderX

75. Nuseibeh BZ, Johns SA, Shih PC, Lewis GF, Gowan TM, Jordan EJ. Co-designing the MOSAIC mHealth app with breast cancer survivors: user-centered design approach. JMIR Form Res 2024 Dec 09;8:e59426 [FREE Full text] [doi: 10.2196/59426] [Medline: 39652855]

Abbreviations

BCW: Behavior Change Wheel
HBM: Health Belief Model
HTMT: heterotrait-monotrait ratio
mHealth: mobile health
SSE: sum of squared errors
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Coristine; submitted 29.Jul.2024; peer-reviewed by Y Yu, A Eisingerich; comments to author 30.Nov.2024; revised version received 10.Jan.2026; accepted 12.Jan.2026; published 10.Feb.2026.

Please cite as:
Wang T, Fan Y, Li Z, Jiao X, Fang Q, Ma J, Lei J
Exploring Self-Management–Based Mobile Health User Typologies and Associations Between User Types and Satisfaction With Key Mobile Health Functions: Comparative Study of Various Fitness and Weight Management App User Types
JMIR Med Inform 2026;14:e64860
URL: https://medinform.jmir.org/2026/1/e64860
doi:10.2196/64860
PMID:

©Tong Wang, Yiwen Fan, Zheng Li, Xiaoyi Jiao, Qichuan Fang, Junhao Ma, Jianbo Lei. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Balancing Privacy and Utility in Child and Adolescent Mental Health Services Research: Retrospective Cohort Study on Synthetic Data Generation

Mounir Haizoune[1], MSc; Bennett L Leventhal[2], MD; Dipendra Pant[3,4], MSc; Øystein Nytrø[3,4], PhD; Kaban Koochakpour[3,4], PhD; Roman A Koposov[5], MD, PhD; Lars Ravn Øhlckers[6], MSci; Norbert Skokauskas[3,7], MD, PhD

[1]Department of Analytics, Western Norway Regional Health Authority, Stavanger, Norway

[2]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, United States

[3]Norwegian University of Science and Technology, Trondheim, Norway

[4]Department of Child and Adolescent Psychiatry, Clinic of Mental Health Care, St. Olav University Hospital Trondheim, Trondheim, Norway

[5]Regional Centre for Child and Youth Mental Health and Child Welfare, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[6]Department of Child and Adolescent Psychiatry and Addiction, Stavanger University Hospital, Stavanger, Norway

[7]Regional Center for Child and Youth Mental Health and Child Welfare, Faculty of Medicine and Health Science, Trondheim, Norway

Corresponding Author:
Norbert Skokauskas, MD, PhD
Norwegian University of Science and Technology
Klostergata 46, 3. etasje
Campus Øya
Trondheim, 7030
Norway
Phone: 1 73558973
Email: norbert.skokauskas@ntnu.no

Abstract

Background: Electronic health records are essential for advancing research aimed at improving clinical outcomes. However, stringent data protection and privacy concerns severely limit the accessibility and use of real clinical data, particularly within Child and Adolescent Mental Health Services (CAMHS) involving vulnerable young individuals. This challenge can be effectively addressed through synthetic data generation, which safeguards individual privacy while facilitating comprehensive analyses of clinical information.

Objective: This study aims to investigate whether hierarchical synthetic data generators (SDGs) can effectively replicate the statistical properties, preserve the utility, and maintain the privacy of real CAMHS clinical data, thereby enabling data sharing and broader access to research-ready datasets.

Methods: This retrospective cohort study used electronic medical record data from 6924 distinct patients from CAMHS in Stavanger, Norway, comprising 7730 referral periods and 58,524 episodes of care. An 80%-20% split was used for training and testing. A hierarchical synthetic data generation model was trained to generate synthetic referral periods and associated episodes of care. Data quality was evaluated using SDMetrics for distribution (Kolmogorov-Smirnov Complement [KSC]/Total Variation Complement [TVC]), correlation (CorrelationSimilarity [CS]), and cardinality (CardinalityShapeSimilarity [CSS]) similarity. Privacy was evaluated using the Anonymeter library to simulate singling out, linkability, and inference reidentification attacks. Utility was assessed using the train synthetic test real (TSTR) pattern, comparing the predictive performance using precision-recall area under the curve [PRAUC] of models trained on synthetic vs real data for classifying the intensity of care.

Results: The hierarchical SDG created highly reproducible synthetic CAMHS data. The average statistical similarity scores were high across all metrics: KSC/TVC at 0.92, CS at 0.77 (intertable CS at 0.75), and CSS at 0.92. The synthetic data also demonstrated a low risk under simulated privacy attacks on a control dataset (n=1546): the average success rate was 6/1546 (0.39%) for singling out and 77/1546 (5%) for multivariate attacks. The average linkability risk was 54/1546 (0.5%), and the highest inference risk for a sensitive variable was 2/1546 (0.12%). The classification model trained on synthetic data (TSTR) produced comparable predictive performance (PRAUC=0.40) to the model trained on real data (PRAUC=0.43) for classifying

XSL•FO
RenderX

the intensity of care (low vs medium or higher). Shapley additive explanations analysis confirmed that the synthetic model's explanations aligned with real-world insights, validating its ability to capture fundamental predictive patterns.

Conclusions: Synthetic data can be used to build trust and promote collaboration among CAMHS researchers by offering access to extensive, representative samples with a low risk of patient identification. This approach expands the breadth of research while safeguarding patient privacy. Effective implementation of synthetic data generation depends on the model's ability to accurately identify and replicate the complex, sequential patterns present in real data.

(JMIR Med Inform 2026;14:e71819) doi:10.2196/71819

KEYWORDS

CAMHS; classification; length of referral; mental health; synthetic data generation

Introduction

The preservation of patient confidentiality is frequently identified as a significant barrier to conducting research with mental health care datasets [1]. Researchers appropriately face strict regulations and ethical considerations that limit access to real patient data. This is further complicated by selection bias [2], digital barriers [3], and high levels of sensitivity of mental health information, which, if disclosed, could lead to additional stigma, discrimination, and other negative consequences for individuals [4].

While these challenges are present across all mental health research, the use of large-scale data from patient records becomes significantly more complex with Child and Adolescent Mental Health Services (CAMHS) data. This is due to a set of problems associated with this vulnerable population. First, the variability in developmental stages among children and adolescents means that symptoms and diagnoses manifest differently across age groups (eg, anxiety in a 5-year-old compared to a 16-year-old), making data standardization and longitudinal analysis inherently difficult. Second, the necessity for age-appropriate interventions and evolving care pathways creates complex, nonlinear data structures that are challenging to model consistently [5]. Perhaps most critically, the ethical considerations surrounding privacy and consent for minors are far more stringent than for adult populations. Researchers must navigate nuanced consent processes involving both parents or guardians and the child's assent, often with varying legal requirements based on age and local jurisdiction [6,7]. This multilayered consent, coupled with heightened safeguarding responsibilities, severely restricts data sharing. Furthermore, due to their inherent nature, which includes stringent privacy regulations and the fragmented structure of care pathways, CAMHS datasets are typically smaller, exhibit greater structural fragmentation, and are consequently more challenging to access large-scale, generalizable research than adult health datasets.

These challenges can be effectively addressed by the generation of synthetic data. Synthetic data not only safeguard individual privacy but also permit comprehensive analyses of large-scale clinical data obtained directly from electronic medical records (EMRs) and other clinical data sources, even in sensitive areas such as CAMHS.

Previous studies have applied synthetic mental health data in data augmentation, privacy preservation, and fairness [8]. The SynthNotes framework, developed by Begoli et al [9], is notable for its ability to produce realistic and high-fidelity synthetic mental health documentation on a large scale. The framework facilitates benchmarking, optimization, and training for biomedical NLP, information extraction, and machine learning systems, all while ensuring strict privacy preservation. To ensure accurate and generalizable research results, synthetic data must accurately represent diverse minority subgroups. This is particularly important in addressing child and adolescent mental health inequities exacerbated by events like the COVID-19 pandemic [10]. Addressing these biases requires expertise in the underlying model and an understanding of fairness concepts like unawareness, demographic parity, and conditional fairness [8]. One approach involves directly de-biasing existing datasets to create "fair data," a process similar to synthetic data generation [11,12]. Alternatively, some researchers view data de-biasing as a "ground-up" generation problem, focusing on developing generative models that inherently incorporate fairness [13].

Predictive models trained on synthetic data for mental health have yielded encouraging findings in numerous studies, highlighting the critical importance of data quality, ethical considerations, and the necessity for ongoing research. A study by Rankin et al [14] used a deep learning approach to predict mental health quality using synthetic data derived from functional network connectivity data from resting-state functional magnetic resonance imaging, achieving high accuracy in classifying mental health quality. Using models trained on synthetic data, Ajith et al [15] combined data from multiple longitudinal mobile sensing studies to predict mental health symptoms, showing improved generalization across different datasets. While these models frequently failed to capture intricate patterns and subtleties, especially within diverse populations [16], their results highlight the potential of synthetic data to enhance predictive modeling in mental health, while addressing privacy and fairness concerns.

While prior synthetic health data studies focused on adult mental health, our research specifically examines how synthetic data generators (SDGs) effectively capture complex patterns and relationships in CAMHS data. We assess the essential trade-offs involving data quality, patient privacy, analytical utility, and optimal synthetic sample size to ensure generated data remains representative and valuable for research.

XSL•FO
RenderX

Methods

Data Source

This study was conducted using a retrospectively collected cohort from the Child and Adolescent Mental Health Department (CAMHD) at Stavanger University Hospital (SUS). The hospital functions as a referral center for about half a million inhabitants in the Rogaland metropolitan area and had an annual admission ranging from 17,000 to 19,000 during the study period. CAMHD at SUS provides assessment, diagnosis, and treatment for young people aged 18 years or younger, experiencing mental health issues such as depression, anxiety, trauma, developmental disorders like autism or attention-deficit/hyperactivity disorder (ADHD), and psychosis.

As detailed in Table 1, the dataset included 6924 distinct patients, 7730 referral periods, and a total of 58,524 episodes of care. Data extraction was performed using the electronic health record system (Distrubuert Informasjons og Pasientdatasystem i Sykehuset [DIPS]; Helse Vest IKT). The system provided access to administrative details like time of admission, department, and clinical personnel, as well as clinical information such as age, sex, and diagnosis for each episode of care.

Each episode is directly linked to a specific referral through a unique identifier. Allowing for comprehensive and continuous tracking of a patient's journey from their initial referral through all subsequent clinical interactions.

XSL•FO
RenderX

Table 1. Overview of the Child and Adolescent Mental Health Services dataset.

Variable	Entity	Description	Descriptive statistics
referralID	Consultation/stay	Unique identifier for each referral	• N/A[a]
Patientage	Referral	Age of the patient at referral	• Mean 11 (SD 4) • Median 12 (IQR 15)
Patientsex	Referral	Patient's sex (0=female; 1=male)	• Unique values=2 • Most frequent=male
referralStartDate	Referral	Date the referral was initiated	• Minimum=2014 • Maximum=2018
referralLength	Referral	The duration of the referral period in days	• Mean 58 (SD 41) • Median 30 (IQR 60)
stayID	Consultation/stay	Unique identifier for each consultation or stay	• N/A
staySartDate	Consultation/stay	Date when the consultation or stay began	• Minimum=2014 • Maximum=2018
StayLength	Consultation/stay	Duration of each stay in minutes	• Mean 58 (SD 41) • Median 30 (IQR 60)
personnelGrp	Consultation/stay	Category of medical personnel involved	• Unique values=4 • Most frequent="psychologist"
stayTypeGrp	Consultation/stay	Grouped type of consultation or stay	• Unique values=5 • Most frequent="treatment"
specialist	Consultation/stay	Indicates if the medical personnel is a specialist (yes/no)	• Unique values=2 • Most frequent="no"
roleGrp	Consultation/stay	Role of the clinical personnel during the consultation or stay	• Unique values=2 • Most frequent="contact"
cond1axisIGrp	Consultation/stay	*ICD-10*[b] Axis I (clinical syndromes)	• Unique values=8 • Most frequent="R40-R46"
cond1axisIIGrp	Consultation/stay	*ICD-10* Axis II (disabilities)	• Unique values=4 • Most frequent=N/A
cond1axisIIIGrp	Consultation/stay	*ICD-10* Axis III (intellectual developmental disorder)	• Unique values=4 • Most frequent=N/A
cond1axisIVGrp	Consultation/stay	*ICD-10* Axis IV (somatic diseases)	• Unique values=4 • Most frequent=N/A
cond1axisVGrp	Consultation/stay	*ICD-10* Axis V (psychiatric disorders and somatic diseases)	• Unique values=6 • Most frequent=N/A
cond1axisVIGrp	Consultation/stay	*ICD-10* Axis VI, global functioning (CGAS[c])	• Unique values=7 • Most frequent=6999

[a]N/A: not applicable.

[b]ICD-10: International Statistical Classification of Diseases, Tenth Revision.

[c]CGAS: Children Global Assessment Score.

Missing Data

The dataset contains *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnostic codes categorized across different axes or levels of detail. The *ICD-10* uses a multiaxial system for mental and behavioral disorders, including Axis I (clinical syndromes), Axis II (disabilities), and Axis III (environmental/lifestyle factors).

XSL•FO
RenderX

A key characteristic of the Axis columns is the high rate of missing data, which indicates "not applicable" rather than an error, as is common in clinical coding. For instance, the first condition in Axis I is nearly complete (0.7% missing), but subsequent codes within the same axis show drastically increased missingness (eg, 6th condition, 99.8% missing). This trend continues across other axes, like Axis II (eg, 3rd condition 99.6% missing). This consistent pattern of missing values signifies that a code was not relevant to a clinical episode, making data imputation inappropriate.

Ethical Considerations

The study is a quality assurance initiative directly building on the results of the "PILOT SYNDATA" project. The project was conducted as a collaboration between CAMHD SUS, the analytical department at SUS, and Helse Vest IKT (the regional IT service delivery company). The project's main purpose was to explore the safe use of synthetic data in mental health care.

The data for the "PILOT SYNDATA" project were collected from DIPS, and Helse Vest IKT performed the deidentification and aggregation procedures, with the resulting anonymized data then being used for synthetic data generation in the next phase. This entire process was conducted under a formal agreement that permits the secondary use of health data for research and quality assurance while ensuring patient privacy and confidentiality. Due to the complete deidentification and aggregation of the data, a waiver was not required to use the data in this study.

The study used a synthetic data generation process that creates entirely new, statistically representative data points. The process is fully compliant with HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) regulations, significantly reducing the risk of an attacker reidentifying individuals or uncovering sensitive patient details from the original dataset.

As this was a retrospective analysis of existing data, no financial compensation was provided to participants. Additionally, it is not possible to identify any individual participants in the figures or supplementary materials. The images are illustrative and use a "real-like path" to represent an anonymized, generalized patient journey.

Synthetic Data Generation

We use HMASynthetiser, a multitable algorithm within the Synthetic Data Vault framework [17]. HMASynthetiser is a class model that implements a "hierarchical modeling algorithm" to analyze data structured in multiple levels, often called nested data. These models are particularly useful when dealing with complex datasets in which observations are grouped at different levels, such as referral periods and episodes of care (Multimedia Appendix 1).

The SDG systematically processes the real CAMHS dataset, examining how information is structured across the referrals and episodes of care tables. It uses a Gaussian copula to map the complex relationships between every data field. A Gaussian Copula [18] is a statistical model used to identify the dependency structure between the nested dataset (referrals/episodes). The approach ensures that the SDG process accurately mirrors the statistical patterns and distributions present in the original patient records, even for variables with nonstandard or highly dependent relationships. In addition, the parameters of the SDG are customized for each individual table in the hierarchy, allowing fine-tuning of the synthetic data generation process. Hierarchical modeling, combined with the statistical method, is essential for accurately capturing the complex dependency structures within CAMHS variables.

Data Modeling

To optimize data utility and patient privacy, the real CAMHS dataset underwent comprehensive preprocessing prior to synthetic data generation. This involved the systematic categorization of variables, such as age and admission type, into broader, analytically meaningful groupings. Specifically, age was discretized into decennial intervals, while variables, such as "performing personnel" and "admission type," were recorded by mapping their original granular values to more generalized, descriptive categories. Additionally, diagnostic Axes I-IV were grouped based on the *ICD-10* hierarchy, and Axis VI (Children's Global Assessment Score) underwent specific binning and categorization to abstract sensitive medical codes.

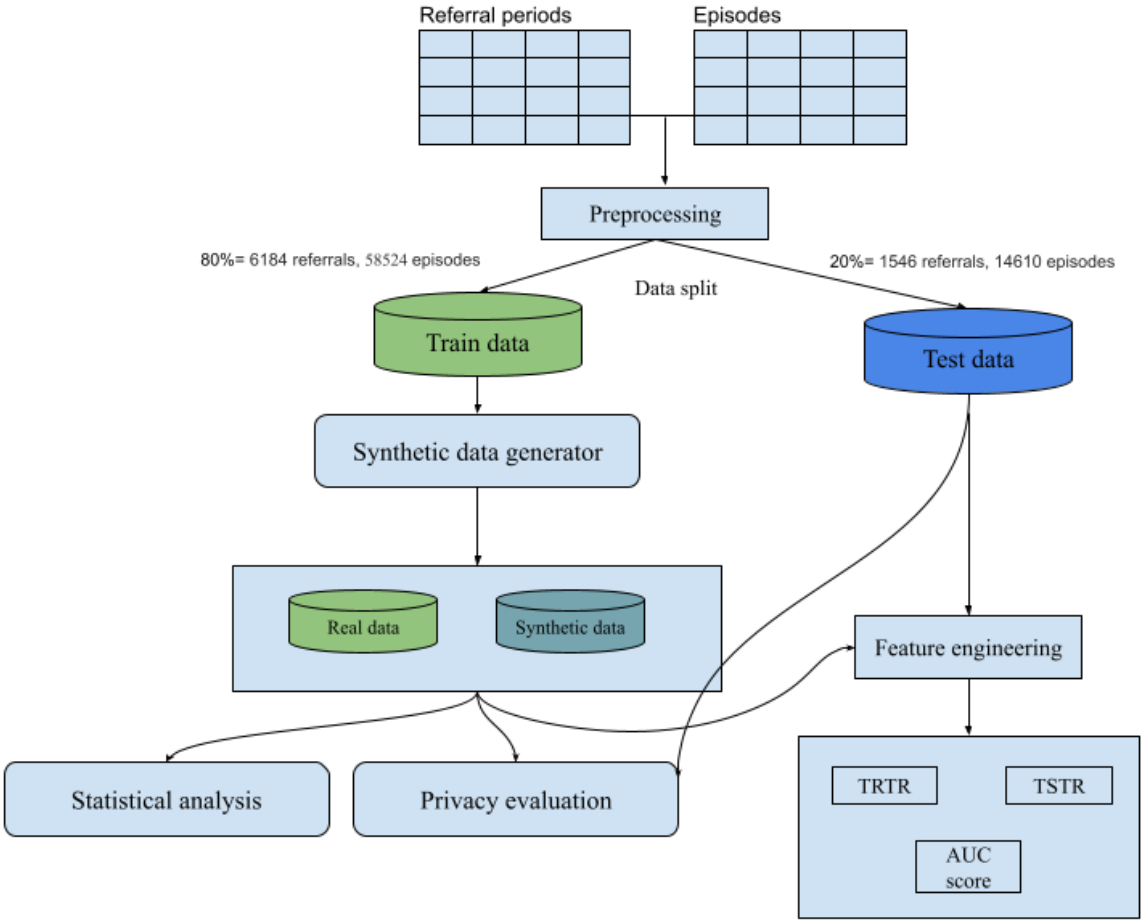
These comprehensive preprocessing steps ensured the data were optimally structured for generating high-fidelity synthetic data and accurately reflected real-world patterns while rigorously upholding patient data protection.

Experimental Setup

Figure 1 displays the experimental setup. The experiments performed in this study followed 3 structures: sample size, privacy risk assessment, and utility assessment.

XSL•FO
RenderX

Figure 1. General study design. AUC: area under the curve; TRTR: train real test real; TSTR: train synthetic test real.



Sample Size

To evaluate the impact of sample size on predictive model performance (Multimedia Appendix 2), 4 distinct training sets (619, 1546, 3092, and 6184 referrals) were created from the real-world CAMHS data. For each of these sizes, a corresponding synthetic dataset was generated to enable a comparison of predictive model performance using both real and synthetic data across various data availability scenarios.

Privacy Risk Assessment

To measure the privacy risk in the generated synthetic data, we used the experimental approach via the Python library Anonymeter [19] (Multimedia Appendix 3). Initially, the attack uses synthetic data to create a series of guesses. The model then evaluates these guesses by comparing them to the values present in the original dataset, thereby assessing the effectiveness of the attack. The privacy model requires the adversary to conduct the same attack on a control dataset (records that were not used in the SDG). By contrasting the success rates of the attack on both training records and control records, a sensitive and significant measure of risk is established. Three types of privacy risk attacks were simulated and analyzed to evaluate the potential for reidentification: singling out, linkability, and inference risks.

Utility Assessment

Prediction tasks represent a robust and frequently used methodology for evaluating synthetic data utility, offering distinct advantages beyond conventional statistical comparisons. The method directly confirms the data's real-world applicability by assessing its capacity to replicate complex, nonlinear relationships for applications such as individualized mental health care planning or treatment pathway optimization. Furthermore, the approach validates the functional equivalence of synthetic data and provides a user-centric, interpretable metric of its inherent value, thereby facilitating the detection of subtle discrepancies overlooked by simpler analytical techniques [20]. To assess synthetic data utility, we formulated a downstream prediction task (Multimedia Appendix 4): classifying patient referral periods by intensity of care. Following Beckers et al [21], "intensity of care" was initially defined as low (<15 sessions/year), medium (15-51 sessions/year), or high (≥52 sessions/year). We extracted primary diagnosis of the first episode, sex, age, and total past referrals as predictor variables. Due to significant skewness (386/7730 high, 2087/7730 medium, and 5874/7730 low), the "medium intensity" and "high intensity" categories were merged into a "medium- or higher-intensity care" group. The combined group was selected to create a sufficiently sized positive class for stable model training, given the constraints of the severely skewed original distribution.

XSL•FO
RenderX

The dataset was then stratified by intensity and split into an 80% training set (6184 referral periods and 58,527 episodes) and a 20% independent test set (1546 referral periods and 14,610 episodes; Multimedia Appendix 5). The stratification ensured balanced distribution and kept all related patient data within either the training or test set; the test set was used solely for evaluation. A random forest model [22] was then trained using synthetic data generated at various scale factors (1, 2, 5, 10). The trained models were evaluated using the Train Synthetic, Test Real evaluation method [23], which involves training a model on synthetic data and testing it on real data.

Model performances were measured using precision-recall area under the curve (PRAUC), a robust metric particularly suited for imbalanced datasets [24,25]. We specifically chose PRAUC over the traditional area under the receiver operating characteristic curve because PRAUC is less susceptible to inflation from the majority (low intensity) class, providing a more accurate and rigorous measure of performance on the positive (medium- or higher-intensity care) class [24]. Results are presented as the median of a stratified 4-fold cross-validation.

Finally, SHAP (Shapley additive explanations) values [26] were computed for both models to assess variable interactions.

Statistical Analysis

To quantify the similarity between real and synthetic data, we used 4 key metrics from the SDMetrics Python library [23]. For numerical column distributions, we used the Kolmogorov-Smirnov Complement (KSC), while the Total Variation Complement (TVC) measured discrete column distributions. The preservation of intercolumn correlation patterns was evaluated using CorrelationSimilarity (CS), and CardinalityShapeSimilarity (CSS) was applied to assess the similarity in cardinality distributions within multitable datasets.

For privacy risk evaluation, we conducted 3 attack-based assessments focused on factual anonymization indicators: singling out determined the potential for identifying individual records; linkability assessed the risk of linking synthetic data back to the real dataset; and an inference attack was used to attempt to deduce sensitive individual information from seemingly nonsensitive data. A low inference risk means it is highly unlikely that an attacker could uncover specific, private details about individual patients from the original dataset.

Qualitative Visualization of Care Pathways

For the qualitative assessment, the specific “real patient path” was selected randomly from the overall real dataset after the synthetic data generation process was completed to ensure the visualization was not conditioned on the selected patient’s attributes prior to generation. This patient, a 12-year-old male diagnosed with “behavioral and emotional disorders” (F90-F98), represents a highly frequent and common clinical profile in our cohort. The corresponding “synthetic path” was then extracted from the synthetic data by identifying a synthetic patient with identical starting attributes (age, sex, and primary diagnosis) to provide a valid, anonymized, and attribute-matched comparison for event sequencing and density.

Results

Similarity and Quality Assessment of Synthetic CAMHS Data

As detailed in Table 2, the SDG produced CAMHS data that demonstrated high statistical similarity to the “real-world” CAMHS dataset. The result was evident across all reported metrics (KSC/TVC score=0.92; CS=0.75; CS (intertable)=0.73; CSS=0.9), with good fidelity remaining consistent regardless of the amount of “real-world” training samples provided to the algorithm.

Figure 2 presents 2 kernel density estimate plots, offering a visual comparison of feature distributions between real-world CAMHS data and synthetic data generated at various sample sizes. The left plot, illustrating “referral length,” reveals that the synthetic data distributions, particularly those derived from larger sample sizes (n=6184 and n=3092), closely align with the real test data’s distribution in the high-density range. While the synthetic data effectively captures the predominant patterns of referral lengths, there are minor deviations observed in the extreme tails, suggesting a slight challenge in perfectly replicating infrequent, very long referrals.

Similarly, the right plot, depicting “age,” demonstrates exceptionally high fidelity. All synthetic age distributions, regardless of the sample size, exhibit remarkable overlap with the real test data’s distribution. The peaks and overall shapes are nearly identical, indicating that the synthetic data generation process is highly successful in preserving the univariate statistical characteristics of patient age.

The observed high scores across the various quality metrics and kernel density estimates provide robust evidence that the synthetic data accurately replicates the original patient information’s characteristics and inherent relationships. Specifically, the significant KSC/TVC scores confirm that individual statistical distributions of each variable within the synthetic dataset demonstrate strong consistency with those observed in the real data, which implies that features like patient age ranges or referral lengths exhibit statistical properties consistent with their real-world counterparts. Furthermore, the high CS scores, including both intratable and intertable relationships, confirm the accurate preservation of complex dependencies and interactions between distinct variables. For example, critical correlational structures, such as the frequent association of specific diagnoses with treatment pathways or demographic profiles in the real data, are accurately maintained within the synthetic dataset. Finally, the high CSS scores confirm that the structural properties of the data, including the diversity of unique values within categorical fields and the overall shape of numerical distributions, are accurately replicated.

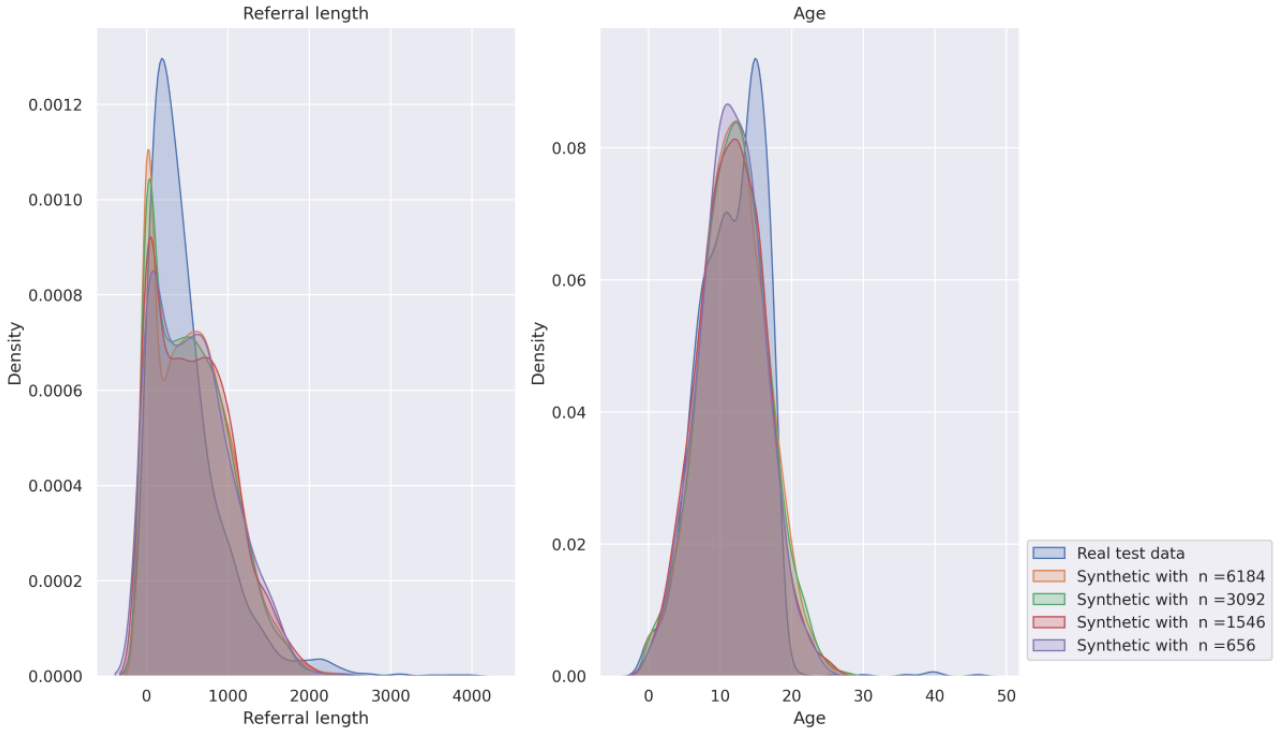
The consistent high scores underscore that the synthetic data not only replicates individual variable characteristics but also captures the intervariable relationships and the overall statistical integrity of the original, sensitive CAMHS patient information. The observed fidelity is critical for establishing synthetic data’s suitability for subsequent analytical and research applications.

XSL•FO
RenderX

Table 2. Average Kolmogorov-Smirnov Complement (KSC)/Total Variation Complement (TVC), CorrelationSimilarity (CS; single and intertable), and CardinalityShapeSimilarity (CSS) metrics from synthetic data evaluation experiments using different synthetic scale factor sizes.

	656 referrals	1546 referrals	3092 referrals	6184 referrals
KSC/TVC	0.91	0.92	0.92	0.92
CS	0.73	0.77	0.75	0.77
CS (intertable)	0.69	0.73	0.73	0.75
CSS	0.87	0.90	0.91	0.92

Figure 2. Visual comparison of the density distribution of the real and the derived synthetic Child and Adolescent Mental Health Services features. Density distributions for 2 selected features, “referral length” and “age,” and their synthetic derivatives are compared.



Synthetic Data Privacy Evaluation

To assess the potential for patient reidentification and information revelation, we performed 3 attack-based privacy evaluations: singling out, linkability, and attribute inference. Table 3 provides an overview of the success rate associated with the 3 simulated privacy attacks using synthetic data on a control real dataset (n=1546), highlighting the success rate achieved with various sample sizes. Specifically, the analysis reveals that both univariate and multivariate singling out attacks yield minimal average success rates, recorded at just 6/1546 (0.4%) and 77/1546 (5%), respectively. This indicates that the likelihood of successfully executing these types of attacks is low, regardless of the volume of data being analyzed. Furthermore, the table presents insights into the risk of linkability when using clinical and demographic variables as auxiliary data. The average risk of linkability is calculated to be 0.5% (8/1546), suggesting that while there is some potential for linking data to individuals, the overall risk remains quite low.

Similarly, inference risk of sensitive variables from patient age and sex was minimal for most attributes (Figure 3), with “admission week number” presenting the highest, yet still low, risk (2/1546, 0.12%).

A low percentage, for instance, a range of 0%-5%, indicates that an attacker’s capacity to deduce private details from the synthetic data is only marginally superior to random chance. To contextualize this, if Figure 3 reveals an inference risk of, for example, 2%, it signifies that an attacker possessing auxiliary nonsensitive information (eg, a patient’s age and gender from an external source) would still have only a 2% likelihood of accurately determining a specific sensitive attribute, such as a rare mental health diagnosis, from the synthetic dataset. The low probability demonstrates the robust privacy preservation achieved, reinforcing the suitability of this approach for sensitive health care data.

The results clearly demonstrate the robust capability of the synthetic data to effectively protect individual privacy. The protection is provided by the synthetic data generation process, which creates entirely new, statistically representative data points rather than simply masking or perturbing original records, thereby breaking any direct link to individual patients. As a result, it becomes highly unlikely for an attacker to reidentify specific individuals or uncover private, sensitive details about

XSL•FO
RenderX

patients from the original dataset, even if they possess some auxiliary information. This significantly mitigates risks associated with reidentification or attribute inference attacks, enabling a secure environment for data sharing.

Table 3. Success rate of singling out, linkability, and inference attacks on control data using various synthetic Child and Adolescent Mental Health Services data samples.

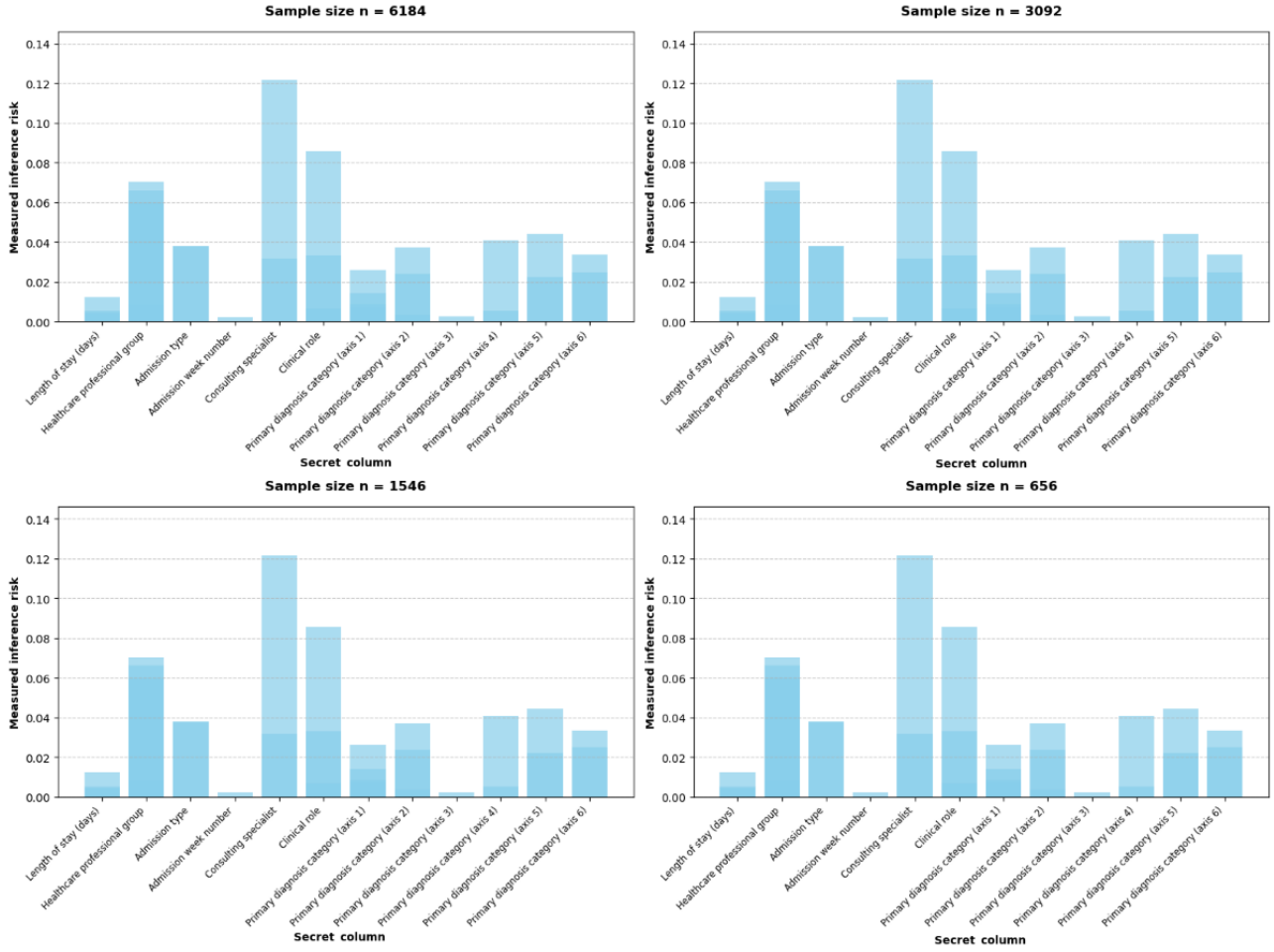
Risk type	6184 referrals, n/N (%)	3092 referrals, n/N (%)	1546 referrals, n/N (%)	619 referrals, n/N (%)
Singling out (univariate)[a] success rate	6/1546 (0.38)	6/1546 (0.38)	6/1546 (0.38)	7/1546 (0.43)
Singling out (multivariate)[b] success rate	46/1546 (3)	93/1546 (6)	62/1546 (4)	108/1546 (7)
Linkability[c] success rate	2/1546 (0.13)	5/1546 (0.34)	11/1546 (0.69)	13/1546 (0.84)

[a]Singling out (univariate) represents the percentage of successful attempts to uniquely identify an individual in the synthetic dataset using a single characteristic (based on 500 attacks). Lower values indicate better privacy.

[b]Singling out (multivariate) represents the percentage of successful attempts to uniquely identify an individual in the synthetic dataset using multiple combined characteristics (based on 100 attacks). Lower values indicate better privacy.

[c]Linkability represents the percentage of successful attempts to link synthetic records back to their original real records (based on 2000 attacks and 10 neighbors). Lower values indicate better privacy.

Figure 3. Inference risk.



Synthetic Data Utility Evaluation

The target variable to be classified (intensity of care) in the utility evaluation was highly unbalanced in the real dataset, with 4111/6136 (67%) categorized as low intensity (<15 sessions annually), 2024/6136 (33%) as medium or higher intensity of care (>15 per year). We use the PRAUC metric to compare the predictive performance of machine learning models trained on real CAMHS data versus those trained on synthetic CAMHS data across different input sizes. Given the positive class prevalence of 33%, the random chance baseline (PRAUC baseline) is 0.33. As Table 4 illustrates, increasing sample size consistently improved predictive performance for both real and synthetic data models. In the most data-scarce setting (619 records), the random forest model's ability to classify medium-

XSL•FO
RenderX

or higher-intensity care patients was limited (real data: average PRAUC=0.39; synthetic data: PRAUC=0.32), However, when we increase the training set size to 6184 records, the model performance improved, with the average PRAUC for real data rising to 0.43 (a 30% increase over baseline) and for synthetic data to 0.40 (a 21% increase over baseline) demonstrating that the synthetic data preserves the analytical utility necessary to exceed the random threshold and capture real-world signals.

Our findings suggest that while privacy safeguards around children's data are paramount and often lead to smaller datasets, the value of more data in building accurate and robust predictive tools cannot be overstated. Simply put, the more comprehensive the data, the better equipped the models are to provide meaningful insights that can ultimately support patient care.

Table 4. Comparison of random forest–based prediction models trained on real, tested on real (TRTR) Child and Adolescent Mental Health Services (CAMHS) data alone versus those trained on synthetic, tested on real (TSTR) CAMHS data.

Sample size	TRTR, PRAUC[a] (range)	TSTR, PRAUC (range)
619 referrals	0.39 (0.34-0.44)	0.32 (0.28-0.38)
1546 referrals	0.40 (0.36-0.47)	0.33 (0.29-0.38)
3092 referrals	0.41 (0.37-0.47)	0.34 (0.30-0.39)
6184 referrals	0.43 (0.38-0.49)	0.40 (0.35-0.45)

[a]PRAUC: precision-recall area under the curve.

Reproducibility of Feature Interpretability (SHAP)

SHAP analysis, presented in Figure 4, provides an insight into how different patient characteristics impact the predicted intensity of care, both when the model is trained on real patient data and when it is trained on synthetic patient data.

The model consistently showed that a lower prevalence of "behavioral and emotional disorders" (F90-F98) led to a higher predicted intensity of care. This finding is a clinically meaningful observation rooted in the structure of CAMHS resource allocation; because F90-F98 diagnoses (eg, ADHD, externalizing difficulties) are highly prevalent and typically managed via lower-intensity, first-line support, their presence is not a strong predictor of high-intensity service use. Therefore, the absence of this common, low-intensity diagnosis in a patient referred for complex care suggests the presence of a less common, more severe, or highly comorbid condition that immediately necessitates the allocation of higher-intensity resources [27].

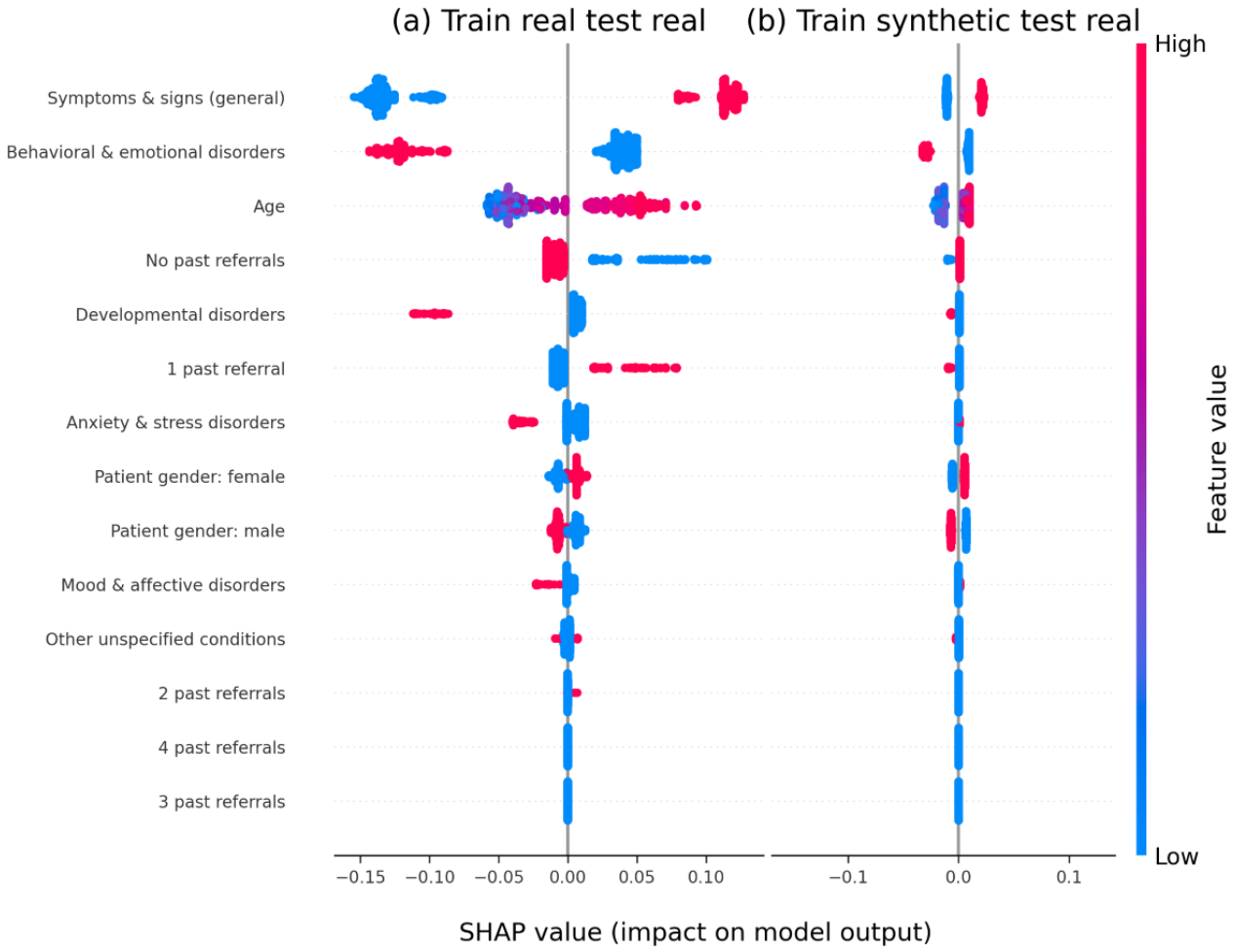
Conversely, a greater presence of "symptoms and signs (general)" (R40-R46) or "developmental disorders" (F80-F89) contributed to a lower predicted intensity of care. Patient age also presented a clear and consistent trend; younger patients were generally predicted to require higher intensity care, while older patients were associated with lower predicted intensity. On the plots, lower values for a feature are typically shown in gray, higher values in black; features increasing predicted intensity extend to the right, while those decreasing it extend to the left.

While the models showed consistent variable impact on the target prediction (intensity of care), we noted a significant difference in the strength of this impact between the synthetic and the real prediction models. The model trained on actual patient data showed a much wider range of impacts from individual features (SHAP values ranging from approximately –0.15 to 0.15), suggesting that real patient characteristics could lead to more pronounced shifts in the predicted intensity of care. In contrast, the model trained on synthetic data displayed a significantly narrower range of influence (approximately –0.03 to 0.03). The "flattening" of the variables' impact on the predicted intensity of care implies that while synthetic data helps the model learn the general direction of how patient characteristics relate to care intensity, it reduces the overall sensitivity of the model to the nuanced, individual variations found in real patient data. This effect, however, can be mitigated by using more advanced synthetic data generation techniques.

Despite the observed reduced variable impact in the synthetic model, the consistency in the direction of SHAP-based explanations remains clinically valuable. It effectively validates synthetic data's ability to capture fundamental predictive patterns in sensitive health care records, making synthetic datasets particularly useful for privacy-preserving exploratory research and facilitating hypothesis generation and validation. While acknowledging the continued need for further efforts to precisely replicate nuanced quantitative impacts for high-stakes clinical predictions, the qualitative fidelity of the result offers an effective tool for initial understanding and safe experimentation in CAMHS health research.

XSL•FO
RenderX

Figure 4. Shapley additive explanations (SHAP) values for intensity of care in (A) train real test real and (B) train synthetic test real models (n=6184).



Synthetic and Real Individual Care Paths

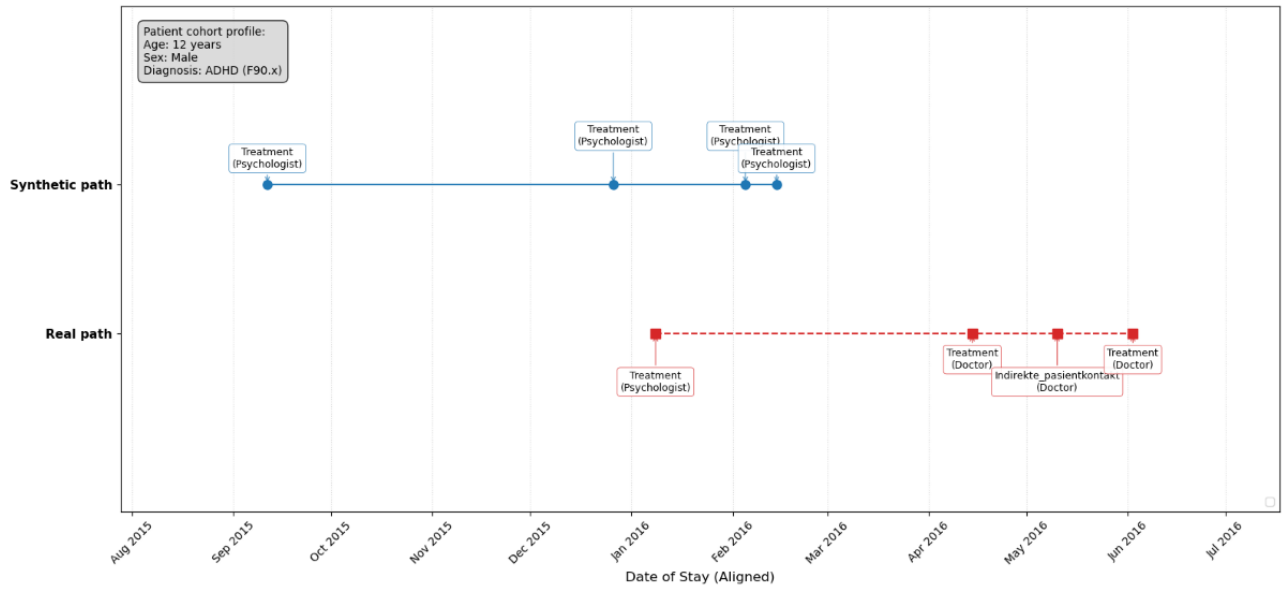
Figure 5 compares 2 distinct care pathways for a 12-year-old male patient diagnosed with "behavioral and emotional disorders" (F90-F98). Both paths are presented on a shared timeline (x-axis) that has been aligned to start at a common point, allowing for direct comparison of event sequencing and density.

The figure serves as a comparative visualization, highlighting potential differences in the flow, types of interventions, and personnel involvement between a synthetic and a more real care path.

While synthetic CAMHS data successfully captures the overarching statistical patterns of the real CAMHS dataset, divergences at the individual patient level offer valuable insights. The differences result from synthetic data representing the most common trajectories, an average derived from the entire dataset that naturally filters out the unique noise and unobserved idiosyncratic factors influencing real-world care. Real CAMHS care is inherently nonlinear and dynamic, shaped by patient responses, unforeseen crises, resource constraints, and patient/family agency. These complexities, including feedback loops and real-world bottlenecks such as waiting lists, are challenging for synthetic data to perfectly replicate at an individual level, leading to the observed divergence.

XSL•FO
RenderX

Figure 5. Comparison of real-like and synthetic care paths for an attention-deficit/hyperactivity disorder (ADHD) patient. The x-axis (Date of stay [Aligned]) shows the progression of time. The y-axis separates the "Synthetic path" (top, solid line) from the "Real path" (bottom, dashed line). Individual markers represent stay events, annotated with stay types and personnel categories. The grey box (top left) indicates shared patient demographics for both trajectories.



Discussion

Principal Findings

This study demonstrates the significant potential of synthetic data to enhance the utility of electronic health care records within CAMHS. Our methodology, using a hierarchical SDG, generated synthetic datasets that exhibited high statistical fidelity to real-world CAMHS data, as confirmed by various quality metrics (KSC/TVC, CS, CSS) and visual density comparisons. While minor quality variations were observed with decreasing sample sizes [28], the results confirm that our approach broadly preserves variable behaviors and dependencies between variables in the generated synthetic datasets [29]. However, identifying the precise minimum real data sample size for generating truly useful and high-fidelity synthetic CAMHS data is complex and highly dependent on the specific research scope and applications.

Models trained on synthetic data demonstrated predictive performance in classifying patient “intensity of care” (low vs medium or higher) that was comparable to models trained on real data. While the absolute predictive power of the model was modest (PRAUC=0.40-0.43 over a 0.33 random baseline), the key finding is the preservation of this utility in the synthetic data, as evidenced by the comparable scores**.**

The SHAP analysis confirmed that the synthetic model’s explainability aligned with real-world insights, underscoring the consistent importance of age and CAMHS diagnosis (eg, anxiety disorders, depression, ADHD, and autism spectrum disorder) in assessing care needs [30,31]. Furthermore, our findings provide strong evidence that synthetic data effectively protects individual privacy. The data generation process makes it highly unlikely for attackers to reidentify patients or uncover sensitive details from the original dataset, a critical requirement given the highly personal nature of CAMHS data.

The implications of high-fidelity, privacy-preserving synthetic CAMHS datasets are profound for children and adolescent mental health research and practice. Their availability removes critical barriers to data sharing and collaborative research, enabling more thorough investigations into mental health aspects across diverse populations [32], especially for vulnerable ones like those served by CAMHS. This capability facilitates a deeper understanding of mental disorder patterns and contributes to the development of more effective, personalized treatment methodologies [33]. Moreover, given that young adolescents often withhold important information from health care providers due to privacy concerns, which hinders effective diagnosis and treatment [34], synthetic data allows researchers to explore critical questions that would be impossible with real, incomplete data. These include understanding the impact of stigma on treatment adherence, evaluating novel personalized interventions, investigating the influence of “hidden” family dynamics, and simulating the effects of systemic barriers on access to care for vulnerable groups.

Furthermore, maintaining highly accurate long-term temporal and sequential relationships across numerous event types remains challenging in synthetic data generation. Most synthetic data quality metrics prioritize aggregate statistical properties over the exact fidelity of unique individual sequences. This means a synthetic dataset can be excellent at the population level but still show microlevel divergence in individual trajectories. Despite these differences, the divergence is valuable. It highlights that synthetic data, while not predicting exact individual pathways, excels at understanding general care models, developing system-level interventions via cohort simulations, and creating privacy-preserving case studies for training. It effectively complements the detailed study of highly personalized real-world clinical journeys.

XSL•FO
RenderX

Future Work

To improve upon the findings of this investigation, future research can focus on enhancing the synthetic data generation process. Conditional generative models, such as variational autoencoders [35] and generative adversarial networks [36], could be trained to produce synthetic data that reflects specific referral reasons or treatment modalities, conditioned on existing basic patient information. These advanced models are adept at learning complex data distributions, leading to higher-quality and more diverse synthetic outputs, which are crucial for capturing the varied presentations within CAMHS.

Beyond static data generation, further research can also explore integrating synthetic data with process mining techniques to enhance the modeling of complex care sequences. Specifically, synthetic process mining offers a powerful avenue for generating and analyzing event logs that mimic real patient journeys, but with full privacy control [37]. This would enable the creation of highly detailed, longitudinal synthetic care pathways, allowing researchers to simulate and analyze "what if" scenarios for process optimization without touching real patient data (eg, assessing the impact of adding a new therapy step or reducing wait times on patient flow). Furthermore, synthetic process mining could also generate diverse and comprehensive training data for predictive models focused on patient trajectories and outcomes, especially for rare events or complex multistage interventions where real CAMHS data are scarce or incomplete. This capability would significantly advance our understanding of care delivery dynamics and improve operational efficiency in CAMHS, ultimately benefiting young people awaiting and receiving support.

Limitations

It is important to recognize that our dataset had limitations in terms of clinical parameters. Our analysis revealed modest predictive power in the classification model (PRAUC approximately 0.43 on real data and 0.40 on synthetic data), necessitating caution in interpreting the absolute predictive accuracy. While the primary goal of this work was to establish the comparable utility between the synthetic and real datasets—which the similar PRAUC scores confirm—the absolute performance remains constrained. This constraint is likely attributable to 2 factors: severe class imbalance in the target variable (necessitating the aggregation of the rare high-intensity category into the 33% positive class) and the inherently limited, aggregated feature set available from administrative EMR. These limitations emphasize that the model should not be used in isolation for clinical prediction but rather that the synthetic data successfully preserves the utility inherent in the original feature space.

Furthermore, our dataset had limitations in terms of clinical parameters. Although we had access to basic data such as sex, age, and clinical primary diagnosis, our dataset lacked detailed information on the specific reason for the referral and the amount of additional care (eg, medication, group sessions, art therapy, physical therapy, and sheltered workshop). These clinical details are crucial for accurately assessing and classifying service intensity. The absence of these detailed clinical parameters may have impacted the accuracy of our analyses and the general applicability of our findings to a broader clinical context. Future research will benefit from a more comprehensive dataset that includes these key clinical parameters.

Regarding fairness evaluation, Norway's highly inclusive public health care system inherently minimizes selection bias. However, we acknowledge the potential for underrepresentation of underserved minorities and rare diagnoses. These issues can be effectively addressed by augmenting the data with synthetic samples of these specific groups and applying fairness measures tailored to the analytical goals of research aimed at serving these minority populations. While the methodology for creating synthetic CAMHS data is highly generalizable across different data environments and health care systems, it is crucial to recognize that the interpretation and applicability of findings derived from such synthetic data would always need to be contextualized within the specific health care system from which the original data was sourced. This underscores the need for localized validation and careful consideration of systemic differences when applying research findings across international borders.

Conclusions

This study provides evidence that synthetic data can be generated from CAMHS EMR that achieves a favorable balance between utility and privacy. The main findings confirm that the hierarchical synthetic data model produces datasets with high statistical fidelity and comparable predictive utility to real data, while simultaneously demonstrating a reduced risk of patient reidentification. This pilot study suggests that synthetic data, developed from standard EMR clinic records, offers a method compliant with HIPAA and GDPR regulations, facilitating clinical and administrative research that was previously unattainable. This advancement creates critical new opportunities for CAMHS research to progress despite confidentiality restrictions.

Acknowledgments

We gratefully acknowledge the invaluable help of the Individualized Digital Decision Assist System project, BUP Helse Stavanger, and Helse Vest IKT, Synnøve Olset, for bringing the collaborators together and providing valuable feedback in preparing this manuscript.

The authors confirm that generative artificial intelligence tools, specifically Gemini by Google, were used for three specific and limited purposes during the preparation of this manuscript and the revision process:

XSL•FO
RenderX

1. Language polishing: minor language refinement and grammatical correction throughout the manuscript text to improve readability and flow.

2. Structural drafting: generating the initial structural outline for the author's response emails to the academic editor during the peer-review process.

3. Content generation: generating the visual representation for Figure 5, based on the data and themes analyzed in this study. The authors take full responsibility for the content, integrity, and scientific accuracy of the research presented.

Data Availability

The underlying dataset can be shared upon reasonable request after signing a data transfer agreement with the receiving institution to ensure the protection of personal information. The code underlying these analyses is available from GitHub [38].

Authors' Contributions

Data curation and analysis: MH, LRØ
Check and investigation: MH, ØN
Methodology: MH
Project administration: MH
Writing—original draft: MH
Review and editing: DP, NS, ØN, BLL, KK, RAK, LRØ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Entity-relationship diagram of the patient data schema. This diagram illustrates the relationship between the "referral" and "stay" entities, detailing the attributes and their data types (ID, datetime, numerical, categorical) for each, along with their respective primary and foreign keys.
[PNG File , 41 KB - medinform_v14i1e71819_app1.png]

Multimedia Appendix 2
Synthetic generation process. We used a constant input for the synthetic data generator HMSynthetiser (ie, a fixed subsample from the real dataset) to output 4 different synthetic CAMHS data. The similarity between each of the datasets was compared to the real training data by performing KSC/TVC, CS, and CSS metrics via the SDMetrics Python package.
[PNG File , 60 KB - medinform_v14i1e71819_app2.png]

Multimedia Appendix 3
Privacy evaluation. The Main attack (1) uses synthetic data to create a series of guesses. Anonymeter evaluates (2) these guesses by comparing them to the values present in the original dataset, to assess the effectiveness of the attack. Anonymeter requires the adversary to conduct the same attack on a control dataset (control attack), which consists of records that were not utilized in the generation of the synthetic data. Any insights the attacker gains regarding the control records from the synthetic dataset cannot be linked to privacy violations, as the design ensures that no information from the control dataset is incorporated into the synthetic dataset. By contrasting the success rates of the attack on both training records and control records, a sensitive measure of risk (3) is calculated.
[PNG File , 35 KB - medinform_v14i1e71819_app3.png]

Multimedia Appendix 4
Train synthetic test real. To evaluate model efficacy, the train synthetic test real (TSTR) approach was used. Two machine learning models were developed: one trained on real data and the other on synthetic data, both targeting patients with referral duration under 18 months to predict Child and Adolescent Mental Health Services (CAMHS) care intensity across three categories: fewer than 15 sessions, 15 or more sessions annually. The latter two categories were combined due to low occurrence of high intensity. Model performance was assessed using a 20% test subset from the real dataset, allowing comparison of synthetic and real data effectiveness. SHAP values were computed to analyze feature significance.
[PNG File , 63 KB - medinform_v14i1e71819_app4.png]

Multimedia Appendix 5

XSL•FO
RenderX

Characteristics of the training (n=6132) and test (n=1531) datasets. This table presents the distribution of demographic, clinical, and diagnostic variables across the training and test cohorts.
[PDF File (Adobe PDF File), 30 KB - medinform_v14i1e71819_app5.pdf]

References

1. Aboujaoude E. Protecting privacy to protect mental health: the new ethical imperative. J Med Ethics 2019;45(9):604-607. [doi: 10.1136/medethics-2018-105313]
2. Iflaifel M, Hall CL, Green HR, Willis A, Rennick-Egglestone S, Juszczak E, et al. Widening participation - recruitment methods in mental health randomised controlled trials: a qualitative study. BMC Med Res Methodol 2023;23(1):211 [FREE Full text] [doi: 10.1186/s12874-023-02032-1] [Medline: 37735627]
3. Berardi C, Antonini M, Jordan Z, Wechtler H, Paolucci F, Hinwood M. Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. BMC Health Serv Res 2024;24(1). [doi: 10.1186/s12913-023-10536-1] [Medline: 38408938]
4. Henderson C, Noblett J, Parke H, Clement S, Caffrey A, Gale-Grant O, et al. Mental health-related stigma in health care and mental health-care settings. Lancet Psychiatry 2014;1(6):467-482. [doi: 10.1016/S2215-0366(14)00023-6] [Medline: 26361202]
5. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, University of Zurich Digital Society Initiative (UZH-DSI) Health Community. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLoS Digit Health 2023;2(10):e0000347 [FREE Full text] [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
6. Waithira N, Mukaka M, Kestelyn E, Chotthanawathit K, Thi Phuong DN, Thanh HN, et al. Data sharing and reuse in clinical research: are we there yet? A cross-sectional study on progress, challenges and opportunities in LMICs. PLOS Glob Public Health 2024;4(11):e0003392. [doi: 10.1371/journal.pgph.0003392] [Medline: 39565766]
7. National Academies of Sciences, Engineering, and Medicine. Frontiers in Massive Data Analysis. Washington DC: The National Academies Press; 2013.
8. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. Entropy 2021;23(9):1165. [doi: 10.3390/e23091165]
9. Begoli E, Brown K, Srinivas S, Tamang S. SynthNotes: a generator framework for high-volume, high-fidelity synthetic mental health notes. 2022 Presented at: IEEE International Conference on Big Data (Big Data); 2018 Dec 10-13; Seattle, WA, USA. [doi: 10.1109/bigdata.2018.8621981]
10. Mori S, Ignat O, Lee A, Mihalcea R. Towards algorithmic fidelity: mental health representation across demographics in synthetic vs. human-generated data. arXiv. Preprint posted online on March 25, 2024 2024 [FREE Full text]
11. van Breugel B, Kyono T, Berrevoets J, van der Schaar M. Decaf: generating fair synthetic data using causally-aware generative networks. 2021 Presented at: Advances in Neural Information Processing Systems 34 (NeurIPS 2021); 2021 Dec 6-14; Online p. 10620-10632.
12. Kamiran F, Calders T. Classifying without discriminating. 2022 Presented at: 2nd International Conference on Computer, Control and Communication; 2009 Feb 17-18; Karachi, Pakistan p. 1-6. [doi: 10.1109/ic4.2009.4909197]
13. Ramachandranpillai R, Sikder MF, Bergström D, Heintz F. Bt-GAN: generating fair synthetic healthdata via bias-transforming generative adversarial networks. J Artif Intell Res 2024;79:1313-1341. [doi: 10.1613/jair.1.15317]
14. Rankin T, Gonzales M. Evaluating predictive models trained on synthetic data for mental health applications. Artif Intell Med 2022;123:102195. [doi: 10.1016/j.artmed.2022.102195]
15. Ajith M, Aycock DM, Tone EB. A deep learning approach for mental health quality prediction using functional network connectivity and assessment data. Brain Imaging Behav 2024;18:630-645. [doi: 10.1007/s11682-023-00801-4]
16. Lee K, Kim H. Machine learning for passive mental health symptom prediction: generalization across different longitudinal mobile sensing studies. PLoS One 2023;18(3):e0266516. [doi: 10.1371/journal.pone.0266516]
17. Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. 2022 Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2016 Oct 17-19; Montreal, QC p. 2016. [doi: 10.1109/dsaa.2016.49]
18. Houssou R, Augustin M, Rappos E, Bonvin V, Robert-Nicoud S. Generation and simulation of synthetic datasets with copulas. arXiv. Preprint posted online on March 30, 2022 2022. [doi: 10.48550/arXiv.2203.17250]
19. Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A unified framework for quantifying privacy risk in synthetic data. arXiv. Preprint posted online on November 18, 2022 2022. [doi: 10.48550/arXiv.2211.10459]
20. Lu Y, Chen L, Zhang Y, Shen M, Wang H, Wang X. Machine Learning for Synthetic Data Generation: A Review. arXiv. Preprint posted online on February 8, 2023 2023. [doi: 10.48550/arXiv.2302.04062]
21. Beckers T, Koekkoek B, Tiemens B, Hutschemaekers G. Measuring the intensity of mental healthcare: Development of the Mental Healthcare Intensity Scale (MHIS). BMC Psychiatry 2023;23(1):377. [doi: 10.1186/s12888-023-04870-w]
22. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019;19(1):281 [FREE Full text] [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]

XSL•FO
RenderX

23. Esteban C, Hyland S, Rätsch G. Real-Valued (MEDICAL) time series generation with recurrent conditional GANs. arXiv. Preprint posted online on June 8, 2017 2017. [doi: 10.48550/arXiv.1706.02633]
24. Kiran A, Kumar S. Synthetic data and its evaluation metrics for machine learning. In: So-In C, Londhe ND, Bhatt N, Kitsing M, editors. Information Systems for Intelligent Systems. Smart Innovation, Systems and Technologies. Singapore: Springer; 2023:433-448.
25. Richardson E, Trevizani R, Greenbaum JA, Carter H, Nielsen M, Peters B. The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns (NY) 2024;5(6):100994. [doi: 10.1016/j.patter.2024.100994] [Medline: 39005487]
26. Lundberg S, Lee S. A unified approach to interpreting model predictions. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4-9; Long Beach p. 4768-4777 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
27. Edbrooke-Childs J, Rashid A, Ritchie B, Deighton J. Predictors of amounts of child and adolescent mental health service use. Eur Child Adolesc Psychiatry 2023;32(11):2335-2342 [FREE Full text] [doi: 10.1007/s00787-022-02063-x] [Medline: 36114311]
28. McDermott MBA, Wang S, Marin N. A closer look at AUROC and AUPRC under class imbalance. 2020 Presented at: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 12-18; Online p. 6753-6763. [doi: 10.52202/079017-1400]
29. Marin J. Evaluating synthetic tabular data generated to augment small sample datasets. arXiv. Preprint posted online on November 19, 2022 2022. [doi: 10.48550/arXiv.2211.10760]
30. Benali F, Bodénès D, Labroche N, de RC. MTCopula synthetic complex data generation using copula. 2021 Presented at: 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP); 2021 Mar 23; Nicosia, Cyprus URL: https://ceur-ws.org/Vol-2840/paper8.pdf
31. Smith J, Doe A. Age and Developmental Considerations in CAMHS. J Child Psychol Psychiatry 2023;64(2):123-134. [doi: 10.1111/jcpp.12345]
32. Brown L, Green P. Gender differences in mental health diagnoses and treatment. J Adolesc Health 2024;58(3):567-578. [doi: 10.1016/j.jadohealth.2024.01.002]
33. Rocks S, Stepney M, Glogowska M, Fazel M, Tsiachristas A. Understanding and evaluating new models of Child and Adolescent Mental Health Services in South-East England: a study protocol for an observational mixed-methods study. BMJ Open 2018;8(12):e024230 [FREE Full text] [doi: 10.1136/bmjopen-2018-024230] [Medline: 30518588]
34. Improvements ahead: how humansAI might evolve together in the next decade. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/12/10/improvements-ahead-how-humans-and-ai-might-evolve-together-in-the-next-decade/ [accessed 2026-01-26]
35. Wan Z, Zhang Y, He H. Variational autoencoder based synthetic data generation for imbalanced learning. 2017 Presented at: 2017 IEEE Symposium Series on Computational Intelligence (SSCI); Honolulu, HI, USA; 2017 Nov 27-Dec 1; Honolulu. [doi: 10.1109/ssci.2017.8285168]
36. Guo J, Jin S, Wang M, Han C. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. NPJ Digit Med 2023;6(1):68. [doi: 10.1038/s41598-023-33049-7]
37. Bullward A, Aljebreen A, Coles A. Process mining and synthetic health data: reflections and lessons learnt. 2022 Presented at: 4th International Conference on Process Mining (ICPM 2022); 2022 Oct 23-28; Bozen-Bolzano p. 4. [doi: 10.1007/978-3-031-27815-0_25]
38. mounirHai/SynthCAMH. GitHub. URL: https://github.com/mounirHai/SynthCAMH [accessed 2026-12-05]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CAMHD: Child and Adolescent Mental Health Department
CAMHS: Child and Adolescent Mental Health Services
CS: CorrelationSimilarity
CSS: CardinalityShapeSimilarity
DIPS: Distrubuert Informasjons og Pasientdatasystem i Sykehuset
EMR: electronic medical record
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
ICD-10: International Statistical Classification of Diseases, Tenth Revision
KSC: Kolmogorov-Smirnov Complement
PRAUC: precision-recall area under the curve
SDG: synthetic data generator
SHAP: Shapley additive explanations
SUS: Stavanger University Hospital

XSL•FO
RenderX

TVC: Total Variation Complement

Edited by A Coristine; submitted 27.Jan.2025; peer-reviewed by W Farr, X Ding; comments to author 12.May.2025; accepted 23.Oct.2025; published 26.Feb.2026.

Please cite as:
Haizoune M, Leventhal BL, Pant D, Nytrø Ø, Koochakpour K, Koposov RA, Øhlckers LR, Skokauskas N
Balancing Privacy and Utility in Child and Adolescent Mental Health Services Research: Retrospective Cohort Study on Synthetic Data Generation
JMIR Med Inform 2026;14:e71819
URL: https://medinform.jmir.org/2026/1/e71819
doi:10.2196/71819
PMID:

©Mounir Haizoune, Bennett L Leventhal, Dipendra Pant, Øystein Nytrø, Kaban Koochakpour, Roman A Koposov, Lars Ravn Øhlckers, Norbert Skokauskas. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Model Generalizability in Medical Artificial Intelligence: Systematic Comparison of Categorical Encoding and Sampling Techniques for Imbalanced Data

Chien-wei Chuang[1,2*]; Chung-Kuan Wu[3,4,5,6*], Prof Dr; Chao-Hsin Wu[1,2]; Ben-Chang Shia[1,2], Prof Dr; Mingchih Chen[1,2], Prof Dr

[1]Graduate Institute of Business Administration, Fu Jen Catholic University, 510 Zhongzheng Rd, Xinzhuang District, New Taipei City, Taiwan

[2]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan

[3]Division of Nephrology, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan

[4]Dialysis Access Management Center, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan

[5]School of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan

[6]Division of Digital Informatics Management, Department of Digital Medicine, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Mingchih Chen, Prof Dr
Graduate Institute of Business Administration, Fu Jen Catholic University, 510 Zhongzheng Rd, Xinzhuang District, New Taipei City, Taiwan

Abstract

Background: Despite the increasing use of machine learning (ML) in clinical research, the early stages of data preparation, especially for structured clinical data, often receive limited methodological scrutiny. These datasets typically contain missing values, complex categorical variables, and imbalanced class distributions, all of which complicate downstream model development and interpretation.

Objective: This study introduces a structured preprocessing framework designed to address common challenges in medical tabular data and to assess how preprocessing choices affect the stability and portability of predictive models across settings.

Methods: We constructed a modular workflow comprising 3 components. First, preprocessing strategies included imputation for missing data, 3 types of categorical encoding (one-hot, frequency, and target), and resampling approaches for class imbalance (Synthetic Minority Over-sampling Technique [SMOTE] and Random Over Sampling Example [ROSE]). Second, 6 classification algorithms were used to evaluate performance patterns, including logistic regression (LGR), decision tree (DT), random forest, XGBoost (XGB), CatBoost (CAT), and light gradient-boosting machine (LightGBM). Third, we assessed cross-dataset portability using 2 datasets with distinct data-generating mechanisms: a registry for patients with end-stage renal disease (ESRD; n=412) and the population-based Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey. For each dataset, we independently cleaned, standardized, encoded, tuned, and evaluated models using the same predefined hyperparameter search space, without cross-dataset feature matching or pooling the area under the ROC curve (AUC) calculations; the complete pipeline was then rerun on BRFSS as an external replication.

Results: One-hot encoding in combination with ROSE yielded the most consistent performance improvements in terms of AUC (0.940) and accuracy (0.932), particularly for classifiers sensitive to class distribution. Notably, ROSE enhanced sensitivity without substantially distorting the original data structure. Feature importance rankings also contributed to model interpretability, and performance trends were largely reproducible in cross-context application.

Conclusions: Our findings suggest that preprocessing decisions often treated as ancillary play a central role in shaping model outcomes, especially in high-variance clinical datasets. The proposed framework offers a reproducible and adaptable tool for aligning data preparation with the unique demands of health care prediction tasks and may serve as a foundation for future efforts to standardize preprocessing in clinical ML workflows.

(JMIR Med Inform 2026;14:e75655) doi:10.2196/75655

KEYWORDS

machine learning; data preprocessing; clinical prediction models; medical informatics; feature engineering

XSL•FO
RenderX

Introduction

Although health care institutions generate vast volumes of health and clinical tabular data on a daily basis, these valuable resources frequently encounter significant bottlenecks when being translated into effective predictive models [1,2]. In practice, suboptimal model performance is often not attributable to the limitations of the algorithms themselves but rather to the lack of consistency and strategic planning in the data preprocessing pipeline [3,4]. Prior studies have demonstrated that various preprocessing techniques such as data imputation, normalization, and feature selection are highly sensitive in their impact on model performance [5,6], and may even introduce bias or lead to overfitting [7]. Consequently, a critical yet insufficiently addressed question arises: Can we establish a generalizable and reproducible data processing framework that assists users in selecting appropriate machine learning (ML) predictive models?

In the domain of medical informatics, ML models have been extensively used for predictive analyses of health and clinical data. However, extant literature has predominantly focused on comparing algorithmic accuracy and model performance [8,9], while offering limited systematic investigation into critical data processing steps such as data cleaning, feature encoding, and handling of class imbalance [10,11]. This oversight is particularly consequential given the inherent characteristics of medical datasets, which frequently exhibit high rates of missing values, numerous categorical variables, and severely imbalanced outcome distributions—factors that significantly influence predictive outcomes depending on the preprocessing choices made [12,13].

Despite the enormous volume of clinical and health data generated daily by health care institutions, transforming these data into effective predictive models remains challenging due to pervasive issues such as high missingness, class imbalance, and the need for robust variable encoding. Although prior studies have examined individual procedures such as comparing one-hot versus target encoding [14,15] and evaluating synthetic data augmentation techniques like the Synthetic Minority Over-sampling Technique (SMOTE) [16], systematic evaluations of the interactive effects among diverse preprocessing workflows are exceedingly scarce, particularly across heterogeneous medical datasets. Furthermore, there is a notable gap in comparative research on encoding strategies (eg, one-hot, frequency, and target encoding) and imbalance correction methods (eg, SMOTE and Random Over Sampling Example [ROSE]), which hampers the establishment of best practices in clinical ML.

This study proposes a comprehensive, reproducible framework specifically for medical tabular data. The framework is built upon three core components: (1) the Data Processing Strategy Layer, which systematically evaluates essential preprocessing techniques, including missing value imputation, variable encoding, and class imbalance correction; (2) the Model Selection and Optimization Layer, which ensures compatibility with a diverse range of supervised learning algorithms; and (3) Cross-Dataset Validation, which tests the framework's transferability and consistency on 2 highly heterogeneous real-world clinical datasets. This design not only streamlines the preprocessing pipeline but also minimizes overlap with subsequent methodological details.

The primary contribution of this work lies in developing and empirically validating an end-to-end data processing framework that transcends the limitations of single-model, single-dataset analyses. By shifting the focus from solely model-centric performance metrics to a holistic methodological architecture, our approach provides both data scientists and clinical researchers with a modular and standardized workflow. This framework is expected to bridge the gap between algorithm development and clinical application, offering robust empirical evidence and actionable guidance for advancing predictive modeling in health care.

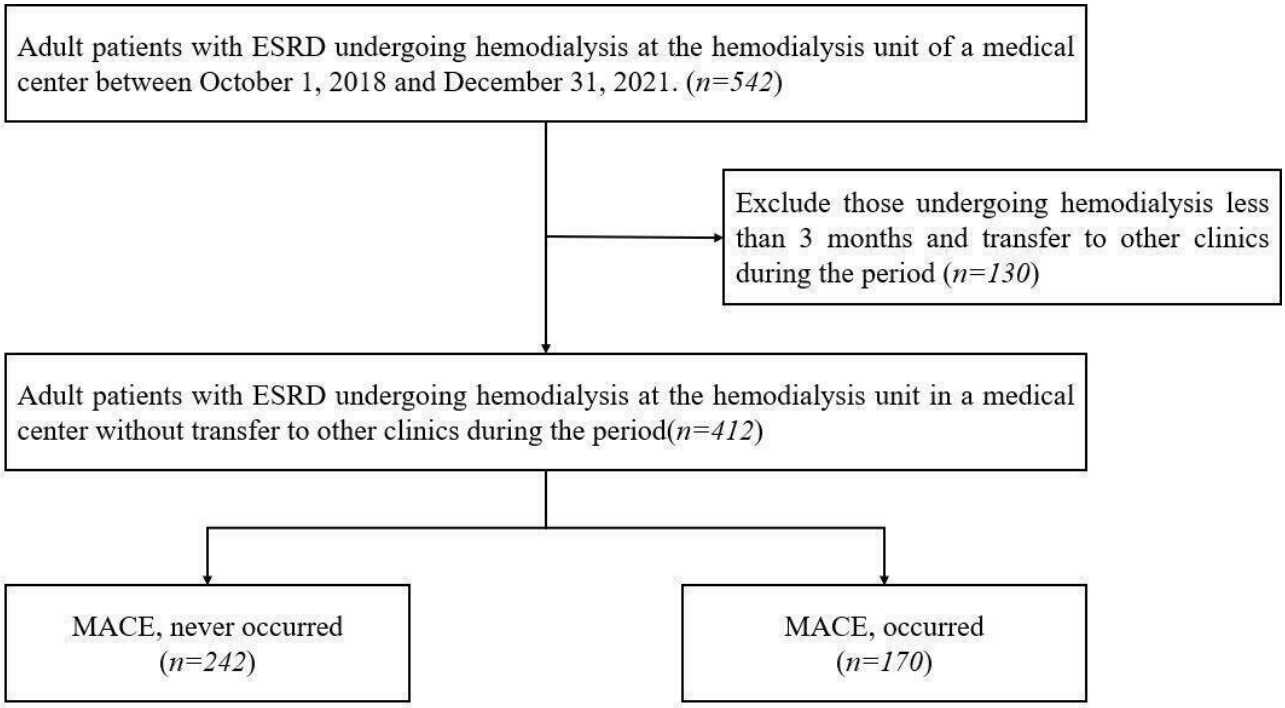
Methods

Data Collection

Initially, we enrolled 542 adult patients with ESRD undergoing hemodialysis at the hemodialysis unit of a medical center between October 1, 2018, and December 31, 2021. Patients who had received hemodialysis for less than 3 months or who were transferred to other clinics during the study period were excluded. After these exclusions, 412 adult patients with ESRD undergoing chronic hemodialysis without transfer remained eligible for analysis. Among them, 242 patients had no occurrence of major adverse cardiovascular events (MACEs), while 170 experienced at least one MACE. The primary objective of this study was to determine the incidence of MACEs in this population. A flowchart of the study participants is presented in Figure 1. The detailed baseline demographic and clinical characteristics of the cohort are summarized in Table 1.

XSL•FO
RenderX

Figure 1. Flowchart of study population selection and major adverse cardiovascular event (MACE) grouping. ESRD: end-stage renal disease; MACE: major adverse cardiovascular event;



XSL•FO
RenderX

Table . Initial demographic and clinical profiles of the research cohort.

Variables	Overall (n=412)	MACE[a]		*P* value
		Never occurred (n=242)	Occurred (n=170)	
Age, mean (SD)	69.19 (12.14)	67.96 (12.59)	70.94 (11.29)	.01
Sex (female), n (%)	192 (46.6)	122 (50.4)	70 (41.2)	.08
AVG (Arteriovenous graft), n (%)	57 (13.8)	31 (12.8)	26 (15.3)	.57
AV cal, n (%)[b]	196 (62.6)	106 (56.1)	90 (72.6)	.005
AR[c], n (%)				.14
None	178 (43.2)	115 (47.5)	63 (37.1)	
Negligible or mild	125 (30.3)	68 (28.1)	57 (33.5)	
Moderate	8 (1.9)	3 (1.2)	5 (2.9)	
AS, n (%)[d]				.03
-	101 (24.5)	56 (23.1)	45 (26.5)	
None	277 (67.2)	172 (71.1)	105 (61.8)	
Negligible or mild	22 (5.3)	12 (5.0)	10 (5.9)	
Moderate	11 (2.7)	2 (0.8)	9 (5.3)	
Severe	1 (0.2)	0 (0.0)	1 (0.6)	
LVH type, n (%)[e]				.02
1	131 (43.8)	72 (39.8)	59 (50.0)	
2	106 (35.5)	63 (34.8)	43 (36.4)	
3	28 (9.4)	24 (13.3)	4 (3.4)	
4	34 (11.4)	22 (12.2)	12 (10.2)	
Comorbidities, n (%)				
DM[f]	198 (48.1)	99 (40.9)	99 (58.2)	.001
Dyslipid	220 (53.4)	130 (53.7)	90 (52.9)	.96
PAOD[g]	111 (26.9)	53 (21.9)	58 (34.1)	.01
Medication, n (%)				
Insulin	85 (20.6)	33 (13.6)	52 (30.6)	<.001
Statin	137 (33.3)	73 (30.2)	64 (37.6)	.14
Antiplatelet	199 (48.3)	82 (33.9)	117 (68.8)	<.001
Calcitriol	173 (42.0)	104 (43.0)	69 (40.6)	.70
No of hypotension episodes, mean (SD)	5.41 (3.26)	5.46 (3.21)	5.34 (3.33)	.71
CXR_AoAC, n (%)[h]				<.001
0	120 (31.4)	90 (40.5)	30 (18.8)	
1	83 (21.7)	49 (22.1)	34 (21.2)	
2	107 (28.0)	49 (22.1)	58 (36.2)	
3	72 (18.8)	34 (15.3)	38 (23.8)	

[a]MACE: major adverse cardiovascular event.

[b]AV cal: aortic valve calculation

[c]AR: aortic regurgitation.

[d]AS: aortic stenosis.

XSL•FO
RenderX

[e]LVH: left ventricular hypertrophy.

[f]DM: diabetes mellitus.

[g]PAOD: peripheral arterial occlusion disease.

[h]CXR_AoAC: chest X-ray for aortic arch calcification.

This study specifically targeted the unique clinical needs and complexities of patients with ESRD by collecting 84 variables. These variables were meticulously selected based on their significant impact on clinical outcomes in patients with ESRD. Demographic data, such as age and gender, were included, and dialysis vintage reflected the duration and history of each patient's dialysis treatment. Additionally, we detailed the anatomical and functional characteristics of the aortic and mitral valves, which are crucial for understanding cardiovascular complications in patients with ESRD. Specifically, we included parameters such as types of arteriovenous access (AVA), mitral valve calcification (MV calc), aortic regurgitation (AR), aortic stenosis (AS), mitral regurgitation (MR), and mitral stenosis (MS).

In terms of cardiovascular health, this study placed particular emphasis on the grading and types of left ventricular hypertrophy (LVH) [17] and the ejection fraction (EF) of the heart, which are critical indicators of cardiovascular health in patients with ESRD. Given the high prevalence and impact of comorbidities such as diabetes mellitus, hypertension, dyslipidemia, coronary artery disease, heart failure, chronic obstructive pulmonary disease, liver cirrhosis, malignancy, arrhythmia, and a history of amputation among patients with ESRD, we included these comorbidities in our analysis.

To better meet the clinical needs of patients with ESRD, we expanded the range of biochemical laboratory data. This included comprehensive assessments of total protein, albumin, liver enzymes (aspartate aminotransferase and alanine aminotransferase), alkaline phosphatase [18], total bilirubin, lipid profiles, glucose levels, complete blood count, iron studies, aluminum levels, postdialysis weight, uric acid, and key electrolytes. To address the specific needs of patients with ESRD, we further measured calcium and phosphate metabolism indicators, such as calcium and phosphate levels, urea kinetics (Kt/V), parathyroid hormone levels, and the calcium-phosphate product, to more effectively manage mineral and bone disorders in these patients.

The medication history was also thoroughly documented, particularly focusing on drugs commonly used in the management of ESRD, such as phosphate binders, calcitriol, and other treatments relevant to the patients' condition.

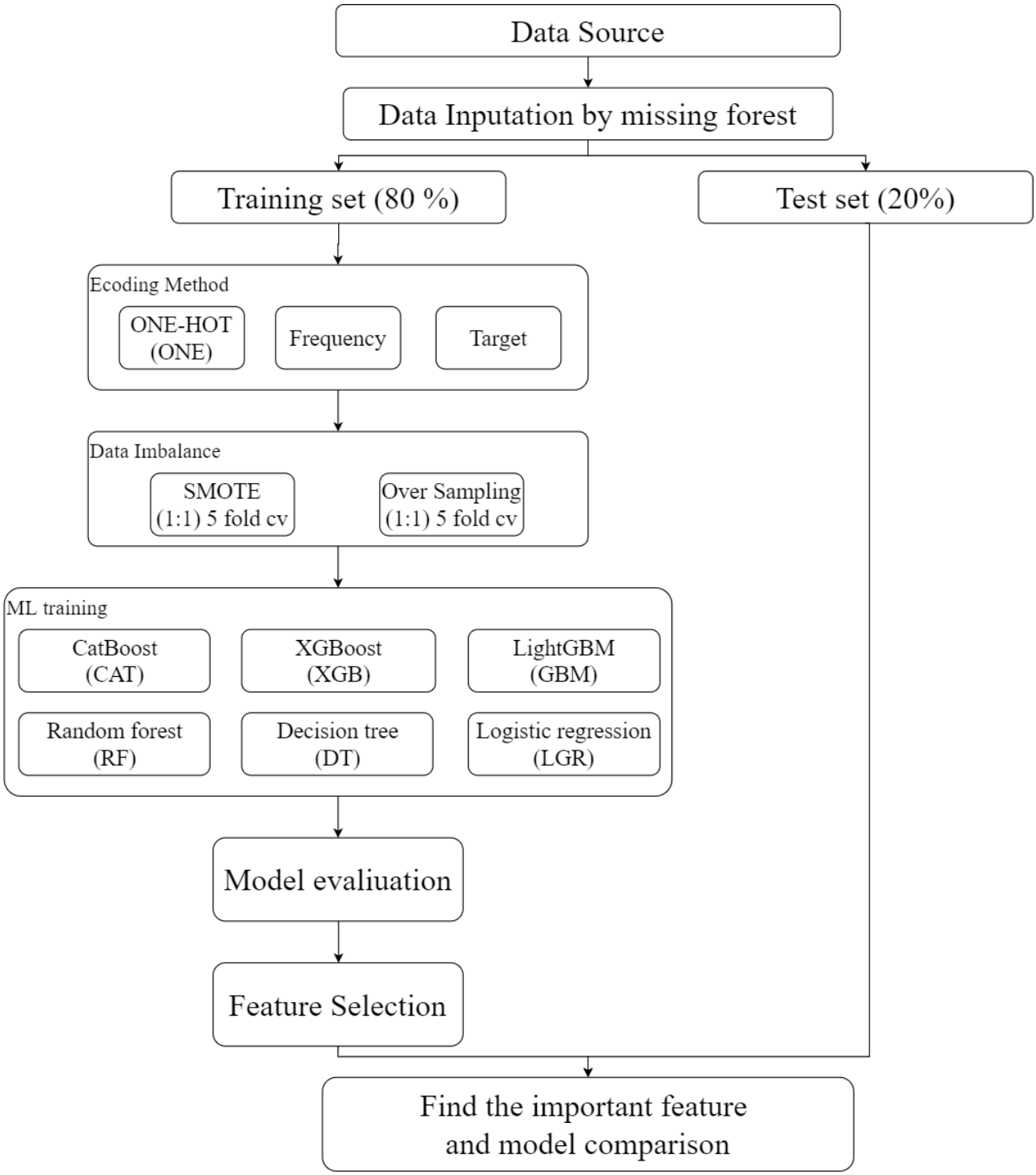
This study included 412 patients. The mean age was 69.19 years, with older patients more likely to experience MACE (70.94 years vs 67.96 years, P=.01). Females comprised 46.6% of the cohort, with a lower proportion in the MACE group, although this difference was not statistically significant (P=.08). Aortic valve calcification was more prevalent in the MACE group (72.6% vs 56.1%, P=.005), and AS as well as certain types of LVH were also associated with the occurrence of MACE. Diabetes mellitus (DM) and peripheral arterial occlusive disease (PAOD) were more common among patients who experienced MACE (DM: 58.2% vs 40.9%; P=.001; PAOD: 34.1% vs 21.9%; P=.008). In terms of medication use, a higher proportion of patients in the MACE group were on insulin (P<.001) and antiplatelet drugs (P<.001). These results provide an overview of the baseline characteristics of the patients and offer important references for improving the accuracy of predictive models.

Data Preparation

The data preprocessing methodology in this study is organized into 2 primary components: variable encoding and data balancing. The overall analytical workflow and preprocessing framework are illustrated in Figure 2. This integrated approach is designed to enhance the quality and representativeness of clinical datasets, thereby improving the robustness and generalizability of subsequent predictive models.

XSL•FO
RenderX

Figure 2. Research analysis process framework diagram. Preprocessing components are fit on training folds only and then applied to validation folds; ROSE or SMOTE are applied to training folds only to prevent leakage.



Data Imputation

Missing data are a common challenge in clinical datasets and can significantly compromise the validity of statistical inferences if not appropriately addressed. In this study, we used a nonparametric multiple imputation strategy using the *missForest* package in R (R Core Team) [19]. This method uses random forest models to iteratively predict missing values based on observed data, effectively capturing complex nonlinear associations and interactions between variables. Unlike simpler imputation techniques such as mean substitution or k-nearest neighbors, missForest has been shown to yield more accurate and less biased estimates in both continuous and categorical variables, particularly in mixed type medical data [15,20]. This approach offers a robust and flexible foundation for downstream ML analysis while preserving the integrity of the original dataset.

Compared to SMOTE, which generates new samples through linear interpolation between neighboring minority class points, ROSE applies a smoothed bootstrap technique. It estimates a kernel density around each minority instance and samples new

XSL•FO
RenderX

points from this local distribution. This approach allows ROSE to preserve the original variance and nonlinear structure of the data, making it particularly effective for clinical datasets where preserving subtle distributional patterns is important.

Imputation models were fit on the training fold only and then applied to the corresponding validation fold within each split. No statistics from the validation fold were used to fit the imputation model. This fold-wise procedure was repeated across all folds to prevent information leakage.

Variable Encoding and Expansion

In clinical datasets, categorical variables are abundant and require conversion into numerical formats to be effectively used in statistical and ML models. Three encoding methods were implemented to address this challenge, each offering a unique balance between preserving information and managing computational complexity. Although all 3 approaches aim to transform qualitative data into quantitative representations, they differ in their operational mechanisms and associated trade-offs.

One-hot encoding converts each categorical variable into a series of binary indicators, where each category is represented by an individual binary feature. This method maintains the inherent nonordinal nature of the original variable but can lead to a substantial increase in dimensionality, particularly when the variable in question has many unique categories. In contrast, target encoding substitutes each category with a statistical summary such as the mean, weighted mean, or smoothed mean of the target variable computed from the training dataset. This not only reduces dimensionality but also encapsulates the predictive relationship between the categorical feature and the outcome variable, although it necessitates careful handling to avoid target leakage [14]. A third approach, frequency encoding, assigns to each category a numerical value based on its relative frequency within the dataset. This method is highly efficient in reducing computational burden and memory usage, as it compresses categorical information into a single continuous variable without imposing any artificial order [21].

To combine the complementary strengths of these methods, a unified encoding strategy was adopted. Specifically, the one_hot function from the *mltools* package in R was applied to perform one-hot encoding, which expanded the original set of 83 variables to 113. This expansion increased the granularity of the dataset, facilitating a more detailed representation of the clinical phenomena under this research. The inclusion of target and frequency encoding provided an additional layer of comparison, enabling an evaluation of their relative performance under conditions of significant missingness and class imbalance. Prior work has demonstrated that, particularly in datasets with high proportions of missing data and imbalanced classes, preprocessing methods based on one-hot encoding can significantly enhance both accuracy and robustness in classification tasks [15].

All categorical encoders were fit on the training fold only and then applied to its validation fold. Target encoding used an out-of-fold smoothing scheme: for each fold, category means and smoothing weights were computed from the training fold and then mapped to the validation fold. No target information from the validation fold was used to compute encodings.

Data Imbalance

Clinical datasets often exhibit imbalanced class distributions, where the minority class, despite its clinical significance, is underrepresented. Such imbalance can lead to biased models that disproportionately favor the majority class. To counteract this, 2 complementary strategies, Random Over-Sampling (ROS) and the SMOTE, were incorporated, along with an evaluation of the ROSE method [22].

The ROSE method uses a bootstrap resampling framework augmented by kernel density estimation to generate synthetic samples for the minority class. This approach avoids the pitfalls associated with simply duplicating minority samples, offering a more nuanced correction of the class distribution. Menardi and Torelli [23] provide an extensive discussion of these resampling techniques, emphasizing their utility in balancing datasets for binary classification tasks. On the other hand, SMOTE, as introduced by Chawla et al [24], synthesizes new minority class instances by interpolating between existing samples. This technique enriches the minority class by generating additional, diverse examples, thereby improving the model's ability to capture the characteristics of rare events. Empirical studies have shown that SMOTE can lead to marked improvements in performance metrics, such as the area under the ROC curve (AUC), though it may overgeneralize when faced with extreme imbalance.

Comparative evaluations in the literature further underscore the respective merits of these techniques. For instance, Kamalov et al [25] found that ROSE, despite its relative simplicity, remains a stable and computationally efficient solution, especially in multi-label contexts. Similarly, investigations by Gnip et al [26] and Nguyen et al [27] have confirmed that both ROSE and SMOTE are effective in mitigating the adverse effects of class imbalance, with the optimal choice being contingent upon the specific characteristics of the dataset and the available computational resources.

Integration of Preprocessing Components

The overall preprocessing workflow was designed to integrate the 2 components variable encoding and data balancing, into a coherent sequence. Initially, missing values were imputed using the random forest approach, ensuring that the dataset was complete and reliable. This was followed by the transformation of categorical variables via the combined encoding strategy, which not only translated qualitative data into numerical features but also expanded the feature set to enhance data granularity. Finally, class imbalance was addressed through the application of ROSE and SMOTE, thereby ensuring that the resulting dataset was well-suited for the development of predictive models.

Each preprocessing step was carefully implemented so that subsequent operations built upon an increasingly refined version of the dataset. The robust imputation stage preserved the original data's distributional properties, while the encoding procedures facilitated the construction of a rich, multidimensional feature space. The balancing techniques further adjusted the dataset to

XSL•FO
RenderX

prevent bias toward the majority class, enabling the models to more effectively capture the subtleties of clinically significant, albeit rare, events.

Methodological Rationale and Literature Justification

The methodological choices articulated above are firmly grounded in this body of literature. The use of a random forest–based imputation method is well-supported by studies demonstrating its efficacy in preserving data structure and ensuring the validity of statistical inferences [15,19]. Similarly, the comparative evaluation of encoding methods draws on prior research that highlights the trade-offs between dimensionality, computational efficiency, and the risk of target leakage [14,21]. Moreover, the advantages of one-hot encoding in contexts marked by high missingness and class imbalance have been substantiated by empirical investigations [15]. The incorporation of data balancing strategies, including ROSE and SMOTE, is equally well-documented, with seminal works establishing their effectiveness in correcting class imbalances [23,24], and more recent studies further validating these methods [25-27].

The data preprocessing methodology presented herein represents a rigorous and systematic approach to overcoming the multifaceted challenges inherent in clinical datasets. By sequentially addressing missing data, encoding categorical variables into a more informative numerical format, and correcting class imbalances, the workflow transforms raw clinical data into a format that is both analytically robust and amenable to predictive modeling. This integrated preprocessing pipeline is instrumental in bridging the gap between the complexities of clinical data and the demands of advanced statistical and ML techniques, ultimately contributing to the development of models that are both reliable and clinically pertinent.

Class rebalancing was applied only to the training portion of each fold. Validation folds preserved the original class distribution. For transparency and reproducibility, we report resampling parameters in Table S4 in Multimedia Appendix 1, including SMOTE and ROSE sampling ratios and the random seed used.

Data Segmentation

k-fold cross-validation is a commonly used resampling technique for evaluating the performance of ML models. Specifically, the dataset is divided into k subsets (folds), and in each iteration, one subset is used as the validation set while the remaining k-1 subsets are used as the training set. This process is repeated k times, with a different subset used for validation each time. This method effectively prevents overfitting and provides a more robust evaluation of the model. It is particularly useful for estimating the generalization error in small datasets [28].

In this study, we used 5-fold cross-validation to obtain an accurate estimate of model performance through multiple splits and evaluations while reducing bias caused by data partitioning.

Robust Model Engineering: Tuning, Processing, and Risk Mitigation

All preprocessing and modeling steps, including cleaning, standardization, encoding, resampling, model fitting, and hyperparameter tuning, were executed separately within each dataset. No features, encoders, parameters, or statistics were transferred across datasets. For each model, hyperparameters were tuned by grid search with 5-fold cross-validation on the training set. A single predefined hyperparameter search space and evaluation criterion, detailed in Table S1 in Multimedia Appendix 1, was used for both ESRD and BRFSS. Grid search and model selection were rerun independently in each dataset so that optimal hyperparameters were learned within ESRD and BRFSS separately and were never reused across datasets.

Model Building and Validation-Phase 1

In Phase 1, we compared alternative encoding and class-imbalance handling strategies for predicting MACE. Traditional logistic regression and 6 ML models were evaluated, including decision trees [29], random forests [30,31], XGBoost [32], CatBoost [33], and LightGBM [34]. All models were trained and evaluated using 5-fold cross-validation [35] within each dataset, following the generic engineering framework described in the previous subsection, that is, grid search hyperparameter tuning with the shared search space in Table S1 in Multimedia Appendix 1 and dataset-specific model selection.

The workflow, including preprocessing, resampling, and hyperparameter tuning, was then rerun independently on the BRFSS 2015 dataset to assess cross-dataset portability of the pipeline rather than to externally validate a single ESRD-trained model. All training was executed on a workstation equipped with an Intel Core i9 10th-generation CPU (3.3 GHz) and 64 GB RAM, using the CPU only. Average runtimes for each pipeline are reported in Tables 2 and 3 to help readers gauge computational cost. In this phase, multiple performance metrics were used, including accuracy, sensitivity, specificity, precision, F_1-score, and AUC. The mean and SD of these metrics were calculated across folds to assess predictive accuracy, robustness, and consistency. The overall workflow is summarized in algorithm S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Evaluation metrics of 5-fold cross-validation using all encoding methods and imbalanced data processing methods.

Encoding, imbalance, and machine learning method	Accuracy		F_1-score		AUC[a]		Runtime (sec)
	Mean (SD)	95% CI	Mean (SD)	95% CI	Mean (SD)	95% CI	
One-hot encoding							
SMOTE[b]							
RF[c]	0.711 (0.048)	0.680 - 0.752	0.660 (0.062)	0.611 - 0.739	0.708 (0.050)	0.658 - 0.757	23.9
CAT[d]	0.706 (0.062)	0.701 - 0.751	0.689 (0.038)	0.653 - 0.726	0.739 (0.044)	0.695 - 0.781	124.0
LightGBM[e]	0.702 (0.043)	0.675 - 0.723	0.694 (0.041)	0.666 - 0.705	0.713 (0.032)	0.670 - 0.764	17.4
ROSE[f]							
RF	0.932 (0.106)[g]	0.783 - 1.000	0.917 (0.131)[g]	0.752 - 1.000	0.938 (0.117)[g]	0.793 - 1.000	23.9[g]
CAT	0.918 (0.094)[g]	0.795 - 1.000	0.890 (0.131)[g]	0.730 - 0.995	0.932 (0.103)[g]	0.788 - 1.000	401.0[g]
LightGBM	0.932 (0.112)[g]	0.759 - 1.000	0.918 (0.137)[g]	0.754 - 1.000	0.940 (0.116)[g]	0.794 - 1.000	19.7[g]
Frequency encoding							
SMOTE							
LGR	0.604 (0.094)	0.433 - 0.660	0.502 (0.052)	0.188 - 0.518	0.565 (0.058)	0.493 - 0.582	0.110
CAT	0.619 (0.079)	0.479 - 0.646	0.600 (0.033)	0.375 - 0.627	0.613 (0.042)	0.498 - 0.692	106.0
LightGBM	0.619 (0.102)	[0.321 - 0.630]	0.490 (0.086)	[0.304 - 0.580]	0.602 (0.069)	0.515 - 0.663	15.0
ROSE							
RF	0.896 (0.158)	0.666 - 1.000	0.867 (0.206)	0.553 - 1.000	0.899 (0.194)	0.623 - 1.000	19.7
CAT	0.886 (0.127)	0.678 - 1.000	0.864 (0.150)	0.694 - 1.000	0.913 (0.140)	0.690 - 1.000	329.0
LightGBM	0.870 (0.145)	0.708 - 1.000	0.825 (0.213)	0.576 - 1.000	0.886 (0.175)	0.707 - 1.000	19.9
Target encoding							
SMOTE							
LGR	0.621 (0.044)	0.324 - 0.684	0.507 (0.102)	0.213 - 0.567	0.576 (0.055)	0.497 - 0.587	0.110
CAT	0.648 (0.038)	0.477 - 0.649	0.600 (0.053)	0.266 - 0.662	0.624 (0.030)	0.492 - 0.687	106.0
LightGBM	0.646 (0.031)	0.464 - 0.638	0.515 (0.182)	0.242 - 0.708	0.605 (0.058)	0.512 - 0.671	14.8
ROSE							
RF	0.882 (0.184)	0.565 - 1.000	0.866 (0.202)	0.553 - 1.000	0.893 (0.204)	0.615 - 1.000	19.4
CAT	0.876 (0.131)	0.641 - 1.000	0.828 (0.214)	0.651 - 1.000	0.900 (0.159)	0.638 - 1.000	285.0
LightGBM	0.891 (0.155)	0.669 - 1.000	0.850 (0.225)	0.602 - 1.000	0.889 (0.178)	0.634 - 1.000	19.5

[a]AUC: area under the ROC curve.

XSL•FO
RenderX

[b]SMOTE: Synthetic Minority Over-sampling Technique.

[c]RF: random forest.

[d]CAT: CatBoost.

[e]LightGBM: light gradient boosting machine.

[f]ROSE: Random Over Sampling Examples

[g]This method has the highest model accuracy and a small standard deviation, indicating that it has a relatively good ability to make judgments

Table . Evaluation metrics of remodeling with the top 15 most important variables, using one-hot encoding and SMOTE or ROSE.

ETL and machine learning methods	Accuracy		F_1-score		AUC[a]		Runtime (sec)
	mean (SD)	95% CI	mean (SD)	95% CI	mean (SD)	95% CI	
SMOTE[b]							
LGR[c]	0.723 (0.045)	0.668 - 0.779	0.680 (0.050)	0.618 - 0.743	0.728 (0.044)	0.673 - 0.783	0.08
DT[d]	0.636 (0.035)	0.592 - 0.680	0.547 (0.073)	0.457 - 0.637	0.622 (0.044)	0.567 - 0.676	2.97
RF[e]	0.699 (0.027)	0.665 - 0.733	0.675 (0.027)	0.642 - 0.709	0.711 (0.020)	0.687 - 0.736	9.79
XGB[f]	0.658 (0.021)	0.632 - 0.684	0.630 (0.052)	0.566 - 0.694	0.670 (0.024)	0.641 - 0.700	2.59
CAT[g]	0.699 (0.044)	0.644 - 0.754	0.703 (0.027)	0.669 - 0.737	0.737 (0.021)	0.711 - 0.763	93.8
LightGBM[h]	0.709 (0.030)	0.671 - 0.747	0.682 (0.045)	0.626 - 0.738	0.717 (0.034)	0.675 - 0.760	14.4
ROSE[i]							
LGR	0.738 (0.030)	0.700 - 0.776	0.699 (0.050)	0.636 - 0.762	0.755 (0.050)	0.694 - 0.817	0.08
DT	0.755 (0.089)	0.644 - 0.866	0.703 (0.126)	0.547 - 0.860	0.750 (0.100)	0.626 - 0.875	2.96
RF	0.891 (0.131)	0.728 - 1.000	0.879 (0.133)	0.713 - 1.000	0.915 (0.133)[j]	0.750 - 1.000	9.05
XGB	0.913 (0.122)[j]	0.762 - 1.000	0.899 (0.138)[j]	0.728 - 1.000	0.906 (0.137)	0.736 - 1.000	3.28
CAT	0.862 (0.073)	0.771 - 0.952	0.831 (0.092)	0.717 - 0.945	0.896 (0.092)	0.782 - 1.000	270[j]
LightGBM	0.906 (0.115)[j]	0.763 - 1.000	0.892 (0.119)[j]	0.745 - 1.000	0.925 (0.112)[j]	0.786 - 1.000	19.2

[a]AUC: area under the ROC curve.

[b]SMOTE: Synthetic Minority Over-sampling Technique.

[c]LGR: logistic regression.

[d]DT: decision tree.

[e]RF: random forest.

[f]XGB: extreme gradient boosting.

[g]CAT: CatBoost

[h]LightGBM: light gradient boosting machine

[i]ROSE: random over sampling example.

[j]This method's model results have the highest AUC and a small standard deviation, indicating a relatively good ability to make judgments.

Performance Estimation and Statistical Testing

Within each dataset, model performance was summarized from 5-fold cross-validation. For each metric (AUC, accuracy, and F_1-score), we calculated 95% CIs from the 5-fold–specific estimates using a 2-sided Student t interval. This procedure corresponds to applying the *t.test* function in R to the vector of fold-level values. Because these metrics are bounded between 0 and 1, upper limits were truncated at 1.0 when necessary, and models with zero variance across folds yield a point interval at the mean.

For pairwise AUC comparisons within the same dataset, we applied the nonparametric DeLong test to out-of-fold predictions pooled across the 5 validation folds, using the *roc.test* function from the *pROC* package. We report 2-sided *P* values; all tests are conducted within the dataset, and we do not compute pooled AUCs or conduct between-dataset hypothesis tests.

Model Refinement and Variable Selection-Phase 2

In the second phase, the AUC [36] was used to evaluate the models from the first phase. For the optimal data processing method and the corresponding best-performing ML model, a detailed scoring and ranking of variables was conducted based on their importance and contribution to the model's predictive performance.

XSL•FO
RenderX

The variables were converted into percentile rankings, with the most important variable assigned a score of 100, while the least important variable was assigned a score of 0. Using the 5-fold cross-validation approach, the scores from 5 iterations were summed and ranked, with the highest possible score being 500. Based on this scoring system, the top 15 highest-ranked variables were selected for remodeling, allowing for a more focused and in-depth analysis of their impact on the outcomes.

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board of Shin-Kong Wu Ho-Su Memorial Hospital (IRB No. 20231101R; approval date: December 14, 2023). The requirement for informed consent was waived by the ethics committee due to the retrospective nature of the study and the use of deidentified data. All procedures were conducted in accordance with the ethical standards of the responsible institutional and national committees on human experimentation and with the principles of the Declaration of Helsinki. The privacy and confidentiality of all participants were strictly protected throughout the study, and no personally identifiable information was disclosed. No compensation was provided to participants.

Results

Phase 1

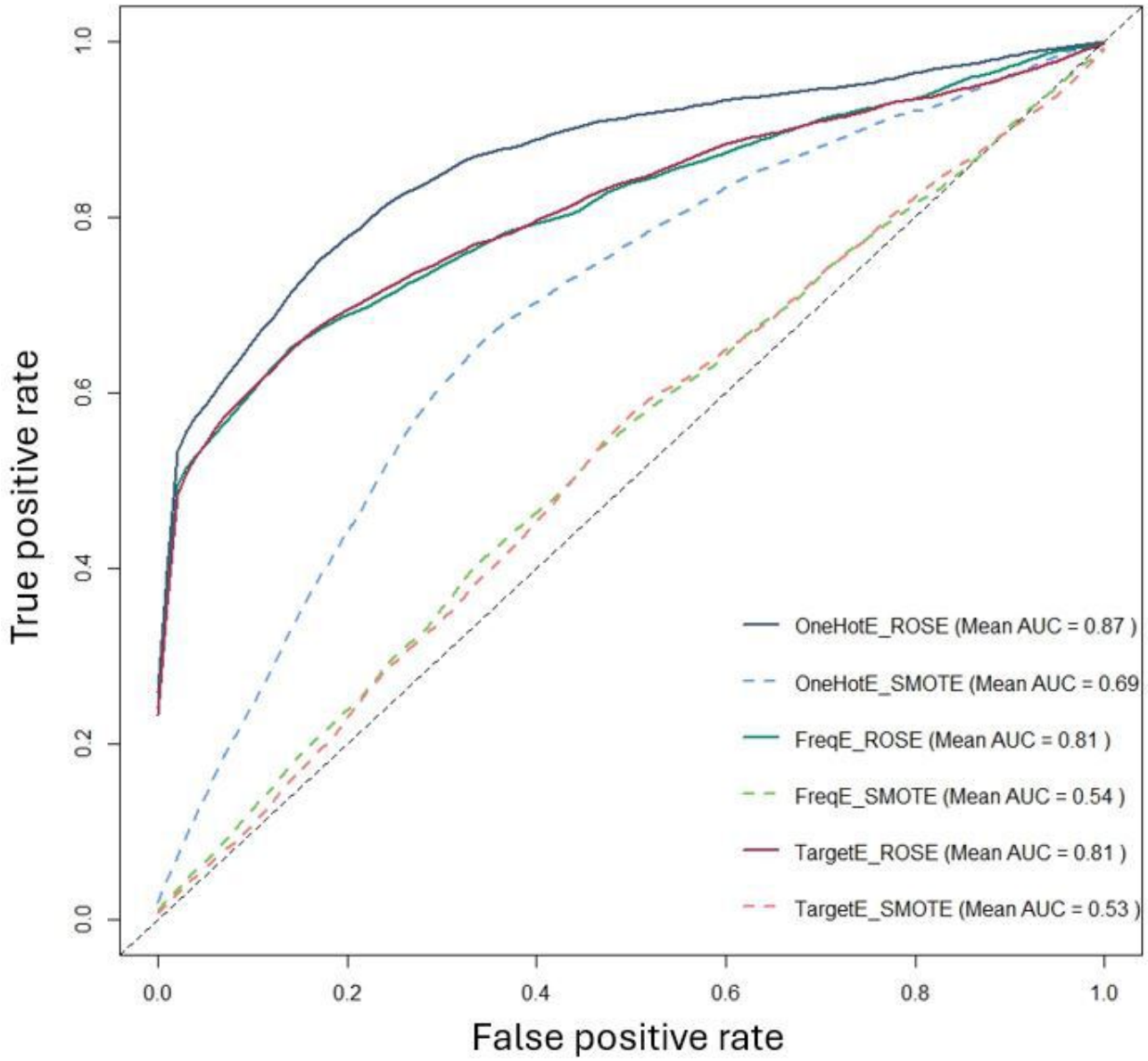
In this phase, we systematically evaluated the predictive performance of multiple ML models, including logistic regression, decision trees, random forests, XGBoost, CatBoost, and LightGBM, using 5-fold cross-validation. The primary objective was to determine the most effective combination of encoding methods (one-hot Encoding, frequency encoding, and target encoding) and data imbalance handling techniques (ROSE and SMOTE) for predicting MACE.

Models were evaluated by accuracy, F_1-score, and AUC with 5-fold cross-validation, and Table 2 reports the fold mean, SD, and t-distribution–based 95% CIs for each combination. Among all pipelines, the OneHotE_ROSE–LightGBM model achieved the best overall performance, with mean accuracy of 0.932 (SD 0.112; 95% CI 0.759 - 1.000), F_1-score of 0.918 (SD 0.137; 95% CI 0.754 - 1.000), and AUC of 0.940 (SD 0.116; 95% CI 0.794 - 1.000). Frequency and target encoding under ROSE also performed strongly (AUC 0.913 and 0.900, respectively), but at a slightly lower level than one-hot encoding in the same setting.

To formally assess differences between preprocessing strategies, we performed pairwise DeLong tests on out-of-fold AUCs, with results summarized in Table S2 in Multimedia Appendix 1. Most ROSE pipelines showed significantly higher AUC than their SMOTE counterparts (all $P \leq .002$), and OneHotE_ROSE in particular was markedly superior to OneHotE_SMOTE. Within the ROSE group, OneHotE_ROSE, FreqE_ROSE, and TargetE_ROSE did not differ significantly from each other ($P \geq .67$), indicating a top-performing cluster. Together with the averaged ROC curves in Figures 3-5, these results support selecting one-hot encoding with ROSE as the primary preprocessing strategy for Phase 2.

XSL•FO
RenderX

Figure 3. Comparison of the average area under the ROC curve (AUC) of different data models. AUC: area under the ROC curve; ROSE: Random Over Sampling Example; SMOTE: Synthetic Minority Over-sampling Technique.

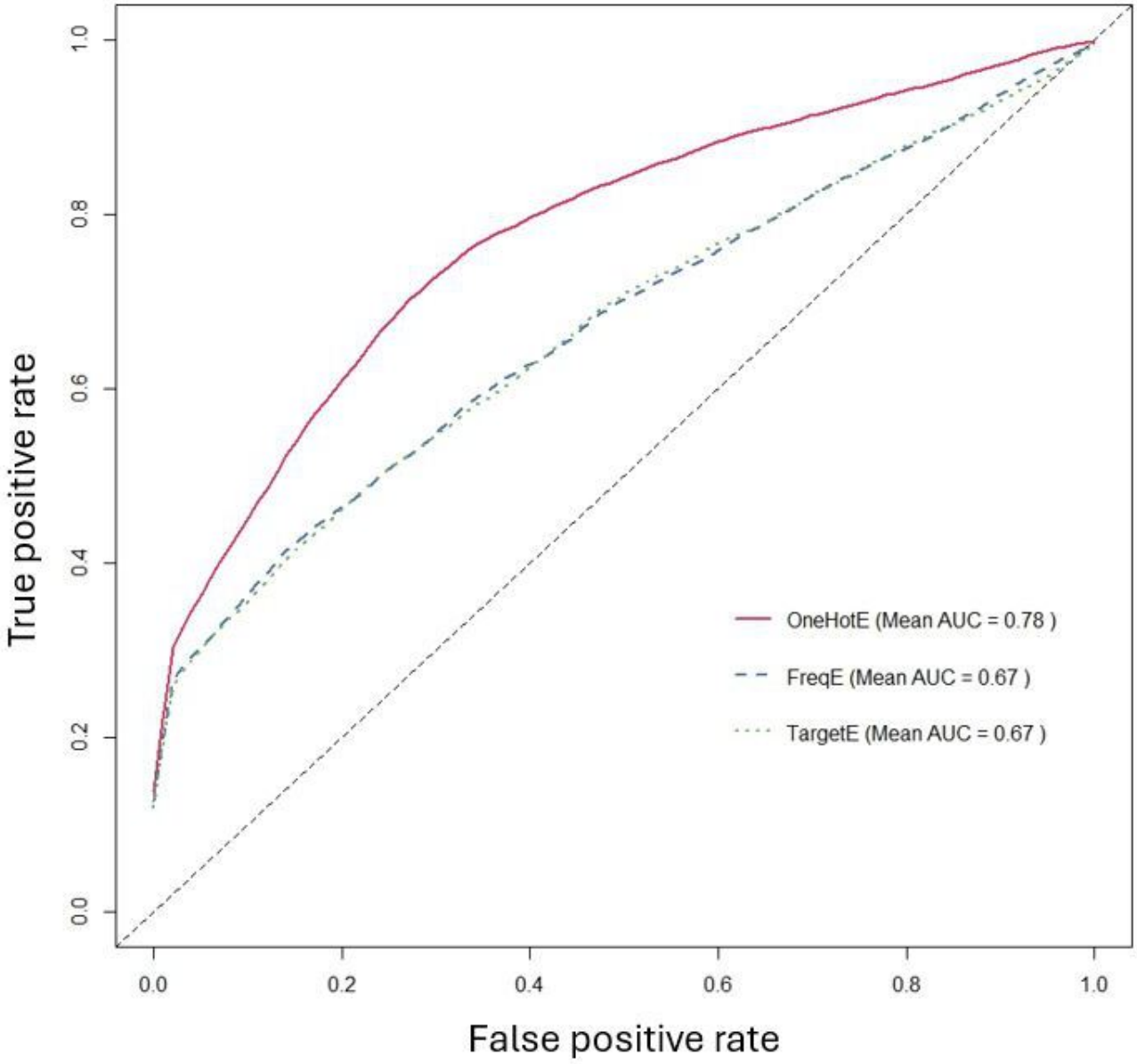


A direct comparison of encoding methods in Figure 4 further confirmed that one-hot encoding was the most effective approach, achieving the highest average AUC (0.78), outperforming both frequency encoding and target encoding (both at 0.67). Meanwhile, Figure 5 illustrates that ROSE significantly outperformed SMOTE, achieving an average AUC of 0.83, compared to 0.59 for SMOTE. This result reinforces the conclusion that ROSE generates more representative synthetic samples, preserves the original data distribution more effectively, and enhances model generalization.

Our findings strongly suggest that one-hot encoding combined with ROSE is the most effective preprocessing strategy for this predictive task. This combination preserves categorical feature integrity while mitigating class imbalance, leading to the most robust and accurate predictive models. In contrast, SMOTE not only struggled to enhance model performance but, in some cases, even contributed to its degradation. These insights guided the selection of the optimal model for further feature importance analysis in Phase 2.

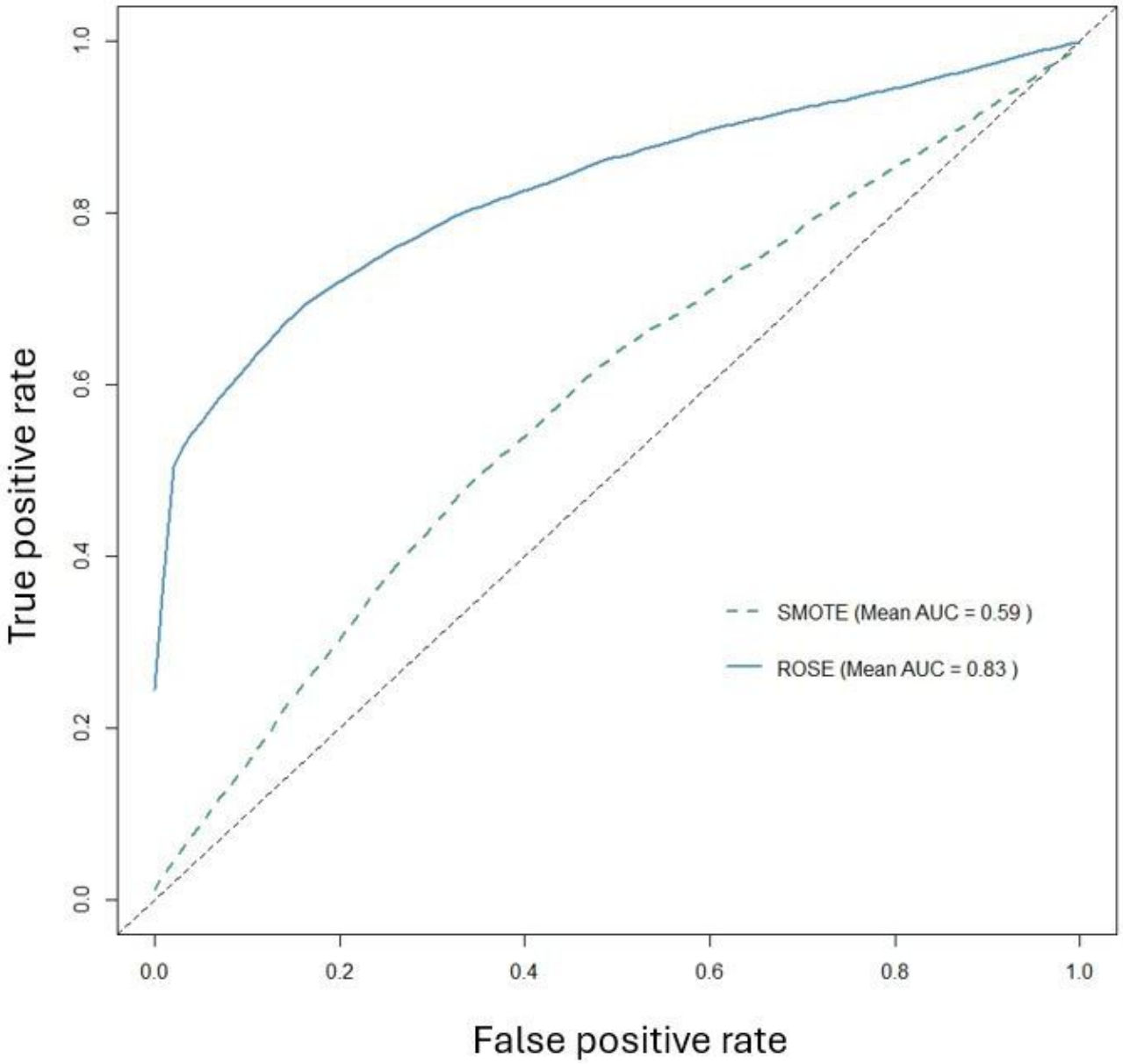
XSL•FO
RenderX

Figure 4. Comparison of the average area under the ROC curve (AUC) of different encoding methods. AUC: area under the ROC curve.



XSL•FO
RenderX

Figure 5. Comparison of the average area under the ROC curve (AUC) of different imbalance methods. AUC: area under the ROC curve; ROSE: Random Over Sampling Example; SMOTE: Synthetic Minority Over-sampling Technique.



Phase 2

Following the identification of the optimal preprocessing strategy in Phase 1, we conducted a detailed analysis of feature importance and remodeled the dataset using the most influential variables. Table 4 presents the top 15 most important variables ranked based on their cumulative scores from 5-fold cross-validation, with antiplatelet, chest X-ray for aortic arch calcification (CXR.AoAC).0, and insulin emerging as the most significant predictors. These variables were selected to refine the models, ensuring a more efficient and interpretable predictive framework.

XSL•FO
RenderX

Table . The top 15 variables ranked by importance after 5-fold cross-validation.

Ranking	Feature	Score[a]
1	Antiplatelet	496
2	CXR.AoAC[b].0	466
3	Insulin	453
4	No.of.hypotension.episodes	435
5	CXR.AoAC.2	434
6	Dyslipid	410
7	DM[c]	392
8	AVA[d]	378
9	AR[e].0	367
10	AV.cal[f]	348
11	PAOD[g]	337
12	LVH[h].type.1	302
13	Calcitriol	299
14	AS[i].0	295
15	Statin	292

[a]The score calculation method is as follows: in the 5-fold cross-validation, the fields are sorted according to their importance, with the highest being 100 and the lowest being 0. The scores from the 5 folds are then summed to obtain the final score.

[b]CXR_AoAC: chest X-ray for aortic arch calcification.

[c]DM: diabetes mellitus.

[d]AVA: aortic valve area.

[e]AR: aortic regurgitation.

[f]AV cal: aortic valve calcification.

[g]PAOD: peripheral arterial occlusive disease.

[h]LVH: left ventricular hypertrophy.

[i]AS: aortic stenosis.

The high-performing ML models from Phase 2 identified a set of correlated features predictive of MACE in patients with ESRD, which can be grouped into 3 main clusters, including structural predictors of MACE, systemic burdens with traditional risk factors, and other indicators of disease severity. The corresponding feature importance profile for the Phase 2 LightGBM model is visualized in Figure S1 in Multimedia Appendix 1. The first group includes aortic arch calcification statuses (CXR.AoAC.0 and CXR.AoAC.2), and concentric LVH (LVH.type.1). These features reflect the model's ability to capture structural markers representative of the broader spectrum of chronic kidney disease–mineral and bone disorder pathology. Moderate AoAC and AV calcification signal cumulative metabolic and inflammatory insults from longstanding phosphate imbalance, high uremic toxin burden, and chronic inflammation [37-39], while the absence of calcification indicates a protective cardiovascular profile to MACE risk. Additionally, concentric LVH has been long recognized as an independent predictor of MACE and all-cause mortality in patients with ESRD [17,40], results from chronic pressure overload, often due to arterial stiffness secondary to vascular calcification. Collectively, these structural changes contribute to hemodynamic consequences such as heart failure and fatal arrhythmias, thereby resulting in higher MACE risk [41].

The second group of predictors comprises DM, dyslipidemia, and PAOD. DM and dyslipidemia are well-established contributors to atherosclerosis and are exacerbated in ESRD [42,43], forming a vicious cycle with both ESRD-specific and traditional risk factors. For example, hyperglycemia promotes the accumulation of advanced glycation end-products due to impaired clearance, leading to endothelial dysfunction, oxidative stress, and inflammation, thereby fueling the atherosclerosis process [44]. PAOD, a direct manifestation of exacerbated systemic atherosclerosis, also shares similar underlying pathophysiology. Despite not being an isolated predictor, PAOD's strong association to higher MACE risk [45] makes it a valuable proxy for assessing a patient's total atherosclerotic burden.

The third group includes antiplatelet, intradialytic hypotension (IDH), and valvular disease status. With a score of 496, antiplatelet use serves as a surrogate marker for clinician-identified ischemic risk and thus MACE vulnerability [46]. Furthermore, the number of IDH episodes ranked fourth, underscoring the hemodynamic stress of dialysis, with

XSL•FO
RenderX

cumulative episodes contributing to myocardial stunning and ischemic damage, thus in turn linked to MACE [47]. Tracking IDH frequency could offer a more comprehensive view of cardiovascular stress rather than focusing on isolated episodes as well as the episode severity. In addition, the absence of valvular diseases such as AS or aortic regurgitation reflects a mitigated risk of LVH and atherosclerosis progression [48]. The Phase 2 model effectively learned patterns of clinical practice—such as medication use and hemodynamic instability—offering real-world applicability and insights that extend beyond static risk markers to isolated hypotensive events.

The models were then retrained with one-hot encoding combined with either ROSE or SMOTE, and their performance was evaluated by accuracy, F_1-score, and AUC with 5-fold cross-validation (Table 3). Among the ROSE-based models, LightGBM and XGBoost achieved the highest discrimination, with LightGBM reaching accuracy 0.906 (SD 0.115; 95% CI 0.763 - 1.000), F_1-score 0.892 (SD 0.119; 95% CI 0.745 - 1.000), and AUC 0.925 (SD 0.112; 95% CI 0.786 - 1.000), and XGBoost reaching accuracy 0.913 (SD 0.122; 95% CI 0.762 - 1.000), F_1-score 0.899 (SD 0.138; 95% CI 0.728 - 1.000), and AUC 0.906 (SD 0.137; 95% CI 0.736 - 1.000). Random forest under ROSE also performed well (AUC 0.915; 95% CI 0.750 - 1.000), confirming that ensemble methods benefited most from the refined feature set.

To place these gains in context, we also examined the traditional logistic regression (LGR) baseline in the same table. LGR trained in only 0.08 s for both ROSE and SMOTE variants, confirming its computational efficiency. However, its discrimination lagged behind the ML models, with AUC values of 0.755 (ROSE) and 0.728 (SMOTE). This contrast highlights a clear trade-off: while LGR offers near-instantaneous runtime, the proposed ML pipeline delivers markedly stronger predictive accuracy, justifying its added complexity for clinical applications.

To validate the statistical significance of these differences, we conducted DeLong tests for all model pairs, with detailed *P* values reported in Table S3 in Multimedia Appendix 1. ROSE-based ensembles (LightGBM, XGBoost, and random forest) achieved significantly higher AUCs than all SMOTE-based models (all $P \leq .022$), while differences among the ROSE ensembles themselves were small and often not statistically significant. In contrast, the logistic regression baseline under ROSE achieved AUC 0.755 (SD 0.050; 95% CI 0.694 - 0.817), clearly lagging behind the gradient boosting models despite its very short training time. These findings confirm that combining ROSE with gradient-boosted ensembles yields the most stable and discriminative performance in the Phase 2 setting.

These findings highlight the robustness of ROSE in preserving critical data characteristics, while the selection of the top 15 features enhanced model efficiency without compromising predictive accuracy. The combination of ROSE with XGBoost and LightGBM yielded the most reliable performance, reinforcing the effectiveness of targeted feature selection and imbalance handling in MACE prediction.

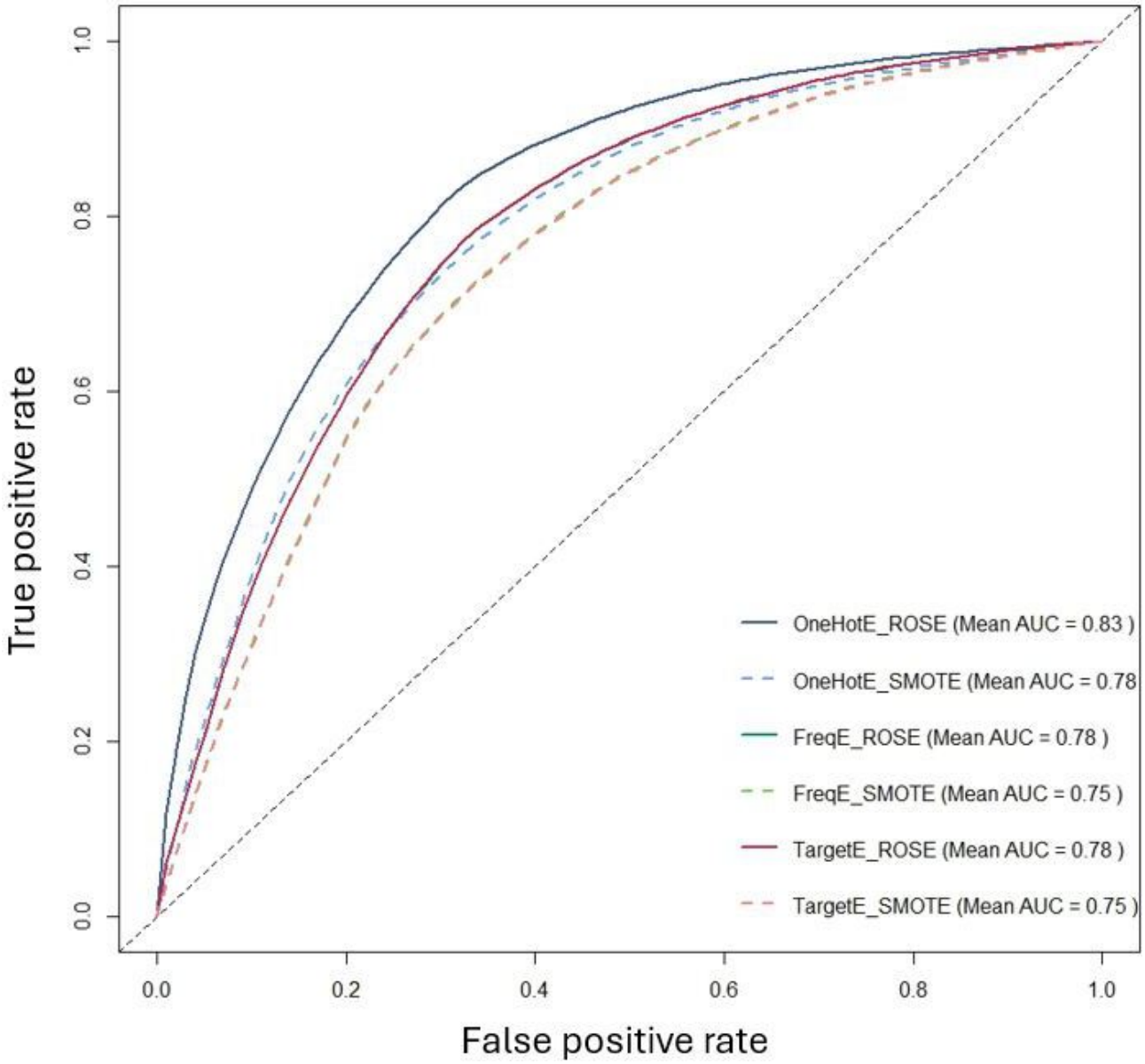
Discussion

Principal Results

To assess cross-dataset portability and reproducibility, we applied the integrated preprocessing and modeling framework to 2 datasets with distinct data-generating mechanisms: an in-house ESRD registry (n=412) and the population-based BRFSS 2015 survey. Cross-dataset comparisons of model performance, encoding strategies, and imbalance-handling methods are summarized in Figures 6-8. For each dataset, the full workflow was executed independently, including missForest imputation, one-hot, frequency, target encoding, ROSE or SMOTE resampling, model tuning, and evaluation, with no cross-dataset feature matching and no pooled AUC computations. The framework yielded improved discrimination and clinically interpretable feature-importance rankings in both settings.

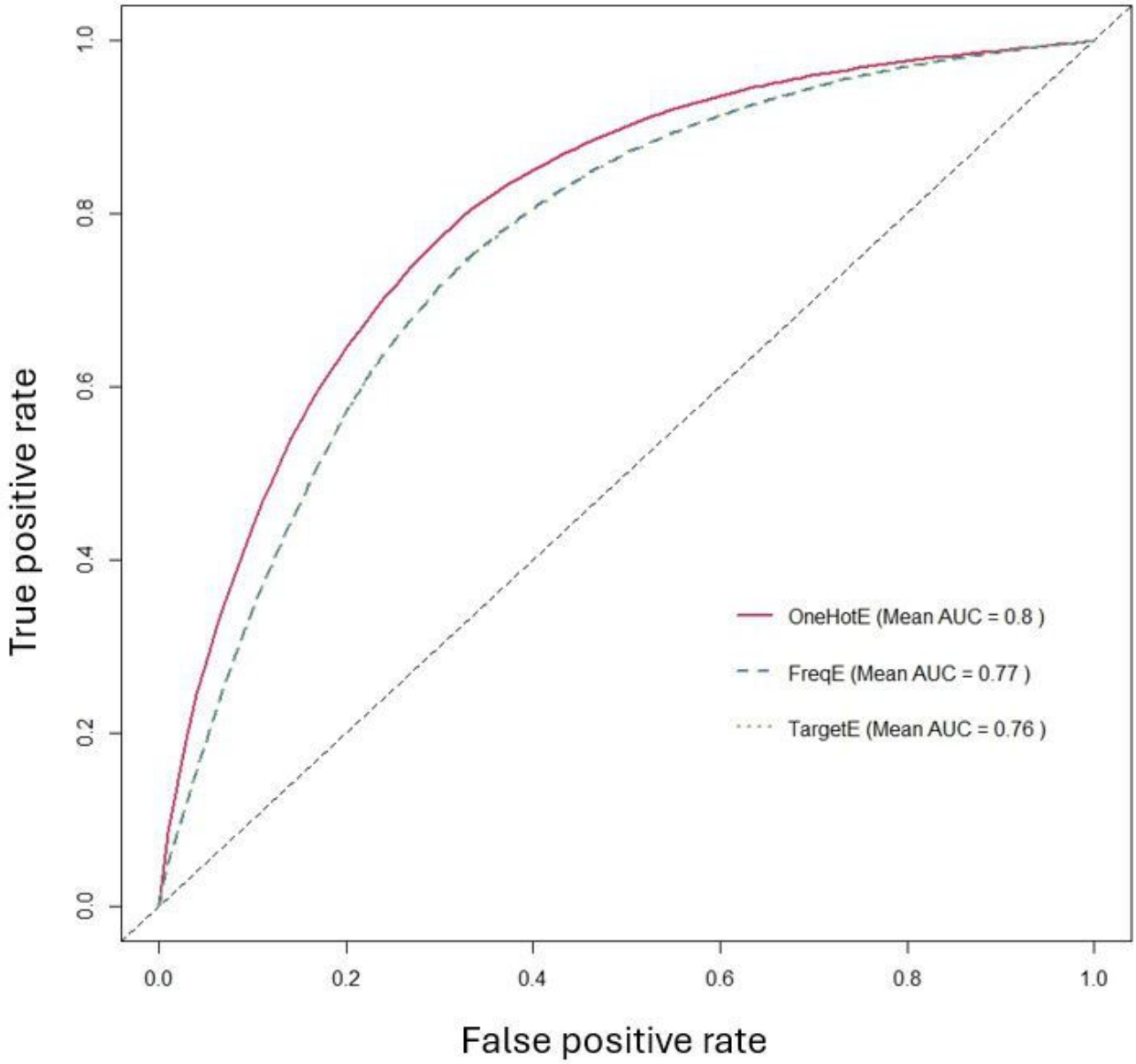
XSL•FO
RenderX

Figure 6. Comparison of the average area under the ROC curve (AUC) of different data models. AUC: area under the ROC curve; ROSE: Random Over Sampling Example; SMOTE: Synthetic Minority Over-sampling Technique.



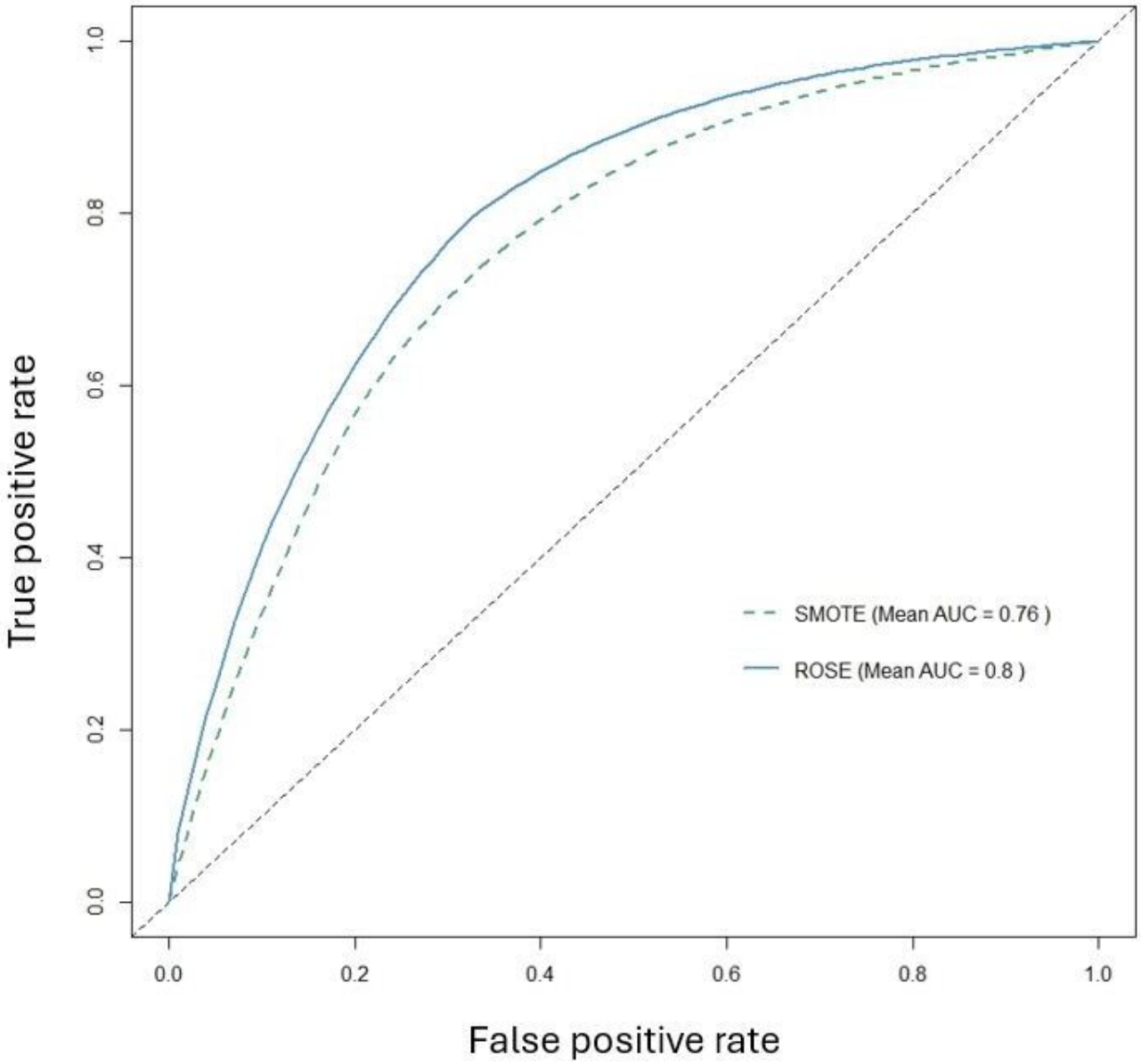
XSL•FO
RenderX

Figure 7. Comparison of the average area under the ROC curve (AUC) of different encoding methods. AUC: area under the ROC curve; ROSE: Random Over Sampling Example; SMOTE: Synthetic Minority Over-sampling Technique.



XSL•FO
RenderX

Figure 8. Comparison of the average area under the ROC curve (AUC) of different imbalance methods. AUC: area under the ROC curve; ROSE: Random Over Sampling Example; SMOTE: Synthetic Minority Over-sampling Technique.



Limitations

Our study has several constraints that warrant consideration. The ESRD cohort includes 412 patients, which limits the capture of rare patterns; multicenter expansion is ongoing to increase diversity. Despite 5-fold cross-validation, nested grid search, and an independent BRFSS run to assess cross-dataset portability, overfitting cannot be fully excluded, so temporal and multi-site validations are planned. Although the workflow performed well for cardiovascular and metabolic endpoints, its applicability to other disease domains remains uncertain and will be explored. By design, the BRFSS analysis did not involve feature harmonization, pooled AUCs, or between-dataset hypothesis tests. Finally, while ROSE outperformed SMOTE in our experiments, optimal resampling and encoding choices are likely dataset dependent, and hybrid strategies and embedding-based representations merit further investigation.

Comparison With Prior Work

In contrast to previous studies that tend to focus on isolated aspects of data preprocessing, our research offers a comprehensive, end-to-end framework that addresses the multifaceted challenges inherent in clinical data, namely high missingness, heterogeneity, and severe class imbalance. While earlier works typically investigated single techniques in a limited scope, our study integrates multiple advanced methods and validates the entire pipeline using cross-dataset comparisons. This holistic approach provides a stronger foundation for both improving predictive performance and facilitating clinical translation, marking a clear advancement over traditional, piecemeal methodologies.

Conclusions

This study introduces a robust, generalizable, and interpretable data preprocessing framework for predicting MACEs among patients with ESRD. Although the improvements in performance metrics are incremental, they directly address key challenges

XSL•FO
RenderX

inherent in clinical data heterogeneity and quality. The proposed pipeline enhances both predictive reliability and model transparency, offering practical value for clinical decision support and operational planning in nephrology care.

While system-level integration was not the primary focus of this work, the pipeline was deliberately designed for seamless adoption in real-world settings. By using routinely available clinical variables and lightweight computational procedures, it can be incorporated into existing analytics environments with minimal architectural adjustment. This design supports efficient deployment across diverse institutions and datasets, as evidenced by the consistent performance observed in both the ESRD registry and the BRFSS dataset. Such scalability reinforces its translational potential for broader clinical and public health applications.

Future work should prioritize cross-context application across more heterogeneous populations and evaluate the integration of deep learning architectures to further enhance predictive accuracy and extend applicability to other high-risk patient groups and chronic disease domains.

Funding

This research was supported by the National Science and Technology Council, Taiwan, under grant NSTC112-2221-E-030-009-MY3; Shin-Kong Wu Ho-Su Memorial Hospital under grant 2024SKHADR038; and Fu Jen Catholic University under grant A0114250. The authors are deeply grateful for the generous support provided by these institutions.

Authors' Contributions

Conceptualization, MC and C-KW; methodology, C-WC and MC; software, C-WC; validation, MC, B-CS and C-HW; formal analysis, C-WC; investigation, C-WC and C-HW; resources, C-KW, B-CS and C-WC; data curation, C-WC, C-KW; writing—original draft preparation, C-WC and C-KW; writing—review and editing, MC; visualization, C-WC; project administration, MC; funding acquisition, B-CS and C-HW All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures including hyperparameter grid search settings, DeLong test *P* value matrices, dataset-specific SMOTE/ROSE settings and random seeds, LightGBM feature importance, and the leakage-safe cross-validation workflow.
[DOCX File, 39 KB - medinform_v14i1e75655_app1.docx]

References

1. Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed 2022 Jun;220:106773. [doi: 10.1016/j.cmpb.2022.106773] [Medline: 35429810]
2. Guo LL, Pfohl SR, Fries J, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Appl Clin Inform 2021 Aug;12(4):808-815. [doi: 10.1055/s-0041-1735184] [Medline: 34470057]
3. Anderson K. The role of data preprocessing in machine learning accuracy for heart disease prediction hybrid models for heart disease prediction: combining neural networks with traditional. 2024 URL: https://www.researchgate.net/publication/386177388_The_Role_of_Data_Preprocessing_in_Machine_Learning_Accuracy_for_Heart_Disease_Prediction_Hybrid_Models_for_Heart_Disease_Prediction_Combining_Neural_Networks_with_Traditional [accessed 2026-03-30]
4. Rahman A. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. Int J Artif Intell 2019;17:44-65 [FREE Full text]
5. Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. JMIR Med Inform 2022 May 31;10(5):e36388. [doi: 10.2196/36388] [Medline: 35639450]
6. Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak 2019 Feb 18;19(1):33. [doi: 10.1186/s12911-019-0747-6] [Medline: 30777059]
7. Chang W, Liu Y, Xiao Y, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics (Basel) 2019 Nov 7;9(4):178. [doi: 10.3390/diagnostics9040178] [Medline: 31703364]
8. Jeong B, Cho H, Kim J, et al. Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data. Diagnostics (Basel) 2020 Jun 18;10(6):415. [doi: 10.3390/diagnostics10060415] [Medline: 32570782]
9. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019 Mar 19;19(1):64. [doi: 10.1186/s12874-019-0681-4] [Medline: 30890124]

XSL•FO
RenderX

10. García-Vicente C, Chushig-Muzo D, Mora-Jiménez I, et al. Evaluation of synthetic categorical data generation techniques for predicting cardiovascular diseases and post-hoc interpretability of the risk factors. Appl Sci (Basel) 2023;13(7):4119. [doi: 10.3390/app13074119]
11. Khushi M, Shaukat K, Alam TM, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 2021;9:109960-109975. [doi: 10.1109/ACCESS.2021.3102399]
12. Xie F, Yuan H, Ning Y, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform 2022 Feb;126:103980. [doi: 10.1016/j.jbi.2021.103980] [Medline: 34974189]
13. Naghipour A, Abbaszadeh Bavil Soflaei MR, Ghader-zefrehei M. Machine learning classifiers and data synthesis techniques to tackle with highly imbalanced COVID-19 data. Comput Knowl Eng 2024;7:51-64. [doi: 10.22067/cke.2024.88940.1121]
14. Pargent F, Pfisterer F, Thomas J, Bischl B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Comput Stat 2022 Nov;37(5):2671-2692. [doi: 10.1007/s00180-022-01207-6]
15. Yu L, Zhou R, Chen R, Lai KK. Missing data preprocessing in credit classification: one-hot encoding or imputation? Emerg Mark Finance Trade 2022 Jan 26;58(2):472-482. [doi: 10.1080/1540496X.2020.1825935]
16. Kivrak M, Avci U, Uzun H, Ardic C. The impact of the SMOTE method on machine learning and ensemble learning performance results in addressing class imbalance in data used for predicting total testosterone deficiency in type 2 diabetes patients. Diagnostics (Basel) 2024 Nov 22;14(23):2634. [doi: 10.3390/diagnostics14232634] [Medline: 39682541]
17. Chao CT, Liao MT, Wu CK. Left ventricular hypertrophy geometry and vascular calcification co-modify the risk of cardiovascular mortality in patients with end-stage kidney disease: a retrospective cohort study. J Atheroscler Thromb 2023 Sep 1;30(9):1242-1254. [doi: 10.5551/jat.63870] [Medline: 36567124]
18. Chao CT, Liao MT, Wu CK. Combinations of valvular calcification and serum alkaline phosphatase predict cardiovascular risk among end-stage kidney disease patients. IJC Heart Vasculature 2024 Oct;54:101505. [doi: 10.1016/j.ijcha.2024.101505]
19. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 Jan 1;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
20. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013 Aug 1;3(8):e002847. [doi: 10.1136/bmjopen-2013-002847] [Medline: 23906948]
21. Uyar A, Bener A, Ciray HN, Bahceci M. A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:6214-6217. [doi: 10.1109/IEMBS.2009.5334548] [Medline: 19964898]
22. Lunardon N, Menardi G, Torelli N. ROSE: a package for binary imbalanced learning. R J 2014;6(1):79. [doi: 10.32614/RJ-2014-008]
23. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Disc 2014 Jan;28(1):92-122. [doi: 10.1007/s10618-012-0295-5]
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR 2002;16:321-357. [doi: 10.1613/jair.953]
25. Kamalov F, Leung HH, Cherukuri AK. Keep it simple: random oversampling for imbalanced data. 2023 Presented at: 2023 Advances in Science and Engineering Technology International Conferences (ASET); Feb 20-23, 2023; Dubai, United Arab Emirates p. 1-4. [doi: 10.1109/ASET56582.2023.10180891]
26. Gnip P, Vokorokos L, Drotár P. Selective oversampling approach for strongly imbalanced data. PeerJ Comput Sci 2021;7:e604. [doi: 10.7717/peerj-cs.604] [Medline: 34239981]
27. Nguyen S, Schumacher P, Olinsky A, Quinn J. Effects of resampling techniques on imbalanced data classification: a new under-resampling method. In: Advances in Business and Management Forecasting: Emerald Publishing Limited; 2021:51-70. [doi: 10.1108/S1477-407020210000014005]
28. Berrar D. Cross-validation. In: Reference Module in Life Sciences: Elsevier; 2019:542-545. [doi: 10.1016/B978-0-12-809633-8.20349-X]
29. Therneau T, Atkinson B, Ripley B, Ripley MB. Package ‘rpart. rpart: Recursive Partitioning and Regression Trees. 2015. URL: https://cran.r-project.org/web/packages/rpart/rpart.pdf [accessed 2026-03-18]
30. Williams GJ. Rattle: a data mining GUI for R. R J 2009;1(2):45. [doi: 10.32614/RJ-2009-016]
31. Breiman L. Random Forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
32. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. r package version 0.4-2. : Tianqi Chen; 2015 URL: https://cran.r-project.org/web/packages/xgboost/index.html [accessed 2026-03-30]
33. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv. Preprint posted online on Oct 24, 2018 URL: https://arxiv.org/abs/1810.11363 [accessed 2026-03-30] [doi: 10.48550/arXiv.1810.11363]
34. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017:30 [FREE Full text] [doi: 10.5555/3294996.3295074]
35. Kuhn M. Building predictive models in R using the caret package. J Stat Soft ;28(5) [FREE Full text]
36. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12:1-8. [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]

XSL•FO
RenderX

37. Shen Y. Pathogenesis and mechanism of uremic vascular calcification. Cureus 2024 Jul;16(7):e64771. [doi: 10.7759/cureus.64771] [Medline: 39026575]
38. Mace ML, Gravesen E, Nordholm A, et al. The calcified vasculature in chronic kidney disease secretes factors that inhibit bone mineralization. JBMR Plus 2022 Apr;6(4):e10610. [doi: 10.1002/jbm4.10610] [Medline: 35434452]
39. Chao CT, Lin SH. Uremic vascular calcification: the pathogenic roles and gastrointestinal decontamination of uremic toxins. Toxins (Basel) 2020 Dec 21;12(12):812. [doi: 10.3390/toxins12120812] [Medline: 33371477]
40. Huang YC, Wu CK. Effect of the geometry and severity of left ventricular hypertrophy on cardiovascular mortality in dialysis patients. Kidney Res Clin Pract 2024 Sep 11. [doi: 10.23876/j.krcp.23.290] [Medline: 39384352]
41. Tomura M, Hamasaki Y, Komaru Y, et al. Prognostic significance of concentric left ventricular hypertrophy at peritoneal dialysis initiation. BMC Nephrol 2021 Apr 16;22(1):135. [doi: 10.1186/s12882-021-02321-1] [Medline: 33863299]
42. Poznyak AV, Sadykhov NK, Kartuesov AG, Borisov EE, Sukhorukov VN, Orekhov AN. Atherosclerosis specific features in chronic kidney disease (CKD). Biomedicines 2022 Aug 27;10(9):2094. [doi: 10.3390/biomedicines10092094] [Medline: 36140195]
43. Lacount S, Tannock LR. Dyslipidemia in chronic kidney disease. In: Endotext 2025. [Medline: 26247091]
44. Curaj A, Vanholder R, Loscalzo J, et al. Cardiovascular consequences of uremic metabolites: an overview of the involved signaling pathways. Circ Res 2024 Mar;134(5):592-613. [doi: 10.1161/CIRCRESAHA.123.324001] [Medline: 38422175]
45. Smilowitz NR, Bhandari N, Berger JS. Chronic kidney disease and outcomes of lower extremity revascularization for peripheral artery disease. Atherosclerosis 2020 Mar;297:149-156. [doi: 10.1016/j.atherosclerosis.2019.12.016] [Medline: 31948675]
46. Su X, Yan B, Wang L, Lv J, Cheng H, Chen Y. Effect of antiplatelet therapy on cardiovascular and kidney outcomes in patients with chronic kidney disease: a systematic review and meta-analysis. BMC Nephrol 2019 Aug 7;20(1):309. [doi: 10.1186/s12882-019-1499-3] [Medline: 31390997]
47. Sars B, van der Sande FM, Kooman JP. Intradialytic hypotension: mechanisms and outcome. Blood Purif 2020;49(1-2):158-167. [doi: 10.1159/000503776] [Medline: 31851975]
48. Stein EJ, Fearon WF, Elmariah S, et al. Left ventricular hypertrophy and biomarkers of cardiac damage and stress in aortic stenosis. J Am Heart Assoc 2022 Apr 5;11(7):e023466. [doi: 10.1161/JAHA.121.023466] [Medline: 35301869]

Abbreviations

AS: aortic stenosis
AUC: area under the ROC curve
BRFSS: Behavioral Risk Factor Surveillance System
CAT: CatBoost
CXR_AoAC: chest X-ray for aortic arch calcification
DM: diabetes mellitus
DT: decision tree
ESRD: end-stage renal disease
IDH: intradialytic hypotension
LGR: logistic regression
LightGBM: light gradient-boosting machine
LVH: left ventricular hypertrophy
MACE: Major Adverse Cardiovascular Events
ML: machine learning
OOF: out-of-fold
PAOD: peripheral arterial occlusive disease
ROS: Random Over-Sampling
ROSE: Random Over-Sampling Examples
SMOTE: Synthetic Minority Over-sampling Technique
XGB: extreme gradient boosting (XGBoost)

XSL•FO
RenderX

Edited by J Klann; submitted 08.Apr.2025; peer-reviewed by P Velmurugan, PC Lin, YW Kao; revised version received 08.Dec.2025; accepted 05.Feb.2026; published 13.Apr.2026.

Please cite as:
Chuang CW, Wu CK, Wu CH, Shia BC, Chen M
Enhancing Model Generalizability in Medical Artificial Intelligence: Systematic Comparison of Categorical Encoding and Sampling Techniques for Imbalanced Data
JMIR Med Inform 2026;14:e75655
URL: https://medinform.jmir.org/2026/1/e75655
doi:10.2196/75655

© Chien-wei Chuang, Chung-Kuan Wu, Chao-Hsin Wu, Ben-Chang Shia, Mingchih Chen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enterocutaneous Fistula–Associated Sepsis and Mortality: Development and Validation of a Multimodal Artificial Intelligence Prediction Model

Hui Li[1*], MM; Jing Chen[1*], MM; Peijun Lin[1], MM; Youmei Pan[1], MM; Yawen Cao[1], MM; Wenfeng Xie[1*], MM

Department of Emergency Medicine, Biomedical InnovationCenter, The Sixth Affiliated Hospital of Sun Yat-sen University, Sun Yat-sen University, Guangzhou, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Wenfeng Xie, MM
Department of Emergency Medicine
Biomedical InnovationCenter, The Sixth Affiliated Hospital of Sun Yat-sen University
Sun Yat-sen University
No. 26, Yuancun Erheng Road
Tianhe District
Guangzhou, Guangdong, 510655
China
Phone: 86 13728010186
Email: xiewfeng@mail.sysu.edu.cn

Abstract

Background: Predicting enterocutaneous fistula (ECF)–associated sepsis and mortality poses significant challenges in digital health care due to the disease's complexity and heterogeneous clinical manifestations. Current approaches that rely on single-modal data or traditional scoring systems often fail to capture the intricate immune-inflammatory dynamics and multisystem involvement in patients with ECF.

Objective: This study aims to develop an artificial intelligence (AI)–driven multimodal fusion model integrating clinical, imaging, and transcriptomic data for early prediction of ECF-associated sepsis and 28-day mortality, addressing the limitations of conventional single-dimensional models.

Methods: This study leveraged publicly available datasets (Medical Information Mart for Intensive Care III [MIMIC-III], electronic Intensive Care Unit [eICU], and The Cancer Genome Atlas) to construct a multimodal framework. Clinical parameters were processed using Extreme Gradient Boosting, abdominal imaging features were extracted via convolutional neural networks, and transcriptomic profiles were analyzed with variational autoencoders. A Transformer-based fusion network was employed for joint prediction and validated through cross-validation and external testing. Key features were identified using Shapley Additive Explanations and Local Interpretable Model-Agnostic Explanations interpretability algorithms, while immune regulatory mechanisms were explored via weighted gene co-expression network analysis.

Results: The multimodal model achieved an area under the curve (AUC) of 0.89 for predicting sepsis and 28-day mortality, outperforming unimodal models (clinical-only model, AUC 0.72, and imaging-only model, AUC 0.78). Critical predictors included Sequential Organ Failure Assessment score, lactate levels, intra-abdominal free fluid on imaging, and immunoregulatory genes (programmed death-ligand 1 [PD-L1] and indoleamine 2,3-dioxygenase 1 [IDO1]). Mechanistic analysis revealed distinct immune reprogramming in patients with sepsis, characterized by increased regulatory T cells and M2 macrophages, along with downregulated cluster of differentiation 8+ (CD8+) T cells.

Conclusions: This multimodal AI model offers an innovative digital solution in medical informatics, enabling precise early risk stratification for ECF-associated sepsis. By integrating multisource data and providing interpretable insights into immune-inflammatory pathways, the model enhances health care quality for patients with ECF and paves the way for personalized intervention strategies.

(JMIR Med Inform 2026;14:e79985) doi:10.2196/79985

XSL•FO
RenderX

KEYWORDS

enterocutaneous fistula; sepsis; multimodal artificial intelligence; immunoregulation; radiomics; transcriptomics

Introduction

Enterocutaneous fistula (ECF) is a severe digestive system complication secondary to abdominal surgery, severe infection, or inflammatory bowel disease [1,2]. Its primary characteristic is the outflow of intestinal contents into the abdominal cavity or onto the body surface through fistulous channels [3,4]. Although uncommon in clinical practice, ECF has attracted significant attention due to its highly complex infection pathways and high mortality rates [5]. Complicated intraabdominal infection (CIAI) and secondary sepsis represent the most critical causes of death in patients with ECF, with rapid disease progression and mortality rates reaching 30%-50% [2,6]. However, current clinical practice lacks effective tools for early risk assessment and intervention in these patients, leading to delayed identification of high-risk cases and postponed interventions, which significantly compromise treatment efficacy and prognostic improvement [7,8].

From a mechanistic perspective, patients with ECF with concurrent abdominal infections exhibit complex immune dynamics involving both structural damage to the intestinal mucosal barrier and bacterial translocation [9,10], accompanied by features of immune dysregulation such as T-cell dysfunction, impaired antigen presentation, upregulation of immunosuppressive factors, and cytokine storms [11]. This immune reprogramming not only exacerbates systemic inflammatory response syndrome but also directly promotes the development of multiple organ dysfunction syndrome [12,13]. Consequently, accurately identifying immune dysregulation states and predicting infection risk in complex clinical scenarios has become a key scientific challenge in infectious diseases and critical care medicine.

Artificial intelligence (AI) and deep learning (DL) have achieved revolutionary progress in medical big data mining in recent years. These technologies demonstrate unprecedented potential, particularly in the joint modeling of multisource heterogeneous data, automated extraction of complex imaging features, and intelligent identification of potential biomarkers [14,15]. As an advanced AI approach, multimodal fusion models integrate clinical indicators, molecular features, and imaging data to construct unified representation spaces, offering novel solutions for precise prediction and exploration of pathological mechanisms in complex diseases [16,17]. This strategy has demonstrated dual advantages in the prediction and interpretation of oncology and stroke.

To further clarify the innovative aspects and methodological positioning of this study, we conducted a systematic review and comparison of recent representative AI-based sepsis prediction studies (Table 1). We summarized their data modalities, algorithmic frameworks, major strengths, and existing limitations to highlight the potential advantages and research gaps of a multimodal fusion strategy for complex infection prediction.

However, AI applications in ECF-associated complex infections and sepsis remain at an early stage. Existing studies predominantly rely on Sequential Organ Failure Assessment (SOFA) scores or single experimental indicators to construct simple classifiers for mortality prediction, lacking systematic integration of immune mechanisms, molecular characteristics, and clinical dynamics. This limitation hinders their ability to meet clinical demands for intelligent stratification of complex diseases. Therefore, there is an urgent need for an AI-based prediction system with strong generalizability, high interpretability, and clear mechanistic relevance to facilitate precise identification and intervention guidance for high-risk individuals.

The novelty of this study is reflected in several key aspects. First, we focus on the highly heterogeneous clinical scenario of ECF-associated CIAI and secondary sepsis, and develop an AI-based risk prediction framework tailored to real-world surgical decision-making rather than relying on generic models derived from unselected intensive care unit (ICU) populations. Second, within a unified modeling architecture, we systematically integrate longitudinal clinical variables, abdominal imaging phenotypes, and transcriptomic features, thereby establishing a closed-loop strategy that simultaneously optimizes predictive performance, model interpretability, and mechanistic insight. Third, the model adopts a modular and scalable multimodal design: while the full trimodal configuration enables mechanistic exploration, a dual-modal version can maintain comparable predictive accuracy in clinical settings where transcriptomic data are unavailable, thus balancing scientific depth with practical deployability. Finally, by combining explainable modeling with network-based bioinformatics analyses, we identify key candidate druggable factors, providing potential translational clues for personalized intervention and targeted therapeutic strategies.

To address these challenges, this study develops a multimodal DL prediction model integrating clinical data, RNA expression profiles, and computed tomography/magnetic resonance imaging (CT/MRI) data. The framework combines multiple submodel architectures, including Extreme Gradient Boosting (XGBoost), convolutional neural networks (CNNs), and variational autoencoders (VAEs), with a transformer-based fusion architecture for multitask learning. The model simultaneously predicts “sepsis risk” and “28-day mortality risk,” while employing interpretability algorithms such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to identify key predictive factors. Further integration of weighted gene coexpression network analysis (WGCNA), differential expression analysis, and immune cell infiltration inference helps decipher immune regulatory network changes and identify core intervention targets, providing a theoretical foundation for mechanistic research and potential therapeutic strategies.

XSL•FO
RenderX

Table 1. Comparative summary of existing artificial intelligence–based sepsis prediction approaches.[a]

Study	Data modality	Core algorithm	Strengths/advantages	Limitations/weaknesses	Relevance to this study
Fleuren et al [7]	Multistudy review (intensive care unit/emergency department/inpatient, clinical data)	Systematic review and meta-analysis	Quantified diagnostic accuracy of various machine learning models for sepsis; established baseline methodology.	High heterogeneity among studies; few real-world deployments.	Serves as a methodological benchmark for AUC[b]/PR[c]-AUC metrics and reporting standards.
Wang et al [18]	Clinical and laboratory (electronic health record)	Random forest and logistic regression	Simple, interpretable, and deployable; used open datasets.	No imaging or molecular data; limited generalization.	Provides a strong single-modality clinical baseline for comparison.
Yan et al [19]	Structured and unstructured clinical text	Natural language processing and machine learning hybrid	Demonstrated that text and structured data improve early detection.	Heterogeneous methods and a lack of a standardized feature space.	Highlights potential to replace transcriptomic data with clinical text for near real-time settings.
Liu et al [20]	Time-series electronic health record (MIMIC-III[d])	XGBoost[e] and temporal feature engineering	Temporal modeling improved AUROC[f]; robust on public datasets.	Still electronic health record–only; lacks imaging/molecular integration.	Serves as a tree-based temporal modeling baseline.
Peng et al [21]	Transcriptomics	Gene signature and risk scoring	High interpretability via immune gene signatures; mortality prediction focus.	Computationally intensive; limited clinical applicability.	Provides a benchmark for molecular-level analysis and model interpretability.
Yang et al [22]	Clinical (multimodel evaluation)	Multiple machine learning algorithms	Broad performance benchmarking across models.	No multimodal integration.	Supplies a comparative reference for overall metric ranges.

[a]This table systematically compares representative artificial intelligence studies on sepsis prediction published in recent years, encompassing both single-modality (clinical, temporal, and transcriptomic data) and multimodal (structured + textual data) modeling strategies. Each study lists its data type, core algorithm, main advantages, and limitations, while the "Relevance to this study" column highlights the relationship and complementary value of each work relative to the multimodal model proposed in this study.

[b]AUC: area under the curve.

[c]PR: precision-recall.

[d]MIMIC-III: Medical Information Mart for Intensive Care III.

[e]XGBoost: Extreme Gradient Boosting.

[f]AUROC: area under the receiver operating characteristic curve.

Methods

Study Patients and Sample Collection

This retrospective study utilized publicly available databases and did not directly involve patient recruitment or intervention. The study population comprised adult patients diagnosed with CIAI, sepsis, or both from the Medical Information Mart for Intensive Care III [MIMIC-III], electronic Intensive Care Unit (eICU), and Gene Expression Omnibus (GEO) databases. Inclusion criteria were (1) confirmed ECF diagnosis; (2) complete clinical follow-up data; (3) imaging findings suggestive of abdominal infection; and (4) availability of contemporaneous laboratory parameters, clinical course documentation, or molecular-level data. Exclusion criteria were (1) incomplete records; (2) severely deidentified data preventing temporal alignment; and (3) significant confounding factors such as underlying immunodeficiency or malignancy. The collected clinical data are presented in Multimedia Appendix 1.

Gene expression data were obtained from publicly available TCGA (The Cancer Genome Atlas) and GEO datasets related to abdominal infections, sepsis, and intestinal barrier dysfunction. Selection criteria required RNA sequencing (RNA-seq) or microarray data types, a sample size of 30 or more, and availability of clinical information or survival data. Imaging data primarily originated from the MIMIC-Chest X-Ray (MIMIC-CXR) and Radiopaedia databases, including CT and MRI scans from patients with ECF or abdominal infection. All data were publicly accessible, fully deidentified, and compliant with ethical standards (see Figure S1 in Multimedia Appendix 2).

For the MIMIC-III and eICU datasets, a 2-step approach was used to identify ECF cases:

- Diagnosis code anchoring: Patients with International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code 569.81 (intestinal fistula, excluding rectal/anal fistula) recorded during hospitalization were selected.
- Supporting evidence within the temporal window (within 72 hours before or after ICU admission; at least one of the following criteria had to be met): (1) surgical or imaging reports containing the keywords "enterocutaneous fistula," "ECF," "entero-cutaneous," "肠皮瘘," or "对外瘘"; (2) fistulography, operative notes, or bedside drainage records explicitly indicating an enteric origin of the fistula; or (3)

XSL•FO
RenderX

nursing or wound care documentation describing bowel contents leaking through a skin tract.

Exclusion criteria were as follows: rectal/anal fistula (ICD-9-CM 565.1) and rectovaginal fistula (ICD-9-CM 619.1). For external cohorts using ICD-10-CM, ECF was mapped to K63.2, while K60.* (anal fistula) and N82.* (female genital tract fistula) were excluded. Postoperative ECF cases were included when K91.89 co-occurred with fistula-related keywords in clinical narratives.

Text mining was as follows: Clinical records were screened using the keyword set {“enterocutaneous fistula,” “ECF,” “entero-cutaneous,” “fistula to skin,” “bowel contents from wound,” “肠皮瘘,” “肠内容物外溢”}. Ambiguous cases were independently reviewed by 2 researchers, and discrepancies were resolved through discussion to achieve consensus. The GEO and TCGA datasets were used solely for mechanistic transcriptomic analysis and were not involved in ECF case identification.

All data used in this study were obtained from publicly accessible, fully deidentified databases (MIMIC-III, eICU, TCGA, and GEO). As this study involved a secondary analysis of anonymized data, ethical review and informed consent were exempt.

Clinical Information and Label Definition

With multidisciplinary support (critical care medicine, infectious diseases, and bioinformatics analysis), clinical labels were systematically defined and categorized for patient data. The primary labels were as follows:

- Infection classification: no infection, localized infection, CIAI, and sepsis.
- Immunophenotype: immune activation, immune dysregulation, and immune suppression.
- Outcome measures: death within 28 days, ICU discharge, and prolonged hospitalization.
- Severity scores: SOFA, quick SOFA, and Acute Physiology and Chronic Health Evaluation II (APACHE II), which were used to quantify acute physiological derangements.

All labels were automatically extracted based on clinical diagnostic criteria and predefined database fields, with manual verification to ensure consistency. The training data were uniformly converted into tensor structures to facilitate subsequent model input and multimodal integration (Figure S2 in Multimedia Appendix 2).

Clinical Variable Preprocessing

The clinical data were derived from diverse sources with heterogeneous variable types, requiring standardized processing. Key steps were (1) *z* score normalization of continuous variables to eliminate scale differences; (2) one-hot encoding of categorical variables to preserve classification information; (3) temporal alignment of time-series indicators (eg, inflammatory markers and organ function scores) with sliding-window feature extraction; and (4) multiple imputation by chained equations (MICEs) for missing values, excluding fields with missing data more than 30%. After standardization and filtering, 125 feature variables were retained, comprising demographic information, organ scores, biochemical indices, and medication/intervention records, collectively forming the clinical data feature space (Figure S3 in Multimedia Appendix 2).

Image Preprocessing

All original CT and MRI images were exported in DICOM (Digital Imaging and Communications in Medicine) format and converted into high-quality PNG/JPEG files (Figure S4 in Multimedia Appendix 2). The preprocessing pipeline involved (1) window width/level adjustment to enhance abdominal structure contrast; (2) Gaussian and median filtering for noise reduction and image stabilization; (3) uniform cropping to 512×512 pixels for standardized resolution; (4) intensity normalization (0-1 range) to minimize scanner-induced pixel value variations; and (5) augmentation, including rotation (+10° or –10°), horizontal flipping, scaling, and brightness perturbation, to improve model robustness (Figure S5 in Multimedia Appendix 2). These operations were implemented using Python libraries (OpenCV, SimpleITK, and Pillow; Python Foundation). All images underwent dual evaluation by radiology experts to exclude blurred, obscured, or ambiguously labeled cases.

Gene Expression Data Processing

The genomic data comprised RNA expression profiles related to intestinal immunity from the TCGA and GEO databases. Data standardization followed this pipeline: (1) expression matrix normalization and batch effect correction using DESeq2 or limma; (2) filtering of low-expression (FPKM [fragments per kilobase of transcript per million mapped reads] <1) and low-variance genes, retaining highly variable gene sets from immune-related pathways; (3) construction of patient × gene matrices for subsequent autoencoder and WGCNA modeling; and (4) dimensionality reduction of RNA expression data to 50 latent variables via an autoencoder for multimodal fusion. All data processing was implemented in R (R Foundation) and Python environments using Bioconductor and scikit-learn packages (Figure S6 in Multimedia Appendix 2).

Data Balancing and Augmentation Strategies

To address class imbalance, the following enhancement and resampling methods were applied to the training set: phenotypic data used the synthetic minority over-sampling technique for minority class augmentation; genomic data incorporated Gaussian noise injection to improve model robustness; imaging data underwent standard data augmentation (rotation/flipping) to simulate clinical complexity; and temporal data employed window slicing and mixed sampling to generate pseudo-sequences. The final unified multimodal input tensors included a clinical variable matrix (N×125), a gene expression matrix (N×50), and an image tensor (N×1×512×512), all formatted for downstream model training.

Model Selection and Construction

This study developed specialized submodels for clinical data, gene expression data, and medical imaging to perform feature extraction and classification prediction. Based on this, a unified multimodal fusion DL framework was established to achieve risk prediction for ECF-associated CIAI and sepsis.

XSL•FO
RenderX

Lightweight tree-based algorithms, including XGBoost and random forest, were selected for clinical data. These algorithms identified key clinical variables affecting prognosis through embedded feature selection and generated stable intermediate prediction features (Figure S7A-B in Multimedia Appendix 2). XGBoost demonstrates excellent robustness and feature interpretability, making it suitable for processing structured medical record data with missing values.

For gene expression data, a VAE model was constructed to learn low-dimensional latent expression features in an unsupervised manner, followed by a t-distributed stochastic neighbor embedding dimensionality reduction algorithm for visualization and feature compression (Figure S7F in Multimedia Appendix 2). The VAE possesses strong nonlinear fitting capabilities, enabling dimensionality reduction while preserving potential distribution patterns of expression data and facilitating subsequent fusion modeling.

A deep CNN was chosen and structurally optimized for imaging data based on 2 backbone networks, residual network 50 (ResNet-50) and EfficientNet-B0, to extract complex anatomical and inflammatory sign features from abdominal CT/MRI images (Figure S7C-E in Multimedia Appendix 2). To enhance model generalizability, a squeeze-and-excitation attention mechanism module was introduced to guide the model to focus on key lesion areas.

The output features from the 3 modalities were fused into a unified feature representation through fully connected layers. A dual-branch prediction module was implemented at the top level to simultaneously accomplish 2 tasks: “sepsis risk” and “28-day mortality risk” (ie, a multitask learning structure). The entire framework was developed on the PyTorch platform, with the fusion module referencing the Transformer Encoder architecture and incorporating a dropout strategy to prevent overfitting.

Model Training Strategy

All models were trained with GPU acceleration using NVIDIA RTX A6000 graphics cards on the PyTorch 2.1 framework. The dataset was split into training, validation, and test sets at a ratio of 7:2:1 to ensure robust generalization performance. The training protocol incorporated the following:

- Optimizer selection: AdamW optimizer with an initial learning rate of 1×10^{-4} and a step learning rate scheduler decay strategy.
- Loss function design: binary cross-entropy (BCE) loss for main task training, with weighted coefficients balancing multitask loss contributions.
- Transfer learning: ImageNet-pretrained weights initialized the imaging module, significantly improving small-sample learning efficiency.
- Early stopping: training terminated when validation loss failed to decrease for 5 consecutive epochs to prevent overfitting.
- Regularization: Dropout (P=0.30) and BatchNorm layers enhanced model stability during training.

To enhance the rigor and reproducibility of the proposed method, this study formalized the multimodal fusion and multitask learning framework as follows:

Let denote the clinical variable matrix, denote the imaging feature tensor, and denote the omics feature matrix.

After encoding via corresponding feature extraction modules (eg, CNN, Transformer, or VAE), we obtain their low-dimensional representations:

$$z_c=f_c(X_c),\ z_i=f_i(X_i),\ z_g=f_g(X_g)$$

The 3 latent features are integrated via a Transformer-based encoder to obtain the fused representation:

$$z_f=\text{Fusion}(z_c,z_i,z_g)=\text{Softmax}(W_c z_c+ W_i z_i+W_g z_g)$$

In the multitask learning framework, the model outputs included the risk of postoperative fistula ($\hat{y}_1$) and the 28-day mortality risk ($\hat{y}_2$), defined as follows:

The loss function was defined as a weighted BCE formulation:

$$L=\lambda_1\cdot\text{BCE}(y_1,\hat{y}_1)+\lambda_2\cdot\text{BCE}(y_2,\hat{y}_2)$$

where λ_1 and λ_2 are task-specific weights, dynamically adjusted according to dataset balance and validation performance.

To evaluate the computational feasibility and clinical deployment potential of the developed multimodal AI model, we analyzed key performance indicators, including the number of parameters, training time, and inference latency. The model contained approximately 3.2×10^6 parameters, with an average training time of about 45 seconds per epoch on a single NVIDIA RTX 3090 GPU, and converged within approximately 1.2 hours. The inference time per sample was 0.18 seconds, meeting the requirement for real-time clinical prediction.

From a theoretical complexity perspective, the core Transformer-based fusion module has a time complexity of $O(N\cdot d^2)$, where N is the number of features and d is the embedding dimension. As the features from each modality were dimensionally reduced before fusion, the overall computational load remained light. The model can be stably executed on a standard high-performance workstation (Intel i9 CPU, 32 GB RAM, single GPU), requiring fewer computational resources than mainstream deep imaging models (eg, ResNet-50).

In summary, the proposed model achieves high predictive accuracy while maintaining excellent computational efficiency and hardware adaptability, supporting scalable deployment in ICU clinical environments.

Model Performance Evaluation

The model’s performance was evaluated on the test set using the following metrics: AUC, which measures discriminative ability; accuracy, which indicates overall classification correctness; sensitivity and specificity, which evaluate identification capability for positive and negative samples; F_1-score, which assesses performance on imbalanced datasets; and the confusion matrix, which analyzes prediction error

XSL•FO
RenderX

distribution. Additionally, to enhance clinical interpretability, SHAP and LIME were employed for explainable analysis of multimodal prediction results, identifying key decision variables, critical regions, and significant genes to strengthen model credibility and application value.

In comparison with potentially similar approaches (Table 2), the trimodal model achieved the best or cobest performance across all evaluated metrics. The dual-modal model (clinical + imaging), while not relying on transcriptomic data, retained approximately 93% of the overall predictive performance, demonstrating its feasibility for clinical deployment.

Table 2. Comparative performance of the proposed approach versus potential analogues (test set).[a]

Method	Modalities	Area under the curve	Accuracy	Sensitivity	Specificity	F_1-score	Precision-recall-area under the curve
Proposed (trimodal)	Clinical + imaging + transcriptomic	0.86	0.879	0.862	0.901	0.851	0.82
Bimodal (clinical + imaging)	Clinical + imaging	0.834	0.862	0.847	0.886	0.836	0.792
Single modal (clinical only)	Clinical	0.811	0.838	0.826	0.869	0.821	0.764
Single modal (imaging only)	Imaging	0.795	0.827	0.812	0.854	0.807	0.751
Single modal (omics only)	Transcriptomic	0.805	0.836	0.82	0.864	0.815	0.758
Logistic regression (clinical feats)	Clinical	0.782	0.818	0.798	0.842	0.792	0.728
Extreme Gradient Boosting (clinical feats)	Clinical	0.803	0.832	0.818	0.86	0.813	0.756
Sequential Organ Failure Assessment score	N/A[b]	0.742	0.792	0.761	0.815	0.747	0.69
Acute Physiology and Chronic Health Evaluation II	N/A	0.721	0.777	0.742	0.807	0.733	0.671

[a]The table compares the proposed trimodal model, the dual-modal version (excluding transcriptomic data), 3 single-modality models (clinical-only, imaging-only, and transcriptomics-only), 2 classical machine learning baselines (Logistic Regression and Extreme Gradient Boosting, both using the same clinical feature set), and traditional clinical scoring systems (Sequential Organ Failure Assessment and Acute Physiology and Chronic Health Evaluation II). All metrics were calculated on an independent test set; the precision-recall area under the curve was included as a complementary measure for imbalanced data. All metrics were calculated on an independent test set. The results show that the multimodal fusion model achieved the best or cobest performance across area under the curve, accuracy, sensitivity, specificity, and F_1-score, while the dual-modal model maintained comparable overall performance with superior clinical deployability.

[b]N/A: not applicable.

Clinical Validation and Model Optimization

To ensure the applicability and generalizability of the model in real-world settings, we designed an external validation framework based on publicly available databases. The external validation data were obtained from the eICU Collaborative Research Database version 2.0 (a multicenter open ICU dataset) and integrated with transcriptomic data from the GEO database (GSE185263 and GSE176307) to construct an independent validation cohort. A total of 134 adult ICU patients were included, with inclusion and exclusion criteria, variable structures, and data modalities kept consistent with the primary training set (MIMIC-III) to ensure the objectivity and independence of the validation results.

During the validation phase, the AI model first underwent transfer inference using the eICU dataset, and its generalization performance was evaluated using metrics such as AUC, sensitivity, specificity, and F_1-score. Subsequently, multimodal feature revalidation was conducted using the GEO transcriptomic samples to confirm the consistency of key immunologic and imaging features across independent datasets. Throughout model optimization, the fusion-layer weighting and feature selection strategies were dynamically adjusted according to the external validation outcomes, thereby improving cross-platform and cross-population prediction stability.

This closed-loop validation and optimization pipeline, built upon publicly accessible databases, facilitates a transition from a purely data-driven AI system to a validation-driven, generalizable AI system, ensuring that the model maintains robust performance and clinical transferability across heterogeneous datasets and diverse clinical environments (Table 3).

XSL•FO
RenderX

Table 3. Baseline characteristics of the external validation cohort (electronic intensive care unit).[a]

Variable	Total (n=134)	Enterocutaneous fistula–related sepsis (n=50)	Nonsepsis (n=84)	*P* value
Age (years), mean (SD)	59.2 (12.9)	60.8 (12.3)	58.3 (13.1)	.31
Male, n (%)	84 (62.7)	33 (66.0)	51 (60.7)	.53
BMI (kg/m^2), mean (SD)	24.0 (3.6)	24.3 (3.9)	23.8 (3.5)	.46
Acute Physiology and Chronic Health Evaluation II score, mean (SD)	17.3 (6.1)	20.1 (5.8)	15.7 (5.9)	.003
Sequential Organ Failure Assessment score, mean (SD)	7.6 (3.2)	9.2 (3.4)	6.7 (2.9)	.001
Lactate (mmol/L), mean (SD)	2.6 (1.3)	3.1 (1.1)	2.3 (1.3)	.004
C-reactive protein (mg/L), mean (SD)	80.2 (30.9)	94.5 (28.4)	71.5 (30.1)	<.001
White blood cells (×10^9/L), mean (SD)	12.2 (4.7)	13.9 (4.9)	11.2 (4.4)	.006
Imaging data available, n (%)	116 (86.6)	46 (92.0)	70 (83.3)	.15
RNA sequencing data available, n (%)	38 (28.4)	16 (32.0)	22 (26.2)	.47
28-Day mortality, n (%)	27 (20.1)	18 (36.0)	9 (10.7)	<.001

[a]The external validation cohort was derived from the publicly available Electronic Intensive Care Unit Collaborative Research Database version 2.0, comprising 134 adult intensive care unit patients, of whom 50 (37.3%) had enterocutaneous fistula–related sepsis. The table summarizes the main demographic, clinical, and laboratory characteristics. Continuous variables are presented as mean (SD), and categorical variables as n (%). Between-group comparisons were performed using independent-sample (unpaired), 2-tailed *t* tests or chi-square tests, depending on the data distribution. *P* values indicate statistical significance at $P<.05$.

Identification of Key Immune Factors

To further investigate the potential immune mechanisms underlying ECF-associated CIAI and sepsis, this study conducted gene-level differential analysis and regulatory network mining based on CIAI- and immune-related datasets from TCGA and GEO. First, DESeq2 and edgeR were used to perform differential expression analysis on transcriptomic data from patients with ECF with infection versus noninfected controls (differentially expressed gene [DEG] screening thresholds: |log2 fold-change| > 1 and adjusted $P<.05$), obtaining a set of candidate genes showing differential expression.

Subsequently, WGCNA was employed to construct expression modules and identify coexpressed gene modules significantly associated with clinical indicators, including sepsis severity, 28-day mortality, and immune scores. Hub genes within these modules were considered core regulatory candidate genes and prioritized for subsequent functional enrichment analysis.

To ensure the robustness of the results, multiple datasets were used to cross-validate candidate gene consistency, and stable core immune factors were screened through intersection and Venn diagram analyses.

Functional Pathway Enrichment Analysis

Functional annotation and pathway enrichment analysis were performed on the DEGs and WGCNA-identified key module genes. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were utilized to investigate biological processes (BP) and signaling pathways related to immune activation, immunosuppression, inflammatory response, cell death, and intestinal barrier function.

The enrichment analysis was implemented in R using the clusterProfiler and enrichplot packages, with visualization of hierarchical structures and significance levels to facilitate the identification of key pathways and functional clusters. These enrichment results subsequently support the interpretation of the molecular basis underlying model predictions and their potential mechanisms.

Immune Cell Infiltration Analysis

This study performed immune cell infiltration analysis using RNA-seq data to complement transcriptomic analysis at the cellular level. The CIBERSORT and xCell algorithms were employed to estimate the relative abundance of major immune cell subsets in each sample (including T cells, B cells, monocytes, macrophages, and neutrophils). This analysis helps elucidate potential relationships between different immune states and infection phenotypes.

Subsequent correlation analysis between immune cell composition and clinical labels explored whether specific immune subset distribution patterns exist for classification or prediction purposes. This strategy may also provide foundational data for identifying potential immune intervention targets.

Intervention Potential Factors and Dynamic Feature Extraction

This study further investigated molecular factors with intervention potential by integrating DEG analysis, coexpression network structure, and immune cell infiltration patterns. Protein-protein interaction (PPI) network analysis of key genes and screening against drug databases (DrugBank and Drug–Gene Interaction Database) identified molecules with known drug-target backgrounds.

XSL•FO
RenderX

Additionally, when sample data permitted temporal grouping, immune dynamic pattern mapping was performed to preliminarily characterize expression trends of key factors during disease progression. These findings may inform future early identification strategies and time-sensitive intervention approaches.

Statistical Analysis

All statistical analyses were performed in R (version 4.2.3), Python (version 3.10), and SPSS (version 26.0) environments, using 2-tailed (unpaired) tests with a significance level of $P<.05$. For clinical and grouping variables, descriptive statistics evaluated baseline characteristic distributions, with normally distributed data presented as mean (SD) and nonnormal data as median (IQR). Between-group comparisons used independent *t* tests or Mann-Whitney *U* tests for continuous variables, and chi-square or Fisher exact tests for categorical variables. Multigroup comparisons employed 1-way analysis of variance or Kruskal-Wallis tests with least significant difference or Bonferroni correction, and repeated-measures analysis of variance was used to analyze longitudinal data. For bioinformatics data, differential expression analysis used DESeq2 or edgeR (thresholds: |log2 fold-change| > 1 and adjusted *P* value <.05, used as predefined screening criteria). GO and KEGG enrichment analyses utilized clusterProfiler (significance assessed by *P* value or false discovery rate). Immune cell deconvolution applied CIBERSORT or xCell, and Pearson or Spearman correlation examined gene-immune cell-clinical label relationships. Missing data were handled using MICE. Normality was assessed using Shapiro-Wilk or Kolmogorov-Smirnov tests, with log transformation or nonparametric tests applied when assumptions were violated. Analyses used pandas, statsmodels, scipy, and ggpubr for statistical computing and visualization.

Ethical Statement

This study used publicly available, fully deidentified datasets (MIMIC-III, eICU, TCGA, and GEO). Under institutional policies and applicable regulations, analyses of such anonymized secondary data do not constitute human research. Therefore, institutional review board approval and informed consent were not required. All data were handled in accordance with applicable data protection and privacy regulations.

Results

Multimodal AI Model Significantly Enhances Prediction Performance for Sepsis and Mortality Risk

This study's multimodal AI fusion prediction model demonstrated marked advantages in identifying high-risk patients with ECF-associated CIAI and sepsis. By integrating clinical indicators, radiomic features, and transcriptomic data, the model achieved prediction performance superior to conventional unimodal models across multiple critical tasks.

In the sepsis prediction task (Figure 1A), the multimodal model generated a receiver operating characteristic curve with an AUC of 0.860, indicating robust discriminative capacity to accurately differentiate between high- and low-risk patient groups. For 28-day mortality prediction (Figure 1B), the model attained an AUC of 0.910, demonstrating equally powerful predictive capability and robustness for short-term prognosis assessment.

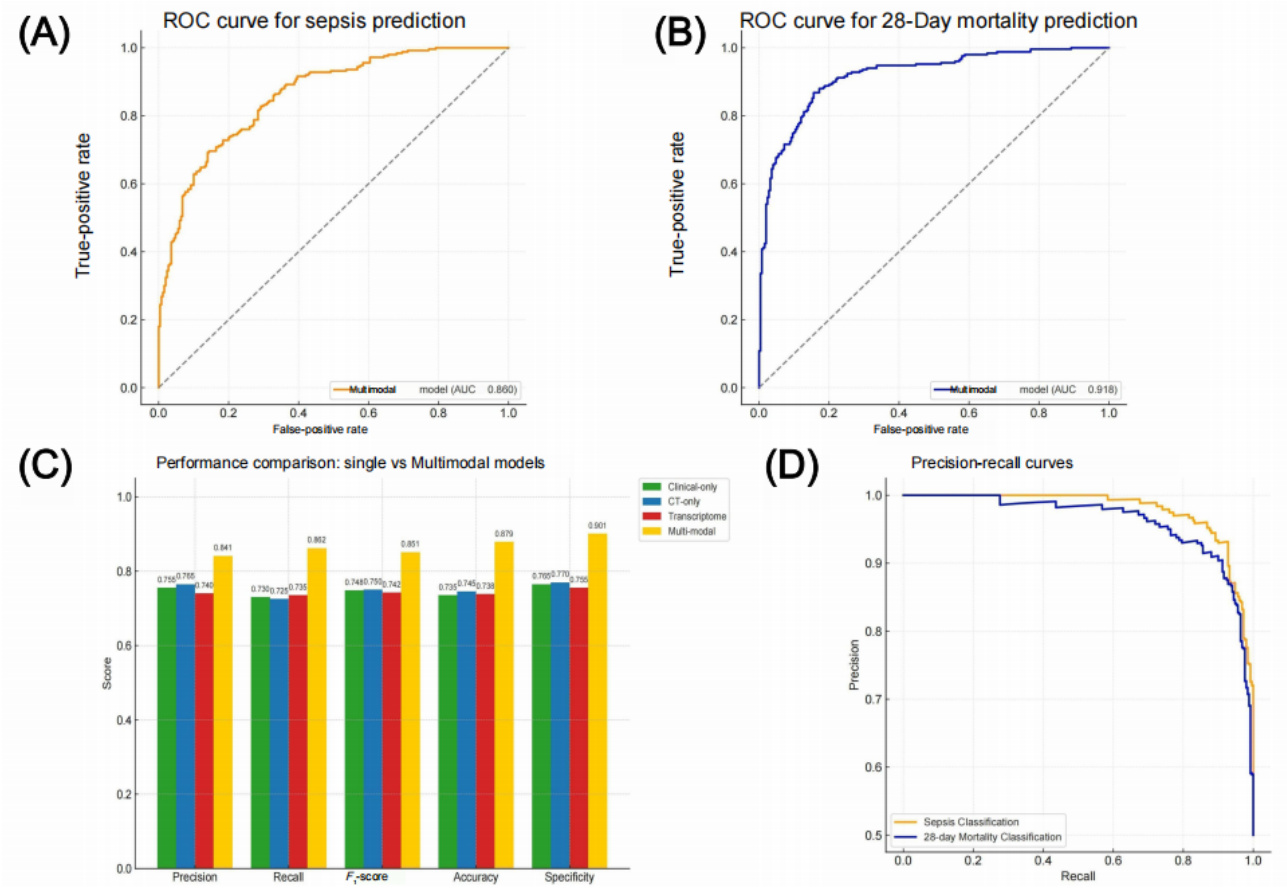
To further validate the practical classification performance across different tasks, we compared our multimodal model with 3 unimodal models (utilizing solely clinical, imaging, or transcriptomic data) across 5 core performance metrics (Figure 1C): precision (positive predictive value), recall (sensitivity), F_1-score, accuracy, and specificity (true-negative rate). Results showed that the multimodal model led across all metrics (precision 0.841, recall 0.862, F_1-score 0.851, accuracy 0.879, and specificity 0.901). Compared with unimodal models, the multimodal approach showed notable improvements in recall and specificity, exceeding the second-best performing model (imaging-based or clinical-based) by nearly 10% points and confirming its capacity for accurate high-risk patient identification while effectively reducing false positives.

Considering the class imbalance in clinical data (particularly for sepsis and mortality cases), we incorporated precision-recall curve analysis to more comprehensively evaluate model performance on imbalanced data (Figure 1D). Results demonstrated that the multimodal model maintained high precision and recall across both prediction tasks, with the precision-recall curve preserving good precision even in high-recall regions. This indicates the model's stability and adaptability for identifying high-risk populations.

In summary, the multimodal fusion model exhibited outstanding overall performance in predicting ECF-associated sepsis and mortality risk, demonstrating high practical utility and clinical translation potential.

XSL•FO
RenderX

Figure 1. Performance evaluation of the multimodal artificial intelligence model for predicting enterocutaneous fistula–associated complicated intraabdominal infection and sepsis. (A) Receiver operating characteristic (ROC) curve for sepsis prediction (area under the curve [AUC] 0.860), demonstrating the model's discriminative ability between positive and negative cases. (B) ROC curve for 28-day mortality prediction (AUC 0.879), indicating the model's capacity to identify high-risk patients. (C) Comparison of multimodal and single-modality models across 5 performance metrics (precision, recall, F1-score, accuracy, and specificity), showing that the multimodal model outperforms all single-modality configurations. (D) Precision-recall curve evaluating the model's robustness under class-imbalanced conditions.



Subgroup Performance Analysis

To evaluate the robustness and fairness of the model across different populations and clinical characteristics, a multidimensional subgroup analysis was conducted. The results demonstrated that the model maintained consistently high predictive performance across subgroups defined by age (<60 years vs ≥60 years), admission type (surgical vs medical patients), and immune phenotype (immune-activated vs immune-suppressed; Table 4). In the age-based analysis, the AUCs were 0.872 and 0.858, with no statistically significant difference (DeLong *P*=.41). For surgical versus medical admissions, the AUCs were 0.866 and 0.881, respectively (*P*=.37). Among immune-activated and immune-suppressed phenotypes, the AUCs were 0.885 and 0.843 (*P*=.29). Furthermore, sensitivity, specificity, and F_1-scores across all subgroups fluctuated within +0.03 or –0.03, indicating strong intergroup consistency and model stability. These findings suggest that the model demonstrates reliable discriminative capability across different age groups, clinical contexts, and immune states, thereby providing evidence for its fairness and potential for clinical generalization.

XSL•FO
RenderX

Table 4. Performance of the multimodal artificial intelligence model across patient subgroups.[a]

Subgroup	n	Area under the curve	Accuracy	Sensitivity	Specificity	F_1-score
Age <60 years	302	0.872	0.881	0.867	0.895	0.852
Age ≥60 years	298	0.858	0.876	0.854	0.884	0.845
Surgical admission	271	0.866	0.878	0.86	0.891	0.848
Medical admission	329	0.881	0.885	0.871	0.896	0.856
Immune-activated phenotype	262	0.885	0.888	0.875	0.902	0.857
Immune-suppressed phenotype	238	0.843	0.869	0.841	0.882	0.836

[a]The subgroup analysis included stratifications by age, admission type, and immune phenotype. All area under the curve values were calculated on the independent test set, and differences between subgroups were assessed using the DeLong test. Performance metrics are reported as mean values. The results indicate that the model maintained consistent predictive performance across different populations, with no statistically significant differences (DeLong test: age, *P*=.41; admission type, *P*=.37; immune phenotype, *P*=.29), suggesting strong fairness and robustness for clinical applications.

Interpretability Analysis Reveals Key Feature Contributions and Decision Logic of Multimodal Model

To enhance the multimodal prediction model's transparency and trustworthiness in clinical practice, this study incorporated 2 mainstream interpretability methods: SHAP and LIME. It conducted both global and local analyses of the model's decision-making logic.

First, the global feature importance ranking based on SHAP analysis (Figure 2A) showed that clinical indicators, including SOFA score, lactate level, white blood cell count, and C-reactive protein (CRP), had the highest mean SHAP values, indicating their greatest contribution to model predictions. Additionally, imaging features such as APACHE II score, intraabdominal free fluid, and intestinal structural collapse demonstrated substantial importance, reflecting the model's ability to focus on key pathophysiological indicators when integrating multimodal information.

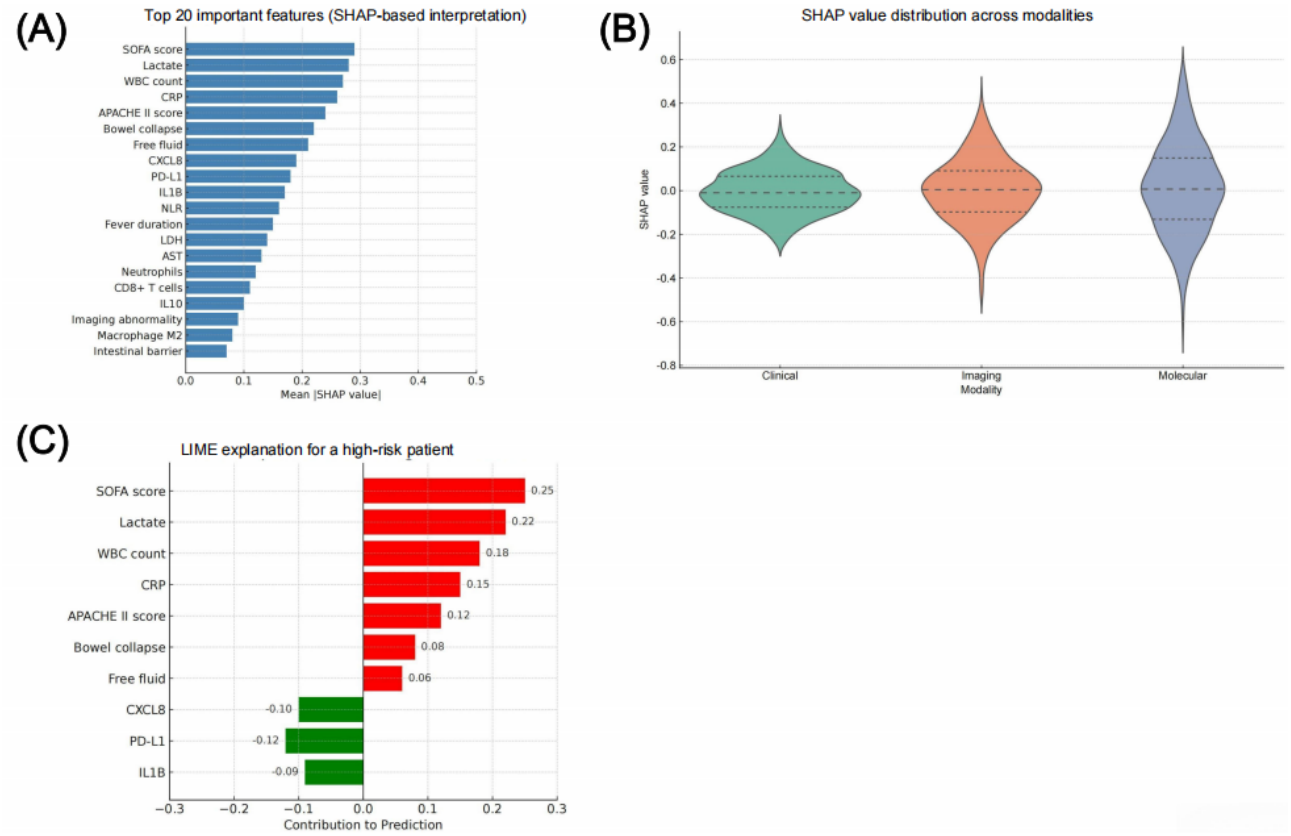
In comparing the contribution distributions across modalities (Figure 2B), the multimodal model exhibited more concentrated and broader SHAP value distributions after fusion, demonstrating enhanced consistency and stability in decision interpretation alongside improved predictive performance. By contrast, unimodal models (clinical or imaging only) showed more dispersed feature influence and less stability.

To further validate the model's reasoning logic at the individual level, LIME analysis was performed on high-risk patient predictions (Figure 2C). Results revealed that SOFA score, lactate concentration, white blood cell count, and CRP level were positive contributing factors that substantially drove high-risk predictions, while molecular indicators such as decreased interleukin 1B (IL1B) expression and downregulated programmed death-ligand 1 (PD-L1) served as negative contributors, helping reduce false alarms. The multimodal features demonstrated complementary synergistic effects, further improving the credibility of individual-level decision-making.

In summary, through global (SHAP) and local (LIME) interpretability mechanisms, the model successfully revealed the contribution pathways of different modal variables in the decision-making process. This enhanced the model's interpretability and provided clinicians with clear references, facilitating the transition from a "black box" to a "transparent box" clinical paradigm.

XSL•FO
RenderX

Figure 2. Global and local interpretability analysis of the multimodal artificial intelligence model. (A) Shapley Additive Explanations (SHAP)–based feature importance ranking showing the top 20 variables contributing to sepsis prediction. The Sequential Organ Failure Assessment (SOFA) score, lactate, white blood cell count (WBC), C-reactive protein (CRP), and Acute Physiology and Chronic Health Evaluation II (APACHE II) score were the major contributors, while imaging and molecular features such as bowel collapse, intraabdominal free fluid, and programmed death-ligand 1 (PD-L1) expression also showed high importance. (B) Distribution of SHAP values across clinical, imaging, and molecular modalities, illustrating the contribution and variability of features within each modality and reflecting the model's stability and consistency in multimodal fusion. (C) Local Interpretable Model-Agnostic Explanations (LIME)–based local interpretability plot for an individual high-risk patient. Red bars indicate positive contributors that increase the predicted risk, whereas green bars denote negative contributors that reduce the risk, revealing the model's decision logic at the individual-case level. AST: aspartate aminotransferase; CXCL8: C-X-C motif chemokine ligand 8; IL: interleukin; LDH: lactate dehydrogenase; NLR: neutrophil-to-lymphocyte ratio.



Differential Expression Analysis Reveals Sepsis-Associated Transcriptomic Alterations

To investigate the molecular mechanisms underlying ECF-associated CIAI and sepsis, we performed systematic bioinformatics analyses of transcriptomic data from public databases (TCGA and GEO).

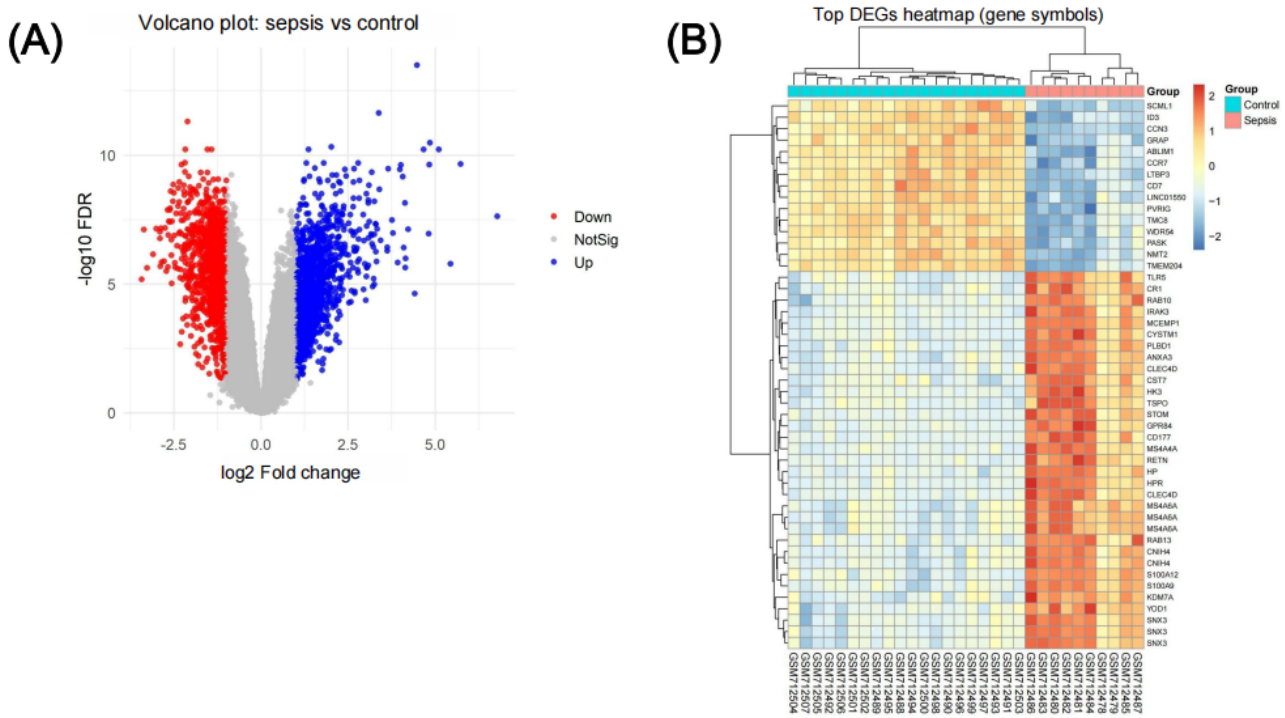
Specifically, we conducted differential expression analysis between high- (sepsis) and low-risk (control) groups using DESeq2 (thresholds: |log2 fold-change| > 1, false discovery rate < 0.05), identifying 2474 significant DEGs (1294 upregulated and 1180 downregulated; Figure 3A; defined by |log2 fold-change| > 1 and false discovery rate (FDR) < 0.05). These genes potentially involve key pathological processes, including inflammation, immune regulation, apoptosis, and gut barrier function.

Hierarchical clustering heatmap analysis (Figure 3B) demonstrated clear separation of expression patterns between sepsis and control groups. In particular, it revealed enrichment of upregulated genes in sepsis samples with strong disease-state correlation. This pattern suggests that these genes may play pivotal roles in sepsis pathogenesis, providing biological insights and potential diagnostic markers.

These findings establish the molecular foundation for subsequent functional enrichment analyses (GO/KEGG) and network construction (WGCNA/PPI), while providing compelling evidence for understanding sepsis transcriptomic signatures.

XSL•FO
RenderX

Figure 3. Differentially expressed gene (DEG) analysis between high- and low-risk patient groups. (A) Volcano plot showing the distribution of DEGs between the sepsis and control groups. Upregulated genes are shown in red, downregulated genes in blue, and nonsignificant genes in gray (selection criteria: |log2 fold-change| > 1, false discovery rate [FDR] <0.05). (B) Heatmap illustrating the expression patterns of the top 50 differentially expressed genes across samples. Red indicates high expression and blue indicates low expression, revealing distinct transcriptional profiles between the sepsis and control groups.



WGCNA Identifies Key Modules and Potential Regulatory Hubs

To further investigate sepsis-associated functional modules and core regulatory genes, we constructed WGCNA based on prior differential expression results, systematically identifying gene modules significantly correlated with clinical traits (Figure 4).

During soft-threshold selection, we determined the power parameter β=7 by analyzing scale-free topology fit and mean connectivity trends (Figure 4A), establishing a similarity matrix that met scale-free network criteria. Dynamic tree-cutting algorithms then clustered genes with similar expression patterns into 21 distinct coexpression modules (Figure 4B), each color-coded.

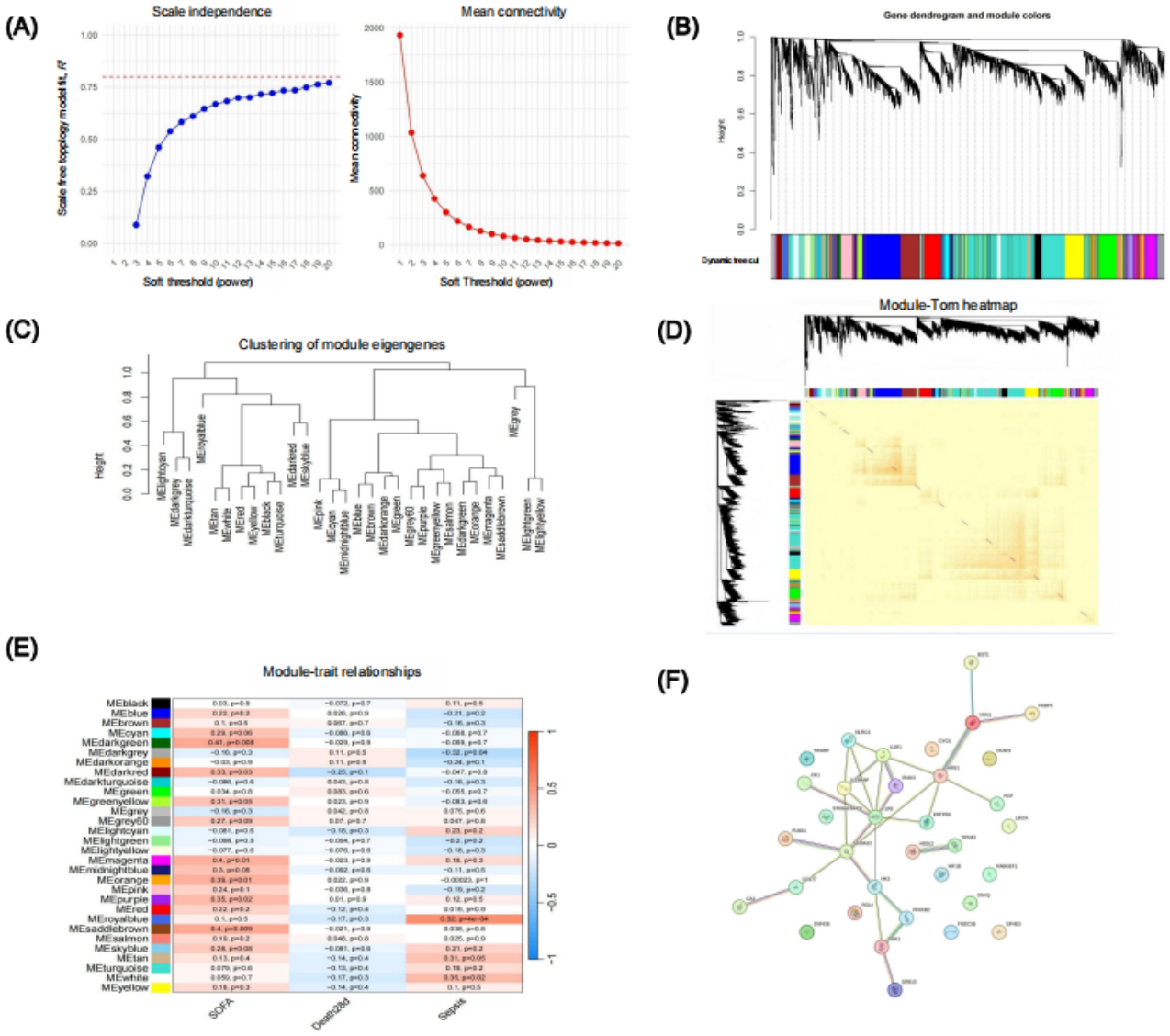
Module eigengene clustering revealed hierarchical relationships among modules (Figure 4C), while a topological overlap matrix heatmap confirmed module structural integrity (Figure 4D). Pearson correlation analysis between modules and clinical traits (SOFA score, sepsis status, and 28-day mortality) identified several associations (Figure 4E): the MEblue module showed a strong positive correlation with sepsis (r=–0.21, P=.20), whereas MEturquoise correlated with SOFA scores (r=0.079, P=.60), suggesting key roles in disease progression and organ dysfunction.

Hub genes with high module membership and clinical trait correlations were selected for PPI network analysis via STRING (Figure 4F). Central hub genes (signal transducer and activator of transcription 3 [STAT3], IL6, C-X-C motif chemokine ligand 10 [CXCL10], TYRO protein tyrosine kinase-binding protein [TYROBP]) demonstrated high connectivity and may serve as regulatory cores mediating inflammatory pathways, immune responses, and cytokine signaling.

Collectively, WGCNA successfully identified sepsis-relevant functional modules and candidate regulatory genes, establishing a multilevel framework connecting expression patterns to functional associations and thereby providing systematic insights into ECF-associated sepsis mechanisms.

XSL•FO
RenderX

Figure 4. Weighted gene co-expression network analysis–based key module identification and core gene network analysis. (A) Results of soft-threshold selection. (B) Gene clustering dendrogram with module identification. (C) Hierarchical clustering of module eigengenes. (D) Topological overlap matrix (TOM) heatmap across modules. (E) Heatmap of module-trait correlations (sepsis, Sequential Organ Failure Assessment [SOFA] score, and 28-day mortality). (F) Protein-protein interaction network of hub genes from key modules, highlighting potential regulatory cores.



Functional and Pathway Enrichment Analysis of DEG

We performed GO and KEGG pathway enrichment analyses on the selected hub genes to investigate the biological functions of DEGs and potential pathways in key modules.

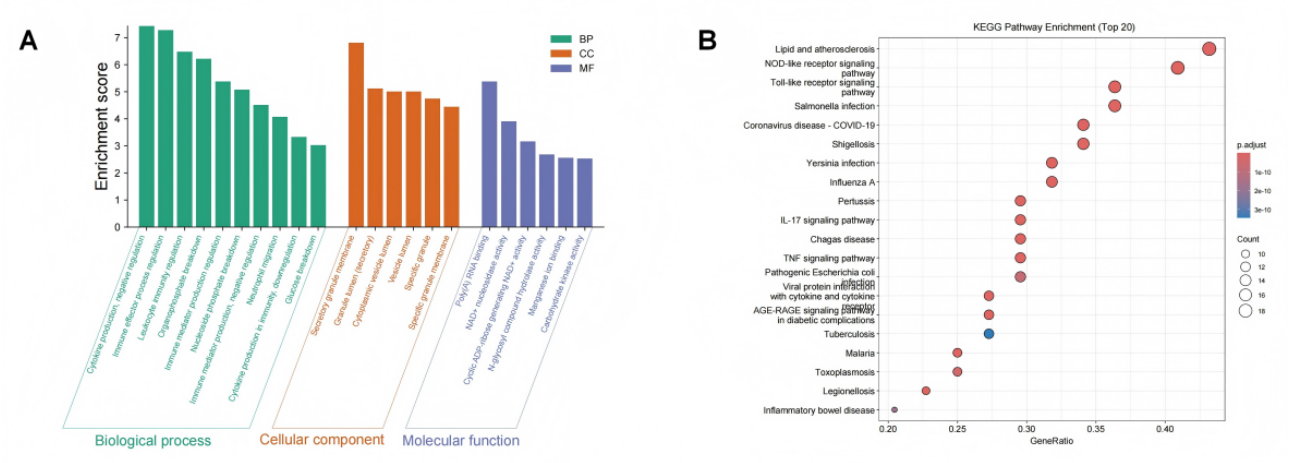
GO analysis revealed significant enrichment of DEGs in BP, including immune response, inflammation mediation, cytokine signaling regulation, leukocyte chemotaxis, and apoptosis regulation (adjusted *P* value <.05). Notably, terms such as "neutrophil chemotaxis," "positive regulation of cytokine production," and "inflammatory response" showed prominent enrichment (Multimedia Appendix 3, green section). For cellular components, enriched terms were primarily associated with "MHC protein complexes," "exosomes," and "plasma membrane regions" (orange), highlighting the importance of intercellular communication and antigen presentation. In molecular function analysis, terms such as "cytokine receptor binding" and "chemokine activity" were significantly enriched (adjusted *P* value <.05, purple), indicating the critical roles of these regulatory molecules in systemic immune dysregulation.

KEGG pathway analysis further demonstrated significant enrichment of DEGs in immune- and inflammation-related signaling pathways (adjusted *P* value <.05; Figure 5), including the nuclear factor kappa B signaling pathway, the Toll-like receptor signaling pathway, the IL-17 signaling pathway, and the tumor necrosis factor signaling pathway, as well as the intestinal immune network for immunoglobulin A production and apoptosis. These findings suggest that ECF-associated complex infections leading to sepsis involve systemic immune response remodeling, with core regulatory pathways centered on innate immune activation and inflammatory cytokine release.

The functional enrichment analysis provides a biological foundation for subsequent mechanistic studies, further supporting transcriptional evidence for inflammatory amplification and immune dysregulation in sepsis pathogenesis.

XSL•FO
RenderX

Figure 5. Gene Ontology (GO) Functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of differentially expressed genes (DEGs). Top 20 KEGG pathway enrichment results, showing significant enrichment in inflammation-related signaling pathways (nuclear factor-kappa B [NF-κB], interleukin [IL]-17, and Toll-like receptor), as well as key regulatory pathways including apoptosis and intestinal mucosal immunity.



WGCNA Identifies Key Modules and Potential Regulatory Hubs

To further investigate immune microenvironment alterations in patients with ECF-associated sepsis, we performed a quantitative analysis of immune cell infiltration patterns using the CIBERSORT and xCell algorithms and examined their correlations with clinical scores and model risk predictions.

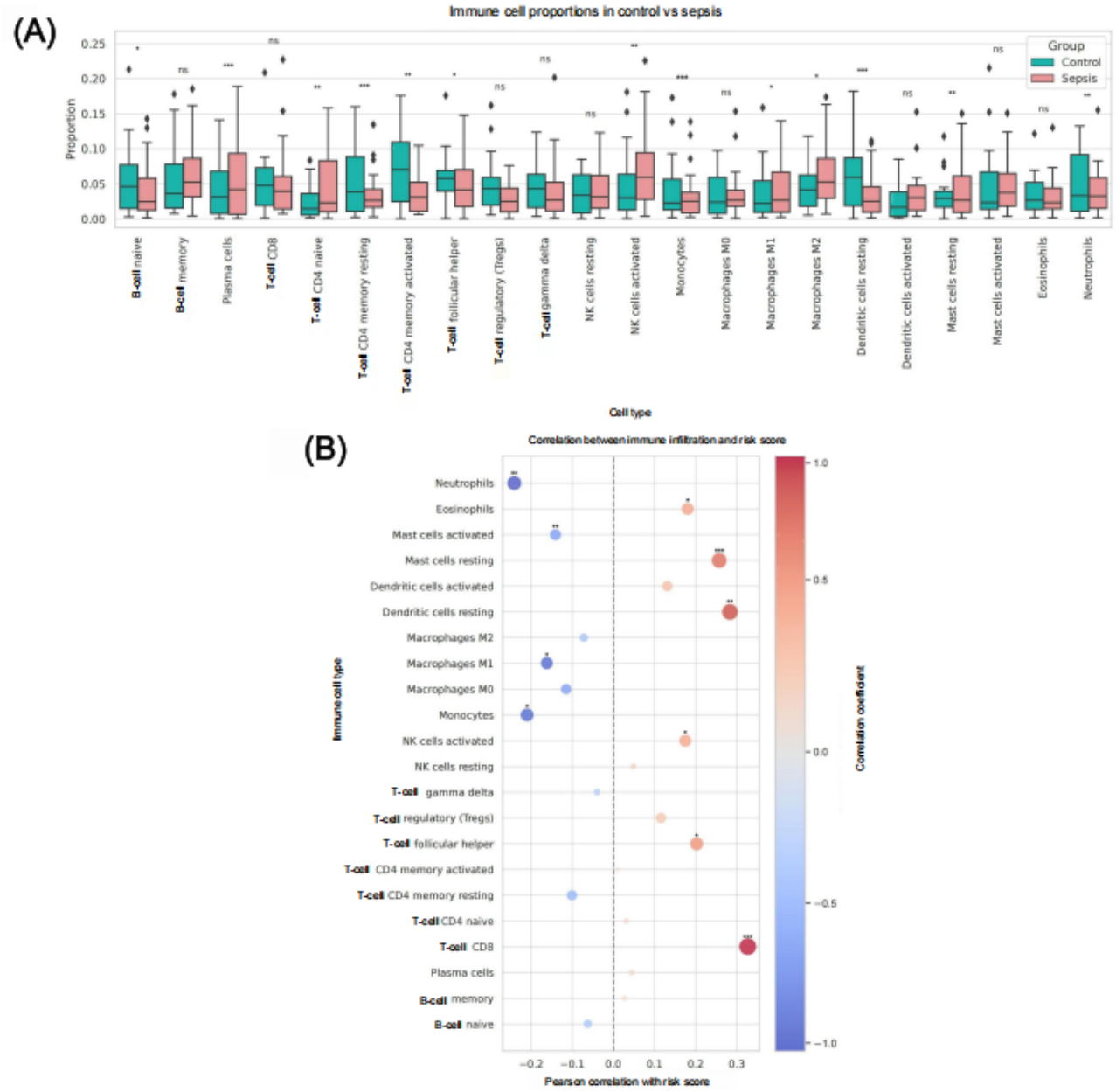
At the population level (Figure 6A), patients with sepsis showed significant changes in key immune cell subsets compared with controls. The sepsis group exhibited markedly increased infiltration of immunosuppressive and proinflammatory cells, including regulatory T cells (Tregs; P<.001), M2 macrophages (P<.001), and neutrophils (P<.001), whereas effector immune cells such as cluster of differentiation 8+ (CD8+) T cells, memory B cells, and activated natural killer cells were consistently reduced. This immunophenotypic heterogeneity suggests a coexisting state of immune activation and suppression in high-risk patients.

Correlation analysis (Figure 6B) revealed that Tregs (r=0.29) and M2 macrophages (r=0.21) were positively correlated with model risk scores, whereas CD8+ T cells (r=–0.38) showed a negative correlation. These findings indicate that these subsets may dynamically regulate disease severity and prognosis beyond their microenvironmental associations.

The immune infiltration profiling demonstrates distinct immunoreprogramming features in high-risk patients with sepsis, characterized by an “effector cell exhaustion + immunosuppressive cell upregulation” pattern that may represent a fundamental mechanism in sepsis progression.

XSL•FO
RenderX

Figure 6. Immune cell infiltration characteristics between sepsis and control groups. (A) Comparison of immune cell subset proportions between the sepsis and control groups. (B) Pearson correlation analysis between immune cell subsets and model risk scores. Statistical significance is denoted as follows: $P<.05$, $P<.01$, and $P<.001$; ns: no significant difference. NK: natural killer; Treg: regulatory T cell.



Integrated Multisource Analysis Identifies Core Candidate Genes With Regulatory Potential and Clinical Translational Value

Building upon the aforementioned DEG analysis, WGCNA construction, and immune cell correlation mining, this study further integrated model interpretability features (eg, SHAP value rankings), hub gene positions, immune factor correlations, and PPI topological properties to identify a set of highly consistent core candidate genes with regulatory potential.

By intersecting DEGs, WGCNA module genes, and known immune-related gene databases, 55 key immune regulatory candidate genes were identified. These genes exhibited outstanding performance in expression patterns, network centrality, and immune cell relationships (Figure 7A). This strategy ensured that the selected molecules were supported by multiple biological and model-based lines of evidence, enhancing their credibility and functional interpretability.

Further, the STRING database was used to construct a PPI network for the candidate genes (Figure 7B). Topological parameters (eg, node degree, betweenness centrality) were used to identify core hub genes. The results revealed that PD-L1 (CD274), indoleamine 2,3-dioxygenase 1 (IDO1), IL1B, STAT1, and CXCL10 were not only located in highly connected regions of the network but were also extensively involved in pathways such as immune checkpoint regulation, inflammatory cytokine release, antigen presentation, and T-cell exhaustion, suggesting central roles in the immune dysregulation mechanisms of sepsis.

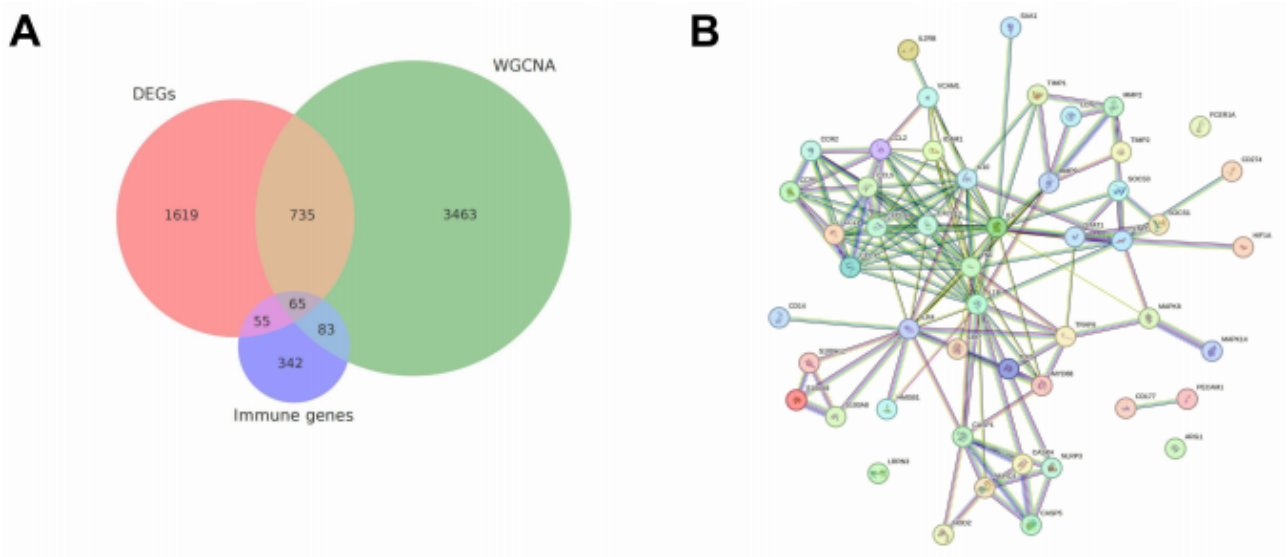
To evaluate the clinical translational potential of these candidate genes, this study further aligned them with public drug databases (eg, DrugBank, Drug-Gene Interaction Database). The results showed that several core factors had known or potential

XSL•FO
RenderX

drug-target backgrounds. The corresponding gene-drug interaction relationships were integrated using high-resolution, interactive visualizations and are provided in Multimedia Appendix 4, enabling comprehensive exploration of complex target networks. For example, PD-L1 (CD274) is the target of multiple immune checkpoint inhibitors (eg, atezolizumab, durvalumab); IDO1 inhibitors (eg, epacadostat) are under clinical investigation for various inflammatory and oncological indications; STAT1 and CXCL10 have also been reported as core regulators of the interferon pathway, with promising drug development prospects.

In summary, these candidate genes play pivotal roles in molecular regulatory networks and exhibit strong drug accessibility, making them potential biomarkers or therapeutic targets for ECF-associated CIAI and sepsis in future clinical applications.

Figure 7. Integrated multisource screening of core candidate genes and their potential drug targets. (A) Venn diagram showing the overlap among differentially expressed genes (DEGs), weighted gene co-expression network analysis (WGCNA) co-expression module genes, and immune-related genes, identifying candidate immune regulatory factors with multidimensional consistency. (B) Protein-protein interaction network constructed using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database, where node size and color represent gene connectivity and module affiliation, highlighting high-degree hub genes such as programmed death-ligand 1 (PD-L1), indoleamine 2,3-dioxygenase 1 (IDO1), signal transducer and activator of transcription 1 (STAT1), and C-X-C motif chemokine ligand 10 (CXCL10).



Sensitivity and Robustness Analysis

To verify the stability and reproducibility of the multimodal AI model's predictive performance, a series of multilevel sensitivity analyses were conducted. First, using a stricter definition of ECF—requiring either 2 or more occurrences of ECF-related diagnostic codes on different dates or a single code combined with text confirmation within 72 hours—cases were reidentified. Under this condition, the model's performance in predicting sepsis and 28-day mortality remained stable, with AUCs of 0.872 and 0.901, respectively.

Second, to evaluate the effect of data partition randomness, the training-validation-testing pipeline was repeated 5 times with different random seeds. The resulting AUC variation across runs was less than 0.02, indicating excellent training stability.

Finally, in the leave-one-modality-out validation, each modality (clinical, imaging, and transcriptomic data) was sequentially excluded. The complete multimodal model consistently outperformed any single-modality configuration, with ΔAUC values ranging from approximately 0.05 to 0.09. A summary of all sensitivity and robustness analyses is presented in Table 5, demonstrating that both the predictive performance and key feature interpretability of the model remained consistent under diverse testing conditions.

XSL•FO
RenderX

Table 5. Sensitivity and robustness analyses of the multimodal artificial intelligence model.[a]

Sensitivity/robustness setting and task	AUC[b], mean (SD)	Accuracy	Sensitivity	Specificity	ΔAUC vs base
Base model (original definition)					
Sepsis prediction	0.860 (0.011)	0.879	0.862	0.901	N/A[c]
28-day mortality	0.910 (0.009)	0.892	0.874	0.914	N/A
Stricter ECF[d] definition (≥2 codes or code + text)					
Sepsis prediction	0.872 (0.010)	0.883	0.868	0.906	0.012
28-day mortality	0.901 (0.013)	0.887	0.867	0.908	–0.009
Random-seed repeat (5 runs)					
Sepsis prediction	0.859 (0.014)	0.877	0.861	0.9	0.001
28-day mortality	0.911 (0.012)	0.89	0.873	0.913	0.001
Leave-one-modality-out: clinical only					
Sepsis prediction	0.802 (0.017)	0.834	0.819	0.861	–0.058
Leave-one-modality-out: imaging only					
Sepsis prediction	0.795 (0.020)	0.827	0.812	0.854	–0.065
Leave-one-modality-out: omics only					
Sepsis prediction	0.811 (0.018)	0.838	0.826	0.869	–0.049

[a]This table summarizes the predictive performance of the multimodal artificial intelligence model under various sensitivity and robustness conditions, including (1) a stricter definition of ECF (≥2 diagnostic codes or 1 diagnostic code confirmed by text within 72 hours), (2) repeated training with different random seeds, and (3) leave-one-modality-out validation. AUC, accuracy, sensitivity, and specificity represent the model's discriminative ability, overall accuracy, sensitivity, and specificity, respectively. ΔAUC indicates the performance change relative to the baseline multimodal model (under the original ECF definition). AUC values are expressed as mean (SD) from 5 independent runs. The results demonstrate that the model maintains stable performance and consistent interpretability across all analytical conditions.

[b]AUC: area under the curve.

[c]N/A: not applicable.

[d]ECF: enterocutaneous fistula.

Discussion

ECF-associated CIAI and sepsis represent a category of highly complex digestive system complications with extremely poor prognosis [23,24], exhibiting mortality rates as high as 30%-50% [4] and posing significant challenges for ICU management and infection control [18,25,26]. However, current clinical practice still lacks a systematic, highly sensitive, and interpretable tool for early identification and risk warning in these patients [24,27]. This unmet need prompted us to develop a multimodal AI prediction model integrating multisource data to accurately identify high-risk patients amid substantial clinical heterogeneity.

In this study, leveraging authoritative databases (MIMIC-III, eICU, TCGA, and GEO), we systematically integrated clinical indicators, abdominal imaging, and transcriptomic data to construct a DL-based multimodal prediction framework. The model demonstrated superior performance (AUC 0.89) across multiple critical tasks, significantly outperforming traditional unimodal approaches while maintaining robust generalizability in complex clinical data environments. Crucially, through interpretability algorithms (SHAP and LIME), we identified key predictive contributors, including SOFA scores, lactate levels, immune cell composition, and critical gene modules, thereby enhancing clinical understanding of model decisions and their practical utility.

Mechanistically, DEG analysis, WGCNA module identification, GO/KEGG enrichment, immune infiltration inference, and PPI network construction revealed a coexisting immunosuppressive and proinflammatory signature in patients with sepsis. Elevated infiltration of Tregs, M2 macrophages, and neutrophils, along with upregulation of immune regulatory genes (*PD-L1*, *IDO1*, *STAT1*, and *CXCL10*), suggested pivotal roles of immune exhaustion and evasion in ECF-associated sepsis. These findings provide biological validation for the model and lay the groundwork for targeted therapy exploration.

In this study, the key genes *PD-L1*, *IDO1*, *STAT1*, and *CXCL10* were identified as forming a biologically coherent immunoregulatory network, whose upregulation reflects the coexistence of immune suppression and inflammation in ECF-associated complex infections. Further mechanistic analyses revealed that this signaling axis is closely linked to intestinal barrier dysfunction and the progression of sepsis. First, PD-L1, as an immune checkpoint molecule, inhibits effector T-cell proliferation and cytotoxicity through its interaction with programmed death-1 (PD-1) in the context of persistent infection and an inflammatory microenvironment, thereby weakening mucosal defense. Studies have shown that blocking PD-L1 in sepsis models markedly reduces intestinal epithelial

XSL•FO
RenderX

permeability, restores tight junction proteins (occludin, claudin-1, zonula occludens-1), and decreases bacterial translocation [28]. Second, IDO1, the rate-limiting enzyme of the tryptophan-kynurenine metabolic pathway, when overactivated, leads to depletion of exogenous tryptophan and accumulation of kynurenine. This process suppresses T-cell activation and induces Treg expansion, deepening immune suppression. Meanwhile, kynurenine and its metabolites affect tight junction homeostasis via the aryl hydrocarbon receptor–mediated signaling pathway, exacerbating mucosal barrier disruption [29,30]. Moreover, STAT1, the core transcription factor of the interferon-γ signaling pathway, is persistently activated under infectious stress, driving monocytes/macrophages toward a mixed M1-M2 polarization phenotype. This results in an imbalanced coexistence of proinflammatory cytokine release and tissue repair. Excessive activation of STAT1 in epithelial cells decreases transepithelial electrical resistance, induces apoptosis, and disrupts tight junction integrity [31,32]. Finally, CXCL10, a critical chemokine downstream of STAT1, robustly recruits T cells and neutrophils. Excessive accumulation of these immune cells in the intestinal mucosa contributes to microcirculatory disturbance and tissue necrosis. Animal studies have demonstrated that CXCL10 blockade significantly improves survival in septic mice and reduces tissue damage and bacterial load [33]. Collectively, the "PD-L1-IDO1-STAT1-CXCL10" axis constitutes a sequential pathological loop in ECF-associated sepsis, progressing from immune checkpoint inhibition → metabolic immune reprogramming → interferon signal amplification → chemotactic inflammatory dissemination. This cascade ultimately results in intestinal barrier disruption, bacterial translocation, and systemic inflammatory amplification, representing a fundamental molecular mechanism underlying persistent complex infections and sepsis progression in patients with ECF.

Unlike previous studies relying solely on clinical scores (SOFA and APACHE II) or single indicators, our "prediction-interpretation-mechanism"-integrated strategy synergized AI modeling with bioinformatics mining, overcoming the "black-box" limitation of traditional models while balancing performance and mechanistic insights, thereby highlighting AI's innovative potential in critical infection management.

Nevertheless, this study still has several limitations: (1) Data sourced from public databases, although representative, may suffer from selection bias and recording quality issues; (2) despite excellent test-set performance, real-world multicenter validation remains insufficient; and (3) causal relationships between immune infiltration and transcriptomic signals require experimental confirmation, as current associations are inferential.

In the future, we plan to undertake the following work to facilitate the translation and implementation of this research: (1) external validation with real-world multicenter ICU cohorts to enhance model robustness; (2) dynamic framework expansion incorporating time-series data for disease progression tracking; (3) high-dimensional data integration (single-cell RNA sequencing and spatial omics) to dissect immune reprogramming features; and (4) development of PD-L1/IDO1-centered intervention prediction submodules to guide personalized therapy.

Although the AI model developed in this study integrates clinical, imaging, and transcriptomic modalities, we fully acknowledge that obtaining transcriptomic data in real-time clinical settings remains challenging. To address this limitation, the framework was designed with inherent modularity and scalability. When retrained using only the clinical + imaging (dual-modality) configuration—after removing the transcriptomic inputs—the model retained approximately 93% of its original predictive performance (AUC 0.83-0.85), demonstrating strong generalizability and adaptability. Further feature-mapping analyses revealed that several routine clinical and inflammatory markers, including CRP, IL-6, neutrophil-to-lymphocyte ratio, and lactate levels, were significantly correlated with immune module embeddings derived from the transcriptomic features of the full model. These findings suggest that such readily accessible biomarkers may serve as biologically interpretable substitutes, preserving the model's explanatory power even in the absence of omics data.

For future applications, we propose a 2-tier deployment strategy: (1) The full 3-modality model will be employed for mechanistic exploration and translational research, elucidating the biological interplay between immune pathways and imaging phenotypes. (2) The lightweight dual-modality version will be integrated into ICU clinical information systems, enabling real-time risk prediction and stratification for ECF-associated complex infections and sepsis. This "research-clinical dual framework" design balances interpretability and deployability, thereby enhancing the model's clinical practicality and multicenter scalability.

Our AI multimodal fusion model offers a novel approach for the early prediction and mechanistic dissection of ECF-associated CIAI. Its combined strengths in performance, interpretability, and biological relevance position it for clinical translation in early risk screening, intervention strategy recommendation, and decision support (Multimedia Appendix 5).

Data Availability

All data generated or analyzed during this study are included in this article or its multimedia appendices. Further inquiries can be directed to the corresponding author.

Authors' Contributions

Conceptualization: HL, JC, WX

XSL•FO
RenderX

Data curation: PL, YP
Formal analysis: JC, PL
Investigation: YP, YC
Methodology: HL, YC
Project administration: WX
Resources: PL, YP
Software: HL, YC
Supervision: WX, YC
Validation: HL, YC
Visualization: HL
Writing – original draft: HL, JC, YC
Writing – review and editing: WX, YC

Conflicts of Interest

None declared.

Multimedia Appendix 1
The collected clinical data.
[PDF File (Adobe PDF File), 117 KB - medinform_v14i1e79985_app1.pdf]

Multimedia Appendix 2
Overview of the multimodal artificial intelligence framework for predicting enterocutaneous fistula–associated complicated intraabdominal infection and sepsis, including model performance evaluation, clinical label definition and multitask learning objectives, and integrated preprocessing pipelines for clinical data, medical imaging (with quality control and multiple preprocessing strategies), and RNA expression profiles with feature standardization and construction.
[DOCX File , 767 KB - medinform_v14i1e79985_app2.docx]

Multimedia Appendix 3
Gene Ontology enrichment analysis results covering 3 categories: biological processes, cellular components, and molecular function.
[PNG File , 102 KB - medinform_v14i1e79985_app3.png]

Multimedia Appendix 4
Sankey diagram mapping candidate genes to known or potential drug targets based on DrugBank and DGIdb annotations. The width of the links represents the strength of gene–drug associations or evidence weight, revealing the potential clinical relevance of PD-L1 and IDO1 as therapeutic targets in immune regulation and anti-infective interventions.
[PNG File , 1391 KB - medinform_v14i1e79985_app4.png]

Multimedia Appendix 5
Graphical Abstract.
[PNG File , 1929 KB - medinform_v14i1e79985_app5.png]

References

1. Taggarshe D, Bakston D, Jacobs M, McKendrick A, Mittal VK. Management of enterocutaneous fistulae: a 10 years experience. World J Gastrointest Surg 2010 Jul 27;2(7):242-246 [FREE Full text] [doi: 10.4240/wjgs.v2.i7.242] [Medline: 21160881]
2. Gefen R, Garoufalia Z, Zhou P, Watson K, Emile SH, Wexner SD. Treatment of enterocutaneous fistula: a systematic review and meta-analysis. Tech Coloproctol 2022 Nov 01;26(11):863-874. [doi: 10.1007/s10151-022-02656-3] [Medline: 35915291]
3. Tuma F, Crespi Z, Wolff C, Daniel D, Nassar A. Enterocutaneous fistula: a simplified clinical approach. Cureus 2020 Apr 22;12(4):e7789 [FREE Full text] [doi: 10.7759/cureus.7789] [Medline: 32461860]
4. Ghimire P. Management of enterocutaneous fistula: a review. J Nepal Med Assoc 2022 Jan 23;60(245):93-100. [doi: 10.31729/jnma.5780]
5. Iglay K, Bennett D, Kappelman M. A systematic review of epidemiology and outcomes of Crohn's disease-related enterocutaneous fistulas. Medicine 2022 Nov 11;101(45):e30963. [doi: 10.1097/md.0000000000030963]
6. Jarczak D, Kluge S, Nierhaus A. Sepsis—pathophysiology and therapeutic concepts. Front Med 2021 May 14;8:628302. [doi: 10.3389/fmed.2021.628302]

XSL•FO
RenderX

7. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020 Jan 21;46(3):383-400. [doi: 10.1007/s00134-019-05872-y]
8. Noori I. Postoperative enterocutaneous fistulas: management outcomes in 23 consecutive patients. Annals of Medicine & Surgery 2021 May 19:66. [doi: 10.1016/j.amsu.2021.102413]
9. Metcalfe C, Kljavin N, Ybarra R, de Sauvage FJ. Lgr5+ stem cells are indispensable for radiation-induced intestinal regeneration. Cell Stem Cell 2014 Mar 06;14(2):149-159 [FREE Full text] [doi: 10.1016/j.stem.2013.11.008] [Medline: 24332836]
10. Cheng H, Guan X, Chen D, Ma W. The Th17/Treg cell balance: a gut microbiota-modulated story. Microorganisms 2019 Nov 20;7(12):583. [doi: 10.3390/microorganisms7120583]
11. Wu Y, Wang L, Ji C, Gu S, Yin Q, Zuo J. The role of α7nAChR-mediated cholinergic anti-inflammatory pathway in immune cells. Inflammation 2021 Jan 06;44(3):821-834. [doi: 10.1007/s10753-020-01396-6]
12. Tang J, Xu L, Zeng Y, Gong F. Effect of gut microbiota on LPS-induced acute lung injury by regulating the TLR4/NF-kB signaling pathway. International Immunopharmacology 2021 Feb;91:107272. [doi: 10.1016/j.intimp.2020.107272]
13. Corcoran S, O?Neill L. HIF1 and metabolic reprogramming in inflammation. Journal of Clinical Investigation 2016;126(10):3699-3707. [doi: 10.1172/jci84431]
14. Ding P, Duan J, Xue L, Liu Y. FCSSL: fusion enhanced contrastive self-supervised learning method for parallel MRI reconstruction. Phys Med Biol 2024 Oct 14;69(20):205018. [doi: 10.1088/1361-6560/ad6d28] [Medline: 39116910]
15. Hu W, Meng X, Bai Y, Zhang A, Qu G, Cai B, et al. Interpretable multimodal fusion networks reveal mechanisms of brain cognition. IEEE Trans Med Imaging 2021 May;40(5):1474-1483. [doi: 10.1109/tmi.2021.3057635]
16. Xiao J, Yu X, Meng F, Zhang Y, Zhou W, Ren Y, et al. Integrating spatial and single-cell transcriptomics reveals tumor heterogeneity and intercellular networks in colorectal cancer. Cell Death Dis 2024 May 10;15(5):326. [doi: 10.1038/s41419-024-06598-6]
17. Ballard J, Wang Z, Li W, Shen L, Long Q. Deep learning-based approaches for multi-omics data integration and analysis. BioData Min 2024 Oct 02;17(1):38 [FREE Full text] [doi: 10.1186/s13040-024-00391-z] [Medline: 39358793]
18. Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, et al. A machine learning model for accurate prediction of sepsis in ICU patients. Front Public Health 2021 Oct 15;9:754348 [FREE Full text] [doi: 10.3389/fpubh.2021.754348] [Medline: 34722452]
19. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. J Am Med Inform Assoc 2022 Jan 29;29(3):559-575 [FREE Full text] [doi: 10.1093/jamia/ocab236] [Medline: 34897469]
20. Liu S, Fu B, Wang W, Liu M, Sun X. Dynamic sepsis prediction for intensive care unit patients using XGBoost-based model with novel time-dependent features. IEEE J Biomed Health Inform 2022 Aug;26(8):4258-4269. [doi: 10.1109/jbhi.2022.3171673]
21. Peng Y, Wu Q, Liu H, Zhang J, Han Q, Yin F, et al. An immune-related gene signature predicts the 28-day mortality in patients with sepsis. Front Immunol 2023 Mar 23;14:1152117 [FREE Full text] [doi: 10.3389/fimmu.2023.1152117] [Medline: 37033939]
22. Yang Z, Cui X, Song Z. Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. BMC Infect Dis 2023 Sep 27;23(1):635 [FREE Full text] [doi: 10.1186/s12879-023-08614-0] [Medline: 37759175]
23. Fathi M, Markazi-Moghaddam N, Ramezankhani A. A systematic review on risk factors associated with sepsis in patients admitted to intensive care units. Australian Critical Care 2019 Mar;32(2):155-164. [doi: 10.1016/j.aucc.2018.02.005]
24. Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med 2020 Nov;48(11):e1091-e1096. [doi: 10.1097/CCM.0000000000004550] [Medline: 32885937]
25. Kim PJ, Attinger CE, Constantine T, Crist BD, Faust E, Hirche CR, et al. Negative pressure wound therapy with instillation: international consensus guidelines update. International Wound Journal 2019 Oct 30;17(1):174-186. [doi: 10.1111/iwj.13254]
26. Jeon K, Suh JK, Jang EJ, Cho S, Ryu HG, Na S, et al. Procalcitonin-guided treatment on duration of antibiotic therapy and cost in septic patients (PRODA): a multi-center randomized controlled trial. J Korean Med Sci 2019;34(14):e110. [doi: 10.3346/jkms.2019.34.e110]
27. Zhang D, Yin C, Hunold KM, Jiang X, Caterino JM, Zhang P. An interpretable deep-learning model for early prediction of sepsis in the emergency department. Patterns (N Y) 2021 Mar 12;2(2):100196 [FREE Full text] [doi: 10.1016/j.patter.2020.100196] [Medline: 33659912]
28. Wu Y, Chung C, Chen Y, Monaghan SF, Patel S, Huang X, et al. A novel role for programmed cell death receptor ligand-1 in sepsis-induced intestinal dysfunction. Mol Med 2016 Oct 25;22(1):830-840. [doi: 10.2119/molmed.2016.00150]
29. Ciorba MA. Indoleamine 2,3 dioxygenase in intestinal disease. Current Opinion in Gastroenterology 2013;29(2):146-152. [doi: 10.1097/mog.0b013e32835c9cb3]
30. Kearns R. The kynurenine pathway in gut permeability and inflammation. Inflammation 2024 Sep 10;48(3):1063-1077. [doi: 10.1007/s10753-024-02135-x]
31. Smyth D, Phan V, Wang A, McKay DM. Interferon-γ-induced increases in intestinal epithelial macromolecular permeability requires the Src kinase Fyn. Laboratory Investigation 2011 May;91(5):764-777. [doi: 10.1038/labinvest.2010.208]

XSL•FO
RenderX

32. Stolzer I, Schickedanz L, Chiriac MT, López-Posadas R, Grassl GA, Mattner J, et al. STAT1 coordinates intestinal epithelial cell death during gastrointestinal infection upstream of caspase-8. Mucosal Immunology 2022 Jan;15(1):130-142. [doi: 10.1038/s41385-021-00450-2]
33. Herzig DS, Luan L, Bohannon JK, Toliver-Kinsky TE, Guo Y, Sherwood ER. The role of CXCL10 in the pathogenesis of experimental septic shock. Crit Care 2014 Jun 02;18(3):R113 [FREE Full text] [doi: 10.1186/cc13902] [Medline: 24890566]

Abbreviations

AI: artificial intelligence
APACHE II: Acute Physiology and Chronic Health Evaluation II
AUC: area under the curve
BCE: binary cross-entropy
BP: biological processes
CD: cluster of differentiation
CIAI: complicated intraabdominal infection
CNN: convolutional neural network
CRP: C-reactive protein
CT: computed tomography
CXCL10: C-X-C motif chemokine ligand 10
CXR: chest x-ray
DEG: differentially expressed gene
DICOM: Digital Imaging and Communications in Medicine
DL: deep learning
ECF: enterocutaneous fistula
eICU: electronic Intensive Care Unit
FPKM: fragments per kilobase of transcript per million mapped reads
GEO: Gene Expression Omnibus
GO: Gene Ontology
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICU: intensive care unit
IDO1: indoleamine 2,3-dioxygenase 1
IL: interleukin
KEGG: Kyoto Encyclopedia of Genes and Genomes
LIME: Local Interpretable Model-Agnostic Explanations
MICE: multiple imputation by chained equation
MIMIC-III: Medical Information Mart for Intensive Care III
MRI: magnetic resonance imaging
PD-1: programmed death-1
PD-L1: programmed death-ligand 1
PPI: protein-protein interaction
RNA-seq: RNA sequencing
SHAP: Shapley Additive Explanations
SOFA: Sequential Organ Failure Assessment
STAT: signal transducer and activator of transcription
TCGA: The Cancer Genome Atlas
Treg: regulatory T cell
TYROBP: TYRO protein tyrosine kinase-binding protein
VAE: variational autoencoder
WGCNA: weighted gene co-expression network analysis
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by A Benis; submitted 02.Jul.2025; peer-reviewed by NC Chen, O Pitsun; comments to author 07.Oct.2025; revised version received 15.Jan.2026; accepted 17.Feb.2026; published 30.Apr.2026.

Please cite as:
Li H, Chen J, Lin P, Pan Y, Cao Y, Xie W
Enterocutaneous Fistula–Associated Sepsis and Mortality: Development and Validation of a Multimodal Artificial Intelligence Prediction Model
JMIR Med Inform 2026;14:e79985
URL: https://medinform.jmir.org/2026/1/e79985
doi:10.2196/79985
PMID:

©Hui Li, Jing Chen, Peijun Lin, Youmei Pan, Yawen Cao, Wenfeng Xie. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing a Thai User Interface Terminology for Systematized Nomenclature of Medicine Clinical Terms Implementation in Primary Care: Cross-Sectional Content Coverage Analysis

Nat Tangchitnob[1], MD; Wanchana Ponthongmak[1], PhD; Boonchai Kijsanayotin[1,2], MD, PhD; Oraluck Pattanaprateep[1], PhD; Sithakom Phusanti[3], MD; Pongsakorn Atiksawedparit[3], MD, PhD; Kamonporn Suwanthaweemeesuk[3], MSc, MD; Jirayus Siangfu[4], MD; Gareth J McKay[5], PhD; John Attia[6], MD, PhD; Ammarin Thakkinstian[1], PhD

[1]Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Rama 6 Road, Ratchathewi, Bangkok, Thailand

[2]Thai Health Information Standards Development Center, Health Systems Research Institute, Ministry of Public Health, Bangkok, Thailand

[3]Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[4]Chulabhorn Hospital, Chulabhorn Royal Academy, Bangkok, Thailand

[5]Centre for Public Health, Queen's University Belfast, Belfast, United Kingdom

[6]Centre for Clinical Epidemiology and Biostatistics, School of Medicine and Public Health, University of Newcastle Australia, Newcastle, Australia

Corresponding Author:
Wanchana Ponthongmak, PhD
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Rama 6 Road, Ratchathewi, Bangkok, Thailand

Abstract

Background: Primary care in Thailand often uses mixed Thai-English free-text documentation for diagnoses and clinical problems, limiting standardization, interoperability, and secondary data use. Clinical terminologies like Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), a comprehensive reference terminology, can bridge this gap through the use of structured clinical data. Developing and mapping a local user interface terminology (UIT) is one of the key strategies for implementing SNOMED CT in real-world clinical settings.

Objective: This study aimed to develop a Thai UIT derived from frequently used terms in real-world primary care practice, map these terms to SNOMED CT concepts, and evaluate the extent of concept coverage.

Methods: Frequently used clinical terms were extracted from outpatient medical records from the family, emergency, and internal medicine departments using a customized tokenization method, N-gram analysis, and expert review. This process yielded 2054 Thai-specific terms. All terms were normalized and mapped to SNOMED CT through manual expert-driven and semiautomated tools. Unmapped terms were subsequently analyzed to identify mapping barriers and solutions.

Results: Of the 2054 Thai-specific terms, 2012 were successfully mapped to 2041 (97.98%) SNOMED CT concepts, including 1781 (85.50%) fully, 123 (5.90%) broader, 56 (2.69%) narrower, 81 (3.89%) inexact mappings, and 42 (2.02%) remained unmapped. Most mappings were one-to-one (1984), with 28 terms mapped to multiple concepts (57), covering 1486 unique SNOMED CT concepts. The remaining 42 unmapped terms were mostly due to culturally specific expressions or concepts not yet represented in SNOMED CT. These were categorized for potential postcoordination, exclusion, or national extension development.

Conclusions: This study demonstrates the feasibility of developing a Thai UIT mapped to SNOMED CT and describes mapping challenges. The resulting UIT enhances semantic clarity in clinical documentation and supports better interoperability, clinical decision-making, and health data analytics within Thailand's health care system.

(JMIR Med Inform 2026;14:e80039) doi:10.2196/80039

KEYWORDS

standards and interoperability; Systematized Nomenclature of Medicine Clinical Terms; SNOMED CT; eHealth infrastructure; clinical terminology; natural language processing; concept coverage

XSL•FO
RenderX

Introduction

Health care provision is a knowledge-intensive industry, but the quality and accessibility of health care data often fail to adequately meet these needs [1]. This is partly due to the data collection process, which commonly takes place in a fast-paced, high-pressure point-of-care environment [2], leading to the generation of data sufficient to meet the immediate clinical tasks but often lacking the completeness, standardization, and structure necessary for secondary use, for example, research, that drives health care improvement [3,4]. This limitation may compromise patient safety, care quality, and the ability to generate actionable insights [2,5]. To harness the full potential of health care data, it is essential to prioritize structured data collection that supports timely and effective use across the health care ecosystem [3,4]. Standardized clinical terminology, structured representations of clinical concepts understandable to both humans and computer systems, is an essential technology improvement required for developing reusable data that enables interoperable health care delivery [6].

Because clinical terminology must be comprehensive and support multilingual use, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) has been widely adopted internationally as a reference terminology standard that is indispensable for standardized information exchange across clinical care and research settings [7]. SNOMED CT adopts a concept-oriented structure; each unique clinical idea is represented by a concept identified by a language-independent, machine-readable unique identifier.

SNOMED CT consists of 3 core components, including concepts, descriptions, and relationships. Concepts represent the underlying clinical meaning; descriptions provide human-readable terms in multiple languages; and relationships define how concepts are linked within the ontology. A single concept can have multiple descriptions, enabling flexible, precise, and multilingual representation of clinical ideas. Relationships form a poly-hierarchical graph structure to express, distinguish, and connect each clinical concept, maximizing practicality while minimizing ambiguity [8]. For example, the concept “Nodule of lung (disorder)” can be described by multiple terms across languages (eg, English, French, and Swedish), and its meaning is defined through relationships as the child of both “Lesion of lung (disorder)” and “Lung mass (finding),” the associated morphology “Nodule (morphologic abnormality),” and the finding site “Lung structure (body structure).” This structure enables consistent clinical data representation and supports semantic interoperability. Successfully implementing SNOMED CT not only improves data quality and standardization of clinical data but also enhances the quality of health care through more efficient data use for both clinical and research applications [9,10]. While SNOMED CT offers a robust, language-independent conceptual framework, it does not automatically capture the variability of clinical expressions used in routine practice, particularly in multilingual or mixed-language documentation environments.

User interface terminology (UIT) refers to the set of terms clinicians use during routine clinical documentation. It consists of user-friendly and context-specific expressions that reflect the natural phrasing and cognitive patterns of clinical practice [11]. UITs are typically designed to be used by a defined user group and certain settings, prioritizing efficiency and usability [12]. Within the Open Medical Record System (OpenMRS; OpenMRS Inc) ecosystem, many implementations rely on Columbia International eHealth Laboratory (CIEL), an openly maintained UIT that serves as a common starting point for clinical data capture in OpenMRS deployments [13]. CIEL supports user-friendly clinical data capture while providing mappings to international reference terminologies (including SNOMED CT) [14], enabling the harmonization of locally meaningful data for use with global standards, facilitating secondary data use, interoperability, and research usability [3,4]. Combining the advantages of both terminologies leverages their complementary strengths, enhancing communication accuracy, data quality, and adoption among clinicians, while providing semantic precision, computability, and cross-system interoperability [10,11]. Together, they facilitate the standardization of a clinical narrative into more structured data [15,16].

Despite its advantages, implementing SNOMED CT presents challenges, particularly in non–English-speaking contexts [7,10,17]. Studies investigating the performance of SNOMED CT in multilingual contexts report 2 important indicators. Concept coverage refers to the proportion of source clinical ideas that can be represented by a concept, while term coverage refers to the proportion of source terms included in the SNOMED CT description, reflecting comprehensiveness of concept and the usability of descriptions, respectively [18]. For example, European-based studies demonstrate varying degrees of success in translation and implementation efforts. A fully translated Swedish version of SNOMED CT achieved good concept coverage (87%) but lower term coverage (47%), reflecting limited alignment with real-world clinical language [19]. In contrast, a German UIT without the use of SNOMED CT reached comparable term coverage (46%) despite the lack of codes in some key semantic areas, such as procedures, devices, or genes [16,19]. These findings suggest that constructing or adopting language-specific UITs, linked to SNOMED CT’s richer semantics and logical structure, is urgently needed and more effective than simply relying on direct translation [10,15].

Thailand became a member of SNOMED International in 2022, signaling a national commitment to adopt SNOMED CT as the reference terminology for addressing the country’s interoperability and health care data quality challenges [20]. Since joining, SNOMED CT codes have been incorporated into the national reporting and reimbursement datasets for diagnostic data. In clinical practice, a few health care institutes have begun integrating SNOMED CT’s abundant descriptions into their diagnosis entry systems. This approach complements the previous process, in which physicians manually selected *International Classification of Diseases (ICD)* codes. However, the potential benefits of adopting SNOMED CT (eg, improved data granularity, semantic interoperability, and enhanced secondary data use) have yet to be fully realized across the country [9,10]. A prior study in Thailand attempted to auto-map

XSL•FO
RenderX

Thai discharge summary terms to *ICD* codes through SNOMED CT descriptions but excluded all non-English characters, limiting its ability to interpret locally used abbreviations and mixed-language expressions common in Thai clinical documentation [21].

This study was conducted with the following 3 main objectives. First, to develop a Thai-language UIT derived from real-world clinical practice. Using a dynamic corpus of clinical texts, we apply a term collection methodology that integrates clinical expertise with natural language processing (NLP) techniques. The resulting UIT is expected to facilitate and support standardized data entry, enhance data quality, and improve user acceptance of SNOMED CT within Thailand's health care system.

Second, to evaluate the content coverage of SNOMED CT, defined as the proportion of clinical ideas in the Thai context that can be represented through international reference terminology.

Third, to perform mapping analysis of partially mapped and unmapped terms in order to identify potential areas where SNOMED CT or the Thai national extension may need further refinement to better represent Thai health care data.

Methods

Data Source

Data used for the development of the UIT were retrospectively retrieved from the chief complaint and present illness sections of electronic medical records (EMRs) at Chakri Naruebodindra Medical Institute, Ramathibodi Hospital, Mahidol University, covering the period from January 2019 to April 2023. Data were obtained from outpatient clinics in the family medicine (FM) and emergency departments, representing primary care settings. To enhance lexical diversity and coverage, additional data were obtained from the high-volume internal medicine (IM) department. Records lacking both the chief complaint and present illness sections were excluded from the study.

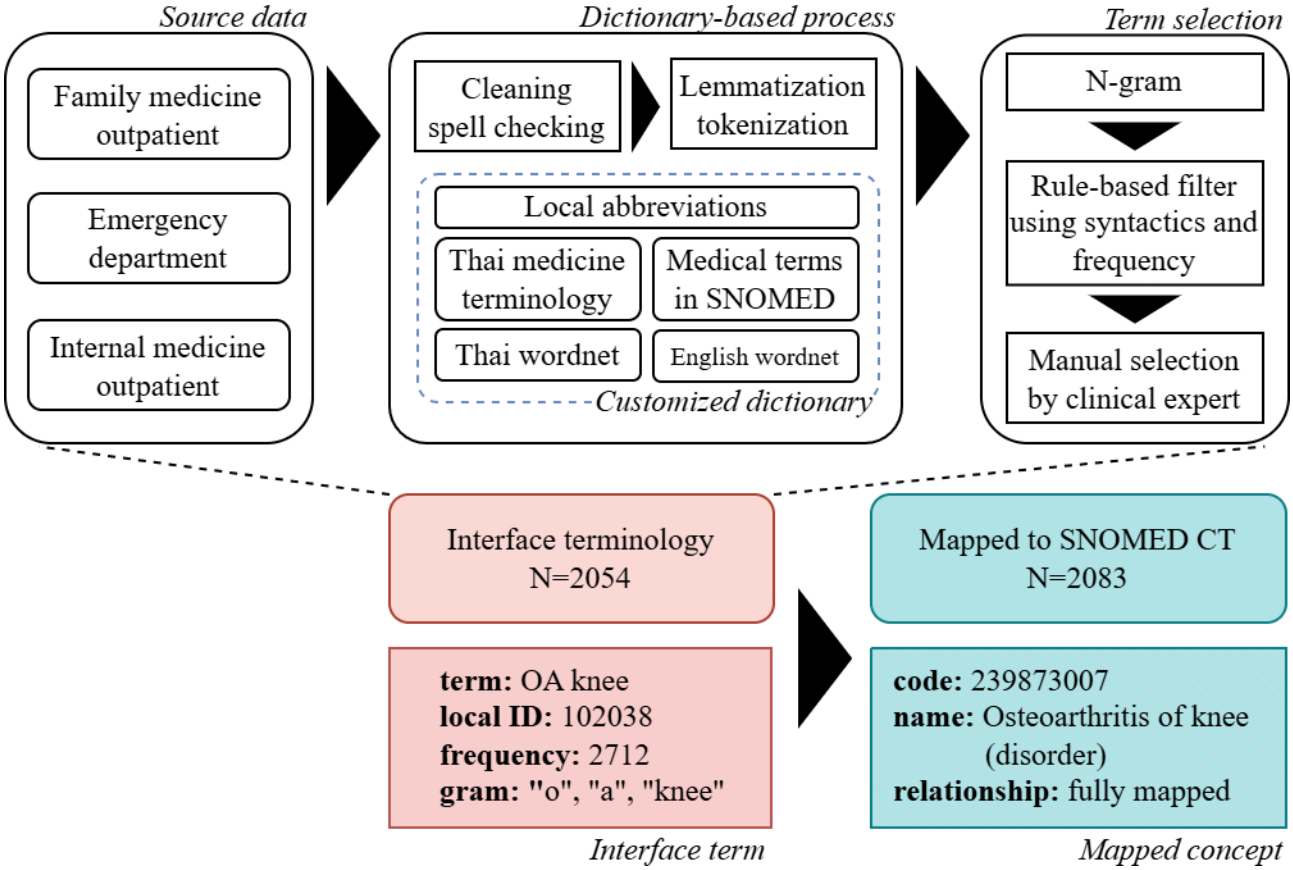
UIT

In this study, UIT was defined as a set of clinically relevant terms used by clinicians to express their clinical ideas within the EMRs. These terms can be presented in either the semistructured or unstructured free-text entries. To qualify as UIT, the terms must contain adequate meaning, clarity, and clinical specificity to support accurate documentation and enable secondary uses, as assessed by medical doctors (eg, " " [chest pain], " " [vomiting blood], " " [ischemic heart disease]), whereas terms that were vague or noninformative (eg, "1 pta" [1 month prior to admission] or "without evidence of") were not considered UIT.

The content of the UIT was harvested from the free-text documents produced during actual patient encounters. Clinician-entered free-text data (chief complaint and present illness) from the EMRs were preprocessed and transformed into clinical terms using NLP techniques, including cleaning, tokenization using our customized method, and N-gram generation. Term frequencies were then calculated, and terms meeting predefined threshold criteria were included in the UIT. The overall process is illustrated as follows (see Figure 1):

XSL•FO
RenderX

Figure 1. Summary overview of the mapping process. Collecting data from all sources (left upper box). Turning free text to N-gram using a customized dictionary from 5 sources (middle upper box). Interface terms selected by term frequency and experts (right upper box). Indexing the interface terms (pink). Mapping to the semantically equivalent SNOMED CT concept (blue). OA: osteoarthritis; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Text Preprocessing and Tokenization

Free-text data were preprocessed using the Python library for Thai NLP (ie, PyThaiNLP version 4.0.2) [22], which normalized texts by applying case normalization, correcting misplaced Thai tone marks and vowels, removing punctuation, and eliminating duplicate entries (eg, repeated sentences from prior visits). This process removed approximately 11.4% of the total characters, improving data consistency and validity. The code used for implementing text preprocessing and tokenization is provided in the GitHub repository [23].

Because Thai is a continuous-script language, which has no word and sentence boundaries (eg, “ ” [ischemic heart disease] consists of 4 lexical units: “ ”(disease), “ ”(heart), “ ”(deficiency), “ ”(blood)), standard Thai tokenizers trained on general Thai corpora perform poorly, particularly for medical text [24]. We therefore used a customized maximum matching algorithm, a greedy approach that segments text by repeatedly choosing the longest possible valid term from a dictionary that matches the current position. The dictionary was expanded to include medical jargon, drug names, and local abbreviations to optimize tokenization accuracy for clinical narratives.

Formally, given an input string S and a dictionary D, at each position j, the algorithm identifies all candidate substrings of S beginning at j that appear in D, and selects the candidate with the maximum length. This substring is then appended to the output sequence of tokens, and the process continues from the next unmatched character until the entire string is segmented.

This left-to-right longest-match strategy enables more accurate tokenization of Thai medical terms.

The custom dictionary D integrated five sources (Figure 1) (1) English WordNet; (2) Thai WordNet; (3) SNOMED CT descriptions (providing English clinical terms and medical jargon); (4) Thai medicine terminology (the national medicine database contains drug ingredients, generic names, trade names, and herbal medicine); and (5) a manually curated abbreviation dictionary (Multimedia Appendix 1), comprising abbreviations with ≥2 capital letters occurring >100 times in the dataset. These abbreviations were reviewed by physicians, validated using 5 random samples of medical records, and mapped to SNOMED CT. This custom dictionary was also used to spell-check before tokenization, ensuring that incorrect or nonstandard spellings were corrected before segmentation (Figure 1).

The N-grams of length 1 - 5, representing continuous sequences of N tokens (or grams), were generated to identify meaningful patterns and construct candidate terms or phrases.

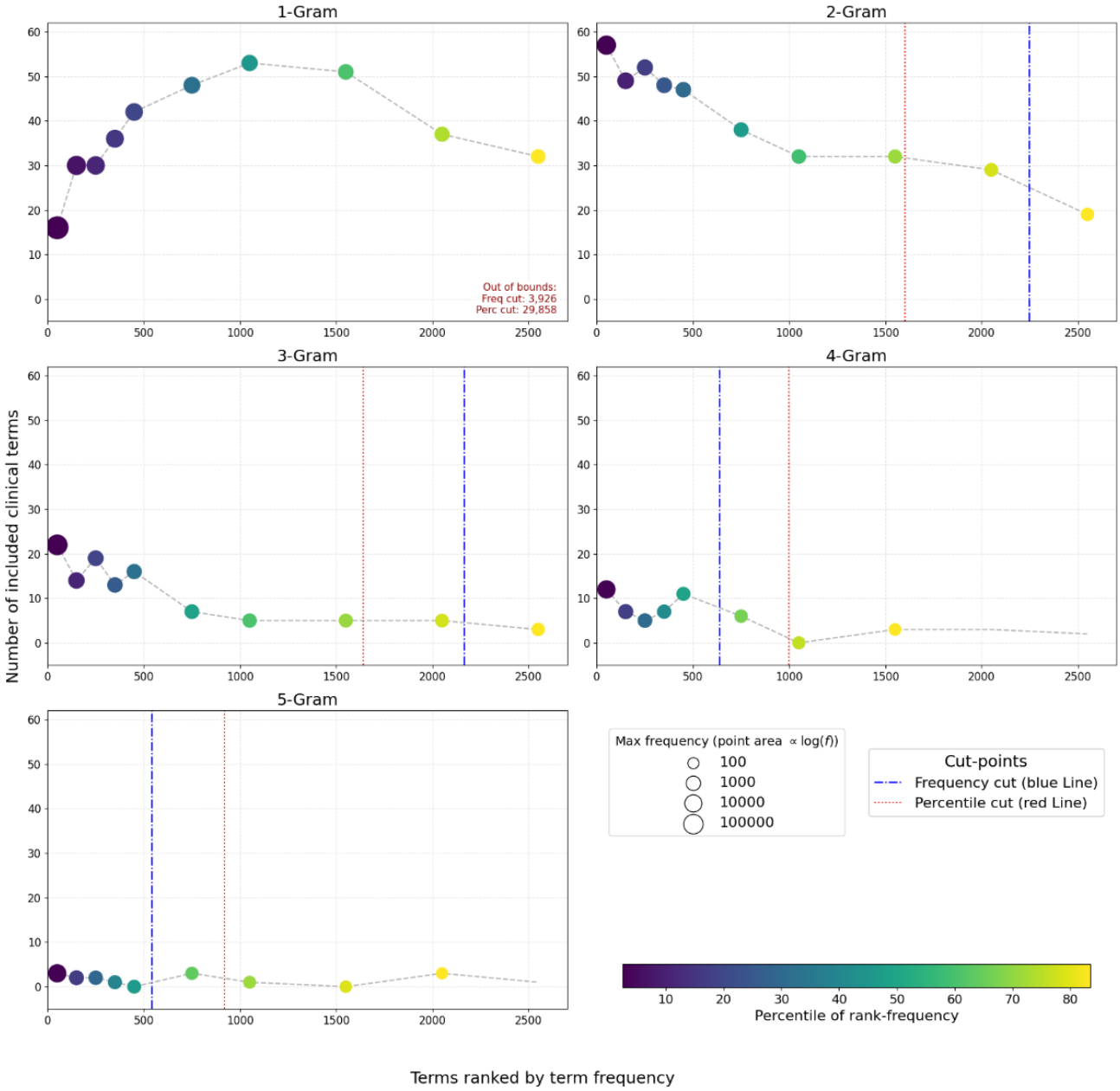
For example, words contained unigram to 5-grams were: “ ” (fever) for unigram, “ ” (abdominal pain) for bigram, “ ” (oral cavity wound) for trigram, “ ” (both legs are swelled) for 4-gram, and “ ” (abnormal bleeding from vagina) for 5-gram. Syntactically irrelevant N-grams (ie, those containing >50% stop words, punctuation,

XSL•FO
RenderX

or numbers; or terms beginning or ending with stop words or prepositions) were excluded. The remaining N-grams were ranked by frequency and carefully reviewed by experienced clinicians. Clinically relevant terms were defined as those that conveyed sufficient meaning for documenting patient conditions.

In the first round, medical doctors reviewed intervals of ranked lists (Multimedia Appendix 1), marked the terms that are clinically meaningful and appropriate for representing as structured data, and selected the terms by applying the following inclusion criteria: (1) frequency rank within the top 70th percentile or (2) absolute frequency ≥300 occurrences, thresholds that reflected diminishing returns in identifying clinically relevant terms (Figure 2). After the second round, 2054 terms were selected from the pool of 10,260 candidate terms for inclusion in the interface terminology, representing commonly used clinical expressions across EMRs.

Figure 2. Numbers of included clinical terms per interval of rank-frequency categorized by the N of N-gram. Blue vertical lines indicate the cut point where term frequency was 300. Red vertical lines indicate the cut point where the percentile of rank-frequency reaches 70.



Mapping to SNOMED CT

Each UIT term was mapped to a semantically equivalent SNOMED CT concept using Snap2SNOMED, an open-source, publicly available mapping tool developed and maintained by SNOMED International. The tool provides key functionalities, including a built-in SNOMED CT browser, versioning, and automated lexical mapping (applicable to English-language terms only). Because automated mapping requires English input, 946 Thai terms were translated into English using Google Translate (Google LLC), and these translated terms were used as source terms for mapping. The automated mapping results were manually reviewed and validated by 2 clinical domain experts (NT and JS). To ensure semantic equivalency to the original Thai terms, 5 clinical usage contexts were supplied to the mappers for each term, selected at random from different patients. Manual mappings were guided by semantic equivalence

XSL•FO
RenderX

criteria recommended in the Snap2SNOMED User Guide, prioritizing alignment for meaning, clinical intent, and hierarchical appropriateness. Disagreements between mappers were resolved through iterative discussion and consensus-based remapping.

Mapping results were classified into three categories, as illustrated in Figure 1: (1) fully mapped: SNOMED CT concepts that fully captured the semantics of the source term (eg, "valvular heart disease" mapped to "Heart valve disorder [disorder]"); (2) partially mapped: SNOMED CT concepts that were close but not fully semantically equivalent to the source term (eg, "home SBP" mapped to the narrower concept "Average home systolic blood pressure [observable entity]"); and (3) unmapped: source terms that lacked sufficient representation in existing SNOMED CT concepts.

A single source term could be potentially mapped to multiple SNOMED CT concepts (referred to as one-to-many mapping), resulting in more mapping pairs than source terms. This arose from one of two scenarios (1) homonyms, terms carried multiple distinct meanings, each represented by different SNOMED CT concepts following guidelines from SNOMED CT managing conventions and (2) ambiguous terms explicitly excluded from SNOMED CT due to insufficient contextual specificity (eg, "fundus" referring to either a hollow organ apex or the eye's interior surface). These ambiguous terms were conservatively classified as unmapped to preserve terminological clarity and precision.

Partially mapped pairs were further classified into broader, narrower, and inexact relationships to quantify mapping quality. The broader referred to the SNOMED CT concept, which encompasses the meaning of the source term. Narrower referred to when the SNOMED CT concept represented only part of the source term's meaning. In cases where semantic equivalence required postcoordination with multiple SNOMED CT concepts, we selected a single precoordinated concept in the same hierarchy as the narrower one. For example, "nodule at right upper lung" was mapped to the broader concept "Nodule of lung (disorder)" rather than combining a finding with "Upper lobe of right lung (body structure)." Inexact referred to mappings where the best available SNOMED CT concept differed in lexical and semantic characteristics but provided a clinically relevant approximation. For example, "over-the-counter drug use" was inexactly mapped to "Patient on self-medication (finding)."

Mapping Analysis

Terms unable to be mapped by mappers were discussed and categorized into either partially mapped or unmapped terms, including ambiguous terms and terms for which no semantic equivalency was found in SNOMED CT. For the terms that can be represented using postcoordination of 2 or more concepts from the same hierarchy, for example, " " (sweating and palpitation), we decided to categorize them as unmapped terms for further analysis. Partially mapped and unmapped terms were analyzed to assess their potential for practical implementation within our research team. Specifically, the analysis examined whether these terms could be adequately represented through SNOMED CT postcoordination or should be proposed for inclusion in an updated national edition of SNOMED CT. This evaluation helped identify specific content areas where concept expansion would better align the terminology with local clinical practice and ensure comprehensive coverage in the national edition.

Ethical Considerations

The study protocol was reviewed and approved by the Ramathibodi Human Research Ethics Committee (Approval number COA.MURA2023/469). The informed consent was waived due to the retrospective nature of the dataset and minimal risk to participants. All personally identifiable information (eg, names, surnames, dates of birth, addresses, hospital numbers, and visit dates) was removed from the dataset prior to analysis to ensure privacy and confidentiality.

Results

Dataset Characteristics

A total of 1,88,468 documents were initially retrieved from the EMRs during the study period. After excluding records with missing data, 1,83,557 documents from 53,142 patients were included in the analysis to develop the interface terminology. Characteristics of the 53,142 patients are described in Table 1. IM patients were the oldest (mean 59.3, SD 16.9), followed by FM (mean 54.4, SD 15.2), and emergency department (mean 43.4, SD 28.0). The percentage of males was highest in emergency (44.3%), followed by IM (39.7%) and FM (36.94%).

At the document level, the IM department contributed the largest volume (1,18,382/1,83,557, 64.5%), followed by the emergency (34,829/1,83,557, 19.0%) and FM departments (30,346/1,83,557, 16.5%). The median number of documents per patient (IQR) was 2 (1-5) for IM, 1 (1-2) for emergency, and 1 (1-3) for FM (Table 1).

Text-level analysis of 1,83,557 documents yielded 14,946,559 total words. IM contributed the majority (10,471,768/14,946,559, 70.1%), followed by emergency (2,750,896/14,946,559, 18.4%) and FM (1,723,895/14,946,559, 11.5%). Median (IQR) words per document were 107 (62-175) for IM, 74 (44-125) for ER, and 60 (38-99) for FM. Thai words were varied by department. Emergency department documents had the highest proportion of Thai words (1,752,756/2,750,896, 63.7%), followed by FM (1,016,914/1,723,895, 59.0%), while IM contained the lowest proportion (4,154,255/10,471,768, 39.7%).

XSL•FO
RenderX

Table . Characteristics of data from the 3 outpatient departments.

Characteristics	FM[a]	ER[b]	IM[c]	Total
Patient-level				
Number of patients (n)	12,128	22,878	29,985	53,142
Age (years), mean (SD)	54.37 (15.21)	43.42 (27.97)	59.31 (16.91)	53.98 (15.42)
Male (sex), n (%)	4480 (36.94)	10,137 (44.31)	11,892 (39.66)	21,821 (41.06)
Document level				
Number of documents (n)	30,346	34,829	1,18,382	1,83,557
Documents per patient, median (IQR)	1 (1-3)	1 (1-2)	2 (1-5)	2 (1-4)
Text-level				
Unique words (n)	20,070	25,055	33,286	41,755
Unique Thai words, n (%)	10,475 (52.19)	12,647 (50.78)	15,216 (45.71)	20,407 (48.87)
Total words (n)	1,723,895	2,750,896	10,471,768	14,946,559
Total Thai words, n (%)	1,016,914 (58.99)	1,752,756 (63.72)	4,154,255 (39.67)	6,923,925 (46.32)
Word per document, median (IQR)	60 (38 - 99)	74 (44 - 125)	107 (62 - 175)	81 (43 - 141)

[a]FM: family medicine.

[b]ER: emergency room.

[c]IM: internal medicine.

Coverage

A total of 2054 frequently used clinical source terms met the inclusion criteria and were integrated into the interface terminology. Of these, 2012 terms were successfully mapped to 2041 SNOMED CT concepts. Most terms were mapped one-to-one (1984 terms), while 28 terms were mapped to 57 multiconcepts through one-to-many mappings, yielding a total of 2083 entries while covering 1486 unique SNOMED CT codes. The final mappings are provided in Multimedia Appendix 2. Mapping characteristics are as follows:

- Language: English (1140 out of 2083 terms [54.73%]) and Thai (943 out of 2083 terms [45.27%]).
- N-gram: unigrams (1098 terms), bigrams (741 terms), trigrams (189 terms), 4-grams (45 terms), and 5-grams (10 terms).
- Equivalency: 1781 (85.50%) fully mapped terms, 123 (5.90%) mapped to broader concepts, 56 (2.69%) mapped to narrower concepts, and 81 (3.89%) inexactly mapped terms (Table 2). The remaining 42 terms (2.02%) could not be mapped due to culturally specific expressions or clinical concepts not yet represented in any SNOMED CT concept.
- Overall, the concept coverage was 85.50%. Combining fully and partially mapped terms yielded an overall concept coverage of 97.98%.

Term coverage is defined as the percentage of local clinical terms that can be accurately mapped to an existing SNOMED CT description. Among 1140 English terms, 656 terms (31.49%) had direct lexical matches with English SNOMED CT descriptions, including simple variations of writing, for example, “wound dressing” versus “Dressing wound.” The remaining 484 English terms (23.24%) did not have an exact lexical match and were still mappable through contextual interpretation, such as local abbreviations (eg, “f/u” → “Follow-up visit [procedure]”), or variation in writing style (eg, “contact tb” → “Exposure to tuberculosis [situation]”). As expected, no Thai terms were considered direct lexical matches because SNOMED CT descriptions are not yet translated into Thai.

Table . Concept coverage of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Thai interface terminology.

Concept equivalency	English	Thai	Total
Fully mapped, n (%)	973 (46.71)	808 (38.79)	1781 (85.50)
Partially mapped, n (%)	140 (6.72)	120 (5.75)	260 (14.48)
Broader	65 (3.12)	58 (2.78)	123 (5.90)
Narrower	38 (1.82)	18(0.86)	56 (2.69)
Inexact	37 (1.78)	44 (2.11)	81 (3.89)
Unmapped, n (%)	27 (1.30)	15 (0.72)	42 (2.02)
Total, n (%)	1140 (54.73)	943 (45.27)	2083 (100)

XSL•FO
RenderX

Mapping Analysis

Mapping analysis of the 260 partially mapped terms revealed linguistic and semantic challenges. All 123 broader (5.90%) and 56 narrower (2.69%) pairs comprised of concepts differed in granularity, where terms belonged to the same semantic domain, but they were at different levels of clinical detail or abstraction. On the other hand, inexact terms arose when clinical intention aligned, but lexical or cultural expression differed. For example, “ ” (sleep with multiple pillows) was used to indicate the situation when a patient cannot sleep lying down due to shortness of breath; it was inexactly mapped to “Orthopnea (finding)” and “ (taking medicine regularly)” and its variations were the patient-reported concept “Drug compliance good (finding).”

We also observed that culture and linguistic characteristics contributed to mismatches. Some Thai expressions convey complex clinical scenarios through idiomatic phrasing, complicating alignment with SNOMED CT. For example, “over-the-counter drug use” in the Thai context often implies access to medications that would require prescriptions in many other countries. After deliberation in our team, although this was mapped to “Patient on self-medication (finding),” this concept does not fully capture its implications within Thailand’s health care system.

Forty-two unmapped terms (2.02%) were further categorized into 3 categories based on their characteristics (see Table 3).

Ambiguous terms (23/42, 54.76%) violated SNOMED CT naming conventions and required additional contexts for appropriate representation (eg, “lobe,” “node,” and “septum”).

Multiple clinical concepts (10/42, 23.81%) represented multiple distinct findings or events (eg, “s1 s2” [heart sounds], “ ” [sweating and palpitation], “ ” [fall and head collision]). Many terms reflect Thai-specific linguistic constructions and could potentially be addressed through postcoordination.

Clinically meaningful terms without SNOMED CT equivalents (9/42, 21.43%). Some terms were (locally) nationally specific (eg, “ ” [fear of heights or a feeling of butterflies in the stomach]) while others were universally applicable but absent from SNOMED CT (eg, “ ” [antipyretic medicine] and “DOI” [date of infection]). The terms were exhaustively searched using both lexical and graph properties of SNOMED CT through the use of SNOMED CT Expression Constraint Language.

Table . Categories for unmapped concepts.

Characteristics	Ambiguous terms	Multiconcept terms	No concept terms
Value, n (%)	23 (54.76)	10 (23.81)	9 (21.43)
Further intervention	Exclusion from terminology or addition of contextual information	Postcoordination or separation to multiple distinct concepts	Creation of new concept
Examples	• Lobe • Node • Septum	• S1S2 (Heart sounds) • (sweating and palpitation) • (fall and head collision)	• DOI (Date of Infection) • (Antipyretic medicine) • (fear of heights or a feeling of butterflies in the stomach)

Discussion

Principal Findings

This study developed a multilingual UIT derived from real-world clinical documents across the FM, emergency, and IM departments and evaluated its alignment with SNOMED CT. The resulting UIT reflects a nearly balanced linguistic composition (54.49% English terms and 45.51% Thai terms). Consistent with the bilingual nature of Thai medical records. SNOMED CT demonstrated strong concept coverage with 85.50% full and 97.98% full plus partial coverage, confirming its suitability as a reference terminology even in the setting where nearly half of the source terms are non-English. However, term coverage remained modest (31.49%) because SNOMED CT descriptions have not yet been translated into Thai, making exact lexical matches possible only for English-language terms.

The source data also showed different patterns in clinical documentation practices. For instance, IM generated the highest volume of text but had the lowest proportion of Thai words, consistent with its more technical and specialty-oriented clinical activities. In contrast, emergency and FM documents contained more Thai narrative language, which may reflect more direct patient-provider interaction, brief encounters, and patient-reported symptoms. These differences emphasize the need for UITs that can accommodate both biomedical language and colloquial Thai expressions.

We also found that only a small proportion of terms remained unmapped (2.02%), underscoring that the majority of local expressions can be integrated into SNOMED CT without major structural changes. These relatively small, unmapped concepts may help identify where efforts should be strategically concentrated at the national level. Moreover, the moderate-term coverage (31.49%), driven entirely by English lexical matches, reinforces the need for localized terminology tools and translation strategies if Thailand aims to scale SNOMED CT adoption nationwide.

Comparison With Previous Work

Despite differences in term extraction methods, particularly our focus on commonly used clinical terms and the use of a more recent SNOMED CT release, our findings generally align with previous evaluations of SNOMED CT concept coverage. Prior studies have reported a strong preference for high coverage in

XSL•FO
RenderX

the English-language international edition (86% fully mapped and 92% partially mapped) and similar strong performance in the fully translated Swedish edition when applied to Swedish clinical datasets (87% fully mapped and 92% partially mapped) [19]. In comparison, our findings showed 85.50% full and 97.98% full plus partial coverage, demonstrating that SNOMED CT remains robust even in bilingual and partially non-English documentation environments.

Comparable evidence has also been observed in low- and middle-income country settings. For example, an evaluation from Argentina reported 83% coverage based on user feedback during real-world implementation [12]. Other specific domains targeting nationwide data collection and surveillance objectives have very high coverage, 99% precoordinated and 100% postcoordinated concept coverage for AIDS-defining illnesses in Kenya [25], and high correspondence for tuberculosis-related terminology, although the latter did not include a formal concept coverage analysis [14].

In Thailand, a prior study mapped Thai discharge summary notes to SNOMED CT but removed non-English content and did not report concept coverage [21]. Our study extends this work by incorporating both English and Thai clinical terms from 3 departments, performing concept coverage, term coverage, and mapping analysis, and examining implications for national UIT development. This provides a more comprehensive and representative assessment of SNOMED CT's applicability for Thai primary care and hospital settings.

Partially mapped terms frequently reflect homonyms (eg, " " meaning both abdominal colic and a cigarette unit), mismatches in granularity (eg, "hypoglycemic symptom" mapped to hypoglycemia [disorder]), or culturally shaped usage patterns. These challenges are consistent with those reported in other multilingual SNOMED CT mapping studies [10,26]. In the Thai language, many biomedical concepts and terms are not readily expressed in Thai. Unlike languages that share linguistic roots (eg, English, French, and Swedish), Thai lacks direct lexical equivalents for some medical terms, creating communication barriers between clinicians, patients, and even health care providers. To ensure mutual understanding, clinicians need to phrase simple questions that patients can comprehend, allowing information to be collected accurately. For example, "compliance" is simply conveyed as "taking medicine regularly" or "never miss a dose," which is reflected in how these terms appeared in the EMRs. This observation has also been reported in other languages, not from the same roots as the English language [27].

Term coverage in our dataset was modest (31.49%), driven entirely by English lexical matches. The observed term coverage is comparable with countries whose SNOMED CT editions are only partially translated, for example, Dutch (35%) and French (39%) [19]. A considerable proportion of our Thai clinical terms were synonymous with globally common English medical jargon, reflecting Thailand's long-standing integration of English terminology in medical education and clinical communication. As shown in Table 1, Thai language use was more prevalent in primary care departments (FM and emergency), while IM documents contained more specialized English terminology, consistent with previous reports [28].

Challenges in Expressing Clinical Concepts in the Thai Context

Our mapping analysis of partially mapped terms highlighted several challenges in representing Thai clinical concepts within SNOMED CT. Many partially mapped terms require determining whether they should (1) be represented through postcoordination with existing SNOMED CT concepts (thereby supporting international interoperability) or (2) be introduced as new concepts in the Thai national extension to preserve locally specific meaning.

Postcoordination can be beneficial for international interoperability but may lose nuanced local semantics. Conversely, creating new concepts preserves the locally used expression but introduces a long-term maintenance burden and requires clear criteria for inclusion. No standardized national guidelines currently exist; decisions must consider clinical relevance, documentation frequency, clinical importance, and anticipated use cases within Thailand's EMR ecosystem. A representative example from our study is "over-the-counter drug use," which frequently appeared in our dataset and reflects a long-standing public health challenge in unsupervised medication access in Thailand [29], which contributes to national issues such as antimicrobial resistance and drug misuse [30]. While "Patient on self-medication (finding)" is the most relevant concept in SNOMED CT, it does not fully represent Thailand's regulatory distinctions and safety concerns. If more granular differentiation is required in Thailand for surveillance or policy purposes, a new child concept under "Patient on self-medication (finding)" may be warranted.

The analysis of unmapped terms also highlighted the role of language-specific and culturally specific expressions that are not directly translatable into English (see Table 3). In our data, 20.93% of unmapped terms reflected culturally unique, ambiguous, or context-dependent meanings. These findings suggested three key strategies for terminology management and strengthening (1) exclude ambiguous terms or add contextual information to the context-dependent terms. For example, the Thai term " " has multiple meanings (fear of heights or a feeling of butterflies in the stomach). Instead of mapping it directly, adding the more specific term " " (giddiness in the heart) mapped to "Palpitations (finding)" can ensure clarity and avoid misclassification. (2) Use postcoordination for multiconcept or idiomatic expressions. This pattern was also found in the English context, for example, the concept "Nausea and vomiting (finding)," which exists as a precoordinated concept within the international edition of SNOMED CT. However, since separated distinct concepts are presented, we suggest mapping to separate entities to preserve legacy information and also interoperability with the international context. (3) Recommend new concepts for national extension or international inclusion for globally relevant but unsupported terms. This should be done with a well-established concept with clear practical implications. One example we found was a local abbreviation, "DOI (Date of Infection)," which we believe has seen an increase in occurrences during the pandemic, particularly

XSL•FO
RenderX

in cases where infectious history has been thoroughly collected. Additionally, some of these represent global clinically relevant concepts insufficiently addressed by SNOMED CT, indicating areas for future expansion [7].

Overall, these challenges highlight the need for national governance structures, clearly defined criteria for new concept creation, and a systematic strategy for translation and localization to ensure accurate, consistent, and interoperable representation of Thai clinical terminology. Regional guidance from the Asia eHealth Information Network (AeHIN) similarly emphasizes that successful and sustainable interoperability depends not only on the technical standards but also on governance, architecture, workforce capacity, and institutional processes [31]. Centralized governance through a national health data dictionary and terminology services, as described in many interoperability architectures, like Open Health Information Exchange (OpenHIE; Linux Foundation Public Health), can standardize terminological assets and reduce fragmentation across implementation sites [32]. Such services can support the controlled use of postcoordination and governed concept creation and enable local expressiveness while preserving semantic consistency and international interoperability. Importantly, these functions are best managed at the terminology service level rather than within individual EMR implementations, thereby reducing duplication and implementation burden. In addition, mappings between shared UITs and reference terminologies or classification systems (eg, the *ICD* family) should be centrally maintained, versioned, and routinely updated. For example, the OpenMRS ecosystem also adopts the similar implementation pattern, where many deployments rely on the CIEL interface terminology, centrally managed and distributed via Open Concept Lab, supporting reuse and reducing divergence across implementations [13].

Strengths and Limitations

To our knowledge, this is the first study in Thailand—and among the few resource-limited countries—to develop a UIT and evaluate both concept- and term-level coverage of SNOMED CT in a real-world clinical setting. A key strength of this study lies in its data source, free-text entries from actual clinical encounters, which ensures that the extracted terms are grounded in real-world usage. Moreover, the terminology development process integrated clinical expertise with NLP techniques in a resource-limited setting, offering valuable insights for similar environments. The final UIT is composed of the commonly used clinical term. Although it is small in numbers, we believe this terminology to be generally applicable for use within Thailand. Although SNOMED CT concept coverage has been examined in a few resource-limited country settings, evidence remains scarce. This study extends the discussion to Southeast Asia by evaluating Thai-language primary care data, constructing a UIT mapped to SNOMED CT, and reporting concept coverage.

Several limitations should be acknowledged. First, source data were collected from a single hospital, which may limit the generalizability of the developed UIT. Consequently, the developed UIT may not fully capture the diversity of expression terms used across Thailand's broader health care system, particularly in regions where distinct dialects or documentation styles prevail. Nevertheless, this study achieved its goal of constructing a realistic UIT to inform national implementation strategies and evaluate the feasibility of adopting SNOMED CT as a national standard. As new clinical terms and concepts emerge, the UTI can be continuously expanded through routine monitoring, domain-expert review, and integration of automated extraction tools to ensure incremental improvement over time. Second, although the mapping and term extraction processes were performed by clinically trained physicians, the subjective nature of manual mapping introduces the potential for interpretation bias or human error. An interannotator agreement score was not assessed due to constraints in available personnel and resources; consequently, only one mapper and one reviewer contributed to the final mapping. Third, the coverage results may be overestimated compared to those reported in previous studies because our term list was developed from commonly used clinical expressions and mapped to a more recent SNOMED CT release. At the same time, the mapping tool Snap2SNOMED does not support postcoordination at the time of this writing, which may lead to a potential underestimation of full SNOMED CT content coverage. However, since postcoordination has not been generally performed in previous studies, the results remain broadly comparable with existing evidence [19].

Recommendation for Future Research

Future research should expand the evaluation of concept and term coverage across hospitals of varying sizes and settings, including northern, northeastern, and southern regions of Thailand, where specific dialects may influence clinical documentation practices. Extending these methods beyond a single institution will help determine the generalizability of the UIT and support efforts toward national standardization.

Automated methods should also be explored. Advanced NLP techniques and large language models may substantially improve the efficiency and accuracy of term extraction, disambiguation, and SNOMED CT mapping, particularly as our team is already beginning to implement these tools in parallel methodological work. Additional studies should examine how postcoordination can be operationalized in real-world EMRs, with the goal of identifying best practices that maintain semantic precision without burdening clinicians during documentation.

Finally, future research should include more diverse document types (eg, operative notes, procedure reports, and specialty consults) and a broader population. Doing such improves characterization of locally used clinical concepts, strengthens coverage estimates, and supports the development of a more comprehensive national UIT that accurately reflects the linguistic and clinical diversity of Thailand's health care system.

Conclusions

In conclusion, SNOMED CT represents most common clinical concepts but lacks sufficient local language coverage. With the interpretation of local clinical terms, a Thai UIT mapped to SNOMED CT can serve as a comprehensive standard for clinical documentation. This hybrid approach can enhance data quality, promote interoperability, and facilitate better use of health care

XSL•FO
RenderX

data, contributing to improved patient care and creating opportunities for international collaboration.

Acknowledgments

This manuscript is a part of the thesis of Nat Tangchitnob, MD, within the international Ph.D. program in Data Science for Healthcare and Clinical Informatics, conducted at the Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.

Funding

The authors received no specific funding for this work.

Data Availability

All terminology data generated or analyzed during this study are included in this published article and its supplementary information files. The developed codes for this study are available at GitHub [23].

Authors' Contributions

Conceptualization: NT, WP, BK, OP, AT
Data curation: SP, PA, KS
Formal analysis: NT, WP, OP, AT
Funding acquisition: OP, AT
Investigation: NT, JS
Methodology: NT, WP, BK, GJM, JA, AT
Project administration: WP, OP
Resources: OP
Software: NT, WP
Supervision: WP, BK, OP, AT
Validation: WP, BK, AT
Visualization: NT
Writing – original draft: NT, WP, OP, JS
Writing – review & editing: NT, BK, SP, PA, KS, GJM, JA, AT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 10 frequency of the most common abbreviations and First-round results of interface term selection.
[DOCX File, 26 KB - medinform_v14i1e80039_app1.docx]

Multimedia Appendix 2
Interface Terminology mapped to SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms).
[XLSX File, 234 KB - medinform_v14i1e80039_app2.xlsx]

References

1. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023 Sep 25;30(10):1730-1740. [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
2. Linder JA, Schnipper JL, Middleton B. Method of electronic health record documentation and quality of primary care. J Am Med Inform Assoc 2012;19(6):1019-1024. [doi: 10.1136/amiajnl-2011-000788] [Medline: 22610494]
3. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 5;22(1):287. [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
4. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010 Mar 1;2010(1-5):1-5. [Medline: 21347133]
5. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag 2013;10(Fall):1c. [Medline: 24159271]
6. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4-5):394-403. [doi: 10.1055/s-0038-1634558] [Medline: 9865037]

XSL•FO
RenderX

7. Millar J. The need for a global language - SNOMED CT introduction. Stud Health Technol Inform 2016;225(683-5):683-685. [doi: 10.3233/978-1-61499-658-3-683] [Medline: 27332304]
8. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: “Ontologies” in today’s biomedical information systems and the limits of OWL. J Biomed Inform 2019;100S:100002. [doi: 10.1016/j.yjbinx.2019.100002] [Medline: 34384571]
9. Thiel R, Birov S, Piesche K, et al. The costs and benefits of SNOMED CT implementation: an economic assessment model. Stud Health Technol Inform 2016;228(441-5):441-445. [doi: 10.3233/978-1-61499-678-1-441] [Medline: 27577421]
10. Vuokko R, Vakkuri A, Palojoki S. Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT) clinical use cases in the context of electronic health record systems: systematic literature review. JMIR Med Inform 2023 Feb 6;11:e43750. [doi: 10.2196/43750] [Medline: 36745498]
11. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13(3):277-288. [doi: 10.1197/jamia.M1957] [Medline: 16501181]
12. Osornio AL, Luna D, Gambarte ML, Gomez A, Reynoso G, de Quirós FGB. Creation of a local interface terminology to SNOMED CT. Stud Health Technol Inform 2007;129(Pt 1):765-769. [Medline: 17911820]
13. Kanter AS, Borland R, Barasa M, et al. The importance of using open source technologies and common standards for interoperability within eHealth: perspectives from the millennium villages project. Adv Health Care Manag 2012;12(189-204):189-204. [doi: 10.1108/s1474-8231(2012)0000012013] [Medline: 22894051]
14. Halland K, Britz K, Gerber A. Investigations into the use of Snomed CT to enhance an OpenMRS health information system. SACJ 2011;47(1):33-45. [doi: 10.18489/sacj.v47i0.14]
15. Schulz S, Rodrigues JM, Rector A, Chute CG. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. Stud Health Technol Inform 2017;245(940-4):940-944. [doi: 10.3233/978-1-61499-830-3-940] [Medline: 29295238]
16. Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an experimental German user interface terminology linked to SNOMED CT. Stud Health Technol Inform 2019 Aug 21;264(153-7):153-157. [doi: 10.3233/SHTI190202] [Medline: 31437904]
17. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthc Inform Res 2021 Jan;27(1):3-10. [doi: 10.4258/hir.2021.27.1.3] [Medline: 33611871]
18. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A model for evaluating interface terminologies. J Am Med Inform Assoc 2008;15(1):65-76. [doi: 10.1197/jamia.M2506] [Medline: 17947616]
19. Miñarro-Giménez JA, Cornet R, Jaulent MC, et al. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform 2019 Mar;123(37-48):37-48. [doi: 10.1016/j.ijmedinf.2018.12.011] [Medline: 30654902]
20. Thailand. SNOMED International. 2025 Oct 22. URL: https://www.snomed.org/members/thailand [accessed 2026-02-27]
21. Isaradech N, Khumrin P. Auto-mapping clinical documents to ICD-10 using SNOMED-CT. AMIA Jt Summits Transl Sci Proc 2021;2021:296-304. [Medline: 34457144]
22. Phatthiyaphaibun W, Chaovavanich K, Polpanumas C, et al. PyThaiNLP: thai natural language processing in python. Presented at: Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023); Jul 13, 2023 URL: https://aclanthology.org/2023.nlposs-1 [accessed 2026-02-05] [doi: 10.18653/v1/2023.nlposs-1.4]
23. Chirurgo/terminology2snomed [internet]. GitHub. URL: https://github.com/Chirurgo/terminology2snomed [accessed 2026-02-05]
24. Damrongkamoltip K, Ruenlek K, Limprasert W, Boonkwan P. Comparing a Thai words segmentation methods in the LST20 dataset. J Comput Creat Technol 2024;2(2):61-70. [doi: 10.14456/jcct.2024.7]
25. Oluoch T, de Keizer N, Langat P, et al. A structured approach to recording AIDS-defining illnesses in Kenya: a SNOMED CT based solution. J Biomed Inform 2015 Aug;56(387-94):387-394. [doi: 10.1016/j.jbi.2015.07.009] [Medline: 26184057]
26. Miñarro-Giménez JA, Martínez-Costa C, Karlsson D, Schulz S, Gøeg KR. Qualitative analysis of manual annotations of clinical text with SNOMED CT. PLoS ONE 2018;13(12):e0209547. [doi: 10.1371/journal.pone.0209547] [Medline: 30589855]
27. Perez-de-Viñaspre O, Oronoz M. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. BMC Med Inform Decis Mak 2015;15 Suppl 2(Suppl 2):S5. [doi: 10.1186/1472-6947-15-S2-S5] [Medline: 26100112]
28. Yusuff KB, Tayo F. Does a physician’s specialty influence the recording of medication history in patients’ case notes? Br J Clin Pharmacol 2008 Aug;66(2):308-312. [doi: 10.1111/j.1365-2125.2008.03202.x] [Medline: 18537964]
29. Cooper RJ. Over-the-counter medicine abuse - a review of the literature. J Subst Use 2013 Apr;18(2):82-107. [doi: 10.3109/14659891.2011.615002] [Medline: 23525509]
30. Chautrakarn S, Khumros W, Phutrakool P. Self-medication with over-the-counter medicines among the working age population in metropolitan areas of Thailand. Front Pharmacol 2021;12(726643):726643. [doi: 10.3389/fphar.2021.726643] [Medline: 34456738]
31. Marcelo AB, Pascual KC, Udayasankaran JG, Allaudin FS, Kijsanayotin B. A framework for regional health information systems interoperability: the asia ehealth information network (AeHIN) experience. In: Roadmap to Successful Digital Health Ecosystems: Elsevier; 2022:399-414. [doi: 10.1016/B978-0-12-823413-6.00005-7]

XSL•FO
RenderX

32. Mamuye AL, Yilma TM, Abdulwahab A, et al. Health information exchange policy and standards for digital health systems in africa: a systematic review. PLOS Digit Health 2022 Oct;1(10):e0000118. [doi: 10.1371/journal.pdig.0000118] [Medline: 36812615]

Abbreviations

AeHIN: Asia eHealth Information Network
CIEL: Columbia International eHealth Laboratory
EMR: electronic medical record
FM: family medicine
***ICD*:** *International Classification of Diseases*
IM: internal medicine
NLP: natural language processing
OpenHIE: Open Health Information Exchange
OpenMRS: Open Medical Record System
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
UIT: user interface terminology

Edited by A Benis; submitted 03.Jul.2025; peer-reviewed by AS Kanter, T Ohlsen; accepted 08.Jan.2026; published 09.Mar.2026.

Please cite as:
Tangchitnob N, Ponthongmak W, Kijsanayotin B, Pattanaprateep O, Phusanti S, Atiksawedparit P, Suwanthaweemeesuk K, Siangfu J, McKay GJ, Attia J, Thakkinstian A
Developing a Thai User Interface Terminology for Systematized Nomenclature of Medicine Clinical Terms Implementation in Primary Care: Cross-Sectional Content Coverage Analysis
JMIR Med Inform 2026;14:e80039
URL: https://medinform.jmir.org/2026/1/e80039
doi:10.2196/80039

© Nat Tangchitnob, Wanchana Ponthongmak, Boonchai Kijsanayotin, Oraluck Pattanaprateep, Sithakom Phusanti, Pongsakorn Atiksawedparit, Kamonporn Suwanthaweemeesuk, Jirayus Siangfu, Gareth J McKay, John Attia, Ammarin Thakkinstian. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Models for Psychiatric Diagnosis Based on Multicenter Real-World Clinical Records: Comparative Study

Maoqian Sun[1,2], MD; Jia Yu[3], MD; Zhuhong Long[2], MD; Yun Yang[2], MD; Tao Xiao[4], MD; Jiaquan Liang[5], MD; Jun Feng[6], MD; Huaili Deng[7], MD; Guoping Huang[2], MD, PhD

[1]Mental Health Center, University-Town Hospital of Chongqing Medical University, Chongqing, China
[2]Sichuan Mental Health Center, The Third Hospital of Mianyang, Mianyang, China
[3]Third People's Hospital of Panzhihua, Panzhihua, China
[4]The Third People's Hospital of Ganzhou, Ganzhou, China
[5]The Third People's Hospital of Foshan, Foshan, China
[6]Lanzhou Third People's Hospital, Lanzhou, China
[7]Shanxi Provincial Mental Health Center, Taiyuan, China

Corresponding Author:
Guoping Huang, MD, PhD
Sichuan Mental Health Center
The Third Hospital of Mianyang
No. 190, East Section of Jiannan Road
Youxian District
Mianyang, 621000
China
Phone: 86 0816 2282249
Email: xyhuanggp@126.com

Abstract

Background: Psychiatric disorders are diagnostically challenging and often rely on subjective clinical judgment, particularly in resource-limited settings. Large language models (LLMs) have demonstrated potential in supporting psychiatric diagnosis; however, robust evidence from large-scale, real-world clinical data remains limited.

Objective: This study aimed to evaluate and compare the diagnostic performance of multiple LLMs for psychiatric disorders using multicenter real-world electronic health records (EHRs).

Methods: We retrospectively analyzed 9923 inpatient EHRs collected from 6 psychiatric centers across China, encompassing all ICD-10 (International Statistical Classification of Diseases, Tenth Revision) psychiatric categories. In total, 3 LLMs—GPT-4.0 (OpenAI), GPT-3.5 (OpenAI), and GLM-4-Plus (Zhipu AI)—were evaluated against physician-confirmed discharge diagnoses. Diagnostic performance was assessed using strict accuracy criteria and lenient classification metrics, with subgroup analyses conducted across diagnostic categories and age groups.

Results: GPT-4.0 achieved the highest overall strict diagnostic accuracy (71.7%) and the highest weighted F1-score under lenient evaluation (0.881), particularly for high-prevalence disorders, such as mood disorders and schizophrenia spectrum disorders. Diagnostic performance varied across age groups, with the highest accuracy observed in older adult patients (up to 79.5%) and lower accuracy in adolescents. Across centers, model performance remained stable, with no significant intercenter differences.

Conclusions: LLMs—especially GPT-4.0—demonstrate promising capability in supporting psychiatric diagnosis using real-world EHRs. However, diagnostic performance varies by age group and disorder category. LLMs should be regarded as assistive tools rather than replacements for clinical judgment, and further validation is needed before routine clinical implementation.

(JMIR Med Inform 2026;14:e77699) doi:10.2196/77699

KEYWORDS

psychiatric disorders; large language models; artificial intelligence; multicenter study; real-world data

XSL•FO
RenderX

Introduction

Mental disorders are a significant public health concern worldwide; they have high prevalence rates, which are continuously increasing each year [1]. About 20% of the global population is estimated to experience mental illness in their lifetime, placing a substantial burden on both society and health care systems [2]. However, there is an acute scarcity of psychiatric professionals. Many countries have fewer than 5 psychiatrists per 100,000 individuals; this results in delays in providing timely, high-quality diagnostic and therapeutic services for patients with mental health disorders, particularly in resource-limited regions, including China [3]. This medical resource shortage further underscores the urgent need to increase the accuracy and efficiency of psychiatric diagnoses. In addition, psychiatric diagnoses are inherently complex and challenging. Complex etiologies and high symptom heterogeneity characterize mental disorders. The symptom descriptions of patients often depend on subjective expression, with physicians' medical records and clinical diagnoses also based on their subjective experience [4]. This subjectivity makes diagnoses susceptible to variation among physicians, leading to inconsistent diagnostic outcomes. Therefore, in psychiatry, there is a crucial need to use modern technologies to increase diagnostic efficiency and accuracy [5].

Recently, the widespread adoption of electronic health records (EHRs) has provided a wealth of unstructured data to support the application of artificial intelligence technologies in medical diagnostics, offering new opportunities for diagnosing psychiatric disorders [6]. Large language models (LLMs), including the GPT series, exhibit powerful capabilities in natural language processing and semantic understanding. They can efficiently process unstructured text and have shown great potential in medical diagnostic assistance [7-10].

At present, studies on the application of LLMs in other medical fields have further confirmed their clinical potential. For example, Hasani et al [11] assessed the performance of GPT-4.0 (OpenAI) in standardizing radiology reports; they reported that it generates high-quality standardized radiology reports and significantly improves the efficiency of report generation and data analysis capabilities. Liu et al [12] compared GPT-4.0 with senior neurosurgeons and reported that the performance of GPT-4.0 in complex medical tasks is comparable with that of experienced experts, demonstrating its value in facilitating the diagnosis of complex cases. In addition, Horiuchi et al [13] reported that GPT-4.0 has demonstrated superior diagnostic accuracy in challenging neuroradiology cases, suggesting that generative artificial intelligence technologies can serve as important supplements to physician cognition in complex diagnostic scenarios. Overall, these studies demonstrate the potential application of GPT-4.0 in multidomain diagnostic tasks.

In psychiatry, LLMs have also begun to garner attention. Because psychiatric diagnoses heavily depend on textual descriptions of patient complaints and medical history, LLMs naturally possess significant application advantages in this field. Previous studies have revealed that ChatGPT-4 (OpenAI) can achieve an accuracy rate of 96.1% in the diagnosis of obsessive-compulsive disorder, significantly outperforming that of clinical psychologists (81.5%) and primary care physicians (49.5%) [14]. In 1 study, researchers used the UTH-BERT model and reported that its diagnostic accuracy for medical history in EHRs (74.3%) surpassed that of psychiatrists with intermediate experience (71.5%) [15]. Despite these advances, prior studies on LLM-based psychiatric diagnosis have several limitations. First, the scope of diseases in previous LLM research has often been narrow, with most studies focusing on a single or a small subset of psychiatric disorders (eg, major depressive disorder and obsessive-compulsive disorder) [16]. In contrast, our study simultaneously included all major *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) psychiatric categories (F00-F99), thereby providing broader coverage than previous studies. Nevertheless, our analysis was conducted at the level of primary diagnoses within each category, without explicitly addressing comorbidity patterns, which remains an important direction for future research. Second, data sources are often single and lack validation in real-world clinical settings. Current studies rely primarily on simulated cases or data from a single institution [17], making it challenging to capture patient characteristics across diverse regions and health care practices. In the Chinese context, several studies have begun to explore the use of LLMs for psychiatric diagnosis and Chinese-language medical records; however, comprehensive, large-scale validation in real-world psychiatric EHRs—especially with direct comparisons between international and Chinese models—remains limited. Third, the diagnostic performance across age groups has not been thoroughly investigated, with most studies focusing solely on overall accuracy without examining differences in diagnostic performance across age groups (eg, adolescents and older adult patients). Symptoms, disease progression, and treatment plans for psychiatric disorders significantly differ across different age groups. Whether LLMs can adapt to these differences remains unexplored.

To address the aforementioned issues, this study optimized 3 key aspects: data sources, disease coverage, and population segmentation. First, we incorporated real-world EHR data from 6 representative provincial and municipal psychiatric centers across different regions of China. By directly using clinical records to evaluate the model, this approach is better aligned with real-world medical settings, filling the gap left by previous studies that lacked real-world data validation in the Chinese language. Second, we covered all psychiatric disorders listed in the *ICD-10* classification (F0-F9), overcoming the limitations of previous studies, which focused primarily on a single or a few disease types. This design allowed us to evaluate the model's generalization ability across a broad range of clinical diagnostic tasks. In addition, we conducted fine-grained age stratification, systematically comparing the model's diagnostic performance across 4 age groups: adolescents, young adults, middle-aged individuals, and older adult patients. This allowed the identification of the model's adaptation differences across different age groups. In summary, our study, based on large-scale, multicenter real-world data, is the first to comprehensively evaluate the clinical potential of LLMs for psychiatric diagnosis in China.

XSL•FO
RenderX

Methods

Research Design and Data Sources

The study data were sourced from 6 provincial and municipal psychiatric centers in China. These centers are located in Sichuan, Gansu, Shanxi, Guangdong, Jiangxi, and other regions, covering Southwest China, Northwest China, North China, South China, and East China, thereby ensuring good regional representation. The dataset included the EHRs of psychiatric inpatients between July 2017 and July 2024. In total, 12,819 cases were initially screened. Among them, 2896 cases were excluded owing to incomplete records or unclear diagnoses. Therefore, 9923 eligible cases were included. The EHR data used in this study consisted exclusively of unstructured narrative text fields, including chief complaints, history of present illness, and psychiatric examination notes. Structured data, such as laboratory test results, medication prescriptions, vital signs, nursing notes, and imaging reports, were not included, as psychiatric diagnosis primarily depends on descriptive symptomatology and mental status examination rather than auxiliary tests. All records included in the model were noted down by attending psychiatrists at the time of admission only. Clinical notes taken during hospitalization were not included, ensuring that the model received only information available at the initial diagnostic encounter and preventing potential temporal information leakage.

Ethical Considerations

The Ethics Committee of the Sichuan Provincial Mental Health Center approved this study (approval 202460). Considering the retrospective nature of the study, the Ethics Committee exempted the need for informed consent from the patients. All EHR data were fully deidentified in accordance with the Health Insurance Portability and Accountability Act safe harbor method (45 CFR 164.514(b)), with direct identifiers (eg, names, addresses, contact details, and hospital IDs) removed before analysis. In addition, because narrative clinical text can contain residual indirect identifiers, a stratified manual audit of 500 randomly sampled EHR narratives (approximately 5% of the dataset) was conducted to further minimize reidentification risk. Among these 500 records, 7 (1.4%) instances of potential indirect identifiers were identified and corrected before model inference. No direct identifiers were detected in the sampled records.

Model queries were transmitted through encrypted API calls (HTTPS/TLS) rather than public chat interfaces. Although no formal data processing agreement was established with the external provider, the institutional review board explicitly approved the use of deidentified clinical narratives for inference on external LLM servers. It was assured that the procedure complied with local data governance regulations and did not include identifiable personal information. Only deidentified free text—without metadata, patient identifiers, or institutional information—was transmitted, and no protected health information left the institution.

The study adhered to relevant international and domestic privacy frameworks, including General Data Protection Regulation Article 46 for cross-border data transfer, China's Personal Information Protection Law, and ISO/IEC 27018 guidelines for the protection of personal data in public cloud environments. Nevertheless, we acknowledge that narrative EHR text has inherent deidentification limitations, and we minimized data transmission to mitigate privacy risks.

Data Preprocessing

To closely replicate real-world clinical settings, the admission diagnosis names were removed from the medical history; no other data were modified, thereby preserving the integrity of the original data [18]. The diagnoses were classified into 11 categories by extracting only the first letter and the 10-digit *ICD-10* codes from Chapter V (Mental and Behavioral Disorders; F0-F9) and other related codes. Specifically, F0 includes organic mental disorders such as dementia; F1 includes substance-related disorders such as alcohol addiction; F2 includes schizophrenia and associated psychotic conditions; F3 includes mood disorders such as depression; F4 includes anxiety and stress-related disorders; F5 includes disorders associated with physiological conditions, such as eating disorders; F6 includes personality and behavioral disorders in adults; F7 includes intellectual disabilities; F8 includes developmental disorders such as autism; and F9 includes behavioral and emotional conditions beginning in childhood, including attention-deficit or hyperactivity disorders [19]. For subgroup analysis, patients were stratified into 4 age groups: adolescents (≤17 years), young adults (18-35 years), middle-aged adults (36-59 years), and older adult individuals (≥60 years). The thresholds at 18 and 60 years follow established conventions in psychiatric and geriatric research [20]. The adult group was further subdivided into young and middle-aged categories to capture differences in onset age, symptom presentation, and comorbidity patterns of major psychiatric disorders across life stages.

Research Tools and Model Setup

In total, 3 LLMs were selected for evaluation: GPT-4.0 and GPT-3.5 (OpenAI) and GLM-4-Plus (Zhipu AI). GPT-3.5 was chosen as a widely used baseline model because of its broad accessibility and extensive application in previous medical natural language processing tasks, providing a reference point for evaluation [21]. GPT-4.0, a more advanced successor in the GPT series, was included to assess performance improvements driven by enhanced semantic understanding and reasoning capabilities [22]. GLM-4-Plus is one of the most advanced Chinese-optimized LLMs [23]. It was selected to examine the adaptability and potential of domestic LLMs in psychiatric diagnosis using Chinese-language EHRs and to compare their performance with that of international mainstream models under localized application scenarios.

In addition, a unified prompt was used: "请根据以下患者信息给出最可能的诊断。" ("Please provide the most likely diagnosis based on the following patient information"). Because the clinical input consisted of Chinese-language EHR narratives, the instruction prompt was also written in Chinese to maintain monolingual consistency and avoid potential cross-lingual prompting effects. This prompt was intentionally selected to ensure methodological consistency and fair comparability across all LLMs, as varying prompt strategies may introduce

XSL•FO
RenderX

model-specific advantages. Zero-shot prompts also reflect a realistic clinical scenario in which clinicians typically expect the model to provide a diagnostic impression directly without additional examples. No English prompt or mixed-language prompt was used at any stage. No prompt engineering or chain-of-thought prompting was applied, as the study aimed to evaluate each model's baseline diagnostic capability under identical input conditions. The preprocessed data included the patient's sex, age, chief complaints, medical history, and psychiatric examination results. The psychiatric examination results included narrative descriptions of key mental status domains, such as appearance and behavior, mood and affect, thought process and content, cognitive function, and insight, which were documented in a standardized format across all participating centers to ensure structural consistency and data comparability. These inputs were provided to the model to generate diagnostic results. All diagnostic analyses by the LLMs were completed between January 1 and January 31, 2025. The discharge diagnosis served as the reference standard and was confirmed by a senior associate chief physician with over 10 years of clinical experience.

Model outputs were evaluated using a 0-2 point scoring system: 2 points for a completely correct diagnosis (exact *ICD-10* match), 1 point for a partially correct diagnosis, and 0 points for an incorrect diagnosis. "Partially correct" was operationally defined as cases in which the predicted diagnosis was clinically close to, but not an exact, *ICD-10* match. Examples included (1) misclassification within the schizophrenia spectrum (eg, schizotypal disorder scored as schizophrenia), (2) assigning a broader category within mood disorders without specifying the exact episode (eg, manic episode scored as mood disorder), or (3) capturing the correct diagnostic domain at a broader level of granularity (eg, generalized anxiety disorder scored as anxiety disorder). Predictions that did not meet these criteria were considered incorrect.

To ensure reliability and minimize bias, multiple 2-rater groups were formed, each comprising 2 attending psychiatrists with more than 5 years of independent clinical experience. Both raters independently scored each case after the model outputs were anonymized and randomized, which blinded them to the source model. Disagreements were resolved by consensus with a senior psychiatrist. Interrater reliability was assessed using Cohen κ coefficient, which demonstrated substantial agreement (Cohen κ=0.82).

Data Analysis

Statistical analyses were conducted using SPSS Statistics (version 27.0; IBM Corp). Descriptive statistics were used to present the demographic characteristics and distributions of disease categories. Model performance was evaluated by comparing the predicted diagnoses with the discharge diagnoses, which served as the reference standard. Accuracy was calculated as the proportion of correct predictions (score=2). In this study, strict evaluation criteria were used as the primary analytic framework. Under this explicit definition, only exact matches between the model-predicted *ICD-10* diagnosis and the discharge diagnosis (score=2) were considered correct, whereas partially correct predictions (score=1) were treated as incorrect.

For the classification performance metrics—namely, precision, recall, and F_1-score—we also applied a lenient evaluation. Under this lenient definition, both exact matches (score=2) and partially correct predictions (score=1) were treated as correct and counted as true positives (TPs). In this framework, TPs were predictions that were exactly or partially consistent with the true diagnosis (score=2 or score=1); false positives (FPs) were predictions for a class when the true diagnosis belonged to another class; and false negatives (FNs) were cases in which the true diagnosis belonged to a class, but the model prediction did not. Predictions with score=0 were always treated as incorrect. These lenient metrics were used exclusively for the classification performance results and are clearly labeled as such. All other analyses throughout the manuscript relied solely on the strict evaluation criteria. In addition, performance metrics, including precision, recall, and F_1-score, were computed under a strict definition: only exact matches between the model-predicted *ICD-10* diagnosis and the discharge diagnosis were considered correct. In this framework, TPs are cases where the prediction exactly matches the true diagnosis; FPs are predictions for a class when the true diagnosis belongs to another class; and FNs are cases where the true diagnosis belongs to a class, but the model prediction does not. Partially correct outputs (score=1) were treated as incorrect predictions and classified as FPs or FNs depending on the prediction-reference mismatch; only score=2 outputs were counted as TPs.

The macro- and weighted-average F_1-scores were calculated to measure overall performance. Because the F8 category contained only 8 cases (0.1% of the dataset), macroaverage F_1-scores were additionally reported, both including and excluding this extremely rare category, to provide a more stable estimate of overall performance. Tests of normality and homogeneity of variance were conducted. The Shapiro-Wilk test indicated that the score data significantly deviated from normality across all model groups, and the Levene test confirmed unequal variances (Table S1 in Multimedia Appendix 1). Therefore, nonparametric rank-sum tests were applied for intergroup comparisons. In addition, Kruskal-Wallis tests were used to examine whether diagnostic accuracy differed across the 6 centers for each model (Table S2 in Multimedia Appendix 1). A P value of <.05 indicated significance. To further compare diagnostic performance across LLMs, an ordinal logistic regression analysis was conducted. Because the diagnostic outcome was an ordered 0-2 score (0=incorrect, 1=partially correct, and 2=correct), a proportional odds model was applied to estimate the likelihood of achieving a higher diagnostic score for each model. GPT-3.5 served as the reference category. Odds ratios (ORs) and 95% CIs were calculated to quantify differences in diagnostic performance. The results of this analysis are presented in Table S3 in Multimedia Appendix 1.

Results

Demographic Characteristics and Disease Category Distribution

In total, 9923 patients (3665, 36.93% males and 6258, 63.07% females) were included in this study. Across age groups, 1881 (18.96%) were adolescents (≤17 years), 2468 (24.87%) were

XSL•FO
RenderX

young adults (18-35 years), 3695 (37.24%) were middle-aged adults (36-59 years), and 1879 (18.94%) were older adults (≥60 years). The disease categories were distributed as follows (Table 1): mood disorders (F3) were the most common, accounting for 36.7% (3646/9923) of the cases, followed by schizophrenia, schizotypal, and delusional disorders (F2, 2470/9923, 24.9%), and neurotic disorders (F4, 1991/9923, 20.1%). The distribution of other disorders was relatively low.

Table 1. Demographic data and clinical characteristics.

Characteristics	Value
Sex, n (%)	
Male	3665 (36.93)
Female	6258 (63.07)
Age (y), mean (SD)	40.5 (20.03)
Age category, n (%)	
Adolescent (≤17 y)	1881 (19)
Young adult (18-35 y)	2468 (24.9)
Middle-aged (36-59 y)	3695 (37.2)
Older adult (≥60 y)	1879 (18.9)
Disorder category, n (%)	
F0 (Organic disorders)	597 (6)
F1 (Substance use disorders)	243 (2)
F2 (Schizophrenia, schizotypal, and delusional disorders)	2470 (24.9)
F3 (Mood disorders)	3646 (36.7)
F4 (Neurotic disorders)	1991 (20.1)
F5 (Eating disorders, etc)	70 (0.7)
F6 (Personality disorders)	86 (0.9)
F7 (Mental retardation)	177 (1.8)
F8 (Autism, etc)	8 (0.1)
F9 (Behavioral disorders)	635 (6.4)

Diagnostic Accuracy of the 3 Models

Overall, GPT-4.0 achieved an overall accuracy of 71.7%, outperforming GLM-4-Plus (69.3%) and GPT-3.5 (68.8%). In terms of category-specific performance, GPT-4.0 exhibited higher overall accuracy than the other models in the F1, F3, and F6 categories. However, in the F2 and F4 categories, the diagnostic accuracy rates of the 3 models were similar, with no significant differences. Among the different age groups, diagnostic accuracy was higher in the older adult group (with an overall accuracy rate of 74.3%-79.5%). Additional analysis using the rank-sum test revealed significant differences between adolescents and middle-aged adults, adolescents and older adults, young adults and middle-aged adults, and young adults and older adults (P<.001). The results are summarized in Tables 2 and 3, and the category-level diagnostic score patterns across the 3 models are visualized in Figure 1.

XSL•FO
RenderX

Table 2. Diagnostic accuracy.

Category	GPT-4.0, n (%)			GPT-3.5, n (%)			GLM-4-Plus, n (%)		
	Correct	Partially correct	Incorrect	Correct	Partially correct	Incorrect	Correct	Partially correct	Incorrect
Disorder category									
F0	394 (66)	152 (25.5)	51 (8.5)	338 (56.6)	143 (24)	116 (19.4)	415 (69.5)	94 (15.7)	88 (14.7)
F1	208 (85.6)	22 (9.1)	13 (5.3)	182 (74.9)	25 (10.3)	36 (14.8)	206 (84.8)	14 (5.8)	23 (9.5)
F2	1906 (77.2)	460 (18.6)	104 (4.2)	1834 (74.3)	531 (21.5)	105 (4.3)	1890 (76.5)	475 (19.2)	105 (4.3)
F3	2738 (75.1)	561 (15.4)	347 (9.5)	2654 (72.8)	489 (13.4)	503 (13.8)	2576 (70.7)	773 (21.2)	297 (8.1)
F4	1347 (67.6)	310 (15.6)	334 (16.8)	1368 (68.7)	308 (15.5)	315 (15.8)	1281 (64.3)	323 (16.2)	387 (19.4)
F5	48 (68.6)	2 (2.9)	20 (28.6)	41 (58.6)	0 (0)	29 (41.4)	50 (71.4)	1 (1.4)	19 (27.1)
F6	59 (68.6)	3 (3.5)	24 (27.9)	20 (23.3)	11 (12.8)	55 (64)	36 (41.9)	0 (0)	50 (58.1)
F7	123 (69.5)	22 (12.4)	32 (18.1)	83 (46.9)	9 (5.1)	85 (48)	131 (74)	6 (3.4)	40 (22.6)
F8	8 (100)	0 (0)	0 (0)	7 (87.5)	0 (0)	1 (12.5)	6 (75)	0 (0)	2 (25)
F9	285 (44.9)	85 (13.4)	265 (41.7)	296 (46.6)	78 (12.3)	261 (41.1)	282 (44.4)	79 (12.4)	274 (43.1)
Age category									
Adolescent (≤17 y)	1148 (61)	367 (19.5)	366 (19.5)	1116 (59.3)	341 (18.1)	424 (22.5)	1010 (53.7)	497 (26.4)	374 (19.9)
Young adult (18-35 y)	1727 (70)	480 (19.4)	261 (10.6)	1638 (66.4)	456 (18.5)	374 (15.2)	1654 (67)	534 (21.6)	280 (11.3)
Middle-aged (36-59 y)	2759 (74.7)	530 (14.3)	406 (11)	2672 (72.3)	551 (14.9)	472 (12.8)	2715 (73.5)	540 (14.6)	440 (11.9)
Older adult (≥60 y)	1482 (78.9)	240 (12.8)	157 (8.4)	1397 (74.3)	246 (13.1)	236 (12.6)	1494 (79.5)	194 (10.3)	191 (10.2)
Total	7116 (71.7)	1617 (16.3)	1190 (12)	6823 (68.8)	1594 (16.1)	1506 (15.2)	6873 (69.3)	1762 (17.8)	1285 (12.9)

Table 3. Comparison of diagnostic scores based on age.

Age category	*P* value		
	GPT-4.0	GPT-3.5	GLM-4-Plus
Adolescent[a] vs young adult[b]	<.001	<.001	<.001
Adolescent vs middle-aged[c]	<.001	<.001	<.001
Adolescent vs older adult[d]	<.001	<.001	<.001
Young adult vs middle-aged	<.001	<.001	<.001
Young adult vs older adult	<.001	<.001	<.001
Middle-aged vs older adult	<.001	.15	<.001

[a]Adolescent: ≤17 years.

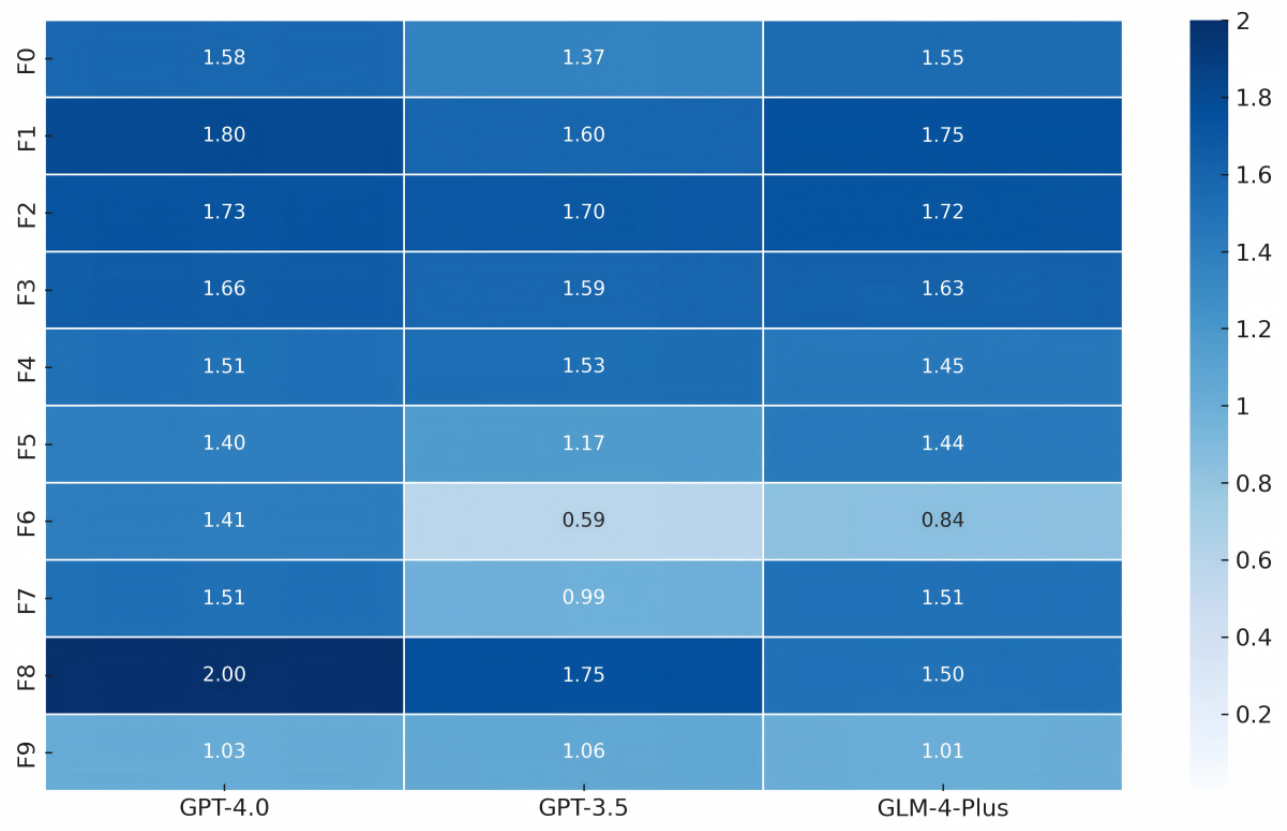
[b]Young adult: 18-35 years.

[c]Middle-aged: 36-59 years.

[d]Older adult: ≥60 years.

XSL•FO
RenderX

Figure 1. Heatmap of the diagnostic scores. The color intensity indicates the model's diagnostic score (on a 2-point scale), with darker colors indicating higher scores. This reflects the diagnostic accuracy of the models in each category.



Center-Level Diagnostic Performance

To evaluate whether model performance varied across institutions, a supplementary center-level analysis was conducted. The diagnostic accuracies of all 3 models demonstrated only minor numerical fluctuations across the 6 centers. Importantly, the Kruskal-Wallis test indicated no statistically significant differences for any of the models (GPT-4.0: P=.52; GPT-3.5: P=.47; and GLM-4-Plus: P=.61). The detailed results for each center are provided in Table S2 in Multimedia Appendix 1.

Analysis of Differences in Scoring Distribution of 3 Models

The rank-sum test was used to compare differences in the scoring distributions of GPT-4.0, GPT-3.5, and GLM-4-Plus across various *ICD-10* categories. We observed that GPT-4.0 had higher scores than GPT-3.5 in the F0 (P<.001), F6 (P<.001), and F7 (P<.001) categories. Furthermore, GLM-4-Plus exhibited differences in its scoring distribution compared with GPT-3.5 in the F0 (P<.001) and F7 (P<.001) categories. Overall, the score distribution of GPT-4.0 was superior to that of both GPT-3.5 (P<.001) and GLM-4-Plus (P<.001). However, the difference in the scoring distribution between GPT-3.5 and GLM-4-Plus was not statistically significant (P=.09). Table 4 summarizes the results.

XSL•FO
RenderX

Table 4. Comparison of the diagnostic scores of the different large language models.

Disorder category	*P* value		
	GPT-4.0 vs GPT-3.5	GPT-4.0 vs GLM-4-Plus	GPT-3.5 vs GLM-4-Plus
F0	<.001	.69	<.001
F1	.002	.67	.008
F2	.02	.60	.08
F3	.002	<.001	.77
F4	.43	.02	.002
F5	.17	.75	.098
F6	<.001	<.001	.14
F7	<.001	.63	<.001
F8	.38	.17	.59
F9	.65	.72	.42
Total	<.001	<.001	.09

Comparison of the Classification Performance of 3 Models

Table 5 summarizes the precision, recall, and F_1-scores of the GPT-4.0, GPT-3.5, and GLM-4-Plus models across various *ICD-10* disease categories. To enhance interpretability, these classification metrics were computed using a lenient evaluation framework, in which both exact matches (score=2) and partially correct predictions (score=1) were treated as correct. GPT-4.0 achieved the highest F_1-scores in the F0 (0.915), F1 (0.926), and F2 (0.929) categories. GLM-4-Plus achieved the highest precision in the F2 category (0.951), whereas GPT-4.0 had a superior recall rate (0.959). Small-sample categories exhibited significant fluctuations, including the F8 category, where GPT-4.0's recall rate reached 1.00, while its precision was only 0.308. This phenomenon reflects class imbalance and overprediction in rare categories. The macroaverage F_1-scores (including F8) were 0.655, 0.410, and 0.712 for GPT-4.0, GPT-3.5, and GLM-4-Plus, respectively. The weighted average F_1-score of GPT-4.0 (0.881) was superior to that of both GPT-3.5 (0.849) and GLM-4-Plus (0.873). To mitigate the instability caused by the extremely small F8 category, macroaverage F_1-scores were also calculated after excluding F8. The exclusion yielded more stable macrolevel estimates: 0.821 for GPT-4.0, 0.745 for GPT-3.5, and 0.806 for GLM-4-Plus.

To further complement this evaluation, we also performed an ordinal logistic regression analysis using the 0-2 scoring system (0=incorrect, 1=partially correct, and 2=correct) as the outcome variable, accounting for its ordinal nature. The results were consistent with the overall performance patterns observed in the main analysis. GPT-4.0 was significantly more likely to achieve a higher diagnostic score compared with GPT-3.5 (OR 1.18, 95% CI 1.11-1.25; *P*<.001). In contrast, GLM-4-Plus showed slightly lower odds of achieving a higher score relative to GPT-3.5 (OR 0.95, 95% CI 0.90-1.01). Although this difference did not reach statistical significance (*P*=.09; Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 5. Precision, recall, and F1-scores of the different large language models (lenient evaluation).

Category[a]	GPT-4.0			GPT-3.5			GLM-4-Plus		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Disorder category									
F0	0.916	0.915	0.915	0.897	0.814	0.853	0.857	0.929	0.891
F1	0.906	0.947	0.926	0.924	0.852	0.887	0.880	0.961	0.919
F2	0.902	0.959	0.929	0.850	0.961	0.902	0.951	0.876	0.912
F3	0.857	0.906	0.881	0.849	0.864	0.856	0.884	0.862	0.873
F4	0.913	0.835	0.872	0.859	0.843	0.851	0.863	0.867	0.865
F5	0.549	0.714	0.621	0.519	0.586	0.550	0.785	0.607	0.685
F6	0.554	0.721	0.626	0.405	0.372	0.388	0.750	0.409	0.529
F7	0.973	0.819	0.890	0.979	0.525	0.684	0.820	0.913	0.864
F8	0.308	1.000	0.471	0.159	0.875	0.269	0.667	0.500	0.571
F9	0.962	0.592	0.733	0.971	0.589	0.733	0.576	0.960	0.720
Macro average	0.653	0.701	0.655	0.436	0.428	0.410	0.730	0.717	0.712
Macro average (excluding F8)	—[b]	—	0.821	—	—	0.745	—	—	0.806
Weighted average	0.888	0.882	0.881	0.859	0.851	0.849	0.882	0.871	0.873

[a]A lenient evaluation framework was used for all metrics in this table. Both score=2 (exact match) and score=1 (partially correct) predictions were considered as correct.

[b]Not applicable.

Discussion

Principal Findings

This multicenter real-world study is the first to systematically evaluate the diagnostic performance of 3 LLMs—GPT-4.0, GPT-3.5, and GLM-4-Plus—across the full spectrum of *ICD-10* psychiatric categories via Chinese psychiatric EHRs. The results demonstrate that GPT-4.0 achieved the highest overall strict diagnostic accuracy (71.7%), followed by GLM-4-Plus (69.3%) and GPT-3.5 (68.8%), showing a clear performance gradient across model generations. In addition to strict accuracy, GPT-4.0 also achieved a higher lenient F_1-score (0.881), indicating performance when partially correct predictions are counted as correct. These complementary metrics highlight different aspects of diagnostic ability under distinct evaluation frameworks.

Consistent with these findings, the ordinal logistic regression provided a quantitative estimate of these differences: GPT-4.0 was significantly more likely to achieve a higher diagnostic score compared with GPT-3.5 (OR 1.18, 95% CI 1.11-1.25; P<.001). In contrast, GLM-4-Plus demonstrated ORs comparable with those of GPT-3.5 (OR 0.95, 95% CI 0.90-1.01), indicating no statistically significant difference in performance between the 2 models. These results further support the interpretation that GPT-4.0 provides a measurable performance advantage, whereas GLM-4-Plus and GPT-3.5 exhibit broadly similar diagnostic capability.

Significant differences were observed across most age groups, with diagnostic accuracy generally increasing with age. Older adult patients exhibited notably better diagnostic performance than adolescents and young adults. For GPT-4.0 and GLM-4-Plus, all pairwise adjacent age-group comparisons were statistically significant (all P<.001). For GPT-3.5, accuracy also increased with age overall; however, the comparison between middle-aged and older adult patients did not reach statistical significance (P=.15). This trend likely reflects several clinical and documentation-related factors: older adults typically present with more stable symptom trajectories, longer disease histories, and more structured and complete EHR documentation, whereas younger patients often exhibit heterogeneous and fluctuating emotional and behavioral symptoms, which increase diagnostic uncertainty.

Across diagnostic categories, GPT-4.0 consistently showed stronger contextual reasoning ability and more accurate extraction of relevant diagnostic cues. It performed particularly well in mood disorders (F3), personality disorders (F6), and several neurotic and stress-related disorders (F4). GLM-4-Plus demonstrated strong performance in schizophrenia spectrum disorders (F2), suggesting an advantage in identifying disorders with clearly structured symptom patterns. GPT-3.5 exhibited the highest rate of misclassification, particularly between disorders with overlapping emotional or behavioral symptoms.

In low-prevalence categories, such as pervasive developmental disorders (F8), all 3 models showed instability in performance. GPT-4.0 achieved perfect recall but exhibited low precision, suggesting overidentification in categories with limited diagnostic cues. These patterns highlight the ongoing challenge

XSL•FO
RenderX

for general-purpose LLMs in identifying rare psychiatric conditions.

Beyond model comparison, this study demonstrates the overall feasibility of applying large-scale language models to real-world psychiatric EHRs covering all *ICD-10* categories. The results indicate that GPT-4.0 can extract diagnostically relevant information efficiently and generate multidimensional diagnostic suggestions, thereby supporting initial clinical screening, reducing physician workload, and improving efficiency in high-volume psychiatric settings. However, LLM-based diagnosis remains dependent on the quality of EHRs, symptom clarity, and physician interpretation, and cannot be used as a substitute for clinical expertise or individualized assessment.

Overall, GPT-4.0 demonstrated the strongest, most stable, and most clinically meaningful diagnostic performance across age groups and psychiatric categories, supporting its potential value as a diagnostic assistance tool within real-world psychiatric workflows.

Center-Level Consistency and Its Implications

Although the 6 centers differed in documentation habits, patient demographics, and narrative styles, all 3 LLMs demonstrated stable diagnostic performance across sites. This consistency suggests that the models rely primarily on core clinical symptom patterns rather than center-specific narrative conventions, which may explain their robustness in real-world settings. The absence of center-level variability also indicates that LLM-based diagnostic support may generalize well across institutions with heterogeneous clinical workflows—a critical requirement for real-world deployment. From an implementation perspective, these findings imply that LLM-assisted diagnostic tools may not require site-specific retraining or local optimization, substantially reducing the technical and operational burden associated with clinical adoption. Collectively, these observations highlight the scalability and practical utility of LLMs in psychiatric diagnostic workflows.

Misclassification Patterns in the F8 Category

The extreme imbalance between recall and precision in the F8 category warrants further clinical interpretation. Although GPT-4.0 successfully identified all true F8 cases, its low precision indicates a substantial number of FP predictions. Upon reviewing these cases, we found that most FPs originated from F9 (behavioral and attention-deficit/hyperactivity disorder–related disorders), with a smaller proportion from F7 (intellectual disabilities). These diagnostic groups share overlapping behavioral and neurodevelopmental characteristics with F8—including impulsivity, irritability, social dysfunction, and developmental impairment—which may lead the model to rely on broad behavioral descriptors when more specific differentiating features are not explicitly documented in the narrative EHR. This overdiagnosis pattern suggests that GPT-4.0 is sensitive to general behavioral dysregulation but is less effective in distinguishing between rare neurodevelopmental disorders with partially overlapping phenotypes. Understanding this tendency toward misclassification is essential for evaluating the model's safety profile when applied to clinically rare categories.

Comparison With Previous Work

The superior diagnostic performance of GPT-4.0 observed in this study is consistent with previous evaluations, which report that newer-generation LLMs outperform earlier versions in clinical reasoning tasks [24-26]. This performance advantage has been attributed to the enhanced ability of GPT-4.0 to capture semantic associations and long-range contextual patterns in complex textual inputs [27]. Previous studies have demonstrated that GPT-4.0 provides higher diagnostic accuracy, better contextual understanding, and more stable reasoning patterns than does GPT-3.5 across diverse clinical applications. Our findings extend this body of evidence to the domain of psychiatric diagnosis via Chinese-language EHRs in real-world settings.

A major distinction between our study and earlier work is the use of multicenter real-world psychiatric EHR narratives rather than simulated vignettes or single-institution datasets. The complexity and variability of real-world clinical documentation allow for a more realistic assessment of LLM behavior in routine psychiatric practice, thereby providing stronger ecological validity.

Consistent with the broader literature, earlier-generation models, such as the GPT-3.5, tend to perform less reliably in disorders characterized by overlapping emotional or behavioral symptoms [24-26]. In our study, this general pattern was reflected in GPT-3.5's tendency to misclassify personality disorders (F6) as mood disorders (F3), likely due to the similarity and ambiguity of symptom descriptions in these categories within real-world psychiatric EHRs. This observation aligns with the known challenges faced by earlier LLMs in distinguishing clinically overlapping symptom domains. In contrast, GLM-4-Plus demonstrated strong performance in schizophrenia spectrum disorders (F2), suggesting that models trained on Chinese corpus characteristics may perform well in categories with clear and structured symptom patterns.

Beyond diagnostic accuracy, prior studies have highlighted the potential role of LLMs as clinical decision support tools capable of improving workflow efficiency and reducing clinician workload in real-world health care settings [28]. In psychiatry, recent evaluations have summarized the diagnostic accuracy of LLMs and emphasized their potential utility in assisting clinicians with complex diagnostic decisions [29]. More broadly, advances in deep learning–based electronic health record analytics have demonstrated the feasibility of extracting clinically meaningful information from narrative EHRs to support timely screening and downstream decision-making processes [30].

Taken together, our findings reinforce the performance advantages of GPT-4.0 observed in previous work and highlight the importance of evaluating both international and domestic LLMs in localized, real-world psychiatric contexts.

Limitations

This study has several limitations. First, although we used multicenter real-world psychiatric EHRs, differences in documentation style—including variability in narrative structure, terminology, completeness of information, and implicit

XSL•FO
RenderX

diagnostic cues—may have influenced model performance. These inconsistencies can alter the prominence, clarity, and temporal linkage of symptom descriptions, potentially affecting diagnostic outputs. Broader validation in institutions with diverse linguistic and cultural backgrounds is needed to assess cross-system and cross-national generalizability more fully [31].

Second, this study adopted a single unified prompting strategy to ensure consistent comparability across models. However, previous research has shown that more advanced prompting techniques can improve diagnostic reasoning. For example, Savage et al [32] reported that diagnostic reasoning prompts enhance LLMs' ability to emulate stepwise clinical reasoning compared with standard prompts. In our setting, a small exploratory pilot study using chain-of-thought prompts on 200 randomly selected cases suggested a modest improvement in exact-match accuracy (about 2-3 percentage points), particularly for mood and anxiety disorder categories. Although this was not a full comparative experiment, the results suggest that the prompting strategy may influence diagnostic performance. Future studies should systematically evaluate zero-shot, few-shot, and chain-of-thought prompt approaches across diagnostic categories and age groups.

Third, the study used a single-label classification framework based on primary discharge diagnoses. While this reflects routine clinical practice, it does not capture psychiatric comorbidities, which are common and clinically important. Multilabel diagnostic frameworks may be necessary to better reflect the real-world complexity of diagnosis.

Fourth, our analysis relied on static EHR text and did not incorporate longitudinal or multimodal clinical information. Real-world psychiatric care involves dynamic symptom progression and treatment trajectories, which cannot be fully captured in static narratives. Integrating time series EHR data, imaging, or laboratory results may increase the diagnostic robustness of LLMs [33].

Fifth, the dataset did not include key socioeconomic variables—such as education level, employment or income status, or urban versus rural residence—which limited our ability to examine potential algorithmic bias or fairness across different social strata. These factors can shape health care access, diagnostic pathways, and linguistic characteristics of clinical documentation, and their absence prevented a systematic assessment of whether model performance varied across sociodemographic groups. Future studies incorporating richer socioeconomic status–related demographic information will be essential for evaluating equity in LLM-assisted diagnostic workflows.

Sixth, although admission diagnoses were removed from model inputs, sections, such as the history of present illness, may still include implicit diagnostic tendencies embedded within physicians' narrative styles. To address this concern, we manually reviewed a subset of correctly and incorrectly classified cases to determine whether the model relied on "giveaway" phrases or diagnosis-specific linguistic cues. We did not identify consistent or systematic patterns—such as psychopharmacological keywords, standardized phrasing, or clinician-specific formulations—that uniquely pointed to a specific diagnosis. When such terms occurred (eg, references to antidepressants or antipsychotics), they were typically nonspecific and appeared across multiple diagnostic categories.

However, certain unmasked psychotropic medications—particularly lithium and clozapine, which function as strong diagnostic proxies—may have allowed the models to infer diagnoses through medication extraction rather than full clinical reasoning, potentially inflating accuracy in specific categories. These observations suggest that while soft leakage cannot be entirely eliminated in narrative EHRs, its impact on model performance in this dataset was likely limited. Future studies should consider masking medication names or evaluating model behavior under medication-removed conditions to better isolate true diagnostic reasoning.

In addition, this study has privacy-related limitations. Although all EHR narratives were deidentified before transmission to external LLM servers, narrative clinical text carries inherent challenges for complete anonymization. Despite multiple safeguards, the possibility of residual identifiers cannot be entirely eliminated, and future research should prioritize on-premise or institution-governed LLM deployment to avoid external data transmission fully. Moreover, because the manual deidentification audit was conducted on a sampled subset of 500 records rather than the full dataset, a small risk of undetected residual identifiers cannot be completely excluded.

Furthermore, the study did not include a human benchmark for comparison, such as the diagnostic accuracy of junior clinicians, admitting psychiatrists, or structured admission diagnoses. Without this comparative context, it is difficult to ascertain whether the observed strict accuracy of 71.7% represents a clinically meaningful performance threshold. Future research should incorporate human-model comparison frameworks or prospective reader studies to more clearly establish the clinical utility of LLM-assisted diagnostic workflows.

Finally, although multicenter data improved external validity, further validation through prospective observational designs or randomized controlled trials is necessary to robustly establish the real-world clinical utility of LLMs in psychiatric diagnostic workflows [31].

Future Directions

Future research should expand the diversity of data sources to evaluate the generalizability of LLMs across different health care systems, linguistic environments, and cultural contexts. The incorporation of longitudinal EHR data, time series symptom trajectories, and multimodal clinical information—such as laboratory tests and imaging—may further enhance the ability of LLMs to capture dynamic clinical changes and improve diagnostic robustness.

Furthermore, more sophisticated strategies warrant systematic investigation. Comparative evaluations of zero-shot, few-shot, and chain-of-thought prompts across different psychiatric categories and age groups may help optimize the integration of LLMs into clinical workflows. Future studies should also explore multilabel diagnostic frameworks to capture psychiatric comorbidities better and reflect the complexity of real-world clinical presentations.

XSL•FO
RenderX

Finally, prospective validation—particularly through observational studies or randomized controlled trials—will be essential for establishing the real-world clinical utility, reliability, and safety of LLM-assisted diagnostic tools in psychiatric practice.

Conclusions

This multicenter real-world study systematically evaluated the diagnostic performance of the GPT-4.0, GPT-3.5, and GLM-4-Plus across all *ICD-10* psychiatric categories using Chinese psychiatric EHRs. GPT-4.0 demonstrated the strongest and most stable diagnostic performance in this study across age groups and diagnostic categories and showed clear advantages in interpreting complex psychiatric narratives. All models performed less reliably in low-prevalence categories, and diagnostic accuracy varied substantially across age groups, with lower accuracy observed in adolescents and young adults.

These findings confirm the feasibility of applying LLMs to real-world psychiatric EHRs and highlight the potential value of GPT-4.0 as a diagnostic support tool. However, the diagnostic outputs of LLMs remain constrained by the quality of documentation, symptom clarity, and the limitations of the models. LLMs should therefore be viewed as complementary tools that can assist clinicians but cannot replace clinical expertise, individualized assessment, or professional decision-making in psychiatric practice.

Acknowledgments

Generative artificial intelligence tools (ChatGPT and OpenAI) were used to assist with language refinement and editing. All content was reviewed and verified by the authors.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to patient privacy, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MS, JY, GH
Methodology: MS, JY
Data curation: MS, JY, ZL, YY, TX
Formal analysis: MS, JY
Investigation: MS, JY, ZL, YY, TX
Writing – original draft: MS, JY, JL
Writing – review & editing: MS, JY, JF, HD, GH
Supervision: GH
Project administration: GH
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Instrument and additional tables.
[DOCX File , 19 KB - medinform_v14i1e77699_app1.docx]

References

1. Freeman M. The world mental health report: transforming mental health for all. World Psychiatry 2022;21(3):391-392 [FREE Full text] [doi: 10.1002/wps.21018] [Medline: 36073688]
2. Murray CJL, GBD 2021 Collaborators. Findings from the global burden of disease study 2021. Lancet 2024;403(10440):2259-2262. [doi: 10.1016/S0140-6736(24)00769-4] [Medline: 38762327]
3. Jacob KS. Mental health services in low-income and middle-income countries. Lancet Psychiatry 2017;4(2):87-89. [doi: 10.1016/S2215-0366(16)30423-0] [Medline: 28063878]
4. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]

XSL•FO
RenderX

5. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
6. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M, et al. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform 2023;11:e44977 [FREE Full text] [doi: 10.2196/44977] [Medline: 37079367]
7. Brown TB, Mann B, Ryder, N, Subbiah, M. Language models are few-shot learners. 2020 Presented at: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 December 6 - 12; Vancouver BC Canada p. 1877-1901.
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023;29(12):2988. [doi: 10.1038/s41591-023-02656-2] [Medline: 37957381]
10. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
11. Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, et al. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol 2024;34(6):3566-3574. [doi: 10.1007/s00330-023-10384-x] [Medline: 37938381]
12. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience 2023;26(9):107590 [FREE Full text] [doi: 10.1016/j.isci.2023.107590] [Medline: 37705958]
13. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol 2024;34(4):779-787. [doi: 10.1007/s00062-024-01426-y] [Medline: 38806794]
14. Kim J, Leonte KG, Chen ML, Torous JB, Linos E, Pinto A, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med 2024;7(1):193 [FREE Full text] [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
15. Otsuka N, Kawanishi Y, Doi F, Takeda T, Okumura K, Yamauchi T, et al. Diagnosing psychiatric disorders from history of present illness using a large-scale linguistic model. Psychiatry Clin Neurosci 2023;77(11):597-604 [FREE Full text] [doi: 10.1111/pcn.13580] [Medline: 37526294]
16. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med 2025;8(1):230 [FREE Full text] [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
17. Reichenpfader D. Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models. 2024 Presented at: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024; 2025 17 December; Torino, Italia.
18. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang C, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inform Decis Mak 2013;13:71 [FREE Full text] [doi: 10.1186/1472-6947-13-71] [Medline: 23842533]
19. First MB, Pincus HA. Classification in psychiatry: ICD-10 v. DSM-IV. A response. Br J Psychiatry 1999 Sep;175:205-209. [doi: 10.1192/bjp.175.3.205] [Medline: 10645319]
20. Zhang H, Ji J. Prevalence of mental disorders in China. Lancet Psychiatry 2019;6(3):188-189. [doi: 10.1016/S2215-0366(19)30034-3] [Medline: 30792113]
21. Perera Molligoda Arachchige AS. Role of ChatGPT 3.5 in emergency radiology, with a focus on cardiothoracic emergencies: proof with examples. iRADIOLOGY 2024;2(5):510-521. [doi: 10.1002/ird3.65]
22. Achiam J, Adler S, Agarwal S. Gpt-4 technical report. arXiv:2303.08774 2024 [FREE Full text]
23. He Y. Chinese simpleqa: A chinese factuality evaluation for large language models. In: arXiv preprint. 2025 Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025 17 December; Vienna, Austria p. 19182-19208. [doi: 10.18653/v1/2025.acl-long.941]
24. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
25. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res 2024;26:e56110 [FREE Full text] [doi: 10.2196/56110] [Medline: 38976865]
26. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr 2024;178(3):313-315. [doi: 10.1001/jamapediatrics.2023.5750] [Medline: 38165685]
27. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligencearly: early experiments with GPT-4. arXiv:2303.12712 2023 [FREE Full text]

XSL•FO
RenderX

28. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. Int J Environ Res Public Health 2021;18(1) [FREE Full text] [doi: 10.3390/ijerph18010271] [Medline: 33401373]
29. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr 2024;100:104168. [doi: 10.1016/j.ajp.2024.104168] [Medline: 39111087]
30. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inform 2018;22(5):1589-1604. [doi: 10.1109/jbhi.2017.2767063]
31. Lotter W, Hassett MJ, Schultz N, Kehl KL, Van Allen EM, Cerami E. Artificial intelligence in oncology: current landscape, challenges, and future directions. Cancer Discov 2024;14(5):711-726. [doi: 10.1158/2159-8290.CD-23-1199] [Medline: 38597966]
32. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024;7(1):20 [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
33. Volinsky-Fremond S, Horeweg N, Andani S, Barkey Wolf J, Lafarge MW, de Kroon CD, et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nat Med 2024;30(7):1962-1973. [doi: 10.1038/s41591-024-02993-w] [Medline: 38789645]

Abbreviations

EHR: electronic health record
FP: false positive
FN: false negative
ICD-10: International Statistical Classification of Diseases, Tenth Revision
LLM: large language model
OR: odds ratio
TP: true positive

Edited by A Coristine; submitted 25.May.2025; peer-reviewed by KH Jung, HJ Kim, J Wang; comments to author 17.Aug.2025; accepted 11.Dec.2025; published 13.Jan.2026.

Please cite as:
Sun M, Yu J, Long Z, Yang Y, Xiao T, Liang J, Feng J, Deng H, Huang G
Large Language Models for Psychiatric Diagnosis Based on Multicenter Real-World Clinical Records: Comparative Study
JMIR Med Inform 2026;14:e77699
URL: https://medinform.jmir.org/2026/1/e77699
doi:10.2196/77699
PMID:41408781

©Maoqian Sun, Jia Yu, Zhuhong Long, Yun Yang, Tao Xiao, Jiaquan Liang, Jun Feng, Huaili Deng, Guoping Huang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptance of Artificial Intelligence in Clinical Practice Among Chinese Physicians: Nationwide Cross-Sectional Survey Using Extended Unified Theory of Acceptance and Use of Technology and Explainable Machine Learning

Xuefei Shi[1], MMed; Zhanxiao Tian[2], Dr Med; Qi Guo[2], MSc; Boxuan Qiao[3], PhD; Xiaolong Wang[1], PhD

[1]Department of Cardiology, Shuguang Hospital, Shanghai University of Traditional Chinese Medicine, 185 Pu'an Road, Huangpu District, Shanghai, China

[2]Research and Development Center, KINGYEE (Beijing) Co., Ltd., Beijing, China

[3]College of Computer Science, Zhejiang University, Hangzhou, China

Corresponding Author:
Xiaolong Wang, PhD
Department of Cardiology, Shuguang Hospital, Shanghai University of Traditional Chinese Medicine, 185 Pu'an Road, Huangpu District, Shanghai, China

Abstract

Background: Artificial intelligence (AI) is rapidly transforming clinical practice, yet empirical evidence on Chinese physicians' acceptance of AI medical tools remains scarce at the national level.

Objective: This study aimed to evaluate the current acceptance of AI medical tools among Chinese physicians, identify key determinants, and elucidate underlying mechanisms using an extended Unified Theory of Acceptance and Use of Technology (UTAUT) and explainable machine learning.

Methods: A nationwide cross-sectional survey was conducted from January to April 2024, recruiting 4024 in-service physicians across 29 provincial-level administrative units in China via stratified random sampling. The questionnaire incorporated 5 UTAUT constructs—performance expectancy, effort expectancy, social influence, facilitating conditions (FC), and a newly introduced "positive impact" dimension. Psychometric properties were validated through exploratory and confirmatory factor analyses. Structural equation modeling assessed direct and moderated effects, with hospital level, professional title, AI familiarity, and future optimism as moderators. Six classification models were compared for predictive performance; balanced random forest was selected, and model interpretability was evaluated using Shapley Additive Explanations (SHAP).

Results: Overall acceptance exceeded 90% across subgroups. Structural equation modeling showed that performance expectancy, social influence, FC, and positive impact significantly and positively predicted physicians' behavioral intention to use AI medical tools. Six negative moderation effects were identified. The random forest achieved 85.6% accuracy and an area under the receiver operating characteristic curve of 0.836; SHAP analysis identified organizational support (FC_HospPromoteAI) as the feature with the highest mean absolute SHAP value, though all effect sizes were modest.

Conclusions: Chinese physicians exhibit high acceptance of AI medical tools, mainly driven by organizational support and perceived clinical benefits. The combined use of extended UTAUT and explainable AI provides actionable insights for targeted AI implementation strategies in health care.

(JMIR Med Inform 2026;14:e85270) doi:10.2196/85270

KEYWORDS

machine learning; physicians; attitude of health personnel; UTAUT; Unified Theory of Acceptance and Use of Technology; SHAP; Shapley Additive Explanations

Introduction

Artificial intelligence (AI) is transforming health care delivery worldwide by enhancing efficiency and quality through automation and intelligent systems [1,2]. Its applications include medical imaging analysis, clinical decision support, and chronic disease management. Convolutional neural networks based on deep learning have demonstrated superior performance in imaging diagnostics [3]. AI medical tools can reduce diagnostic time by approximately 30% and improve accuracy by about 10%, particularly in oncology and cardiovascular medicine [4]. In recent years, large language models such as ChatGPT, Med-PaLM (Google LLC), and domestic systems like DeepSeek and Doubao (ByteDance Ltd.) have shown significant potential

XSL•FO
RenderX

in generating medical records, translating scientific literature, supporting clinical decision-making, and facilitating physician-patient communication [5]. In China, the adoption of AI tools for imaging analysis and decision support has steadily increased, especially in tertiary hospitals [6]. However, concerns about algorithmic bias, data privacy, and accountability remain key barriers to widespread adoption, making physicians' acceptance a critical factor in successful implementation [7,8].

Globally, physicians' attitudes toward AI medical tools are complex. In Western countries, about half of physicians recognize the potential of AI to improve diagnostic efficiency, but many express concerns about privacy and ethics [9,10]. In the United Kingdom, some physicians have refused to use AI imaging tools due to insufficient training [11]. Algorithmic bias, such as racially skewed diagnostic errors, has been shown to undermine trust, particularly when AI-generated recommendations come from large language models [12]. A small-scale Japanese survey indicated that high learning costs also discourage adoption of AI medical tools [13]. Health system differences further influence acceptance: physicians in predominantly private health care systems in Western countries generally enjoy higher autonomy, more resources, and shorter working hours (about 40 h/wk) [14], whereas in China's predominantly public health care system, physicians often work more than 50 hours/week [15]. Existing Chinese studies are mostly regional or small-scale, lack nationally representative data, and rarely integrate explainable machine learning approaches [5,16].

The technology acceptance model identifies perceived usefulness and perceived ease of use as core drivers of adoption [17]. The Unified Theory of Acceptance and Use of Technology (UTAUT) expands this framework by adding social influence (SI), facilitating conditions (FC), and risk perceptions [18]. However, most UTAUT-based studies are conducted in Western health care contexts and have not addressed the unique challenges of China's public health care system, such as high workloads and reduced autonomy [15]. Furthermore, the positive impacts (PIs) of AI—such as improving efficiency, enhancing diagnostic accuracy, and strengthening physician-patient relationships—have not been systematically incorporated into the UTAUT framework.

Therefore, this study aimed to assess the current acceptance of AI medical tools among Chinese physicians, identify key influencing factors and potential moderating effects, and apply explainable machine learning to uncover the complex mechanisms underlying physicians' behavioral intentions (BIs) toward AI adoption.

Methods

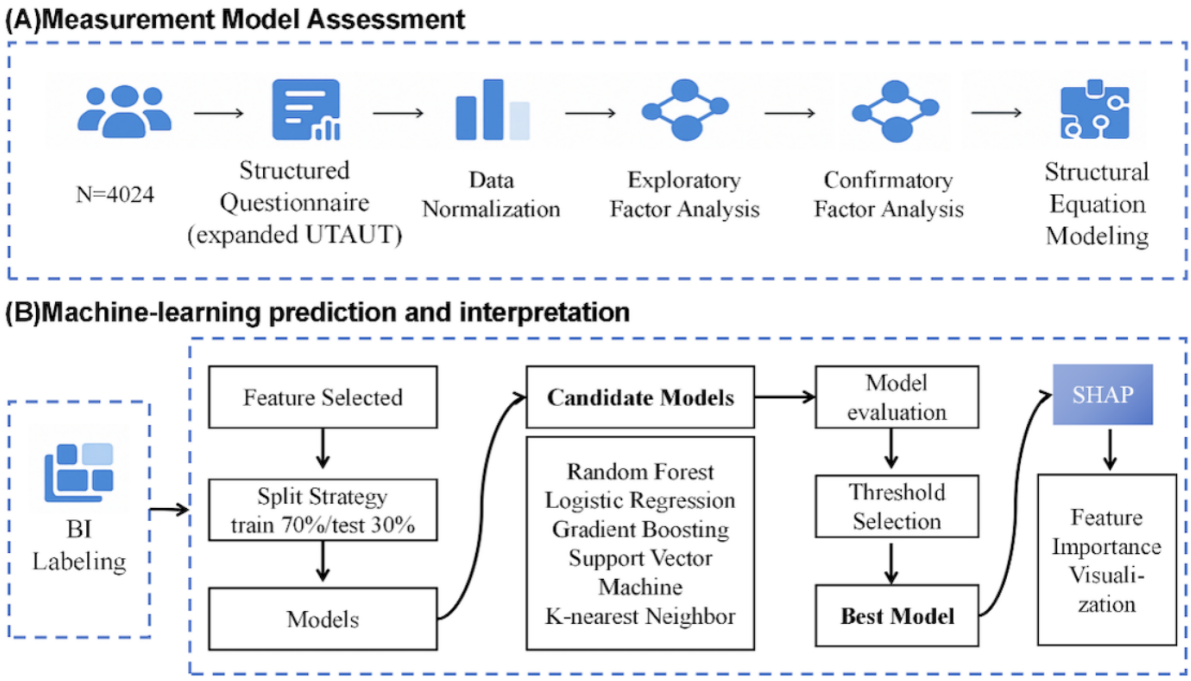
Study Design and Participants

This study used a nationwide cross-sectional survey (January-April 2024) to assess Chinese physicians' acceptance of AI medical tools and related determinants. The questionnaire was developed by adapting established scales from the UTAUT [17]. It included 4 original constructs—performance expectancy (PE), effort expectancy (EE), SI, and FC—and an additional "PI" construct. The PI items were derived from the concept of "perceived usefulness" [18] and contextualized to capture specific perceived clinical benefits. The conceptual distinction is clarified as follows: PE items focused on task-level efficiency gains (eg, "AI medical tools can improve health care efficiency"), whereas PI items captured broader clinical and patient-oriented outcomes (eg, "AI can help improve patients' clinical outcomes"; "AI can improve diagnostic accuracy"). While PE reflects expectations about operational performance improvements, PI emphasizes the perceived benefits for clinical quality and patient welfare.

Structured questionnaires were distributed online via Medlive, which is the largest leading professional academic platform for physicians in China, possessing a comprehensive database of verified real-name physicians. To ensure representativeness, we used a stratified random sampling strategy based on 2 key dimensions: geographic location (31 provincial-level administrative regions in mainland China) and hospital classification (tertiary, secondary, and primary institutions). The sampling aimed to achieve nationwide coverage across all provinces rather than predetermined numerical quotas for each stratum. The platform randomly pushed survey invitations to eligible physicians within these strata. A total of 4725 physicians clicked the survey invitation link, of whom 4030 (85.3%) completed all required items. After excluding 6 responses with inconsistent data patterns, 4024 valid questionnaires were retained for analysis, ultimately covering 29 of 31 provincial-level regions. The overall study framework is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of study design and analysis workflow. (A) Psychometric and structural modeling. (B) Machine learning prediction and interpretation. BI: behavioral intention; SHAP: Shapley Additive Explanations; UTAUT: Unified Theory of Acceptance and Use of Technology.



Measurement Model Assessment

The instrument comprised 32 items across 6 dimensions: BI, PE, EE, SI, FC, and PI. BI was assessed using 6 dichotomous items (1=willing, 0=unwilling). PE, PI, SI, and FC each contained 4 items scored on a 4-point Likert scale (1=strongly disagree to 4=strongly agree). All construct scores were linearly transformed to a 0-100 scale for comparability across dimensions.

Exploratory factor analysis was conducted using principal component analysis with Varimax rotation; factors with eigenvalues greater than 1 were retained. Sampling adequacy and sphericity were evaluated using the Kaiser-Meyer-Olkin statistic and Bartlett's test, respectively [16]. Confirmatory factor analysis was then performed within a structural equation modeling (SEM) framework, and model fit was evaluated using root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI) according to established cutoffs [17]. Internal consistency (Cronbach α), composite reliability (CR), and average variance extracted (AVE) were computed for each construct.

BI was measured using 6 dichotomous (yes or no) items. We acknowledged that weighted least squares mean and variance-adjusted or robust diagonally weighted least squares estimators would be more appropriate for categorical indicators. However, given that the primary analysis used maximum likelihood (ML) estimation consistently across all constructs, we reported the ML results while noting this methodological consideration.

SEM estimated the direct effects of PE, EE, SI, FC, and PI on BI while controlling for hospital level and professional title [18]. Moderation effects were examined for hospital level, AI familiarity, and future optimism by standardizing variables, creating product-term interactions, and entering them hierarchically into regression models [19]. Multicollinearity was evaluated using variance inflation factors.

Machine Learning Prediction and Interpretability

To capture potential nonlinear and high-dimensional associations, 6 supervised classifiers were compared: logistic regression (LR), support vector machine, random forest (RF), gradient boosting (GB), extreme gradient boosting, and *k*-nearest neighbors [20-25]. The binary outcome was defined as high acceptance (BI score ≥ 83.3, corresponding to "yes" responses on 5 or more of the 6 items; n=3622, 90%) versus moderate or low acceptance (BI score <83.3; n=402, 10%). Data were partitioned into training (n=2817, 70%) and testing (n=1207, 30%) sets using stratified sampling to preserve class proportions. Five-fold stratified cross-validation on the training set was used for model selection and hyperparameter tuning. For tree-based models, key hyperparameters (max_depth, n_estimators, min_samples_split) were optimized via grid search.

Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F_1-score. Given class imbalance, precision-recall AUC was additionally considered for robustness [26]. Decision thresholds were optimized using the Youden index to balance sensitivity and specificity [27].

The balanced RF was selected for Shapley Additive Explanations (SHAP) interpretation for three reasons: (1) it achieved the highest accuracy (85.6%) and sensitivity (87.3%) among all models, which is critical for identifying physicians with potential implementation barriers; (2) tree-based SHAP values provide additive, locally accurate explanations that are well-suited for identifying feature interactions; and (3) the

XSL•FO
RenderX

class-balancing mechanism addresses the 9:1 class imbalance, producing more reliable feature importance estimates for the minority class. While LR and GB achieved equivalent or higher AUC (0.840), we prioritized sensitivity and interpretability for this exploratory analysis. SHAP was applied to quantify each feature's contribution to the predicted probability of high AI acceptance, providing global importance rankings (mean absolute SHAP values) and local interpretability (summary plots) [28].

Statistical Analysis

Categorical variables were summarized as counts and percentages, and continuous variables were reported as medians with IQRs. Group comparisons used the chi-square test or Fisher exact test for categorical variables and the Wilcoxon rank-sum test for continuous variables, as appropriate. Two-sided *P* values <.05 were considered statistically significant; when multiple pairwise comparisons were performed, Bonferroni correction was applied. All statistical analyses were conducted in Python (version 3.10.0) using standard libraries (pandas, NumPy, SciPy, and statsmodels) for inference and scikit-learn for model-related procedures.

Ethical Considerations

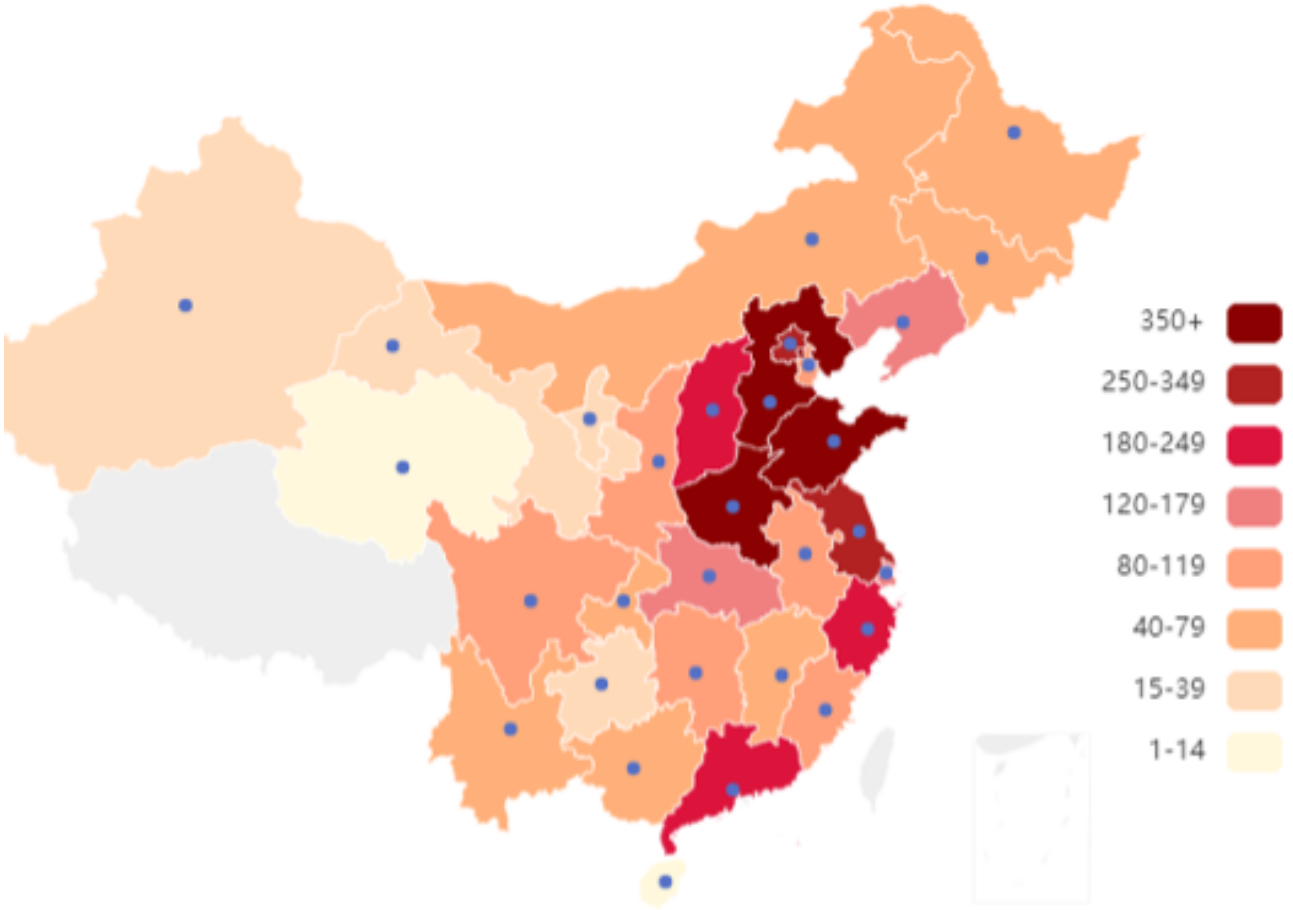
The study was conducted in accordance with the Declaration of Helsinki. Prior to data collection, all participants were presented with an electronic informed consent statement detailing the study objectives, its voluntary nature, and the privacy measures. Participants were required to confirm their agreement by clicking a mandatory "I Agree" button to proceed to the questionnaire. As the study involved a noninterventional, anonymous survey of professionals, formal institutional review board approval was waived in accordance with the operational guidelines of the data collection platform (Medlive), provided that strict data privacy and anonymity protocols were enforced. All participants provided informed consent prior to data collection. No financial compensation or incentives were provided to participants for their involvement in this study.

Results

Sample Characteristics

A total of 4024 in-service physicians from mainland China were included. As shown in Figure 2, the sample covered 29 provincial-level regions, with most participants located in East China (n=1143, 28.4%) and North China (n=978, 24.3%). Tertiary hospitals accounted for 3114 (77.4%), followed by secondary hospitals (n=717, 17.8%) and primary care institutions (193, 4.8%). Regarding professional title, 2933 (72.9%) held senior titles (chief or associate chief), 1040 (25.8%) held intermediate titles, and 51 (1.3%) were junior or ungraded.

Figure 2. Geographic distribution of physicians across China (n=4024). Provinces are shaded by the number of participants, with darker colors indicating higher counts. A total of 29 provinces were covered in the survey. Blue dots represent provincial capital cities.



XSL•FO
RenderX

Measurement Properties of the Questionnaire

The AVE for BI was 0.40, slightly below the conventional threshold of 0.50. This lower AVE is consistent with the strong ceiling effect observed in BI responses, where over 90% (n=3621) of participants indicated willingness to use medical AI across all 6 dimensions (range: 90.6% - 96.4%). Despite the lower AVE, BI demonstrated good internal consistency through CR (0.80) and Cronbach α (0.80), with all standardized factor loadings exceeding 0.59.

As detailed in Table S1 in Multimedia Appendix 1, exploratory factor analysis indicated adequate sampling and factorability (Kaiser-Meyer-Olkin=0.898; Bartlett test χ^2=47,985.41; *P*<.001). Based on eigenvalues greater than 1 and scree plot inspection, 6 factors were extracted, explaining 62.8% of the total variance, which is consistent with the extended UTAUT specification. All standardized factor loadings exceeded 0.40 (Table S1 in Multimedia Appendix 1), and internal consistency was acceptable to good across constructs (mean Cronbach α >0.78).

Confirmatory factor analysis further supported the 6-factor model with good fit (RMSEA=0.043; CFI=0.937; TLI=0.930). CR exceeded 0.80 for all constructs, and AVE was greater than 0.50 for every construct except BI (AVE=0.478), indicating satisfactory convergent validity (Table 1). Discriminant validity was supported, as the square root of each construct's AVE on the diagonal was greater than the corresponding interconstruct correlations (Table 2).

Table . Convergent validity of constructs for Chinese physicians' acceptance of artificial intelligence medical tools. All factor loadings are standardized.

Construct	Cronbach α	CR[a]	AVE[b]	Average loading
EE[c]	0.894	0.915	0.642	0.766
PE[d]	0.818	0.830	0.549	0.729
PI[e]	0.792	0.861	0.608	0.700
SI[f]	0.828	0.851	0.588	0.739
FC[g]	0.852	0.849	0.587	0.774
BI[h]	0.788	0.846	0.478	0.691

[a]CR: composite reliability.
[b]AVE: average variance extracted.
[c]EE: effort expectancy.
[d]PE: performance expectancy.
[e]PI: positive impact.
[f]SI: social influence.
[g]FC: facilitating conditions.
[h]BI: behavioral intention.

Table . Discriminant validity for Chinese physicians' acceptance of artificial intelligence medical tools.

Construct	EE[a]	PE[b]	PI[c]	SI[d]	FC[e]	BI[f]
EE	*0.766*					
PE	0.321	*0.730*				
PI	−0.063	0.195	*0.702*			
SI	0.203	0.598	0.122	*0.742*		
FC	0.179	0.590	0.202	0.536	*0.778*	
BI	0.118	0.363	0.156	0.327	0.438	*0.633*

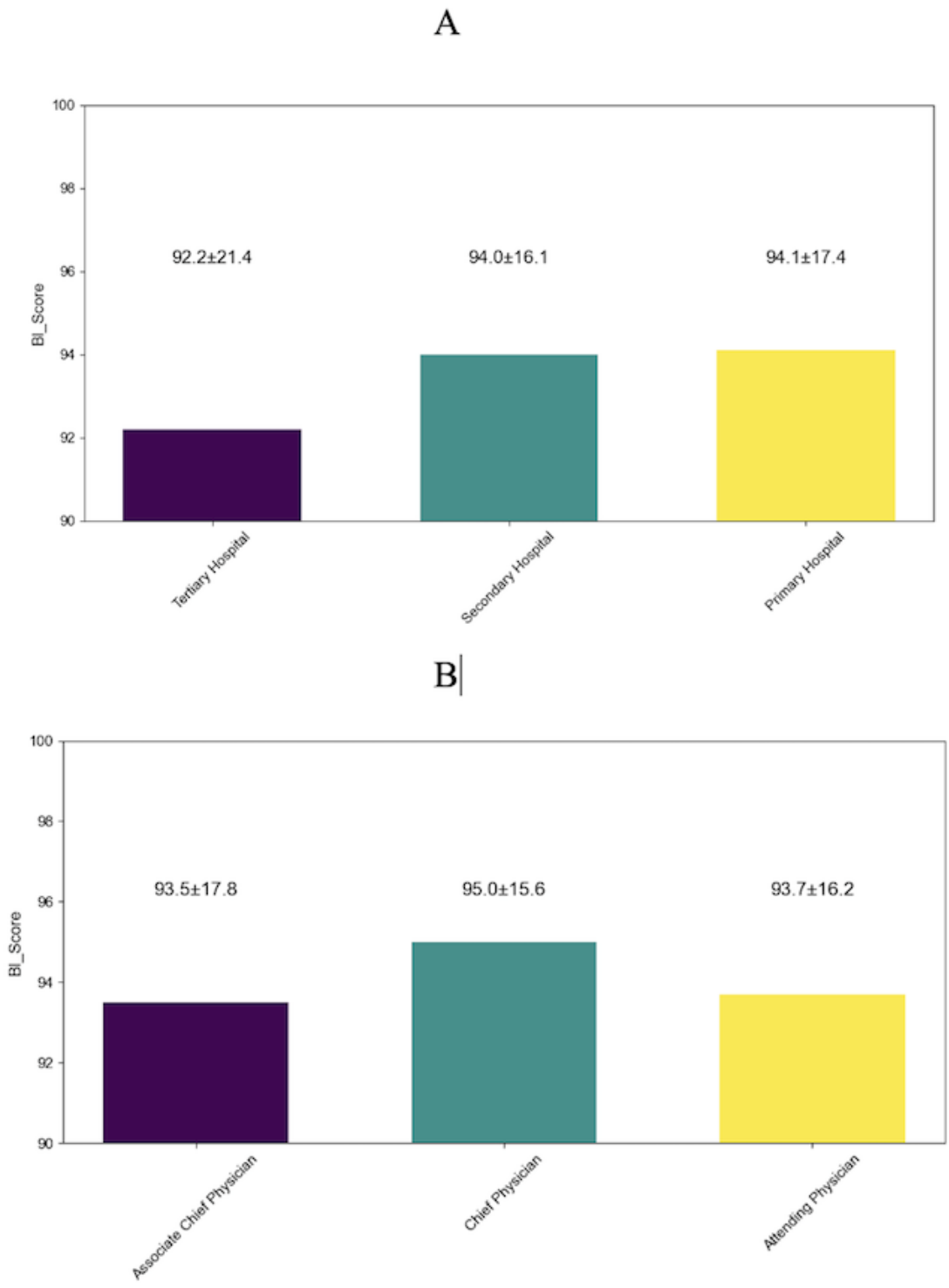
[a]EE: effort expectancy.
[b]PE: performance expectancy.
[c]PI: positive impact.
[d]SI: social influence.
[e]FC: facilitating conditions.
[f]BI: behavioral intention.

XSL•FO
RenderX

Acceptance Across Subgroups

As shown in Figure 3, overall willingness to adopt AI medical tools exceeded 90% (3622/4024). Descriptive analysis of the aggregated BI score revealed a negative skewness of −3.55 and a kurtosis of 13.91, statistically confirming the nonnormal distribution associated with the ceiling effect. For inferential analyses, categories with value n less than 50 were excluded (private hospitals, n=29; clinics, village health posts, or outpatient departments, n=11; junior title, n=46; ungraded, n=5). One-way ANOVA indicated no significant between-group differences by hospital level and a borderline difference by professional title that became nonsignificant after Bonferroni correction.

Figure 3. Overall acceptance of artificial intelligence medical tools across subgroups: (A) by hospital level and (B) by professional title. Categories with n<50 were excluded from inferential tests. Bars indicate the proportion reporting willingness (BI=1). BI score was computed as the unweighted arithmetic mean of the 6 items and linearly transformed to a 0 - 100 scale. BI: behavioral intention.

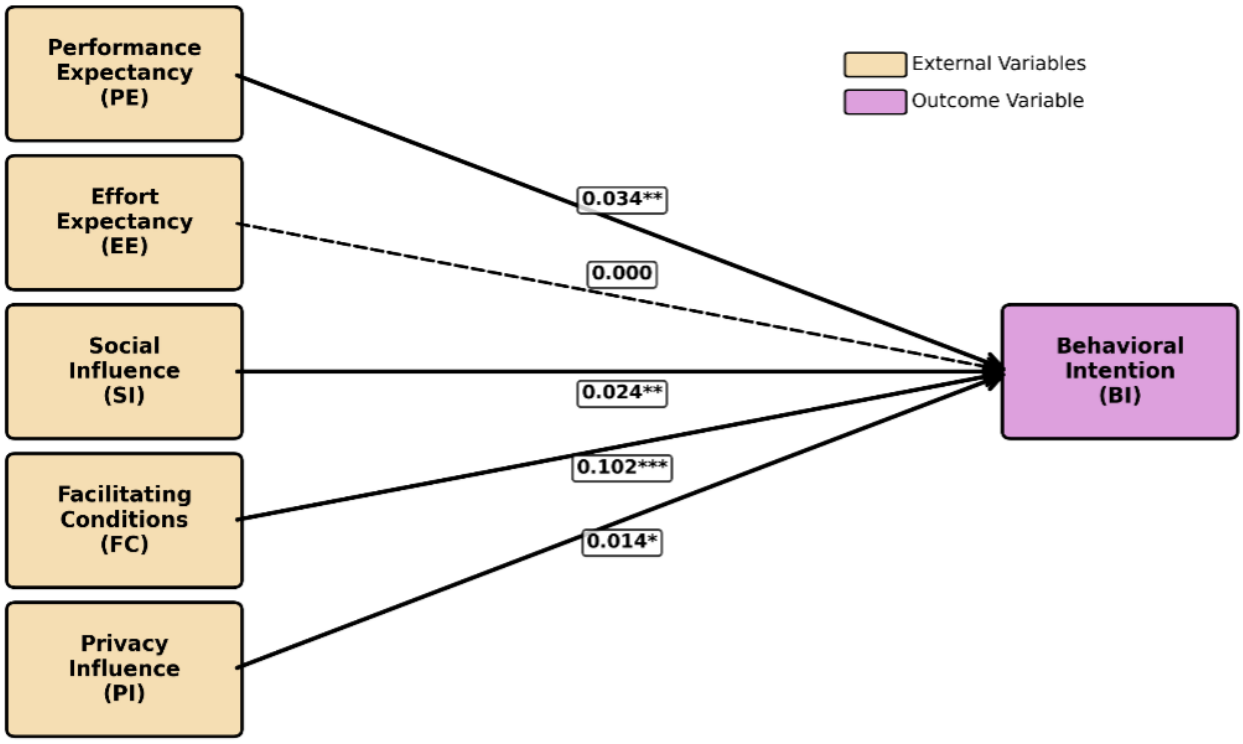


XSL•FO
RenderX

Structural Model Results

The extended UTAUT structural model demonstrated a good fit to the data (RMSEA=0.043; CFI=0.948; TLI=0.941). As shown in Figure 4, PE (β=.03, 95% CI 0.014-0.055; *P*=.001), SI (β=.02, 95% CI 0.010-0.039; *P*<.001), FC (β=.10, 95% CI 0.085-0.119; *P*<.001), and PI (β=.01, 95% CI 0.000-0.028; *P*=.047) were positive predictors of BI, whereas EE was not significant (β≈0, 95% CI −0.014 to 0.014]; *P*=.99). The model explained 21.6% of the variance in BI (R^2=0.216).

Figure 4. Structural path model based on the extended Unified Theory of Acceptance and Use of Technology framework. Solid lines indicate statistically significant paths, dashed lines indicate nonsignificant paths. Path coefficients are standardized β values, with significance levels denoted as *P*<.05 (*), *P*<.01 (**), and *P*<.001 (***). External variables (performance expectancy, effort expectancy, social influence, facilitating conditions, positive impact) are shown in beige and the outcome variable (BI) is shown in purple (R^2 for BI=0.216). BI: behavioral intention.



Moderation Analysis

To examine whether background variables altered the primary UTAUT paths, product-term interactions were created using standardized latent variables for the following moderators: hospital level (W5), professional title (W7), AI familiarity (N1), and future optimism (N5). Hierarchical regression identified 6 negative interaction effects (*P* values <.05), indicating attenuation of the corresponding main effects at higher levels of the moderator (Table 3). Variance inflation factors ranged from 1.06 to 1.99 (mean 1.54), suggesting no multicollinearity concerns. Other tested interactions were not statistically significant.

XSL•FO
RenderX

Table . Moderation effects on Unified Theory of Acceptance and Use of Technology paths (standardized coefficients).

Interaction (moderator × predictor)	β	*P* value	Direction
EE[a] × W5[b]	−.054	.004	Negative
EE × N1[c]	−.037	.03	Negative
FC[d] × N1	−.048	.01	Negative
EE × N5[e]	−.043	.01	Negative
FC × N5	−.074	<.001	Negative
PI[f] × N5	−.054	.001	Negative

[a]EE: effort expectancy.

[b]W5: hospital level.

[c]N1: artificial intelligence familiarity.

[d]FC: facilitating conditions.

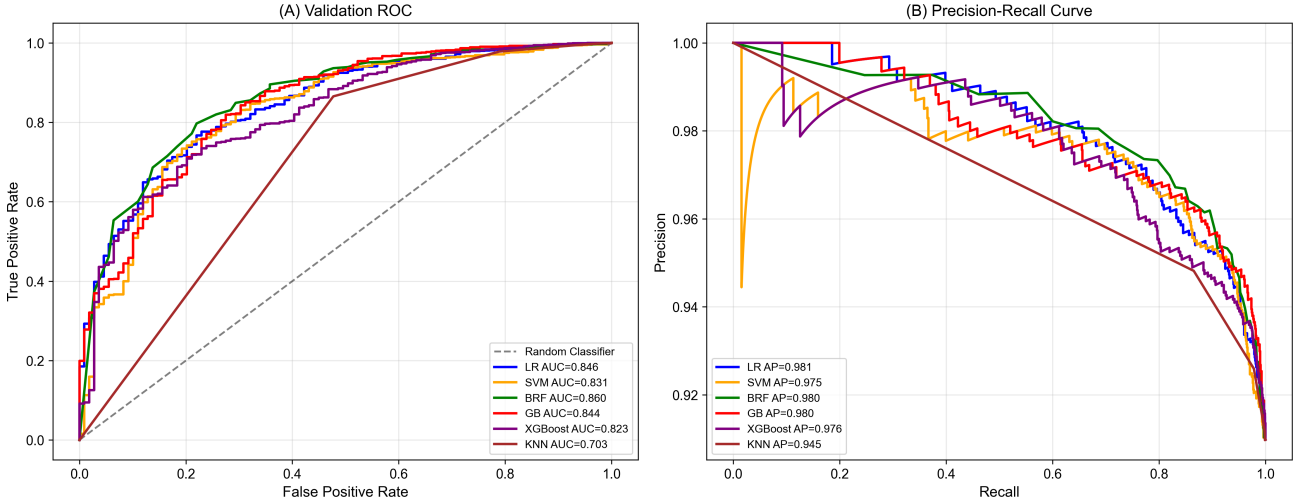
[e]N5: future optimism.

[f]PI: positive impact.

Explainable Prediction Analysis

Given the class imbalance in the dataset, the decision thresholds for each model were optimized according to the Youden index to achieve a balance between sensitivity and specificity [29]. Figure S1 in Multimedia Appendix 2 shows the results of threshold tuning analysis. As shown in Table 4, the LR model (AUC 0.840) and GB model (AUC 0.840) demonstrated superior predictive performance for physicians' acceptance of AI medical tools compared with other ML models. Unadjusted (default) threshold results for all models are provided in Table S3 in Multimedia Appendix 1. Figure 5 shows the receiver operating characteristic and precision or recall curves for the 6 evaluated models. LR achieved the highest specificity (0.877) and average precision, whereas RF yielded the highest accuracy (0.856) and sensitivity (0.873).

Figure 5. Performance comparison of 6 machine learning models. (A) ROC curves showing that LR and gradient boosting achieved the highest AUC (0.840), followed closely by random forest (0.836). (B) Precision-recall curves. Based on the balance between accuracy (0.856) and sensitivity (0.873) shown in Table 4, the BRF model was selected for the subsequent interpretability analysis. AP: average precision; AUC: area under the receiver operator characteristic curve; BRF: balanced random forest; GBDT: gradient boosting decision tree machine; KNN: *k*-nearest neighbor; LR: logistic regression; ML: machine learning; ROC: receiver operating characteristic; SVM: support vector machine; XGBoost: extreme gradient boosting.



XSL•FO
RenderX

Table . Comparison of various machine learning models' performance.

Model	AUC[a]	Accuracy	Sensitivity	Specificity	F_1-score	Youden index
RF[b]	0.836	0.856	0.873	0.685	0.917	0.558
LR[c]	0.840	0.682	0.663	0.877	0.791	0.539
GB[d]	0.840	0.794	0.799	0.740	0.876	0.539
XGBoost[e]	0.822	0.752	0.749	0.781	0.846	0.529
SVM[f]	0.806	0.745	0.742	0.781	0.841	0.523
KNN[g]	0.685	0.835	0.870	0.479	0.905	0.350

[a]AUC: area under the receiver operating characteristic curve.

[b]RF: random forest.

[c]LR: logistic regression.

[d]GB: gradient boosting.

[e]XGBoost: extreme gradient boosting.

[f]SVM: support vector machine.

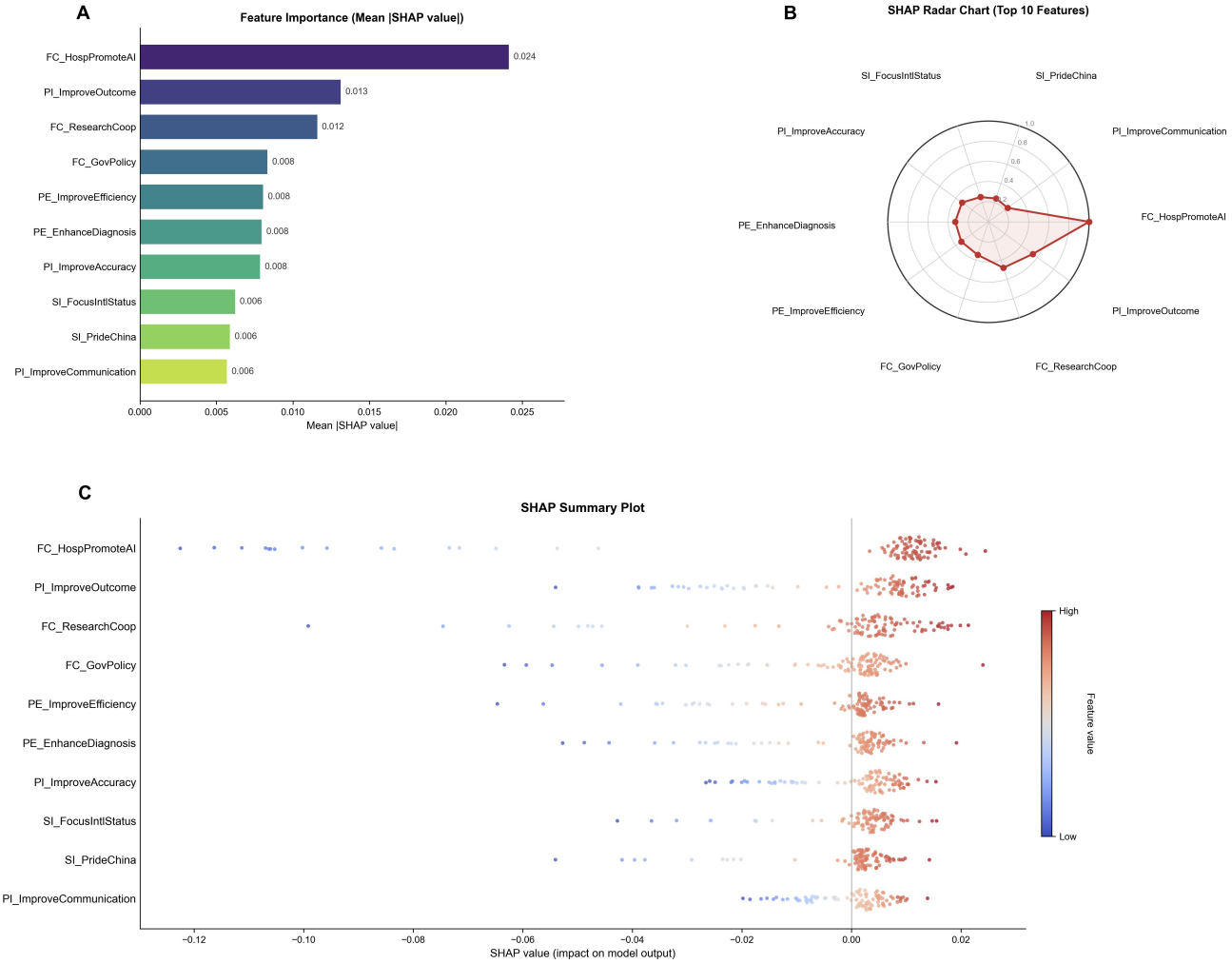
[g]KNN: *k*-nearest neighbor.

Among all tested machine learning models, GB and LR achieved the highest performance with AUC of 0.840, followed by RF (AUC=0.836) and extreme gradient boosting (AUC=0.822). The detailed performance metrics for all models under unadjusted (default) thresholds are provided in Table S3 in Multimedia Appendix 1.

The SHAP algorithm was applied to interpret the prediction of high acceptance of AI medical tools among physicians. Figure 6A shows the feature importance ranking with the top 10 predictors. FC_HospPromoteAI had the highest mean absolute SHAP value (0.024), though all effect sizes were modest, consistent with the ceiling effect in the outcome variable. Figure 6B presents the radar chart illustrating the relative importance of FC_HospPromoteAI compared to other variables. Figure 6C shows the SHAP summary plot, indicating that higher scores in organizational support (FC_HospPromoteAI) and perceived clinical benefits increased the predicted probability of high acceptance, while lower scores reduced it.

XSL•FO
RenderX

Figure 6. SHAP-based interpretability analysis of the optimal prediction model. (A) Mean absolute SHAP values ranking the top 10 most influential features for predicting physicians' acceptance of artificial intelligence (AI) medical tools. (B) SHAP radar chart illustrating the relative importance of the top 10 features. (C) SHAP summary plot showing the distribution and direction of feature impacts. Higher perceived organizational support and clinical benefits of AI (red points) were associated with an increased probability of high acceptance, whereas lower scores (blue points) decreased the predicted acceptance. FC_HospPromoteAI: Hospitals should actively promote the use of AI medical tools; PI_ImproveOutcome: AI can help improve patients' clinical outcomes; FC_ResearchCoop: Research institutions should strengthen AI-related research and collaboration; FC_GovPolicy: Government should introduce policies to promote AI in health care; PE_ImproveEfficiency: AI medical tools can improve health care efficiency; PE_EnhanceDiagnosis: AI medical tools can enhance physicians' diagnostic ability; PI_ImproveAccuracy: AI can improve diagnostic accuracy; SI_FocusIntlStatus: I pay attention to China's international status in medical AI; SI_PrideChina: I am proud of China's rapid development in medical AI; PI_ImproveCommunication: AI can help improve communication between physicians and patients. FC: facilitating conditions; PE: performance expectancy; PI: positive impact; SHAP: Shapley Additive Explanations; SI: social influence.



Decision curve analysis indicated that the GB classifier provided greater net benefit than both treat-all and treat-none strategies across most clinically plausible threshold probabilities (approximately 0.40 to 0.95; Figure S2 in Multimedia Appendix 2). In practice, this suggests that the model can help exploratorily identify characteristics associated with lower acceptance intentions, providing preliminary insights for targeted intervention design rather than definitive screening criteria. For individuals with these characteristics, administrators may consider initiating targeted interventions—specifically, reinforcing organizational support and demonstrating clinical evidence (as highlighted by SHAP)—rather than applying a one-size-fits-all training approach.

Discussion

Principal Findings

This study examined Chinese physicians' acceptance of AI medical tools using an extended UTAUT framework combined with explainable machine learning. Acceptance was uniformly high (>90%) across subgroups, revealing a pronounced ceiling effect. In the structural model, PE, SI, FC, and PI were positive predictors of BI, whereas EE was not significant. Six negative moderation effects indicated that higher hospital level, greater AI familiarity, and stronger future optimism attenuated several main effects. In predictive analyses, the RF provided the best overall trade-off after threshold optimization, while LR and GB attained the highest AUC. SHAP interpretation ranked organizational support ("hospitals should actively promote AI") and perceived clinical benefits ("AI improves outcomes" and

XSL•FO
RenderX

"AI enhances diagnosis") as the most influential features driving high acceptance.

While several path coefficients in the structural model were statistically significant yet relatively small in magnitude (PE: β=.03; SI: β=.02; PI: β=.01), these effects merit careful interpretation in context. First, the strong ceiling effect in BI (>90% acceptance across all dimensions) naturally constrains the range of variance available for prediction, mathematically limiting potential effect sizes. Second, in the presence of strong FC (β=.10, the largest predictor), the incremental contributions of PE and SI may be attenuated, as physicians already possess sufficient organizational support and resources to adopt AI regardless of individual perceptions. Third, from a practical significance perspective, even small standardized coefficients can translate to meaningful behavioral differences at population scale: a 0.03 increase in acceptance probability across China's 4 million physicians represents over 100,000 additional potential AI adopters. Thus, while the effect sizes appear modest statistically, they reflect substantively important patterns within a highly accepting physician population operating under strong institutional support.

The strong ceiling effect in BI merits discussion. Rather than merely a psychometric limitation, this pattern reflects a substantive finding: Chinese physicians demonstrate remarkably high acceptance of medical AI technology across diverse applications (automated medical record generation, chronic disease management, diagnostic assistance, personalized treatment planning, medical imaging analysis, and surgical assistance). This high acceptance rate aligns with China's national strategy to integrate AI into health care systems and suggests favorable conditions for medical AI implementation in Chinese health care settings.

Over 90% of physicians are willing to accept AI medical tools. Such a ceiling effect—wherein nearly all respondents express positive intent—demands careful interpretation. While high acceptance could reflect genuine enthusiasm driven by clinical needs and prior exposure to digital technologies, it may also result from social desirability bias or limitations in the measurement scale that restrict variance [30]. In contrast, studies from North America and Europe often report moderate acceptance levels. Recent multicountry surveys indicate that only 40% - 60% of physicians express readiness to integrate AI into routine practice [31,32]. Factors underlying this gap likely extend beyond technological maturity. In China, structural pressures—such as high patient volumes, long working hours, and relatively constrained staffing—may foster a pragmatic openness toward tools that promise efficiency gains. Government policy has also played a pivotal role [33]. However, a persistent challenge in behavioral research is the "attitude–behavior gap" [30]. High stated willingness does not guarantee sustained or effective use in practice. For example, follow-up studies in the United Kingdom found that despite favorable attitudes toward AI-based diagnostic support, actual use rates remained low due to workflow incompatibility and medico-legal concerns [34]. Thus, while this study's findings provide optimism, they also highlight the need for longitudinal monitoring to assess whether intentions translate into ongoing adoption.

Within the extended UTAUT framework, 4 predictors—PE, SI, FC, and PI—emerged as statistically significant, whereas EE did not. *PE* consistently ranks as a central driver in technology adoption models. Physicians in our sample may perceive AI as capable of enhancing diagnostic accuracy, reducing manual workload, and improving patient outcomes, aligning with global findings that perceived clinical benefit is the most influential determinant of adoption [33,35]. This is particularly relevant in specialties, such as radiology and pathology, where AI applications have already demonstrated time-saving and accuracy-enhancing capabilities. *SI* reflects the role of peer networks, professional hierarchies, and institutional culture. In the Chinese health care system, hierarchical structures and strong alignment with administrative directives may amplify the influence of respected colleagues or hospital leadership [32]. In environments where senior physicians endorse AI tools, junior colleagues may be more inclined to adopt, regardless of personal reservations. FC captured organizational readiness, including training, IT infrastructure, and institutional support. The high predictive value of FC in this study aligns with recent organizational readiness research, which identifies supportive environments as critical enablers of AI deployment [36,37]. In resource-constrained settings, the absence of infrastructure and technical assistance often outweighs individual willingness as a barrier. The inclusion of *PI* as an extended UTAUT construct reflects an emerging recognition that clinicians also evaluate the societal and ethical implications of technology. Perceptions that AI could enhance health care equity, reduce medical errors, or improve doctor-patient relationships may bolster acceptance, particularly in systems under strain [38]. This dimension resonates with a broader shift in digital health evaluation frameworks toward sociotechnical models. The nonsignificant role of EE is noteworthy. In classical technology acceptance model and UTAUT studies, ease of use often exerts a strong influence on BI. Its absence here may signal several possibilities. First, the usability of current AI medical tools may have reached a level where interface complexity is no longer a major barrier for physicians already accustomed to electronic medical records. Second, the survey population consisted largely of physicians in tertiary hospitals (3115/4024, 77.4%), who may have greater access to IT support, thus minimizing perceived effort barriers [39].

The moderation analysis revealed 6 significant negative interaction effects, indicating that certain background factors attenuated the influence of main predictors. To better illustrate these patterns, simple slope analyses were conducted for the key moderators (Figure S3 in Multimedia Appendix 2). The results suggest a "substitution effect" mechanism. Specifically, higher hospital levels and greater AI familiarity reduced the impact of FC and SI on BI. Physicians in high-level hospitals or those familiar with AI likely possess sufficient internal resources and knowledge, making them less dependent on external organizational support or peer norms to form adoption intentions. In contrast, physicians with low familiarity (as shown by the steeper slopes in Figure S3A in Multimedia Appendix 2) rely heavily on external FC to bridge the gap in their own capability. Similarly, future optimism dampened the effects of PE. As shown in Figure S3B in Multimedia Appendix 2, physicians with high optimism maintain high adoption intentions

XSL•FO
RenderX

regardless of the immediate performance gains, suggesting their internal belief acts as a buffer. Conversely, physicians with low optimism are more pragmatic; their willingness to adopt is highly sensitive to the tangible PE of the tools. This contrasts with the "wait-and-see" stance observed in previous studies, where anticipation of superior next-generation systems delays current uptake [40]. Instead of delaying adoption, our data suggests that high optimism provides a strong internal drive that substitutes for the need for immediate performance guarantees.

Methodological Synergy: Integration of SEM and Machine Learning

This study integrates SEM and ML to provide complementary perspectives on AI adoption. The SEM analysis established that FC (β=.10) was the strongest UTAUT predictor of BI. Correspondingly, SHAP analysis identified the item-level feature FC_HospPromoteAI (hospital organizational promotion of AI) as the most influential predictor of high acceptance (mean |SHAP|=0.024). This convergence between latent construct effects and item-level importance provides actionable specificity: interventions should prioritize institutional promotion and leadership endorsement rather than diffuse FC. Furthermore, SHAP revealed that physicians in secondary hospitals with moderate AI familiarity may respond differently to organizational support compared to tertiary hospital counterparts, suggesting the value of subgroup-specific implementation strategies. The complementary use of theory-driven SEM and data-driven ML thus validates the extended UTAUT framework while identifying concrete targets for management intervention.

Policy Implications: Evidence-Based Ethics and Governance

Our empirical findings suggest 3 policy priorities directly informed by the data. First, the dominance of FC_HospPromoteAI in SHAP analysis indicates that institutional endorsement is more influential than individual-level factors, supporting policies that mandate hospital-level AI governance committees and leadership commitment before deployment. Second, the negative moderation effect of AI familiarity on FC→BI (β=−.024) suggests that as physicians gain experience, their reliance on organizational support diminishes. This pattern has governance implications: early-stage implementation requires strong institutional backing and training, while mature adoption may benefit from peer-driven knowledge sharing and reduced top-down mandates. Third, the consistently high acceptance (>90%) but modest effect sizes suggest that the primary challenge is not physician willingness but rather structural readiness—policies should therefore prioritize infrastructure investment, workflow integration, and liability clarification over persuasion campaigns. Legal frameworks must clarify that AI serves as a decision support tool, with final clinical responsibility remaining with the physician.

Limitations

This study has several limitations. First, while our sample achieved broad geographic coverage (29 of 31 provincial-level regions), it was skewed toward tertiary hospitals (77.4% vs 52.1% nationally) and physicians with senior titles (72.9% vs approximately 35% nationally). This distribution reflects two factors: (1) the user composition of the Medlive platform, which predominantly serves physicians in larger academic medical centers and (2) inherent regional differences in physician density across China. While we considered weighting adjustments, we opted to retain the raw data to avoid statistical instability from heavily weighting the smaller primary care subsample. These patterns may limit generalizability to primary care settings, junior physicians, and underrepresented regions. Future studies should use targeted recruitment strategies to ensure adequate representation of community health centers and rural clinics, where AI implementation may face distinct resource and infrastructure challenges.

Second, the AVE for the BI construct was 0.40, falling below the conventional 0.50 threshold. This reflects the variance constraint imposed by the dichotomous nature of the items and the pronounced ceiling effect (over 90% acceptance) observed in the data. However, the construct demonstrated high internal consistency (CR=0.80; Cronbach α=0.80). Furthermore, sensitivity analyses comparing factor loadings between ML and diagonally weighted least squares estimators showed minimal differences (mean loading difference=0.027, max=0.048; Table S4 in Multimedia Appendix 1), indicating that the choice of estimator did not substantively affect parameter estimates and the measurement model is robust.

Third, the generic framing of "AI medical tools" may overlook differences between specific application types. Finally, while explainable machine learning added predictive depth, it was not applied to longitudinal adoption data, which could further refine implementation strategies.

Conclusion

Chinese physicians report an exceptionally high willingness to adopt AI medical tools, with organizational facilitation emerging as a key driver. Combining behavioral modeling with explainable ML highlights the relative importance of institutional readiness, though effect sizes are modest due to the pronounced ceiling effect. This underscores the need for context-specific strategies to translate intention into sustained, safe, and effective use.

Acknowledgments

No generative AI tools were used at any stage during the preparation, writing, editing, or revision of this manuscript. All content, including text, analysis, references, and citations, was originally created by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the entire manuscript.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional restrictions, as the data were obtained from the company's internal prescribing and system backend records, but are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey instrument items based on the extended Unified Theory of Acceptance and Use of Technology constructs and sociodemographic variables.
[DOCX File, 28 KB - medinform_v14i1e85270_app1.docx]

Multimedia Appendix 2
Threshold optimization for 6 classifiers and decision curve analysis for the primary model.
[DOCX File, 346 KB - medinform_v14i1e85270_app2.docx]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature New Biol 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
3. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018 May;24(5):539-540. [doi: 10.1038/s41591-018-0029-3] [Medline: 29736024]
4. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018 Nov;15(11):e1002686. [doi: 10.1371/journal.pmed.1002686] [Medline: 30457988]
5. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
6. Liao X, Yao C, Jin F, Zhang J, Liu L. Barriers and facilitators to implementing imaging-based diagnostic artificial intelligence-assisted decision-making software in hospitals in China: a qualitative study using the updated Consolidated Framework for Implementation Research. BMJ Open 2024 Sep 10;14(9):e084398. [doi: 10.1136/bmjopen-2024-084398] [Medline: 39260855]
7. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
8. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
9. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
10. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
11. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
12. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 2019 Jan;49(1):15-21. [doi: 10.1002/hast.973] [Medline: 30790315]
13. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019 Jul;20(7):938-947. [doi: 10.1016/S1470-2045(19)30333-X] [Medline: 31201137]
14. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med (Lausanne) 2022;9:990604. [doi: 10.3389/fmed.2022.990604] [Medline: 36117979]
15. Scheetz J, Rothschild P, McGuinness M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep 2021 Mar 4;11(1):5193. [doi: 10.1038/s41598-021-84698-5] [Medline: 33664367]

XSL•FO
RenderX

16. Iyanna S, Kaur P, Ractham P, Talwar S, Najmul Islam AKM. Digital transformation of healthcare sector. What is impeding adoption and continued usage of technology-driven innovations by end-users? J Bus Res 2022 Dec;153:150-161. [doi: 10.1016/j.jbusres.2022.08.007]
17. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 2019 Jan;112(1):22-28. [doi: 10.1177/0141076818815510] [Medline: 30507284]
18. Li H, Wu J, Gao Y, Shi Y. Examining individuals' adoption of healthcare wearable devices: an empirical study from privacy calculus perspective. Int J Med Inform 2016 Apr;88:8-17. [doi: 10.1016/j.ijmedinf.2015.12.010] [Medline: 26878757]
19. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
20. Venkatesh V, Thong J, Xu X. Unified theory of acceptance and use of technology: a synthesis and the road ahead. J Assoc Inf Syst 2016;17(5):328-376. [doi: 10.17705/1jais.00428]
21. Dwivedi YK, Hughes L, Ismagilova E, et al. Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manage 2021 Apr;57:101994. [doi: 10.1016/j.ijinfomgt.2019.08.002]
22. Park CW, Seo SW, Kang N, et al. Artificial intelligence in health care: current applications and issues. J Korean Med Sci 2020 Nov 2;35(42):e379. [doi: 10.3346/jkms.2020.35.e379] [Medline: 33140591]
23. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
24. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
25. Patel VL, Kannampallil TG. Human factors and health information technology: current challenges and future directions. Yearb Med Inform 2014 Aug 15;9(1):58-66. [doi: 10.15265/IY-2014-0005] [Medline: 25123724]
26. Zhang X, Liu S, Chen X, Wang L, Gao B, Zhu Q. Health information privacy concerns, antecedents, and information disclosure intention in online health communities. Inf Manag 2018 Jun;55(4):482-493. [doi: 10.1016/j.im.2017.11.003]
27. Guo X, Han X, Zhang X, Dang Y, Chen C. Investigating m-Health acceptance from a protection motivation theory perspective: gender and age differences. Telemed J E Health 2015 Aug;21(8):661-669. [doi: 10.1089/tmj.2014.0166] [Medline: 25919800]
28. Maillet É, Mathieu L, Sicotte C. Modeling factors explaining the acceptance, actual use and satisfaction of nurses using an electronic patient record in acute care settings: an extension of the UTAUT. Int J Med Inform 2015 Jan;84(1):36-47. [doi: 10.1016/j.ijmedinf.2014.09.004] [Medline: 25288192]
29. Bozan K, Parker K, Davey B. A closer look at the social influence construct in the UTAUT model: an institutional theory based approach to investigate health IT adoption patterns of the elderly. Presented at: 2016 49th Hawaii International Conference on System Sciences (HICSS); Jan 5-8, 2016. [doi: 10.1109/HICSS.2016.391]
30. Venkatesh V. Paradigm changing metaverse: future research directions in design, technology adoption and use, and impacts. Decis Support Syst 2025 Jan;188:114307. [doi: 10.1016/j.dss.2024.114307]
31. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
32. Heuer M, Reindl-Spanner P, Prommegger B, Krcmar H. Building bridges: fostering AI adoption in primary care. Presented at: Proceedings of the 45th International Conference on Information Systems (ICIS 2024 TREO Papers); Dec 15-18, 2024 URL: https://aisel.aisnet.org/treos_icis2024/16/ [accessed 2026-04-04]
33. Mooghali M, Stroud AM, Yoo DW, et al. Trustworthy and ethical AI-enabled cardiovascular care: a rapid review. BMC Med Inform Decis Mak 2024 Sep 4;24(1):247. [doi: 10.1186/s12911-024-02653-6] [Medline: 39232725]
34. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742. [doi: 10.2196/39742] [Medline: 36626192]
35. Emma L. Explainable AI for high-stakes decision making in healthcare. ResearchGate. Preprint posted online on Jul, 2024 URL: https://www.researchgate.net/publication/387538513_Explainable_AI_for_High-Stakes_Decision_Making_in_Healthcare [accessed 2026-04-07]
36. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
37. Lipton ZC. The mythos of model interpretability. Queue 2018 Jun;16(3):31-57. [doi: 10.1145/3236386.3241340]
38. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017 URL: https://dl.acm.org/doi/10.5555/3295222.3295230#purchase-access [accessed 2026-03-19]
39. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016. [doi: 10.1145/2939672.2939778]
40. Shmueli G. To explain or to predict? Statist Sci 2010;25(3):289-310. [doi: 10.1214/10-STS330]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
AVE: average variance extracted
BI: behavioral intention
CFI: comparative fit index
CR: composite reliability
EE: effort expectancy
FC: facilitating conditions
GB: gradient boosting
LR: logistic regression
ML: maximum likelihood
PE: performance expectancy
PE: performance expectancy
PI: positive impact
RF: random forest
RMSEA: root mean square error of approximation
SEM: structural equation modeling
SI: social influence
TLI: Tucker-Lewis index
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Benis; submitted 04.Oct.2025; peer-reviewed by F Amakye, X Ding; revised version received 06.Jan.2026; accepted 17.Feb.2026; published 16.Apr.2026.

Please cite as:
Shi X, Tian Z, Guo Q, Qiao B, Wang X
Acceptance of Artificial Intelligence in Clinical Practice Among Chinese Physicians: Nationwide Cross-Sectional Survey Using Extended Unified Theory of Acceptance and Use of Technology and Explainable Machine Learning
JMIR Med Inform 2026;14:e85270
URL: https://medinform.jmir.org/2026/1/e85270
doi:10.2196/85270

© Xuefei Shi, Zhanxiao Tian, Qi Guo, Boxuan Qiao, Xiaolong Wang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication at https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

An Intelligent Interactive Management Platform for Rheumatoid Arthritis Care: Real-World Observational Study

Ziyun Zhang[1,2*], BSN; Xinyue Zhang[1,3*], MSc; Li Zhang[1,2], BSN; Mingyan Du[1,2], BSN; Lijuan Zhang[1,2], BSN; Siyu Yang[1*], MSc; Xiaoli Cai[1*], BSN; Shengchao Hou[4*], PhD

[1]Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1095 Jiefang Road, Qiaokou District, Wuhan, Hubei, China

[2]Department of Rheumatology and Immunology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[3]School of Nursing, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[4]Library, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[*]these authors contributed equally

Corresponding Author:
Siyu Yang, MSc
Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1095 Jiefang Road, Qiaokou District, Wuhan, Hubei, China

Abstract

Background: Effective postdischarge management is essential for maintaining disease control and improving long-term outcomes in rheumatoid arthritis (RA). Digital health technologies, particularly intelligent management platforms, provide new opportunities for continuous care and self-management in real-world settings.

Objective: This study aimed to develop a nurse-led artificial intelligence (AI)–assisted interactive management platform for postdischarge RA care and examine its association with patient outcomes in a real-world clinical setting.

Methods: A single-center, multicampus, real-world observational study with a quasi-experimental framework was conducted at Tongji Hospital between November 2024 and October 2025. Participants were allocated to either a platform-based management group (campus 1; n=184) or a comparison group receiving routine postdischarge care (campus 2; n=157). The platform enabled remote monitoring through a smartphone app, which collected information on joint symptoms, fatigue, medication adherence, laboratory results, and emotional status, with data uploaded to a secure cloud server. The AI system analyzed data in real time and alerted health care providers to abnormalities. Nurses and health coaches delivered personalized education and interventions. Primary outcomes were changes in Disease Activity Score in 28 Joints and Health Assessment Questionnaire–II scores at discharge and the 6-month follow-up. Secondary outcomes included medication adherence measured as the proportion of prescribed doses taken and patient satisfaction assessed via a 5-point Likert-scale questionnaire at 6 months.

Results: A total of 341 patients with RA completed the 6-month follow-up. Disease Activity Score in 28 Joints values decreased significantly in the platform-based management group at 6 months (mean 5.04, SD 1.38 at discharge to mean 4.27, SD 1.20 at 6 months; P<.001), showing a greater reduction than in the comparison group (mean 4.86, SD 1.19 at discharge to mean 4.59, SD 1.36 at 6 months; within-group P=.03; between-group P=.02). Health Assessment Questionnaire–II scores improved significantly in the platform-based management group (mean 2.45, SD 0.75 at discharge to mean 1.67, SD 0.56 at 6 months; P<.001), while the comparison group showed no significant change (P=.12; between-group P<.001). Medication adherence was higher in the platform-based group (166/184, 90.2% vs 125/157, 79.6%; χ_1^2=7.6; P=.006). Overall health care satisfaction was also higher in the platform-based group (176/184, 95.7% vs 120/157, 76.4%; χ_1^2=27.8; P<.001), with a greater proportion reporting to be “very satisfied” in this group (150/184, 81.5% vs 59/157, 37.6%).

Conclusions: The intelligent interactive management platform markedly enhanced disease control, functional status, medication adherence, and satisfaction among patients with RA after discharge. This nurse-led, AI-assisted system represents a promising model for improving outpatient RA management.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500115743; https://www.chictr.org.cn/hvshowproject.html?id=291630&v=1.0

(JMIR Med Inform 2026;14:e90784) doi:10.2196/90784

XSL•FO
RenderX

KEYWORDS

rheumatoid arthritis; intelligent interactive platform; disease activity; medication adherence; patient satisfaction; artificial intelligence; AI

Introduction

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease with a global prevalence of approximately 1% to 2% [1]. It predominantly affects middle-aged and older adults, with higher prevalence among women aged 40 to 60 years [1,2]. RA is frequently complicated by extra-articular manifestations, including cardiovascular disease, pulmonary dysfunction, neurological involvement, gastrointestinal disturbances, and malnutrition [3]. Although no curative therapy exists, timely and standardized treatment can effectively control disease activity, relieve symptoms, and delay disability [4]. Long-term management commonly involves pharmacological therapy, intra-articular injections, physical rehabilitation, regular follow-up visits, and surgical intervention when indicated [5]. Nevertheless, despite ongoing treatment, many patients experience varying degrees of joint deformity and functional impairment, substantially limiting their daily activities and compromising overall well-being [6].

Given the chronic and progressive nature of RA, effective self-management is essential for maintaining patients' quality of life and delaying disease progression [7]. Conventional nursing models primarily focus on inpatient care, whereas postdischarge management often relies heavily on patient self-management, leading to suboptimal outcomes [8]. Successful RA management requires patients to develop both the confidence and the necessary skills to manage their condition independently [9]. Key components of self-management include self-monitoring of symptoms and disease flares, setting achievable goals, adopting behavioral modifications, adhering to medication, adjusting physical activity levels, and enhancing problem-solving abilities [3]. Robust self-management practices have been shown to improve adherence, reduce disease activity, and slow progression of RA [10,11]. Therefore, innovative and sustainable strategies are needed to enhance postdischarge self-management and continuity of care.

The rapid proliferation of mobile technologies has transformed health service delivery worldwide [12]. As mobile apps become increasingly embedded in routine activities, the delivery of health care through digital platforms has emerged as an inevitable evolution in line with technological advancement [13]. In the context of RA management, mobile health (mHealth) technologies, including smartphones, wearable devices, and disease-specific apps, are playing an increasingly pivotal role [12,14]. These tools enable real-time health data collection, thereby facilitating early detection of disease fluctuations. They empower both patients and health professionals by enhancing disease monitoring, improving medication adherence, and enhancing timely personalized interventions [15]. Such innovations improve the efficiency of disease management by reducing unnecessary hospital visits and optimizing health care resource allocation. Furthermore, instant feedback mechanisms reinforce patients' self-management capabilities and promote active engagement in their own care [16].

Building on our institution's "internet hospital" infrastructure and standardized RA management protocols, we developed an intelligent, nurse-led interactive management platform to support postdischarge care for patients with RA. This study aimed to evaluate the real-world effectiveness of this platform by comparing patients managed through platform-based follow-up to those receiving routine postdischarge care across different hospital campuses under unified clinical standards.

Methods

Study Design and Population

This study used a single-center, multicampus observational design with a quasi-experimental framework conducted at the Department of Rheumatology and Immunology, Tongji Hospital, between November 2024 and October 2025. Tongji Hospital consists of multiple hospital campuses operating under unified clinical guidelines and standardized treatment protocols for RA. An intelligent interactive follow-up platform was first implemented at campus 1 as part of routine postdischarge care, whereas campus 2 had not yet adopted the platform at the time of the study. This natural difference in platform availability allowed for the formation of a platform-based management group (campus 1) and a comparison group receiving usual postdischarge care (campus 2) without individual-level assignment by the research team. Eligible participants were adults (aged ≥18 years) diagnosed with RA according to the 2010 American College of Rheumatology and European Alliance of Associations for Rheumatology classification criteria, receiving standardized RA therapy, and capable of using a smartphone or having a caregiver assist with platform access. Exclusion criteria included inability to comply with follow-up, severe psychiatric or cognitive disorders, visual or hearing impairment precluding platform use, severe systemic comorbidities (eg, advanced cardiac, hepatic, or renal disease), or discontinuation of RA treatment during the follow-up period.

Group Allocation and Blinding

Participants were allocated to study groups based on hospital campus rather than through randomization. Patients discharged from campus 1 were included in the platform-based management group, whereas those discharged from campus 2 constituted the comparison group receiving routine postdischarge care. This campus-based allocation reflected real-world clinical implementation and helped avoid ethical concerns related to unequal access to a newly introduced digital management platform. Given the nature of the platform-based management model, blinding of patients and health care providers was not feasible. However, outcome assessors and data analysts were blinded to group allocation to reduce assessment and analysis bias.

XSL•FO
RenderX

Technical Design and Platform-Based Management Model

Platform Overview

The intelligent interactive platform for RA is a multidisciplinary, nurse-led, and patient-centered digital health management system established within the smart health care ecosystem of Tongji Hospital. Designed to enhance routine outpatient care, the platform incorporates artificial intelligence (AI) and is operated by a specialized team of rheumatology nurses, with additional support from health coaches trained in rheumatology, nutrition, and psychological support.

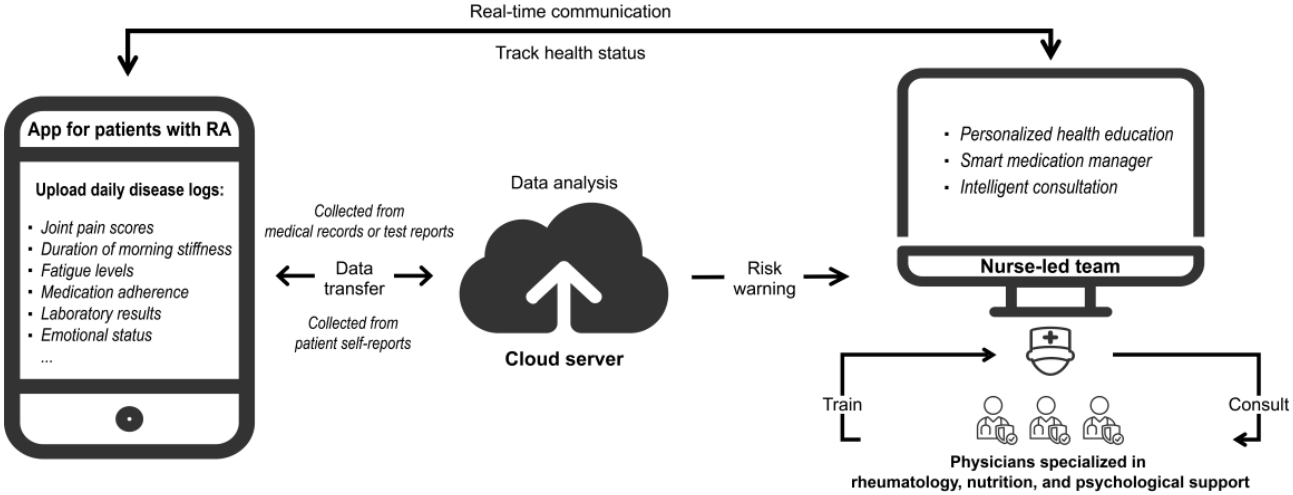
The primary objective of the platform is to extend continuous care beyond hospital discharge. Timely responses to self-reported symptoms and system-generated alerts support individualized guidance and facilitate timely clinical decision-making during routine follow-up. It also facilitates complication monitoring, self-care education, and psychological support, helping patients manage their condition more effectively during home-based recovery. Integrated communication tools ensure that patients have continuous access to their care team.

Technical Architecture and Core Functions

Overview

The platform was developed collaboratively by the hospital's Information Center and technology partners. Its architecture comprises 3 core components: a patient-facing mobile app, a web-based dashboard for nurses and health coaches, and a cloud-based AI data processing system (Figure 1).

Figure 1. Framework of the intelligent interactive platform for rheumatoid arthritis (RA).



Data Input and Processing

Patients use the mobile app to submit health logs either daily or as needed based on their symptom fluctuations. These logs include joint pain scores measured using a visual analog scale from 1 to 100, duration of morning stiffness, fatigue levels, medication adherence, laboratory results, emotional status, and sleep quality. While the "Smart Medication Manager" provides automated daily notifications for drug intake, other health logs are recorded voluntarily by patients without daily reminders to avoid excessive burden. All data are securely encrypted and transmitted to the cloud server. For image-based data, such as medical records or test reports, a deep learning–based optical character recognition system automatically extracts key information, which is manually verified and structured into the database.

AI-Driven Clinical Decision Support

An embedded AI engine continuously analyzes patient data in real time. Unlike traditional rule-based systems that rely on static, predefined thresholds, our AI engine uses a multilayered approach. A deep learning–based optical character recognition system processes unstructured data from patient-uploaded images (eg, external laboratory reports). In addition, the engine uses longitudinal data processing to monitor trends in disease activity and medication adherence over time. When abnormalities are detected, such as elevated disease activity, decreased medication adherence, or signs of relapse, the system immediately generates alerts for the clinical care team through the dashboard even if individual parameters have not yet exceeded traditional laboratory triggers. Nurses and health coaches can then respond promptly via the app, providing timely follow-up support and facilitating triage for further outpatient or inpatient care if necessary (Multimedia Appendix 1).

Personalized Education and Self-Management Support

Beyond reactive support, the platform provides proactive, personalized health education. On the basis of individual patient assessments and preferences, the care team delivers guidance on diet, exercise, and psychological well-being. This is facilitated by an integrated intelligent education module that pushes customized content to patients (Multimedia Appendix 2). A "Smart Medication Manager" feature allows patients to log their prescribed oral medications and receive automated reminders to enhance adherence (Multimedia Appendix 3). Furthermore, an "Intelligent Consultation" module serves as a knowledge base for frequently asked questions related to RA care, covering topics such as appointments, insurance, follow-ups, and common rheumatological concerns, thereby improving accessibility and potentially reducing unnecessary clinic visits (Multimedia Appendix 4).

XSL•FO
RenderX

Pilot Feasibility Assessment

Prior to the main study, an internal, unpublished pilot test was conducted to evaluate the platform's feasibility and support its user-centered design. The pilot involved 25 patients with RA over a 4-week period. During this phase, the platform demonstrated high levels of engagement. Specifically, the patient self-assessment completion rate rose from 68.0% (17/25) to 88.0% (22/25). To ensure that the intervention was not solely investigator driven, we collected qualitative feedback from participants via telephone interviews and structured satisfaction surveys. Users identified several areas for improvement, such as the need for a more intuitive user interface for joint symptom logging and more flexible medication reminder settings. On the basis of the feedback, the platform underwent iterative refinements, including optimizing user interface responsiveness and streamlining the data entry process before the formal study commenced. While traditional clinical measures such as inflammatory markers and disease activity scores did not show statistically significant changes in the short-term pilot, promising improvements were observed in behavioral outcomes, including follow-up completion rates, platform log-in frequency, and consultation response times, indicating good usability and feasibility for real-world implementation.

Data Collection and Study Procedure

Between November 7, 2024, and April 20, 2025, eligible patients with RA who were clinically stable and scheduled for discharge were consecutively recruited. Baseline data were collected prior to discharge. Participants were allocated to study groups based on hospital campus, reflecting routine clinical implementation of the intelligent interactive platform. All participants were followed up on for 6 months after enrollment, and the final follow-up assessment was completed on October 20, 2025.

During hospitalization, patients in both groups received standardized nursing care and health education in accordance with unified clinical protocols across campuses. This included RA-related disease education, psychological support, medication management, symptom monitoring, rehabilitation guidance, complication prevention, routine follow-up planning, and discharge counseling.

After discharge, patients in the comparison group received routine postdischarge follow-up care, which consisted of periodic assessments of disease status, emotional well-being, medication adherence, and rehabilitation activities. Follow-up was conducted using conventional approaches, including scheduled outpatient visits, telephone consultations, or home visits, during which health care professionals provided individualized guidance and addressed patient concerns as needed.

In addition to routine care, patients in the platform-based management group were supported through the platform. One to 2 days prior to discharge, designated rheumatology nurses introduced the platform to eligible patients using standardized instructions and obtained written informed consent. Nurses then provided individualized, hands-on guidance on platform access, account setup, and creation of a personal health record. Patients were guided through key functional modules to ensure independent use after discharge. For patients without access to a smartphone or those unfamiliar with mobile apps, family members or caregivers were instructed on how to operate the platform on the patient's behalf to support ongoing disease management.

Outcome Measures

Overview

Primary outcomes were disease activity and functional status assessments. Secondary outcomes included medication adherence and health care satisfaction. These outcomes were assessed to examine differences between groups rather than to evaluate causal effects of the platform. Disease activity and functional status were evaluated both at discharge and the 6-month follow-up, whereas secondary outcomes were assessed only at 6 months as many patients were treatment naive at baseline or satisfaction could not be measured until follow-up.

Disease Activity

The Disease Activity Score in 28 Joints (DAS28) was used to assess RA disease activity [17]. It combines a count of 28 swollen and tender joints, a laboratory marker of inflammation (either C-reactive protein or erythrocyte sedimentation rate), and a patient global health assessment. Higher DAS28 scores indicate more severe disease activity.

Functional Status

Functional disability was measured using the Health Assessment Questionnaire–II (HAQ-II) [18]. The HAQ-II is a validated instrument that assesses difficulty in performing activities of daily living across 8 domains (dressing, arising, eating, walking, hygiene, reach, grip, and common activities). It contains 10 items, each scored on a scale from 0 to 3 (0="without difficulty"; 3="unable to do"). The total score is the average of the category scores, with higher scores indicating greater functional impairment. The HAQ-II showed good internal consistency in this study (Cronbach α=0.83), consistent with its original validation (Cronbach α=0.88) [18].

Medication Adherence

Medication adherence was assessed based on participants' self-reported medication intake during the follow-up period. Adherence was operationally defined as the proportion of prescribed doses reportedly taken. Consistent with commonly used adherence definitions in chronic disease management, participants were classified as having good adherence if they reported taking 85% or more of the prescribed medication regimen and poor adherence otherwise [19].

Health Care Satisfaction

Patient satisfaction with care was assessed using a department-developed questionnaire. Responses were recorded on a 5-point Likert scale ("very satisfied," "satisfied," "neutral," "dissatisfied," and "very dissatisfied"). The satisfaction rate was calculated as the percentage of participants reporting being "very satisfied" or "satisfied" among the total respondents. The scale's content validity index was 0.92 as evaluated by a panel of 5 rheumatology experts, and the questionnaire demonstrated good reliability in this study (Cronbach α=0.86).

XSL•FO
RenderX

Data Analysis

Statistical analyses were performed using R (version 4.4.1; R Foundation for Statistical Computing). Continuous variables were described as means and SDs for normally distributed data or medians and IQRs for nonnormally distributed data. Normality was assessed using the Shapiro-Wilk test. Variance homogeneity was evaluated using the Levene test. Between-group comparisons were performed using 2-tailed Student *t* test or Welch *t* test for normally distributed variables and the Mann-Whitney *U* test (Wilcoxon rank-sum test) for nonnormally distributed variables. Categorical variables were compared using the chi-square test or Fisher exact test, as appropriate. All *P* values were 2 sided and are reported as exact values in accordance with JMIR Publications' statistical reporting guidelines.

Ethical Considerations

This study was approved by the Medical Ethics Committee of Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology (TJ-IRB202411023). The study was conducted in accordance with the Declaration of Helsinki. Participation was voluntary, and written informed consent was obtained from all participants prior to enrollment. Participants retained the right to withdraw from the study at any time without penalty or explanation. All collected data were deidentified and stored securely in compliance with the hospital's scientific research data management standards. Deidentification was performed using a key-encoded system, with reidentification keys securely stored by the study team. No financial incentives were offered to participants. The trial was prospectively registered at the Chinese Clinical Trial Registry (ChiCTR2500115743). This study is reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Checklist 1).

Results

Demographic Characteristics

Between November 7, 2024, and April 20, 2025, a total of 459 patients with RA who met the inclusion criteria were identified at our hospital. Of these 459 patients, 401 (87.4%) consented to participate in the study, including 206 (44.9%) from campus 1 (platform-based management group) and 195 (42.5%) from campus 2 (comparison group). Following baseline assessment, 10.2% (21/206) of the patients from campus 1 and 20.0% (39/195) of the patients from campus 2 were lost to follow-up. The higher retention rate in the platform-based group may reflect the impact of continuous digital engagement. Ultimately, 89.3% (184/206) of the patients from campus 1 and 80.5% (157/195) of the patients from campus 2 completed the 6-month follow-up, resulting in a total of 341 (74.3%) participants included in the final analysis.

The baseline characteristics of the participants are summarized in Table 1. The mean age of the participants was 57.41 (SD 13.45) years in the platform-based management group and 55.47 (SD 13.57) years in the comparison group. Female individuals comprised most of both groups, accounting for 76.1% (140/184) and 78.3% (123/157) in the campus 1 and campus 2 groups, respectively. RA duration was similar between the 2 groups (median 4, IQR 2-12 years vs median 4, IQR 2-10 years). Educational levels, residence, smoking and alcohol consumption, BMI, RA disease duration, comorbidities including hypertension and diabetes, and RA-related complications did not differ significantly between the groups.

XSL•FO
RenderX

Table . Demographic characteristics of the participants (N=341).

Characteristic	Platform-based management group (n=184)	Comparison group (n=157)	χ^2 (*df*)	*P* value
Age (years)[a], mean (SD)	57.41 (13.45)	55.47 (13.57)		.19
Sex, n (%)			0.133 (1)[b]	.72
Male	44 (23.9)	34 (21.7)		
Female	140 (76.1)	123 (78.3)		
Educational level, n (%)			5.324 (3)[b]	.15
Primary school	42 (22.8)	44 (28.0)		
Lower secondary school	71 (38.6)	56 (35.7)		
Upper secondary school	45 (24.5)	43 (27.4)		
Tertiary education	26 (14.1)	11 (7.0)		
Residence, n (%)			1.511 (1)[b]	.22
Rural	121 (65.8)	93 (59.2)		
Urban	63 (34.2)	64 (40.8)		
RA[c] disease duration (years)[d], median (IQR)	4 (2-12)	4 (2-10)		.54
BMI (kg/m^2)[d], median (IQR)	20.69 (19.33-22.79)	20.58 (19.23-22.77)		.78
Smoking, n (%)			0.030 (1)[b]	>.99
Yes	20 (10.9)	18 (11.5)		
No	164 (89.1)	139 (88.5)		
Alcohol consumption, n (%)			0.139 (1)[b]	.83
Yes	26 (14.1)	20 (12.7)		
No	158 (85.9)	137 (87.3)		
Hypertension, n (%)			3.269 (1)[b]	.07
Yes	61 (33.2)	38 (24.2)		
No	123 (66.8)	119 (75.8)		
Diabetes, n (%)			1.319 (1)[b]	.25
Yes	39 (21.2)	26 (16.6)		
No	145 (78.8)	131 (83.4)		
RA-related complications[e], n (%)			0.052 (1)[b]	.89
Yes	95 (51.6)	79 (50.3)		
No	89 (48.4)	78 (49.7)		

[a]Student *t* test. Age: t_{339}=1.321.

[b]Chi-square test.

[c]RA: rheumatoid arthritis.

[d]Mann-Whitney *U* test (Wilcoxon rank-sum test). RA disease duration: *W*=15,003. BMI: *W*=14,695.

[e]Primarily comprising interstitial lung disease, osteoporosis, and femoral head avascular necrosis.

Changes in DAS28 Scores

Changes in disease activity over the 6-month follow-up are shown in Table 2. In the comparison group, the mean DAS28 score decreased from 4.86 (SD 1.19) at discharge to 4.59 (SD 1.36) at 6 months, with a statistically significant within-group change (*P*=.03). In the platform-based management group, the mean DAS28 score decreased from 5.04 (SD 1.38) at discharge to 4.27 (SD 1.20) at 6 months, with a larger within-group change (*P*<.001). Between-group comparison at 6 months indicated a significantly greater reduction in DAS28 scores in the

XSL•FO
RenderX

platform-based management group than in the comparison group (P=.02).

Table . Disease Activity Score in 28 Joints values of the participants. t_{339}=–1.293 and P=.20 at discharge and t_{339}=2.286 and P=.02 at the 6-month follow-up. Disease Activity Score in 28 joints: range 0-9.4; higher scores indicate more activity.

Group	At discharge, mean (SD)	6-month follow-up, mean (SD)	Paired *t* test (*df*)	*P* value
Comparison group	4.86 (1.19)	4.59 (1.36)	2.172 (156)	.03
Platform-based management group	5.04 (1.38)	4.27 (1.20)	7.404 (183)	<.001

Changes in HAQ-II Scores

At discharge, HAQ-II scores were comparable between the comparison group and the platform-based management group (P=.79; Table 3). At the 6-month follow-up, the comparison group showed no significant change from baseline (P=.12), whereas the platform-based management group exhibited a significant reduction (P<.001). Between-group comparison at 6 months indicated a significantly greater improvement in functional status in the platform-based management group than in the comparison group (P<.001).

Table . Health Assessment Questionnaire–II scores of the participants. t_{339}=–1.293 and P=.20 at discharge and t_{339}=2.286 and P=.02 at the 6-month follow-up. Disease Activity Score in 28 joints: range 0-9.4; higher scores indicate more activity.

Group	At discharge, mean (SD)	6-month follow-up, mean (SD)	Paired *t* test (*df*)	*P* value
Comparison group	2.47 (0.60)	2.39 (0.58)	1.565 (156)	.12
Platform-based management group	2.45 (0.75)	1.67 (0.56)	14.529 (183)	<.001

Changes in Medication Adherence

On the basis of the predefined adherence criterion (≥85% of prescribed doses reportedly taken), 90.2% (166/184) of the participants in the platform-based management group exhibited good adherence compared with 79.6% (125/157) in the comparison group (χ_1^2=7.6; P=.006) at the 6-month follow-up.

Changes in Health Care Satisfaction

Table 4 presents the health care satisfaction levels of participants in both groups. The overall satisfaction rates were 76.4% (120/157) in the comparison group and 95.7% (176/184) in the platform-based management group (P<.001). Notably, a greater proportion of participants in the platform-based management group reported being very satisfied compared to the comparison group (150/184, 81.5% vs 59/157, 37.6%, respectively).

Table . Health care satisfaction of the participants.

Group	Very satisfied, n (%)	Satisfied, n (%)	Neutral, n (%)	Dissatisfied, n (%)	Very dissatisfied, n (%)	Satisfaction rate (participants who were very satisfied or satisfied), n (%)[a]
Comparison group (n=157)	59 (37.6)	61 (38.9)	24 (15.3)	13 (8.3)	0 (0.0)	120 (76.4)
Platform-based management group (n=184)	150 (81.5)	26 (14.1)	5 (2.7)	3 (1.6)	0 (0.0)	176 (95.7)

[a]χ_1^2=27.8; P<.001.

Discussion

In this quasi-experimental study conducted in real-world clinical settings, we demonstrated that a nurse-led, AI-assisted intelligent interactive platform significantly improved disease activity, functional status, medication adherence, and satisfaction in patients with RA over a 6-month follow-up period. Specifically, participants in the platform-based management group showed greater reductions in DAS28 scores and more substantial improvements in HAQ-II scores compared with those receiving routine postdischarge care. These findings highlight the potential of digital health interventions to enhance real-world RA management beyond conventional postdischarge care.

Our results align with and expand on evidence from large, multicenter studies showing that digital health apps can contribute to disease control in RA. In a pragmatic randomized clinical trial involving more than 2000 patients with RA, patients using a digital health app to assess patient-reported outcomes had higher rates of achieving a DAS28 based on C-reactive protein score of 3.2 or less at 6 months compared with controls, reflecting improved disease control with digital monitoring and clinician intervention [20]. Additionally, smaller prospective

XSL•FO
RenderX

studies have reported clinically meaningful improvements in disease activity and other patient-reported outcomes following mHealth interventions for patients with inflammatory arthritis, underscoring the potential for mobile platforms to complement standard therapies [21]. There is systematic evidence further suggesting that asynchronous mHealth tools, including smartphone apps, web apps, and SMS text message reminders, are associated with desirable outcomes in RA self-management and health behaviors, although the heterogeneity of existing studies highlights the need for more rigorous designs [22]. Moreover, systematic reviews of digital health interventions for arthritis broadly support the role of these tools in enhancing medication adherence and quality of life, with most randomized controlled trials showing positive contributions of digital health in arthritis populations [23]. Together, these data contextualize our findings and indicate that digital tools, particularly those integrating real-time monitoring and clinician feedback, may improve RA outcomes beyond traditional follow-up models.

The mechanisms underlying observed improvements likely reflect multiple synergistic components of the platform. First, continuous data capture and AI-driven alerts enabled early identification of disease activity fluctuations, facilitating prompt clinical decision-making and adjustment of treatment plans consistent with treat-to-target strategies [24]. Second, personalized patient education modules and self-management support improved disease understanding, reduced anxiety-induced nonadherence, and reinforced positive health behaviors [25,26]. Third, integrated communication tools and nurse-led remote engagement fostered sustained patient-health care team interaction, thereby empowering patients and enhancing self-management [26-28], which are factors strongly associated with improved adherence in chronic disease management models. Collectively, this multimodal approach may underpin the greater improvements in functional status, symptom control, and engagement observed in the platform-based management group.

Importantly, enhanced functional outcomes (as measured using the HAQ-II) in the platform-based management group suggest that early and sustained control of inflammatory activity may translate into preserved musculoskeletal function and reduced disability progression. Remote access to rehabilitation guidance, individualized exercise recommendations, and rehabilitation education likely reduced care barriers for patients with mobility limitations or geographic challenges, expanding continuous care delivery beyond periodic clinic visits [14,25,28]. These effects are consistent with those observed in prior research [13,16] emphasizing the value of remote monitoring and patient engagement in optimizing long-term functional outcomes in chronic musculoskeletal diseases.

The platform significantly improved both medication adherence and patient satisfaction, which is consistent with findings from prior digital health interventions targeting chronic disease management [7,12,28]. Traditional outpatient follow-up is frequently constrained by logistical and geographical barriers, especially in rural or underserved areas, leading to delays in care and reduced compliance [9]. To address these challenges, the platform's integrated "Medication Manager" provided automated medication reminders, individualized dosage guidance, and symptom reporting functions, thereby enabling timely clinical interventions in response to adverse events. Previous studies have provided evidence that such digital tools can decrease medication omissions and enhance treatment persistence in patients with RA and similar chronic conditions [10,23,29]. Moreover, the nurse-led, patient-centered platform empowered patients through shared decision-making, self-monitoring, and goal setting, which contributed to improved self-management. Personalized care plans and ongoing professional support not only reinforced adherence but also enhanced patient and nurse satisfaction with the care process [30,31]. These findings highlight the effectiveness of technology-assisted, nurse-led interventions in promoting satisfaction alongside adherence in remote care [32].

However, several limitations should be acknowledged. The quasi-experimental, campus-based allocation does not fully replicate randomized assignment, although this design reflects real-world implementation and enhances external validity. Participant and provider blinding was not feasible, which might introduce reporting bias; however, outcome assessors and analysts were blinded to mitigate this risk. The high-frequency data collected by the platform, such as daily pain and stiffness scores, were not included in this statistical analysis. Future research should leverage these real-time metrics to better characterize disease trajectories. A notable attrition rate was observed during the 6-month follow-up. Preliminary observations indicated that older patients with limited digital literacy or those with lower treatment compliance were more likely to drop out. These individuals represent a vulnerable population that might benefit the most from digital support but faces significant barriers to technology adoption. The lack of detailed demographics for these participants may limit the generalizability of our results regarding digital health equity. While the 6-month follow-up period was adequate to observe significant clinical improvements, longer follow-up is warranted to evaluate sustained disease control, radiographic progression, long-term disability outcomes, and cost-effectiveness. Future multicenter implementation studies with diverse patient populations, combined with qualitative research to explore user preferences, will be valuable to further validate and optimize digital care strategies for RA.

In conclusion, our study demonstrates that a nurse-led, AI-assisted intelligent interactive platform significantly improved clinical and patient-centered outcomes in postdischarge RA care. By facilitating continuous monitoring, timely clinical action, personalized education, and sustained engagement, this digital intervention offers a scalable model to enhance disease control and patient empowerment in routine rheumatology practice.

XSL•FO
RenderX

Acknowledgments

The authors sincerely thank the patients who participated in this research.

Funding

This study was supported by the Hubei Provincial Natural Science Foundation (2025AFB835), Tongji Hospital Research Fund (2023D45), University-Industry Collaborative Education Program (230905329045253), and Scientific Research Innovation Fund of the Hubei Preventive Medicine Association (2025SWGKY487).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZZ, XZ, SY, XC
Data curation: ZZ, Li Zhang, SH
Formal analysis: XZ, SY
Funding acquisition: ZZ, SY
Investigation: Li Zhang, MD
Methodology: ZZ, XZ, Li Zhang, Lijuan Zhang, XC, SH
Project administration: ZZ, XZ, MD, Lijuan Zhang
Resources: ZZ, Lijuan Zhang, XC, SH
Software: SH
Supervision: SY, XC
Validation: SH
Visualization: XZ, SY
Writing—original draft: ZZ, XZ
Writing—review and editing: SY, XC, SH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Intelligent interactive platform for rheumatoid arthritis (English- and Chinese-language versions of the web interface).
[DOCX File, 1974 KB - medinform_v14i1e90784_app1.docx]

Multimedia Appendix 2
The integrated intelligent education module for rheumatoid arthritis (RA; partial interface in the Chinese-language version and all health education topics related to RA in the English-language version).
[DOCX File, 1335 KB - medinform_v14i1e90784_app2.docx]

Multimedia Appendix 3
“Smart Medication Manager” in the platform for rheumatoid arthritis (English- and Chinese-language versions of the web interface).
[DOCX File, 4297 KB - medinform_v14i1e90784_app3.docx]

Multimedia Appendix 4
The “Intelligent Consultation” module in the platform for rheumatoid arthritis (English- and Chinese-language versions of the web interface).
[DOCX File, 1758 KB - medinform_v14i1e90784_app4.docx]

Checklist 1
STROBE checklist.
[DOCX File, 32 KB - medinform_v14i1e90784_app5.docx]

References

1. Zhang Z, Gao X, Liu S, et al. Global, regional, and national epidemiology of rheumatoid arthritis among people aged 20–54 years from 1990 to 2021. Sci Rep 2025 Mar 28;15(1):10736. [doi: 10.1038/s41598-025-92150-1] [Medline: 40155668]

XSL•FO
RenderX

2. Finckh A, Gilbert B, Hodkinson B, et al. Global epidemiology of rheumatoid arthritis. Nat Rev Rheumatol 2022 Oct;18(10):591-602. [doi: 10.1038/s41584-022-00827-y] [Medline: 36068354]
3. Ramos-Petersen L, Reinoso-Cobo A, Ortega-Avila AB, et al. A clinical practice guideline for the management of the foot and ankle in rheumatoid arthritis. Rheumatol Int 2024 Aug;44(8):1381-1393. [doi: 10.1007/s00296-024-05633-1] [Medline: 38850327]
4. He J, Maharaj AB. Highlights of advancement in rheumatoid arthritis research and clinical practice. Best Pract Res Clin Rheumatol 2025 Mar;39(1):102040. [doi: 10.1016/j.berh.2025.102040] [Medline: 39961734]
5. Cubberley C, Maharaj A. Global RA treatment recommendations: an update from the various international societies. Best Pract Res Clin Rheumatol 2025 Mar;39(1):102019. [doi: 10.1016/j.berh.2024.102019] [Medline: 39496528]
6. Fahmi A, MacBrayne A, Humby F, Curzon P, Marsh W. Dynamic Bayesian network models for self-management of chronic diseases: rheumatoid arthritis case-study. Comput Biol Med 2025 May;189:109909. [doi: 10.1016/j.compbiomed.2025.109909] [Medline: 40043419]
7. Hu H, Xu A, Wang Z, Gao C, Wu X. Self-management behaviours in rheumatoid arthritis patients: what role do health beliefs play? Int J Nurs Pract 2025 Feb;31(1):e13320. [doi: 10.1111/ijn.13320] [Medline: 39834097]
8. Oh H, Suh CH, Kim JW, Boo S. mHealth-based self-management program for patients with rheumatoid arthritis: a pilot randomized controlled study. Nurs Health Sci 2024 Dec;26(4):e13187. [doi: 10.1111/nhs.13187] [Medline: 39498761]
9. Rodríguez-Sánchez-Laulhé P, Luque-Romero LG, Blanquero J, et al. A mobile app using therapeutic exercise and education for self-management in patients with hand rheumatoid arthritis: a randomized controlled trial protocol. Trials 2020 Sep 10;21(1):777. [doi: 10.1186/s13063-020-04713-4] [Medline: 32912305]
10. Rodríguez Sánchez-Laulhé P, Luque-Romero LG, Barrero-García FJ, et al. An exercise and educational and self-management program delivered with a smartphone app (CareHand) in adults with rheumatoid arthritis of the hands: randomized controlled trial. JMIR Mhealth Uhealth 2022 Apr 7;10(4):e35462. [doi: 10.2196/35462] [Medline: 35389367]
11. Shao JH, Yu KH, Chen SH. Effectiveness of a self-management program for joint protection and physical activity in patients with rheumatoid arthritis: a randomized controlled trial. Int J Nurs Stud 2021 Apr;116:103752. [doi: 10.1016/j.ijnurstu.2020.103752] [Medline: 32928503]
12. Mollard E, Michaud K. Self-management of rheumatoid arthritis: mobile applications. Curr Rheumatol Rep 2020 Nov 26;23(1):2. [doi: 10.1007/s11926-020-00968-7] [Medline: 33244621]
13. Müskens WD, Rongen-van Dartel SAA, Vogel C, Huis A, Adang EMM, van Riel P. Telemedicine in the management of rheumatoid arthritis: maintaining disease control with less health-care utilization. Rheumatol Adv Pract 2021;5(1):rkaa079. [doi: 10.1093/rap/rkaa079] [Medline: 33688619]
14. Magnol M, Eleonore B, Claire R, et al. Use of eHealth by patients with rheumatoid arthritis: observational, cross-sectional, multicenter study. J Med Internet Res 2021 Jan 29;23(1):e19998. [doi: 10.2196/19998] [Medline: 33512320]
15. Barlas N, Barlas SB, Basnyat S, Adalier E. Telemedicine in rheumatoid arthritis: a review of the PubMed literature. Mediterr J Rheumatol 2023 Mar 31;34(1):16-23. [doi: 10.31138/mjr.34.1.16] [Medline: 37223586]
16. Srikesavan C, Bryer C, Ali U, Williamson E. Web-based rehabilitation interventions for people with rheumatoid arthritis: a systematic review. J Telemed Telecare 2019 Jun;25(5):263-275. [doi: 10.1177/1357633X18768400] [Medline: 29669470]
17. Aletaha D, Ward MM, Machold KP, Nell VPK, Stamm T, Smolen JS. Remission and active disease in rheumatoid arthritis: defining criteria for disease activity states. Arthritis Rheum 2005 Sep;52(9):2625-2636. [doi: 10.1002/art.21235] [Medline: 16142705]
18. Maska L, Anderson J, Michaud K. Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment Questionnaire II (HAQ-II), Improved Health Assessment Questionnaire (Improved HAQ), and Rheumatoid Arthritis Quality of Life (RAQoL). Arthritis Care Res (Hoboken) 2011 Nov;63 Suppl 11(S11):S4-S13. [doi: 10.1002/acr.20620] [Medline: 22588760]
19. Lee JK, Grace KA, Foster TG, et al. How should we measure medication adherence in clinical trials and practice? Ther Clin Risk Manag 2007 Aug;3(4):685-690. [Medline: 18472991]
20. Li C, Huang J, Wu H, et al. Management of rheumatoid arthritis with a digital health application: a multicenter, pragmatic randomized clinical trial. JAMA Netw Open 2023 Apr 3;6(4):e238343. [doi: 10.1001/jamanetworkopen.2023.8343] [Medline: 37058302]
21. Fedkov D, Berghofen A, Weiss C, et al. Efficacy and safety of a mobile app intervention in patients with inflammatory arthritis: a prospective pilot study. Rheumatol Int 2022 Dec;42(12):2177-2190. [doi: 10.1007/s00296-022-05175-4] [Medline: 36112186]
22. Seppen BF, den Boer P, Wiegel J, et al. Asynchronous mHealth interventions in rheumatoid arthritis: systematic scoping review. JMIR Mhealth Uhealth 2020 Nov 5;8(11):e19260. [doi: 10.2196/19260] [Medline: 33151161]
23. Gajadhur K, Patil R, Anadachee AP, Kashyap A, Banerjee I. Effectiveness of digital health intervention in enhancing medication adherence among arthritis patients: a systematic review of randomized controlled trials. Nepal J Epidemiol 2025 Sep 1;14(4):1362-1373. [doi: 10.3126/nje.v14i4.77347] [Medline: 40919069]

XSL•FO
RenderX

24. De Angelis G, Brosseau L, Davies B, King J, Wells GA. The use of information and communication technologies by arthritis health professionals to disseminate a self-management program to patients: a pilot randomized controlled trial protocol. Digit Health 2018 Dec 17;4:2055207618819571. [doi: 10.1177/2055207618819571] [Medline: 30574341]
25. de Thurah A, Stengaard-Pedersen K, Axelsen M, et al. Tele-health followup strategy for tight control of disease activity in rheumatoid arthritis: results of a randomized controlled trial. Arthritis Care Res (Hoboken) 2018 Mar;70(3):353-360. [doi: 10.1002/acr.23280] [Medline: 28511288]
26. Richter JG, Nannen C, Chehab G, et al. Mobile app-based documentation of patient-reported outcomes - 3-months results from a proof-of-concept study on modern rheumatology patient management. Arthritis Res Ther 2021 Apr 19;23(1):121. [doi: 10.1186/s13075-021-02500-3] [Medline: 33874994]
27. de Punder YMR, Hendrikx J, den Broeder AA, Valls Pascual E, van Riel PL, Fransen J. Should we redefine treatment targets in rheumatoid arthritis? Low disease activity is sufficiently strict for patients who are anticitrullinated protein antibody-negative. J Rheumatol 2013 Aug;40(8):1268-1274. [doi: 10.3899/jrheum.121438] [Medline: 23729803]
28. Knudsen LR, Lomborg K, Hauge EM, Zangi HA, de Thurah A. The WebRA study: opportunities and challenges in digital patient education from the perspective of patients with rheumatoid arthritis: a qualitative study. Patient Educ Couns 2023 Nov;116:107969. [doi: 10.1016/j.pec.2023.107969] [Medline: 37672918]
29. Subdar S, Hoens AM, White K, et al. An environmental scan and appraisal of patient online resources for managing rheumatoid arthritis flares. J Rheumatol 2024 Jun 1;51(6):577-586. [doi: 10.3899/jrheum.2023-1025] [Medline: 38490667]
30. Sparks JA, Harrold LR, Simon TA, et al. Comparative effectiveness of treatments for rheumatoid arthritis in clinical practice: a systematic review. Semin Arthritis Rheum 2023 Oct;62:152249. [doi: 10.1016/j.semarthrit.2023.152249] [Medline: 37573754]
31. Knudsen LR, Lomborg K, de Thurah A. Design and development of an e-learning patient education program for self-management support in patients with rheumatoid arthritis. PEC Innov 2021 Oct 31;1:100004. [doi: 10.1016/j.pecinn.2021.100004] [Medline: 37364010]
32. Primdahl J, Bremander A, Hendricks O, et al. Development of a complex Interdisciplinary Nurse-coordinated SELf-MAnagement (INSELMA) intervention for patients with inflammatory arthritis. BMC Health Serv Res 2024 Jan 17;24(1):87. [doi: 10.1186/s12913-023-10463-1] [Medline: 38233834]

Abbreviations

AI: artificial intelligence
DAS28: Disease Activity Score in 28 Joints
HAQ-II: Health Assessment Questionnaire–II
mHealth: mobile health
RA: rheumatoid arthritis
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Benis; submitted 04.Jan.2026; peer-reviewed by S Butler, S Mohanty; accepted 06.Mar.2026; published 02.Apr.2026.

Please cite as:
Zhang Z, Zhang X, Zhang L, Du M, Zhang L, Yang S, Cai X, Hou S
An Intelligent Interactive Management Platform for Rheumatoid Arthritis Care: Real-World Observational Study
JMIR Med Inform 2026;14:e90784
URL: https://medinform.jmir.org/2026/1/e90784
doi:10.2196/90784

© Ziyun Zhang, Xinyue Zhang, Li Zhang, Mingyan Du, Lijuan Zhang, Siyu Yang, Xiaoli Cai, Shengchao Hou. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying and Analyzing Bot-Generated Responses in Online Health Care Surveys: Methodological Study

Emily Hamovitch[1], MSW, MPH; Kaileah McKellar[1], MPH, PhD; Walter P Wodchis[1,2], MA, PhD

[1]Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, Toronto, ON, Canada

[2]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada

Corresponding Author:
Emily Hamovitch, MSW, MPH
Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, Toronto, ON, Canada

Abstract

Background: The increasing reliance on online surveys for collecting patient-reported feedback for health care research has led to growing concerns over fraudulent responses generated by bots. These automated responses threaten data integrity by fabricating survey results, distorting statistical analyses, and potentially misguiding policy decisions. Addressing this issue is critical for maintaining the validity of research findings that inform health care practice and policy.

Objective: This study aimed to develop a robust set of criteria for identifying bot-generated responses in online health care surveys and to examine how these responses impact data quality. We then explored differences in survey results between probable human and suspected bot respondents in a survey assessing patient-reported outcome measures and patient-reported experience measures within a geographic region in Ontario, Canada.

Methods: A survey was conducted from July to October 2023 using Research Electronic Data Capture (REDCap; Vanderbilt University), distributed with a generic link via email, and later shared on social media. The survey collected data on health care use, patient experiences, health outcomes, digital health care engagement, and demographics. A 3-tier classification system was developed to detect bot responses based on predefined "red flags," including duplicate open-ended responses, inconsistent demographic reporting, identical timestamps, and location discrepancies. Quantitative analysis included chi-square tests to assess differences between probable human and suspected bot responses and Spearman correlation tests to examine relationships among health care indicators.

Results:: Analysis included 1154 responses, with 58% (n=668) classified as suspected bot-generated. The most frequent suspected bot-identification criterion was duplicated open-ended responses (293/668, 44%). Chi-square tests revealed statistically significant differences (P<.05) between suspected bots and probable humans across most survey items. Suspected bots demonstrated response patterns concentrated in the middle of Likert scales, whereas probable humans were more likely to select extreme values. Correlation analyses showed that expected relationships between key health indicators (eg, depression symptoms) were present in probable human responses but reversed in suspected bot-generated data, highlighting the potential for compromised validity in unfiltered survey datasets.

Conclusions: The findings underscore the necessity of implementing bot prevention and detection methods in online health care surveys to preserve data integrity. Failure to do so risks distorting research conclusions, particularly in health equity studies where demographic misclassification may bias results. The study highlights effective bot detection strategies, including open-text analysis, timestamp evaluation, and geographic validation, and recommends integrating these techniques into survey design. As bots continue to evolve, ongoing advancements in bot prevention and detection will be crucial to maintaining the reliability of digital health research.

(JMIR Med Inform 2026;14:e73622) doi:10.2196/73622

KEYWORDS

bots; online surveys; data integrity; patient-reported outcome measures; patient-reported experience measures; health care research; digital health; survey validity

Introduction

Patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) are invaluable tools for understanding and improving health care from the patient's perspective. PROMs assess health status, functional abilities, symptoms, and quality of life directly from patients, often through self-reported questionnaires, providing insights into the impact of interventions and overall care quality [1-3]. PREMs, on the other hand, capture patients' perceptions of their health care experience, focusing on whether key processes

XSL•FO
RenderX

occurred and how care was delivered [2]. These measures are essential for implementing person-centered care, as they enable health care providers to align their practices with patient needs and experiences [4]. PROMs and PREMs have also been incorporated into care routines to support quality improvement initiatives [4]. By prioritizing patient-reported data, health care systems can address inequities, promote patient-centered and value-based care, and ensure that care delivery aligns with the priorities and needs of diverse patient populations [4].

A common approach to collecting PROMs and PREMs is through online surveys. Online surveys have become an increasingly prominent tool for research, particularly in social and behavioral health fields, due to their efficiency, cost-effectiveness, and ability to overcome challenges faced by traditional data collection methods such as face-to-face, mail, or telephone surveys [5,6]. These challenges include declining response rates, rising costs, and the need for social distancing during events like the COVID-19 pandemic [5,7-10]. Internet-based platforms such as Qualtrics (Qualtrics International) Inc., SurveyMonkey (Momentive Inc.), and Research Electronic Data Capture (REDCap; Vanderbilt University) have enabled researchers to collect data remotely, making it possible to reach large general populations [5,7,10]. These tools reduce the resources required for data collection, minimize participant burden, and offer greater anonymity, encouraging honest disclosure of personal or sensitive information [7,10]. Additionally, the limited availability of a sampling frame with accurate contact information has necessitated the use of distribution methods involving generic survey links and promotion through online platforms and social media. Unfortunately, recruiting participants through generic links creates susceptibility to data quality issues [11].

While early guidance on online surveys for health research warned of the risk of hoax responses [12], the increasing use of online surveys has been accompanied by a growing concern over fraudulent responses, particularly from bots - applications that automatically complete surveys – producing insincere data and inaccurate results [13]. These bots are capable of mimicking human behavior by selecting multiple-choice answers and generating plausible open-ended responses, making them difficult to detect. Bots can submit hundreds of fabricated responses within hours, potentially overwhelming researchers with invalid data and causing significant financial loss if participant incentives are misallocated [8]. Their presence threatens data quality, which can distort statistical analyses and compromise research findings [10]. Researchers, evaluators, and program analysts also face challenges in identifying fraudulent responses due to the increasing sophistication of bots and the anonymity of online participants [14]. Despite the implementation of data protection mechanisms to prevent bot responses, such as reCAPTCHA (Google LLC), bot programmers easily bypass these safeguards using virtual private networks or servers [7]. As bots become more prevalent and capable, addressing this issue is critical to ensuring data integrity and the validity of research findings that inform policy and practice.

This study adds to this literature by proposing a structured 3-tier classification system that integrates multiple indicators of responses by suspected bots, offering a robust and transparent protocol for identifying suspected bot responses. Additionally, this study goes beyond merely identifying suspected bot-generated responses in online surveys; it systematically examines how suspected bots influence response patterns in a health care experience survey. While existing research has focused on detecting fraudulent responses, fewer studies have examined how suspected bot-generated data may distort patient-reported experience and outcome measures in health care contexts. Our study extends this work by analyzing discrepancies between suspected bot and probable human responses in a survey of PROMs and PREMs. Such differences provide an assessment of potential bias and the relative importance of detection.

Methods

Survey

This study was undertaken within a broader evaluation of Ontario Health Teams (OHT), focusing on patient experiences and outcomes. A patient survey was implemented to support planning and evaluation of integrated care for attributed populations in Ontario, Canada. Consent was obtained from all respondents included in the analysis. The survey was intended for patients from a large OHT, gathering data on PREMs and PROMs. Questions included Likert scale items on health care use, health outcomes, accessing care, having someone to count on, being heard, managing health, safety, care transitions, digital health care use, demographics; as well as an open-text section to provide additional feedback. Six timestamps were programmed throughout the survey at multiple points to track survey progress. Data were collected between July 2023 and October 2023 using REDCap. REDCap is a web-based platform designed for research data capture, widely used in academic and clinical studies. The survey was distributed through the OHT's email lists and social media platforms. Social media distribution occurred primarily through Facebook (Meta), which was the OHT's main platform. Posts containing the survey link and QR code were shared periodically between July and October 2023, and a paid Facebook advertisement was implemented in early October to boost participation. Participants were invited to enter a draw for a $25 CAD (US $18.62) gift card as an incentive. Entry into the draw was voluntary and completed through a separate form linked on the survey's thank-you page, allowing respondents to opt out and ensuring that contact information was stored separately from survey responses. A copy of the survey is provided in Multimedia Appendix 1.

Suspected Bot Identification

Upon reviewing the qualitative responses and timestamp data, the research team identified clear evidence of suspected bot-generated responses in the survey. Many responses were duplicative, contained overly formal or generic language, or were nonsensical. The response patterns, combined with irregular timestamp data, confirmed the presence of suspected bots. As a result, the team developed specific criteria to systematically identify and remove suspected bot-generated responses from the dataset.

XSL•FO
RenderX

Criteria to identify suspected bots were developed using the following themes: generic and duplicative open-ended responses; concerns with location as determined by postal code (including postal codes that were invalid or nonsensical); timing irregularities; identical timestamps to other participants; illogical demographic reporting; and discrepancies in reporting health care use. In addition, a "bot takeover day" (October 10 onward) was included as a criterion, reflecting the sudden influx of suspicious responses observed on or after that date.

To identify bots in the survey data, we developed a 3-tier classification system based on specific "red flags" (concerning or questionable features or responses in the data). This approach builds on prior research that has typically relied on individual indicators or sets of criteria considered in parallel to flag fraudulent responses [5,7-9]. By combining multiple indicators into a structured, tiered framework, our system increases the robustness and transparency of suspected bot identification. Tier 1 criteria identified a respondent as a suspected bot if any single criterion occurred on its own (eg, duplicate open-ended responses that were exact full-text matches, not brief "No"-type replies, invalid postal codes, or identical timestamps). Tier 2 classified a respondent as a suspected bot if 2 or more criteria from its list were observed. For example, one tier 2 criterion included having a valid postal code outside of the region where participants received health care services. This criterion was not sufficient on its own to categorize a respondent as a suspected bot, as some participants may legitimately live elsewhere and receive care in the area where the participants were sampled. However, when combined with other criteria, it strengthened the likelihood of being a suspected bot-generated response. Finally, tier 3 flagged respondents as suspected bots if any tier 3 criterion was combined with at least 2 additional criteria from either tier 2 or tier 3, resulting in a total of 3 or more issues. One Tier 3 indicator reflected a sudden influx of suspicious responses observed on October 10, 2023 ("bot takeover day"). This indicator was not used as a standalone criterion and had minimal influence on classification: more than 99% of suspected bot responses were already identified using Tier 1 and/or Tier 2 criteria alone, with only 3 additional responses flagged through the use of this indicator. Demographic characteristics such as race or ethnicity, gender, and sexual orientation were not used as criteria for suspected bot identification. Demographic variables were analyzed only after classification to describe group-level differences between suspected bots and probable humans.

One criterion for identifying suspected bots was the presence of open-text responses that were deemed to be either (1) generic or informal, or (2) had a similar sentence structure to other responses. To evaluate this more subjective criterion, 2 reviewers (LC and SU) independently coded open-ended responses based on these criteria. Responses were categorized as suspected bots if reviewers agreed they met either criterion. When reviewers disagreed on whether open-ended responses should be classified as generic or informal, a third member of the research team (KM) provided the final determination. Interrater reliability between the 2 independent coders was substantial (Cohen κ=0.63, P<.001).

Although this study applied the identification criteria for a survey implemented on the REDCap platform, the 3-tier system is designed to be adaptable across different survey platforms, including Qualtrics, SurveyMonkey, and other online tools. The criteria used (such as duplicated open-text responses, timestamp irregularities, and geographic inconsistencies) are not platform-specific.

Suspected Bot Response Analysis

To examine differences in survey responses between probable humans and suspected bots, bar charts were created to visually compare response patterns. Chi-square tests were conducted to identify significant differences in the proportions of responses between the 2 groups. For select-all-that-apply items, including "Who did you receive care from?," responses were summarized into a mutually exclusive categorical variable prior to analysis. Respondents selecting more than 1 option were classified as having received care from "multiple providers," while those selecting a single option were classified according to that provider type.

To account for multiple comparisons across survey items, a Bonferroni correction was applied (adjusted α=.05/59=0.00085) to control for Type I error. We quantified the magnitude of group differences by calculating effect sizes for each chi-square test. Cramer V was computed for all categorical comparisons, along with 95% CIs.

Additionally, Spearman correlation tests were conducted to evaluate significant relationships between 2 health indicators, analyzed separately for probable humans and suspected bots, to identify patterns that align with or deviate from expected human perception and behavior. Two sets of correlations were examined: (1) the correlation between "little interest or pleasure in doing things" and "feeling down, depressed, or hopeless," and (2) the correlation between self-rated health and the frequency of interaction with the health system. The first pair represents the 2 core items of the Patient Health Questionnaire–2 (PHQ-2), a validated depression screener in which these items are strongly and positively correlated with one another [15]. The second pair was selected because extensive literature has shown that poorer self-rated health is associated with increased health care use, including higher rates of hospitalization and outpatient visits [16,17]. These 2 pairs, therefore, provide theoretically justified validity checks.

Ethical Considerations

This study received approval from the University of Toronto Research Ethics Board (#38072). All procedures were conducted in accordance with institutional ethical standards and the principles of the Declaration of Helsinki. Participants provided informed consent electronically prior to beginning the survey. The consent form outlined the purpose of the study, the voluntary nature of participation, potential risks and benefits, and data management procedures. Participants were informed that they could decline to answer any question or withdraw at any time prior to survey submission. To protect privacy and confidentiality, survey responses were collected anonymously. Any contact information provided for entry into the incentive draw was submitted through a separate form and stored

XSL•FO
RenderX

separately from survey data. Data were stored on secure, password-protected institutional servers accessible only to authorized members of the research team.

Results

Frequency of Suspected Bot Identification

Table 1 illustrates the final criteria for identifying suspected bots and the frequency with which each criterion was met among respondents ultimately classified as suspected bots or probable humans. While some probable human respondents exhibited isolated red flags, they did not meet the necessary combination of criteria required for suspected bot classification. After removing survey responses where substantive questions were not answered (n=231), 1154 survey respondents remained. Among them, 58% (n=668) were categorized as suspected bots, and 42% (n=486) were categorized as probable humans. Table 1 illustrates the number of respondents who met the eligibility criteria for each identification criterion. Overall, 529 survey respondents met the criteria for Tier 1, 570 respondents met the criteria for Tier 2 (136 of whom did not meet the criteria for Tier 1), and 551 respondents met the criteria for Tier 3 (3 of whom did not meet the criteria for Tier 1 or Tier 2).

XSL•FO
RenderX

Table . Frequency of each suspected bot identification criteria. Percentages within each tier sum to greater than 100% because a single respondent may meet multiple identification criteria.

Criteria	Assessed categories		
	Suspected bot (n)	Probable human (n)	Respondents identified as suspected bots who met criteria, n (%)
Tier 1			
Had an identical open-ended question response identical to at least one other respondent (Duplicate open-ended responses that were exact full-text matches, not brief "No"-type replies)	293	0	(293/668) 44%
Open text is generic, informal, or similar sentencing structure	49	0	(49/668) 7%
Duplicate invalid postal codes that were completed on the same day	189	0	(189/668) 28%
Duplicate out-of-province postal codes (excluding Quebec) that were completed on the same day	152	0	(152/668) 23%
Tier 2			
Identical open-ended question marked as "no" that also fell within the same day and similar time frame	80	0	(80/668) 12%
2 or more survey respondents entered their surveys within 30 seconds of one another (but not identical), and this is the case in 3 or more timestamps	340	2	(340/668) 51%
Had an identical timestamp (identical hours, minutes, and seconds) with at least one other respondent on the following timestamps:			
First timestamp	114	0	(114/668) 17%
Second timestamp	93	0	(93/668) 14%
Third timestamp	75	1	(75/668) 11%
Fourth timestamp	80	0	(80/668) 12%
Fifth timestamp	69	2	(69/668) 10%
Sixth timestamp	75	0	(75/668) 11%
Survey was completed between 1 AM and 5:00 AM Eastern Time (as identified on any of the timestamps)	154	7	(154/668) 23%
Time spent on survey is 3 minutes or less (among respondents who completed the survey)	42	11	(42/668) 6%
Invalid postal code (gibberish or numeric)	189	0	(189/668) 28%
Valid postal code but outside of Ontario	153	0	(153/668) 23%
Valid Ontario postal code but outside of the region of the primary care center in which respondents were recruited	165	5	(165/668) 25%

XSL•FO
RenderX

Criteria	Assessed categories		
	Suspected bot (n)	Probable human (n)	Respondents identified as suspected bots who met criteria, n (%)
Discrepancy in reporting ED[a] visits	347	55	(347/668) 52%
Between the ages of <18 - 44 years but reported living in a retirement home or long-term care facility.	123	1	(123/668)18%
Tier 3			
Bot takeover day of October 10 or after	618	94	(618/668) 93%
Postal code from neighboring province	55	3	(55/668) 8%

[a]ED: emergency department.

In tier 1, the most commonly reported “red flag” was having an identical open-ended response to another survey respondent. In tier 2, the most common ’red flag’ was a discrepancy in reporting an emergency department visit, where respondents indicated they had received care from the emergency department in 1 question but later reported that they had not been to the emergency department in the past 12 months. There was also a high frequency of respondents completing their surveys within similar timeframes (completing 3 or more of the same sections within 30 s of another respondent).

Differences in Responses Between Suspected Bots and Probable Humans

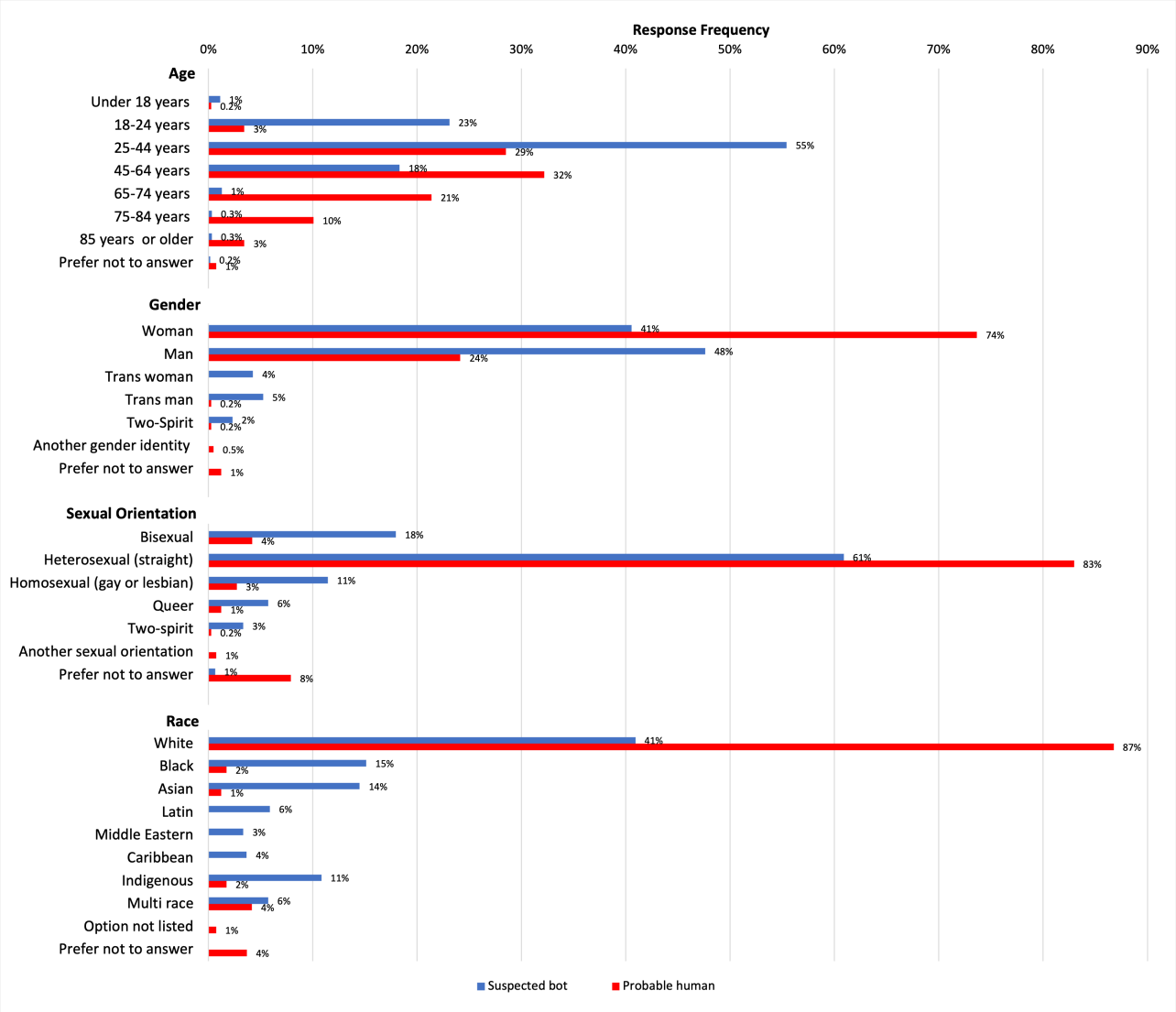
Overall, probable humans had a higher average percentage of missingness on all questions (mean 13% [for probable humans vs mean 5% for suspected bots).

Chi-square tests revealed significant differences ($P<.05$) in response distributions between probable humans and suspected bots for 55 out of 59 survey questions. After Bonferroni correction for multiple comparisons (adjusted $\alpha=.00085$), 50 of 59 valid comparisons remained statistically significant (Table S1 in Multimedia Appendix 2). The magnitude of effect sizes ranged from Cramer V=0.01 (“When you left the hospital were you provided with easy to follow instructions on whom to contact if you had a question about your treatment or if your condition became worse?”) to Cramer V=0.72 (“In general, do your health care providers encourage you to bring someone with you to your appointments?”), with a mean Cramer V of 0.31 (Table S1 in Multimedia Appendix 2). Among the 9 survey questions that showed no statistically significant difference between probable humans and suspected bots, most items were binary (yes or no) questions that followed branching logic and were therefore presented only to a subset of respondents. The remaining non-significant items were measured on a 5-point response scale, where probable humans tended to provide slightly less positively skewed responses (for example, being more likely to select experiencing “slight pain or discomfort” rather than “no pain or discomfort”).

In terms of demographics (Figure 1), suspected bots more often reported younger ages (under 18–44 years) and being male, while probable humans more often reported being female. Both groups most frequently identified as heterosexual and White, but suspected bots more often selected minority sexual orientations (eg, bisexual, queer, and two-spirit) and racial or ethnic categories (eg, Black, Asian, Latin, and Indigenous).

XSL•FO
RenderX

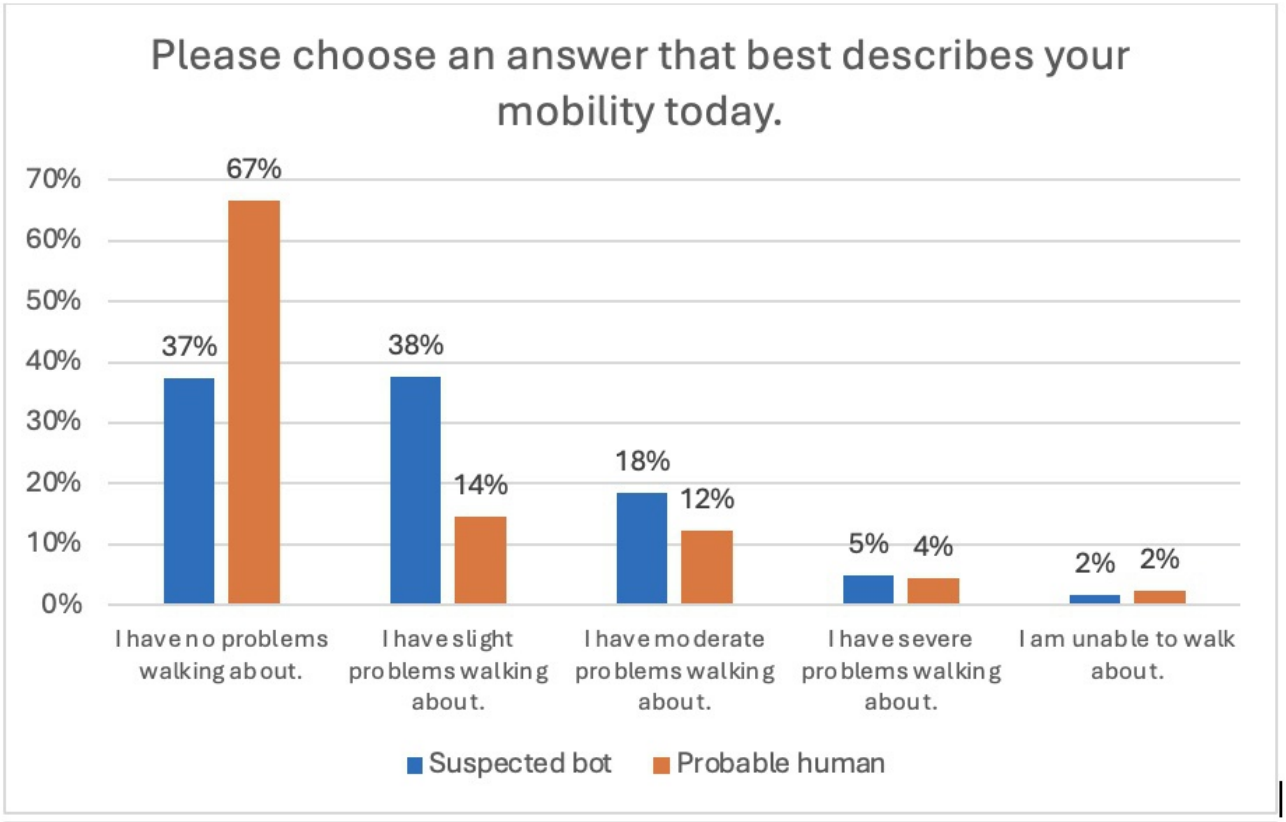
Figure 1. Distribution of demographics for suspected bots versus probable humans.

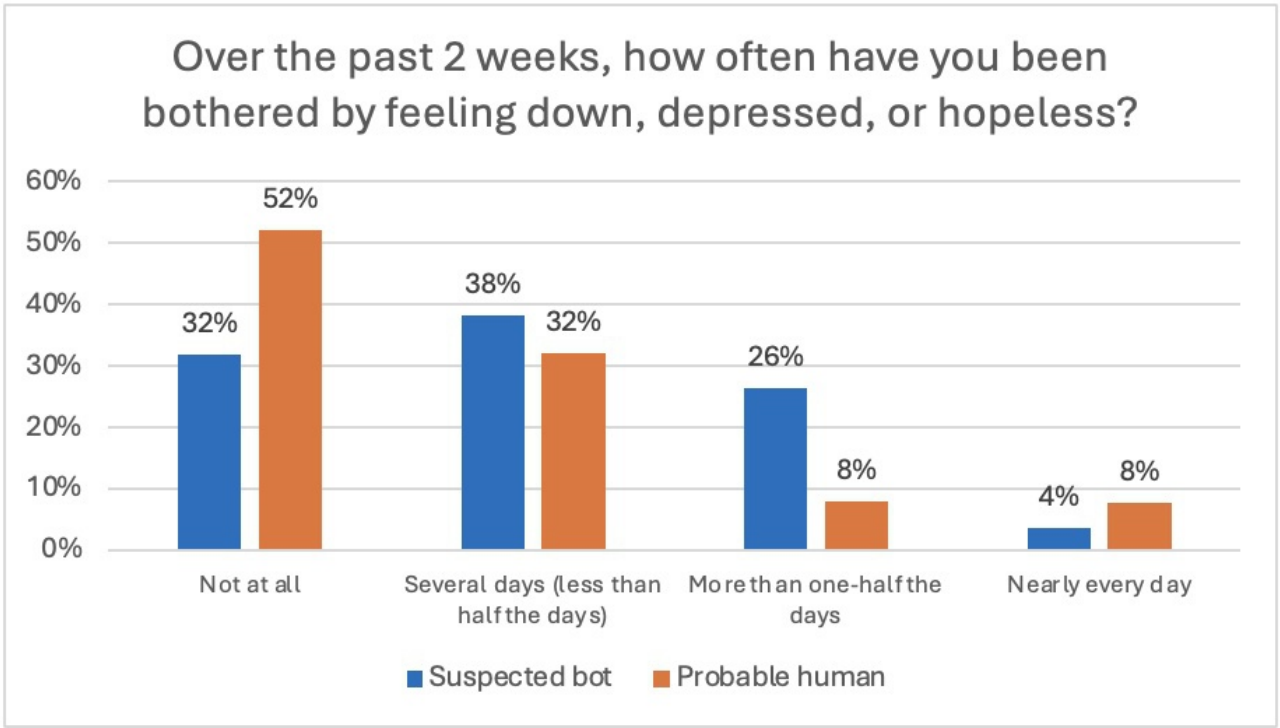


Visualizations (Figures 2 and 3) showed that suspected bots tended to cluster at mid-scale responses, while probable humans more often selected extreme values. For PROMs, probable humans were more likely to report the most favorable outcomes (eg, “no problems walking about”) and the least favorable (eg, “nearly every day” feeling depressed), whereas suspected bots gravitated toward midpoints (eg, “slight problems,” “several days”). Similar patterns appeared in PREMs, where suspected bots more frequently selected “somewhat” responses (eg, “somewhat easy” to access supports, “somewhat too long” to get care), while probable humans showed a greater spread across extremes.

XSL•FO
RenderX

Figure 2. Distribution of probable human and suspected bot responses for patient-reported outcome measures.





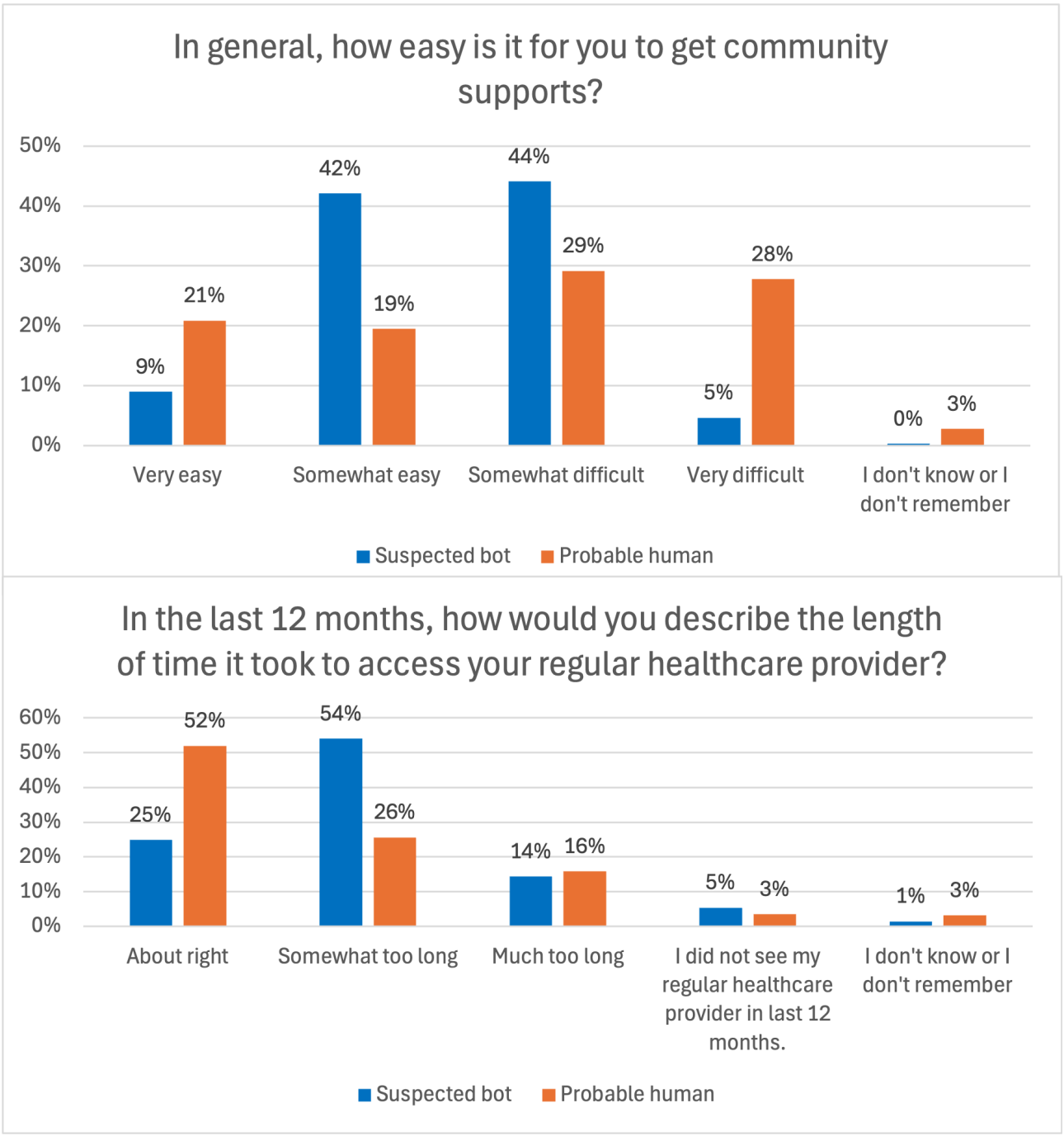
The survey also assessed health service use, including emergency department visits, hospitalizations, and specialist care in the past year. Suspected bots more frequently reported using the emergency department and being hospitalized compared to probable humans. When asked yes or no questions about these services (such as whether an emergency department visit could have been managed by a regular provider, whether their provider was informed after hospital discharge, or whether a specialist had relevant medical information), suspected bots were more likely to respond "yes" than probable humans. This was consistent with other patterns, such as reporting "somewhat confident" when asked about managing health post-emergency department visit, while probable humans were more likely to select "very confident" or "not at all confident."

The survey included 1 nominal question, which asked about the main reason participants had not looked at their medical records online (for those who indicated this to be the case). There did not appear to be any trend or pattern for the distribution of

XSL•FO
RenderX

responses. Ironically, a substantially higher percentage of suspected bots (compared to probable humans) indicated that a reason for not having looked at medical records online was not having reliable access to the internet.

Figure 3. Distribution of probable human and suspected bot responses for patient-reported experience measures.



Correlations

Spearman correlations indicated significant correlations between the indicators of having "little interest or pleasure in doing things" and "feeling down, depressed, or hopeless," both of which are indicators of depression. For probable humans, the Spearman correlation was 0.48 (P<.001), indicating a moderate positive correlation between the 2 variables, whereas for suspected bots, the Spearman correlation of −0.33 (P<.001) indicated a moderate negative correlation between the 2 variables.

When analyzing the correlation between self-rated health ("how do you rate your own health?") and frequency of interacting with the health system ("how often do you interact with the health system?"), a Spearman correlation of −0.07 for probable humans indicated a weak negative correlation; however, this was not significant (P=.15). For suspected bots, the Spearman correlation was 0.06, indicating a weak positive correlation; however, this was also not significant (P=.15).

XSL•FO
RenderX

Discussion

Principal Findings

The results of this study demonstrate the imperative need for researchers, evaluators, and program analysts in the health care field to recognize and address the potential threats posed by bots in online survey research. Failure to detect and remove bot-generated responses can lead to compromised data quality and incorrect inferences. Implementing robust data screening strategies is important for reducing fraudulent responses and improving the likelihood that valid data are retained. These responses respect the integrity of genuine participants whose experiences deserve to be accurately reported.

Implications of Including Bot Data in Health Care Surveys

Our findings extend prior work on bot response patterns by demonstrating how these differences manifest in PROMs and PREMs, with implications for health care survey research. Across multiple survey domains—health outcomes, health care access, and experiences—probable humans were more likely than suspected bots to select the most favorable outcomes (or in some cases, the least favorable response options). In contrast, suspected bots exhibited a tendency to select responses clustered around the middle of the scale. As a result, the presence of bot-generated data could dilute the strength of human responses, skewing average scores toward the midpoint and diminishing the ability to detect meaningful trends. This issue extends beyond scaled responses; even in nominal questions, such as the reason for not accessing medical records online, suspected bots disproportionately selected specific response options, such as lacking reliable internet access. This discrepancy suggests that bot-generated data may introduce artificial patterns in categorical variables, leading to inaccurate estimates of barriers to health care access and potentially misinforming policy decisions. This distortion may undermine construct validity, as the data fail to capture real-world experiences. Additionally, probable humans demonstrated a higher level of missingness compared to suspected bots. This likely reflects natural variation in survey completion behavior (such as skipped or unanswered questions). In contrast, the lower level of missingness among suspected bots may indicate automated completion rather than a marker of higher data quality. This difference underscores the importance of evaluating missing data patterns alongside other indicators when assessing data integrity.

Beyond response patterns, the inclusion of bot-generated data introduces a significant risk of bias, particularly in research focusing on health equity. The demographics of suspected bot responses differed notably from those of probable human participants. These patterns may reflect random bot response patterns inflating less common categories; however, this pattern may be partially driven by the ordering of the race or ethnicity response options. In the survey, several minority categories appear in the middle of the list, whereas “White” appears at the bottom. Given the tendency of suspected bots to select middle-of-the-scale responses, the positioning of race or ethnicity response options likely contributed to the higher selection of minority categories among bots. We also acknowledge the possibility that some marginalized respondents were misclassified as suspected bots. This discrepancy poses a major threat to the validity of studies that examine health care use and outcomes among specific demographic subgroups. In equity-focused research, where accurate representation of marginalized populations is crucial, failing to filter out bots could lead to misleading conclusions and inappropriate policy recommendations.

The impact of suspected bot interference is further underscored when examining the relationship between key indicators of depression. Among probable human respondents, the correlation between “having little interest or pleasure in doing things” and “feeling down, depressed, or hopeless” aligned with established patterns in the literature, reflecting a logical association between these 2 related constructs. However, among suspected bot-generated responses, this relationship was unexpectedly reversed, indicating a negative correlation that does not align with theoretical expectations. Notably, a moderate negative correlation suggests a systematic response pattern (such as contradictory response selection) rather than a pattern consistent with genuine symptom co-occurrence. This inconsistency suggests that suspected bot responses do not accurately reflect underlying psychological constructs, reinforcing the need to identify and eliminate suspected bots from survey research. The inconsistency also provides face validity to our bot-detection algorithm.

Bot Detection Strategies

Several methods to identify suspected bots that were used in this study align with those used in similar research. These include: analyzing open-ended responses for nonsensical, duplicative, or inconsistent data [5,8,10,18]; tracking survey completion times [5,7,8,13,19,20]; flagging surveys that were completed at unusual times of the day [8]; and identifying suspicious patterns related to zip codes and geographic location [20]. We applied all applicable bot detection strategies for which we had data available. There are additional bot detection techniques that were not used in this study but have been used in prior research. These include: checking participants' responses to questions that are asked in 2 ways [8,14,19], (such as “how old are you?” and “what y were you born?”); using IP addresses to check for responses that were completed in the same household or very close to one another in relation to the time the survey was submitted [5,14]; and examining email addresses for suspicious wording or similar content [8,20]. Another strategy involves external validation of participant information, such as verifying names, addresses, phone numbers, and birth dates to ensure responses are unique and legitimate [21]. Investigators can also cross-reference participants’ details using publicly available databases, social media, or professional networking sites to further confirm authenticity, although ethical considerations must be weighed when implementing such approaches [21]. While these strategies may help reduce fraudulent responses, they can also create additional burdens and reduce anonymity for legitimate participants.

Among the bot detection strategies used in this study, some were particularly effective in identifying suspected bot-generated responses. Notably, open-ended responses provided valuable

XSL•FO
RenderX

insight, as suspected bot-generated answers often contained repetitive, illogical, or off-topic content across multiple responses. Unusual quantitative responses alone can be difficult to interpret, as they might seem plausible despite being artificially generated. However, open-ended responses can provide valuable context that may help in assessing the authenticity of data. Unlike structured response options, which bots can randomly or systematically select, qualitative responses require coherence and reasoning, making them a powerful tool for distinguishing between genuine participants and automated responses. That said, recent advances in generative artificial intelligence mean that open-text answers are becoming increasingly sophisticated and may no longer be a fully reliable safeguard against fraud [22]. Geographic inconsistencies, such as invalid or duplicate postal codes, or locations outside the intended survey region, were also strong indicators of suspected bot activity, although individual respondents may misreport to preserve anonymity. Additionally, anomalies in time-related metrics, including extremely short survey completion times, completion at unusual hours, or multiple surveys submitted at identical timestamps, further helped differentiate suspected bots from probable human respondents. Based on these findings, we recommend collecting postal or ZIP codes and incorporating multiple timestamps within surveys to enhance suspected bot detection. While no single strategy is universally effective across all surveys, we recommend using a combination of detection methods tailored to the specific survey and target population.

While several strategies are outlined to detect bots, it is also important to prevent bots from entering surveys in the first place. Tools such as Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) have been developed to offer an additional level of security against bot infiltration [23]. However, research has shown that these tools can be bypassed, as bots are increasingly programmed to recognize and circumvent such challenges [7,20,24]. Similarly, honeypot questions (hidden prompts designed to engage only automated respondents) have also proven ineffective, as more sophisticated bots can now avoid these traps [8,20]. Indeed, researchers who have used this method have reported that bots did not provide responses to these questions [8]. Given the growing complexity of automated survey fraud, it is essential for technology companies to continue developing advanced bot detection solutions. Survey platforms such as REDCap and Qualtrics should integrate these technologies to automate bot prevention, reducing the burden on researchers and ensuring more consistent data quality measures.

Recommendations

It is recommended that bot detection strategies be implemented at multiple stages in the research process. This includes prior to data collection (eg, using targeted recruitment with specific distribution lists), during data collection (eg, monitoring data daily to identify potential fraud), and following data collection (eg, implementing a systemic check of potential fraud) [18]. Finally, increasing awareness among researchers about the impact of bots in survey research is crucial. We recommend that training in bot prevention and detection be incorporated into research methods and survey design coursework to equip researchers with the necessary skills to safeguard data integrity. There will continue to be a value and need to reach out with surveys to broad groups of public and enabling survey research in this context will continue to be important.

As internet-based research (and survey research in particular) continues to expand, there is a growing opportunity for researchers across disciplines to contribute to the development of comprehensive bot detection methodologies. Moving forward, greater efforts should be directed toward developing universal technologies and initiatives for bot detection. Standardized approaches that can be seamlessly integrated across different research platforms will help safeguard the validity of online data collection and ensure that digital survey research remains a trusted tool for generating meaningful insights.

Limitations

While this study offers valuable insights, several limitations must be acknowledged. Since bot detection methods were applied to a single case study, their effectiveness has yet to be evaluated across different datasets, which may limit generalizability. Additionally, the data were collected within one region of Ontario, and the characteristics of the probable human responses, and their differences from suspected bot responses, may not be representative of broader populations. As response patterns may vary across settings, researchers working in other regions or with different populations should assess whether similar differences between suspected bot and probable human data are observed. However, the bot identification framework is adaptable, as criteria, such as postal or ZIP code validation, open-text analysis, and completion times can be tailored to reflect local contexts. Differences observed between probable human and suspected bot responses may also partially reflect differences between recruitment channels, such as respondents reached through email lists versus social media postings. Survey respondents recruited via open online platforms may differ systematically in demographics or response behaviors, independent of bot activity. Although some degree of misclassification is therefore possible, our detection framework was deliberately conservative and assumed responses were human unless sufficient evidence across multiple indicators suggested otherwise. Despite using various suspected bot detection strategies, there is also potential for undetected sophisticated bots, which could still impact the dataset's validity. Furthermore, the study was conducted at a specific point in time, and as automated response technologies evolve, the effectiveness of these detection methods may diminish. The study relied on assumed response patterns to identify suspected bots, but without absolute verification, such as IP tracking or manual validation, there remains some uncertainty in classification. Finally, while we observed that suspected bot responses reversed an expected positive correlation between depression-related items, this pattern was not incorporated into our 3-tier classification framework. Such a pattern would align conceptually with a Tier 3 criterion, as it represents a complex and context-dependent anomaly that would require corroboration from additional indicators to support classification. Future research should examine whether anomalies in response patterns, in addition to structural indicators, can serve as reliable markers of fraudulent responses.

XSL•FO
RenderX

Conclusions

This study provides criteria for identifying bots in online surveys. In comparing survey results from the suspected bots and probable humans, identified by the developed criteria, we found important differences in response patterns that would jeopardize data validity if bots were not correctly identified. By incorporating robust data validation protocols early in the research process, researchers can minimize bias and prevent the inclusion of inauthentic data, strengthening the reliability of findings. This study should be interpreted as a cautionary case rather than a prescriptive template. Without verified contact lists or other forms of participant validation, open online surveys will remain vulnerable to contamination despite the use of screening methods.

Acknowledgments

Artificial intelligence tools were used to assist with technical tasks such as formatting tables and supporting code refinement during data analysis. All analytic decisions, code implementation, result verification, and manuscript content were developed and reviewed by the authors.

Funding

This research was supported by a grant from the Ontario Ministry of Health to the Health System Performance Network (#694). The funders had no role in data analysis, decision to publish, or preparation of the report.

Data Availability

The datasets that were created and analyzed during this study are not publicly available, as public data sharing was not included in the original research ethics approval and participant consent form. Deidentified data may be made available for research purposes upon reasonable request to the corresponding author, subject to institutional approvals and data-sharing agreements, and provided that participant confidentiality can be maintained.

Authors' Contributions

EH and KM contributed to the conceptualization and methodology of the study, curated the data, and developed and implemented the coding criteria to identify suspected bots (with oversight from WW). EH completed the analysis and drafted the manuscript. KM and WW provided critical revision and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ontario Health Team Patient Survey (full survey instrument used in the study).
[DOCX File, 1252 KB - medinform_v14i1e73622_app1.docx]

Multimedia Appendix 2
Chi-square test results comparing survey responses between probable humans and suspected bots, with Bonferroni-adjusted *P* values and Cramer V effect sizes.
[DOCX File, 98 KB - medinform_v14i1e73622_app2.docx]

References

1. Gibbons C, Porter I, Gonçalves-Bradley DC, et al. Routine provision of feedback from patient-reported outcome measurements to healthcare providers and patients in clinical practice. Cochrane Database Syst Rev 2021 Oct 12;10(10):CD011589. [doi: 10.1002/14651858.CD011589.pub2] [Medline: 34637526]
2. Donald EE, Whitlock K, Dansereau T, Sands DJ, Small D, Stajduhar KI. A codevelopment process to advance methods for the use of patient-reported outcome measures and patient-reported experience measures with people who are homeless and experience chronic illness. Health Expect 2022 Oct;25(5):2264-2274. [doi: 10.1111/hex.13489] [Medline: 35411709]
3. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Education 2017 Apr;17(4):137-144. [doi: 10.1093/bjaed/mkw060]
4. Casaca P, Schäfer W, Nunes AB, Sousa P. Using patient-reported outcome measures and patient-reported experience measures to elevate the quality of healthcare. Int J Qual Health Care 2023 Dec 26;35(4):mzad098. [doi: 10.1093/intqhc/mzad098] [Medline: 38113907]
5. Thompson AD, Utz RL. Online surveys: lessons learned in detecting and protecting against insincerity and bots. Qual Quant 2025 Feb;59(S1):23-39. [doi: 10.1007/s11135-024-01973-z]

XSL•FO
RenderX

6. Singh S, Sagar R. A critical look at online survey or questionnaire-based research studies during COVID-19. Asian J Psychiatr 2021 Nov;65:102850. [doi: 10.1016/j.ajp.2021.102850] [Medline: 34534919]
7. Griffin M, Martino RJ, LoSchiavo C, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022;56(4):2841-2852. [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
8. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. TQMP 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
9. Dewitt J, Capistrant B, Kohli N, et al. Addressing participant validity in a small internet health survey (the restore study): protocol and recommendations for survey response validation. JMIR Res Protoc 2018 Apr 24;7(4):e96. [doi: 10.2196/resprot.7655] [Medline: 29691203]
10. Guy AA, Murphy MJ, Zelaya DG, Kahler CW, Sun S. Data integrity in an online world: demonstration of multimodal bot screening tools and considerations for preserving data integrity in two online social and behavioral research studies with marginalized populations. Psychol Methods 2024 Sep 9. [doi: 10.1037/met0000696] [Medline: 39250292]
11. Pozzar R, Hammer MJ, Underhill-Blazey M, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020 Oct 7;22(10):e23021. [doi: 10.2196/23021] [Medline: 33026360]
12. Eysenbach G, Wyatt J. Using the Internet for surveys and health research. J Med Internet Res 2002;4(2):E13. [doi: 10.2196/jmir.4.2.e13] [Medline: 12554560]
13. Chen AT, Komi M, Bessler S, Mikles SP, Zhang Y. Integrating statistical and visual analytic methods for bot identification of health-related survey data. J Biomed Inform 2023 Aug;144:104439. [doi: 10.1016/j.jbi.2023.104439] [Medline: 37419375]
14. Xu Y, Pace S, Kim J, et al. Threats to online surveys: recognizing, detecting, and preventing survey bots. Soc Work Res 2022 Nov 14;46(4):343-350. [doi: 10.1093/swr/svac023]
15. Dadfar M, Lester B. Psychometric characteristics of Patient Health Questionnaire-2 (PHQ-2) in Iranian psychiatric open access outpatients. Austin J Psychiatry Behav Sci 2017;4(1) [FREE Full text]
16. Xu F, Johnston JM. Self-rated health and health service utilization: a systematic review. Int J Epidemiol 2015 Oct;44(suppl_1):i180-i180. [doi: 10.1093/ije/dyv096.267]
17. Tamayo-Fonseca N, Nolasco A, Quesada JA, et al. Self-rated health and hospital services use in the Spanish National Health System: a longitudinal study. BMC Health Serv Res 2015 Nov 4;15(1):492. [doi: 10.1186/s12913-015-1158-8] [Medline: 26537822]
18. Johnson MS, Adams VM, Byrne J. Addressing fraudulent responses in online surveys: insights from a web ‐ based participatory mapping study. People and Nature 2024 Feb;6(1):147-164. [doi: 10.1002/pan3.10557]
19. King-Nyberg B, Thomson E, Morris-Reade J, Borgen R, Taylor C. The bot toolbox: an accidental case study on how to eliminate bots from your online survey. J Soc Thought 2023;7 [FREE Full text]
20. Goodrich B, Fenton M, Penn J, Bovay J, Mountain T. Battling bots: experiences and strategies to mitigate fraudulent responses in online surveys. Applied Eco Perspectives Pol 2023 Jun;45(2):762-784. [doi: 10.1002/aepp.13353]
21. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to “fraudsters” in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133. [doi: 10.1111/jlme.12200] [Medline: 25846043]
22. Pinzón N, Koundinya V, Galt RE, et al. AI-powered fraud and the erosion of online survey integrity: an analysis of 31 fraud detection strategies. Front Res Metr Anal 2024;9:1432774. [doi: 10.3389/frma.2024.1432774] [Medline: 39687573]
23. Loebenberg G, Oldham M, Brown J, et al. Bot or not? Detecting and managing participant deception when conducting digital research remotely: case study of a randomized controlled trial. J Med Internet Res 2023 Sep 14;25(5):e46523. [doi: 10.2196/46523] [Medline: 37707943]
24. Lebrun B, Temtsin S, Vonasch A, Bartneck C. Detecting the corruption of online questionnaires by artificial intelligence. Front Robot AI 2023;10:1277635. [doi: 10.3389/frobt.2023.1277635] [Medline: 38371744]

Abbreviations

CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart
OHT: Ontario Health Team
PHQ-2: Patient Health Questionnaire-2
PREM: patient-reported experience measure
PROM: patient-reported outcome measure
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by A Coristine; submitted 07.Mar.2025; peer-reviewed by C Falk, N Pinzón, SP Uppu; revised version received 20.Jan.2026; accepted 21.Jan.2026; published 05.Mar.2026.

Please cite as:
Hamovitch E, McKellar K, Wodchis WP
Identifying and Analyzing Bot-Generated Responses in Online Health Care Surveys: Methodological Study
JMIR Med Inform 2026;14:e73622
URL: https://medinform.jmir.org/2026/1/e73622
doi:10.2196/73622

© Emily Hamovitch, Kaileah McKellar, Walter P Wodchis. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Chatbot Outreach in Value-Based Preventive Care: Retrospective Analysis

Jincy Chacko[1,2*], DHI; Yuane Jia[3*], PhD; Kimon Stathakos[1*], MBA; Kathleen Mazza[1*], PhD; Doran Kim[1*], MSM; Michael Diefenbach[1*], PhD; Lorena Carlo[2*], PhD

[1]Northwell Health, New Hyde Park, NY, United States

[2]Department of Health Informatics, School of Health Professions, Rutgers University, Piscataway, NJ, United States

[3]Department of Interdisciplinary Studies, School of Health Professions, Rutgers University, Newark, NJ, United States

[*]all authors contributed equally

Corresponding Author:
Jincy Chacko, DHI
Northwell Health
2000 Marcus Avenue, Suite 300
New Hyde Park, NY, 11042-1069
United States
Phone: 1 9863946
Email: chackoj.informatics@gmail.com

Abstract

Background: As health care delivery shifts toward value-based care, proactive strategies to close preventive care gaps are essential. However, patient engagement remains suboptimal due to logistical, behavioral, and socioeconomic barriers. Traditional outreach methods, such as phone calls, emails, and postal mail, have long been used, but emerging digital approaches, such as chatbot-based messaging, offer potential advantages in scalability and personalization. Their comparative effectiveness, however, remains underexplored.

Objective: This study aimed to evaluate the effectiveness of chatbot outreach compared with traditional communication methods (phone, email, mail, and multichannel) in promoting compliance with preventive screenings and wellness visits defined by the Healthcare Effectiveness Data and Information Set and the Centers for Medicare & Medicaid Services guidelines.

Methods: This retrospective study evaluated patient outreach campaigns conducted from 2021 to 2023 across an integrated health system in New York. The final analytic sample included 50,145 care gaps from 41,959 eligible participants, predominantly female (29,989/50,145, 60%), White (31,857/50,145, 64%), with mean age ranging from 49.36 to 72.81 years over the study period. All participants were residents of New York state, and 81% (40,553/50,145) maintained an active relationship with a primary care provider during the participation year. Outreach modalities included automated chatbot SMS text messages, nonautomated phone calls, and organization-led email or mail campaigns. Participant data were enriched with social vulnerability scores to account for community-level disadvantages. Exposure was defined as the outreach method (chatbot, phone, email, mail, or multichannel), with assignment based on engagement history and operational protocols. The primary outcome was care-gap closure or compliance with identified measure gaps annually. Logistic regression and chi-square analyses examined associations between outreach method, patient demographics, primary care physician relationship, social vulnerability index (SVI), and compliance.

Results: Phone outreach consistently achieved higher compliance than chatbot or multichannel outreach across most groups and years. Chatbot messages outperformed phone calls only in diabetes care in 2023 (odds ratio [OR] 1.81, 95% CI 1.48-2.21; $P<.001$). Primary care physician continuity remained a strong predictor of gap closure, especially in primary care (ORs ranged 1.36-2.61; $P<.001$). Higher SVI quartiles were associated with lower compliance in blood pressure, cancer care, and diabetes care groups; however, primary care outcomes showed higher odds of compliance in the third quartile of SVI, contradicting the typical linear-deprivation narrative. Women, Hispanic or Latino individuals, and Asian patients demonstrated higher odds of compliance in some groups and years.

Conclusions: Outreach modality is an important, modifiable factor in preventive care adherence. While phone-based outreach remains the most effective overall approach, chatbot-based strategies may have targeted applications in digitally engaged

XSL•FO
RenderX

populations such as the diabetic group. Segmented, equity-informed outreach strategies that integrate technology, patient preferences, and primary care continuity are essential to achieving high-impact, scalable outcomes in value-based care settings.

(JMIR Med Inform 2026;14:e81370) doi:10.2196/81370

KEYWORDS

chatbot outreach; digital health; health communications; health equity, preventive care gaps; value-based care

Introduction

Background and Rationale

The development of health care delivery in the United States over the past decade has increasingly centered on value-based care (VBC), which emphasizes quality, outcomes, and cost-effectiveness over service volume. Promoted by the Centers for Medicare & Medicaid Services (CMS), the value-based preventive care concept represents a focused application of the VBC framework that prioritizes proactive prevention and early disease detection, aligning financial incentives with population health goals [1,2]. This approach is particularly relevant for managing chronic diseases and addressing preventive care needs in large, diverse patient populations [3].

Despite widespread consensus on the benefits of preventive services such as annual wellness visits, cancer screenings (CXs), diabetes care (DB), and blood pressure (BP) monitoring, patient adherence remains suboptimal [4-6]. Barriers to engagement include socioeconomic challenges, low health literacy, lack of awareness, and fragmented communication between care providers [7,8]. Health systems have historically used outreach methods such as telephone calls, emails, and postal mail to notify and engage patients with gaps in preventive care. However, these approaches can be resource-intensive and vary in their reach and effectiveness, particularly among hard-to-engage populations [9].

Emerging digital outreach tools, including automated chatbots and SMS text messaging, may offer a promising alternative for improving engagement at scale. These technologies can support bidirectional communication, personalize patient interactions, and reduce administrative burdens [10]. Several public health campaigns, particularly during the COVID-19 pandemic, have demonstrated the feasibility of chatbot-based outreach in promoting vaccine uptake and disseminating timely health information. However, the use of these tools for broader preventive care engagement, particularly in chronic disease management and CX, remains underexplored [11].

Based on a previous systematic review conducted on this topic by Chacko et al [11], the current state indicates a limited body of evidence on chatbot outreach beyond immunization programs, highlighting a need for further research into their effectiveness, cost-efficiency, and equity impact. Studies reporting on outreach mechanisms for preventive care have varied in their implementation approaches and study methodologies. Therefore, drawing a generalizable conclusion is not possible. It is uncertain if benefits of chatbot outreach apply uniformly across various demographic and clinical groups. Moreover, the integration of these tools into large health systems workflows may present unique challenges that remain underexplored in the literature [11].

Objectives

To address these gaps, this study aims to evaluate the impact of emerging technology, such as chatbot-based outreach, on preventive screening uptake compared to conventional methods, such as outreach via phone calls and email communications. The authors posited that the use of chatbots in preventive care will significantly increase the uptake of screening for annual wellness visits, hypertension, diabetes, and CXs compared to traditional methods of communication. Conducted at a large integrated delivery network in New York, the study focused on patients identified as having care gaps using the Healthcare Effectiveness Data and Information Set (HEDIS) and CMS preventive screening criteria. This study builds upon recent findings that suggest automated outreach tools may offer scalable and equitable solutions for preventive health engagement [11]. By evaluating their effectiveness across multiple screening domains, which included CXs, DB, and hypertension screening, along with primary care (PC) measures such as annual wellness visits, the study contributes to the growing body of evidence informing digital outreach strategies in VBC-aligned care delivery. Findings from this evaluation may help guide future investments in digital health infrastructure and support the implementation of targeted, technology-enabled outreach programs.

Methods

Study Design

This study retrospectively evaluated data collected from VBC outreach activities conducted over a 3-year period, from 2021 to 2023. Members identified by their insurance carrier as having care gaps, attributed to the health system, were targeted for outreach. Outreach was conducted using phone calls, web-based chatbots, email messages, mailed letters, and sometimes a combination of these methods. The strategy for outreach method allocation was dependent on operational capacity and technology adoption. During year 1 (2021), due to reduced operational capacity following the 2020 pandemic year, new technology was piloted to aid in outreach. The subsequent years' attribution strategy was adjusted to maintain a balance between outreach methods with increased volume of gaps, capacity in the workforce, and efficiencies found in technology adoption. Multichannel outreach involves the sequential use of 2 or more methods within the same measurement year when initial outreach attempts fail to result in engagement. Due to the inability to attribute the outcome to any single method alone, outreach using multiple methods was grouped into 1 category. Outreach method assignments followed operational protocols

XSL•FO
RenderX

and capacities that varied each year, rather than being randomized. Therefore, this study is exploratory in nature, focused on understanding trends that may drive empirical decisions about the modality of outreach in the future.

Data Sources

Participant demographic information was collected from the health system's electronic medical record (EMR). Outreach process measures and final compliance data were collected 3 months post each study year from the respective operational teams performing outreach. These data included an indicator for participant relationship with primary care provider (PCP) within the health system and the total number of care gaps for the given year per participant. Deidentification was completed before the data were shared with researchers. The social vulnerability index (SVI) was acquired from the 2022 dataset compiled by the Centers for Disease Control and Prevention [12]. The census tract level SVI was mapped to participant zip code with the aid of cross-mapping from the Department of Housing and Urban Development, accessed in June 2024 [13].

Inclusion Criteria

Participant eligibility for the identified gap measures was a primary inclusion criterion. Clinical criteria for eligibility are industry-defined guidelines (electronic Clinical Quality Measures), maintained by CMS [14], or HEDIS specifications provided by the National Committee for Quality Assurance [15], with changes published each year publicly. These define the required age ranges, diagnosis inclusion, and completion intervals for preventive services such as wellness visits, CX, BP, and DB. The other important criterion determining inclusion in the final analysis was having at least one outreach attempt using a phone, chatbot, email, or letter during the measurement year. As the study focused on understanding the efficiency of outreach methods, the population that received no outreach was outside the scope of this study.

Exclusion Criteria

Preventive measure groups and individual measures with an insufficient sample size (n<10) were excluded due to the need for valid statistical and comparative testing. Records with unknown outcomes (ie, unverified or missing documentation of care gap closure) or missing demographic variables (age, gender, race, ethnicity, and address [needed for SVI grouping]) were excluded. For example, if there was no known address for the patient documented, it would not be possible to map an SVI and include the participant in the stratified analysis, and therefore, such participants' data were removed from all study datasets before any analysis was conducted. Process measures such as documented patient engagement steps (eg, message opened or call answered) were collected and used to define successful outreach delivery. Unsuccessful outreach deliveries (eg, in case of erroneous contact information) were excluded.

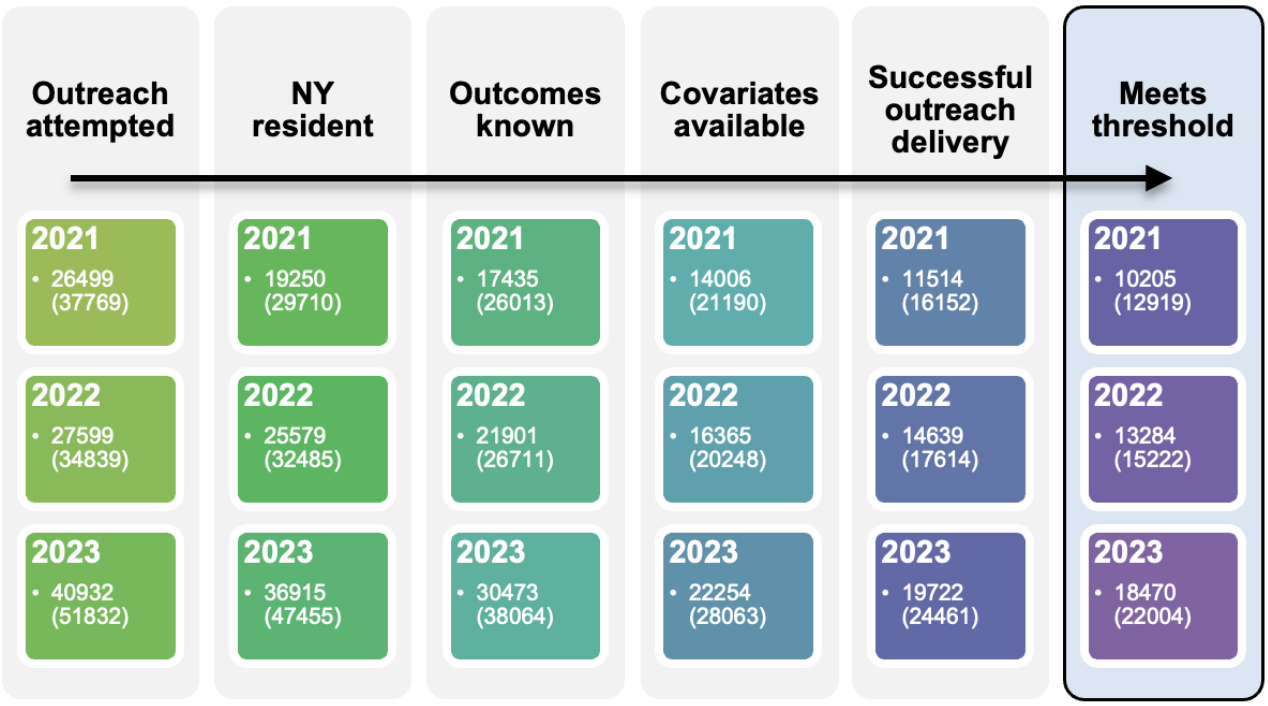
The dataset remained representative of the population served by the health system and met eligibility criteria for each measure throughout this process. Four measure groups remained eligible for analysis after the applied exclusions: PC, BP, CX, and DB. Multimedia Appendix 1 provides the distribution of participants and measure gaps within each group.

Data Management

Deidentified data were aggregated at the patient-gap-year granularity using Microsoft Power BI [16]. Individual preventive measures were grouped into care domains (eg, PC, CX, etc) based on clinical relevance and alignment with HEDIS and CMS quality measurement categories. The grouping was determined by the study team using established measure definitions to facilitate domain-level comparisons. The case of multiple SVIs mapped to a single participant due to geographic overlap in the data was resolved by assigning the SVI from the most populous county to the participant, which is a common approach in health care research. Figure 1 shows the flow of participants and gaps volume through the applied exclusions.

XSL•FO
RenderX

Figure 1. Participant and measure gap flow diagram for each year in the study, listing participant volume first and gaps (n) in parentheses. NY: New York.



Study Variables

The primary independent variable, also referred to as the outreach method or intervention, was coded as 1 of the 4 single-channel methods (phone, chatbot, email, and mail) or multichannel to indicate any combination of these methods. Multichannel methods were grouped because the sample sizes in individual groups did not meet the statistical validity requirements for comparison.

The primary outcome or dependent variable evaluated was the completion of or compliance with the identified measure, measured 3 months following each study year (eg, March 2022 for the study year of 2021). It was documented separately for each year of the study, 3 months post end of year, and represented as a binary variable, indicated by 1 for "yes" and 0 for "no." The primary outcome was collected as part of the operational process of verifying completion of or compliance with the measure gap within the reporting period. This may include a verification by patient report, EMR validation, or a confirmation by payor data sharing. The authors collected this information retrospectively from data reported to governing agencies for each year.

Demographic data, including age, gender, race, ethnicity, and the SVI, were used to study covariates that could potentially impact the outcomes of outreach. Additionally, covariates included the presence of an active relationship with a PCP (defined as having a designated PCP of record within the health system and at least 1 documented encounter with that provider or associated practice within the preceding 18 months) and the burden of unmet care (defined as the total count of gaps for each participant per year).

Among the continuous variables, age was capped at 89 years to prevent reidentification. The count of gaps variable was used to assess potential imbalances in record weighting within the grouped measure files, as the datasets were analyzed at a gap level (n=1 gap per participant) rather than at a participant level. Tables S1-S4 in Multimedia Appendix 2 demonstrate the range and mean of the count of gaps per year for each measure group.

Categorical variables included demographic factors, gender, race group, ethnicity, socioeconomic factors, SVI, and active PCP relationship. Gender was coded into male/female as collected in the EMR. Race was grouped into White, Black or African American, Asian, and other. The other race group included Pacific Islander, Alaskan, and Native Americans, as well as patient self-reported "other" race identity. Ethnicity was only collected as a bivariate containing options for Hispanic and Latino or not Hispanic or Latino. Each participant was assigned an SVI quartile (1-4), where quartile 1 was the group with the least social vulnerability, and quartile 4 was the group with the most social vulnerability.

Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics for Windows (version 29.0) [17]. The dataset was divided into measure-group files for each year. Analysis was performed for each measure group, with BP as an exception because it contained only 1 measure. Textbox 1 provides a list of subgroups (measures) within each measure group that were included in the final analysis. Descriptive statistics and contingency tables verified adequate cell counts (n≥10) for bivariate analyses per measure group, and any variables not meeting the threshold were excluded from final analysis for the respective measure group [18,19]. Notably, SVI quartile 2 often did not meet the valid n threshold, as was the case when the outreach method was email only. These were excluded from statistical analysis. This conservative approach ensured the robustness of the reported associations while maintaining data

XSL•FO
RenderX

integrity across all measure groups. Effect sizes are reported as odds ratios (ORs) with 95% CI and *P* values. Categorical variables were converted into dummy variables using reference coding (eg, phone outreach, male, White, non-Hispanic/Latino, SVI quartile 1, and no active PCP as baseline categories) before logistic regression analysis.

Textbox 1. Measure groups and subgroups 2021-2023.

Primary care

- Annual wellness visit
- Pediatric well-care visit

Blood pressure (single measure only)

Cancer screening

- Breast cancer screening
- Cervical cancer screening
- Colorectal cancer screening

Diabetes care

- Hemoglobin A1c control <8.0% in diabetes
- Hemoglobin A1c poor control >9.0% in diabetes
- Hemoglobin A1c testing in diabetes
- Nephropathy assessment
- Retinal eye examination
- Kidney health evaluation for participants with diabetes

Collinearity was assessed between included variables, which led to the exclusion of the insurance group from the final analysis due to its high correlation with age in this dataset. Chi-square tests identified preliminary predictors among the categorical variables, after which a 2-step backward stepwise logistic regression was performed. Step 1 entered all demographic variables and PCP relationship. Step 2 introduced outreach modality to quantify its independent association with the final compliance outcome, that is, gap closure after controlling significant covariates in the previous step. A chi-square test was also conducted between the PCP relationship variable and outreach methods to examine any existing associations before both were included in the regression model. Statistical significance was set at α=.05.

Ethical Considerations

The Northwell Institutional Review Board reviewed this study protocol and approved it on August 15, 2024 (HSRD24-0151), with a waiver of consent requirements in accordance with research involving deidentified secondary data. No patient contact or intervention occurred as part of the study, and all data were fully anonymized prior to analysis. To minimize analytic bias, eligibility and exclusion criteria were applied consistently using standardized definitions from HEDIS and CMS electronic Clinical Quality Measures.

Results

Participant Characteristics and Descriptive Statistics

After applying exclusions, the analytic dataset contained 12,919, 15,222, and 22,004 care gaps for 2021, 2022, and 2023, respectively, bringing the overall volume of gaps to 50,145. Gap volumes increased by approximately 18% (15,222/12,919) in 2022 and 45% (22,004/15,222) in 2023, with PC accounting for 47% (23,481/50,145) of gaps, CX 31% (15,335/50,145), BP 12% (5891/50,145), and DB 11% (5438/50,145). A total of 81% (40,553/50,145) of gaps came from participants who had an active relationship with the PCP during the participating year. Women predominated in all groups except diabetes in 2021-2022, and the cancer cohort showed an expected skew due to female-targeted measures. Among the 50,145 gaps, participants mostly belonged to the White racial group at 64% (31,857), followed by other 18% (8812), Black at 12% (6230), Asian at 6% (3246), and Hispanic/Latino at 16% (8199). The mean age for the PC group declined from 72.8 (SD 11.6) years in 2021 to 60.2 (SD 27.1) years and 49.4 (SD 30.5) years in 2022 and 2023, respectively, resulting from the gradual inclusion of the pediatric subgroup. All other measure groups had a mean age range from 56.4 to 66.3 years. The average gaps per participant were below 2 for all measures except DB, which had 2.9 gaps per participant in 2021, 2 in 2022, and 2.2 in 2023. See Multimedia Appendix 2 (Tables S1-S4) for detailed distribution for all 3 years.

Outreach included phone, chat, email, and multichannel approaches. Email was only valid for PC in 2021 and DB in 2023. Multichannel cohort met criteria in PC and CX across all years and DB and BP in 2023. Outreach shifted from phone-dominated in 2021 to greater chatbot and multichannel use by 2023. Multimedia Appendix 2 displays this trend with detailed values for each group.

XSL•FO
RenderX

Compliance Trends and Bivariate Associations

Multimedia Appendix 3 indicates the compliance rate for all 3 years across the measure groups. All measure groups indicate the highest compliance rate in 2021 and a subsequent decrease in 2022. While other measure groups have increased compliance in 2023, possibly indicating a postpandemic stabilization, compliance in the BP group continues to decrease this year. This decrease, however, is relative, as BP continues to be the group with the highest compliance rate each year (1042/1486, 70% to 1440/2550, 56%), while CX is the group with the lowest compliance rate (1698/4264, 40% to 1600/7656, 21%).

Across all measure groups, the outreach method showed a significant association with compliance. The PC group demonstrated the strongest relationship in 2021 with a large effect size (χ^2_3=1767.8; *P*<.001; V=0.57). Though this association persisted for all 3 years, its magnitude decreased, showing a moderate effect size in 2022 (χ^2_2 =220.3; *P*<.001; V=0.16) and 2023 (χ^2_2=409.5; *P*<.001; V=0.21). A significant but weak association was observed for the DB group in 2021 (χ^2_1=10.1; *P*=.001; ϕ=0.08) and a moderate effect size emerged in 2023 (χ^2_3=41.3; *P*<.001; V=0.14). The CX group showed a significant association in 2022 (χ^2_2=14.1; *P*<.001; V=0.10) and 2023 (χ^2_2=70.2; *P*<.001; V=0.10), though effect sizes remained weak throughout. BP had significant but weak associations between intervention and outcome for all 3 years. All significant chi-square test results are listed in Table 1, along with the analytic sample size. A complete report on all variables is available in Multimedia Appendix 4.

XSL•FO
RenderX

Table 1. Significant associations between covariates and outcome.

Measure and variable	Year	n	Chi-square (*df*)	*P* value	Effect size (ϕ or V)	Interpretation
Primary care						
Active PCP[a] relationship	2021	5415	270.8 (1)	<.001	ϕ=0.22	Moderate
Active PCP relationship	2022	8445	14.1 (1)	<.001	ϕ=0.04	Weak
Active PCP relationship	2023	9544	265.7 (1)	<.001	ϕ=0.17	Weak
Ethnicity	2022	8445	41.4 (1)	<.001	ϕ=0.07	Weak
Ethnicity	2023	9544	46.8 (1)	<.001	ϕ=0.07	Weak
Gender	2021	5415	8.7 (1)	.003	ϕ=0.04	Weak
Gender	2022	8445	8.7 (1)	.003	ϕ=0.03	Weak
Gender	2023	9544	5.4 (1)	.02	ϕ=0.02	Weak
Intervention	2021	5415	1767.8 (3)	<.001	V=0.57	Large
Intervention	2022	8445	220.3 (2)	<.001	V=0.16	Moderate
Intervention	2023	9544	409.5 (2)	<.001	V=0.21	Moderate
Race group	2021	5415	8.8 (3)	.03	V=0.04	Weak
Race group	2022	8445	54.9 (3)	<.001	V=0.08	Weak
Race group	2023	9544	20.5 (3)	<.001	V=0.05	Weak
SVI[b] quartile	2021	5415	25.2 (2)	<.001	V=0.07	Weak
SVI quartile	2022	8445	10.4 (2)	.006	V=0.04	Weak
D iabetes care						
Active PCP relationship	2021	1737	29.9 (1)	<.001	ϕ=0.13	Weak
Active PCP relationship	2022	1470	7.2 (1)	.007	ϕ=0.07	Weak
Gender	2022	1470	11.1 (1)	<.001	ϕ=0.09	Weak
Intervention	2021	1737	10.1 (1)	.001	ϕ=0.08	Weak
Intervention	2023	2216	41.3 (3)	<.001	V=0.14	Moderate
SVI quartile	2023	2216	18.6 (2)	<.001	V=0.09	Weak
Cancer screening						
Active PCP relationship	2021	4252	8.2 (1)	.004	ϕ=0.03	Weak
Active PCP relationship	2023	7571	4.4 (1)	.04	ϕ=0.02	Weak
Ethnicity	2022	3407	15.4 (1)	<.001	ϕ=0.07	Weak
Ethnicity	2023	7571	6.1 (1)	.01	ϕ=0.03	Weak
Gender	2021	4252	35.0 (1)	<.001	ϕ=0.09	Weak
Gender	2022	3407	5.1 (1)	.02	ϕ=0.04	Weak
Intervention	2022	3407	14.1 (2)	<.001	V=0.10	Weak
Intervention	2023	7571	70.2 (2)	<.001	V=0.10	Weak
Race group	2022	3407	22.7 (3)	<.001	V<0.01	Weak
SVI quartile	2023	7571	17.0 (2)	<.001	V=0.05	Weak
Blood pressure						
Active PCP relationship	2021	1484	6.8 (1)	.009	ϕ=0.07	Weak
Active PCP relationship	2022	1852	135.7 (1)	<.001	ϕ=0.27	Moderate
Gender	2022	1852	10.4 (1)	.001	ϕ=0.08	Weak
Intervention	2021	1484	17.0 (1)	<.001	ϕ=0.11	Weak
Intervention	2022	1852	41.0 (1)	<.001	ϕ=0.15	Weak

XSL•FO
RenderX

Measure and variable	Year	n	Chi-square (*df*)	*P* value	Effect size (ϕ or V)	Interpretation
Intervention	2023	2542	58.1 (2)	<.001	V=0.03	Weak
SVI quartile	2021	1484	14.2 (2)	<.001	V=0.10	Weak
SVI quartile	2022	1852	21.7 (3)	<.001	V=0.11	Moderate
SVI quartile	2023	2542	34.3 (2)	<.001	V=0.12	Weak

[a]PCP: primary care provider.

[b]SVI: social vulnerability index.

The PCP relationship was significantly associated with preventive care gap closure across several measure groups. In the PC group, a moderate association was observed in 2021 (χ^2_1=270.8; *P*<.001; ϕ=0.22), with significance maintained through 2023 (*P*<.001; ϕ=0.17). In the BP group, associations were significant in 2021 (χ^2_1=6.8; *P*=.009; ϕ=0.07) and stronger in 2022 (χ^2_1=135.7; *P*<.001; ϕ=0.27) but not sustained in 2023. The DB group followed a similar pattern, showing small but significant associations in 2021 (χ^2_1=29.9; *P*<.001; ϕ=0.13) and 2022 (χ^2_1=7.2; *P*=.007; ϕ=0.07), which was not sustained in 2023. The CX group showed inconsistent results, with weak associations in 2021 (χ^2_1=8.2; *P*=.004; ϕ=0.03) and 2023 (χ^2_1=4.4; *P*=.04; ϕ=0.02).

The association between SVI and compliance showed dynamic patterns across years. SVI demonstrated a consistent association with compliance in the BP group for all 3 years, peaking in 2022 with the strongest effect size (χ^2_3=21.7; *P*<.001; V=0.11). In contrast, the CX group (χ^2_2=17.0; *P*<.001; V=0.05) and the DB group (χ^2_2=18.6; *P*<.001; V=0.09) showed this association in 2023 only, both with small effect sizes. The same association weakened over time in the PC group, with significant, weak associations observed in 2021 (χ^2_2=25.2; *P*<.001; V=0.07) and 2022 (χ^2_2=10.4; *P*=.006; V=0.04), and no significant association detected in 2023.

Race was significantly associated with compliance in the PC group across all years, but effect sizes were negligible (Cramer V ranged from 0.04 to 0.08). A similar near-zero association was observed for the CX group (V=0.004) in 2022 only. Ethnicity demonstrated small but consistent associations with PC compliance in 2022 (χ^2_1=41.4; *P*<.001; ϕ=0.07) and 2023 (χ^2_1=46.8; *P*<.001; ϕ=0.07). In the CX group, ethnicity was associated in 2022 (χ^2_1=15.4; *P*<.001; ϕ=0.07) with a small effect, and in 2023 with a weaker effect (χ^2_1=6.1; *P*=.01; ϕ=0.03). No significant associations with race or ethnicity were observed in the DB or BP groups.

A significant but small and diminishing association between gender and compliance was noted in the PC group across all 3 years, with effect sizes ranging from 0.02 to 0.04. For the BP and DB groups in 2022 only, gender was significantly associated with compliance with weak effect sizes. Gender associations in the CX group are not reported because most measures in this group specifically targeted women.

Regression Analysis of Compliance Predictors

After adjusting for other significant predictors, the outreach method remained significantly associated with compliance. Compared with phone call outreach, chatbot outreach consistently reduced the odds of gap closure across PC, BP, and CX measures, with a 30%-65% decrease in likelihood, as shown in Figures 2-4. An exception was observed in the DB group in 2023, where chatbot outreach outperformed phone calls (OR 1.81, 95% CI 1.48-2.21; *P*<.001). Multichannel campaigns were generally less effective than phone calls and chatbot outreach, except in the CX group in 2021, where multichannel outreach increased the odds of gap closure compared with phone calls (OR 2.12, 95% CI 1.01-4.44; *P*=.047). This was likely a reflection of the relatively smaller sample size that year, whereas subsequent years had greater representation of multichannel outreach, yielding narrower CIs.

XSL•FO
RenderX

Figure 2. Odds ratio and 95% CI by significant predictors and measure group in 2021. BP: blood pressure; CX: cancer screening; DB: diabetes care; OR: odds ratio; PC: primary care; PCP: primary care provider; SVI: social vulnerability index.

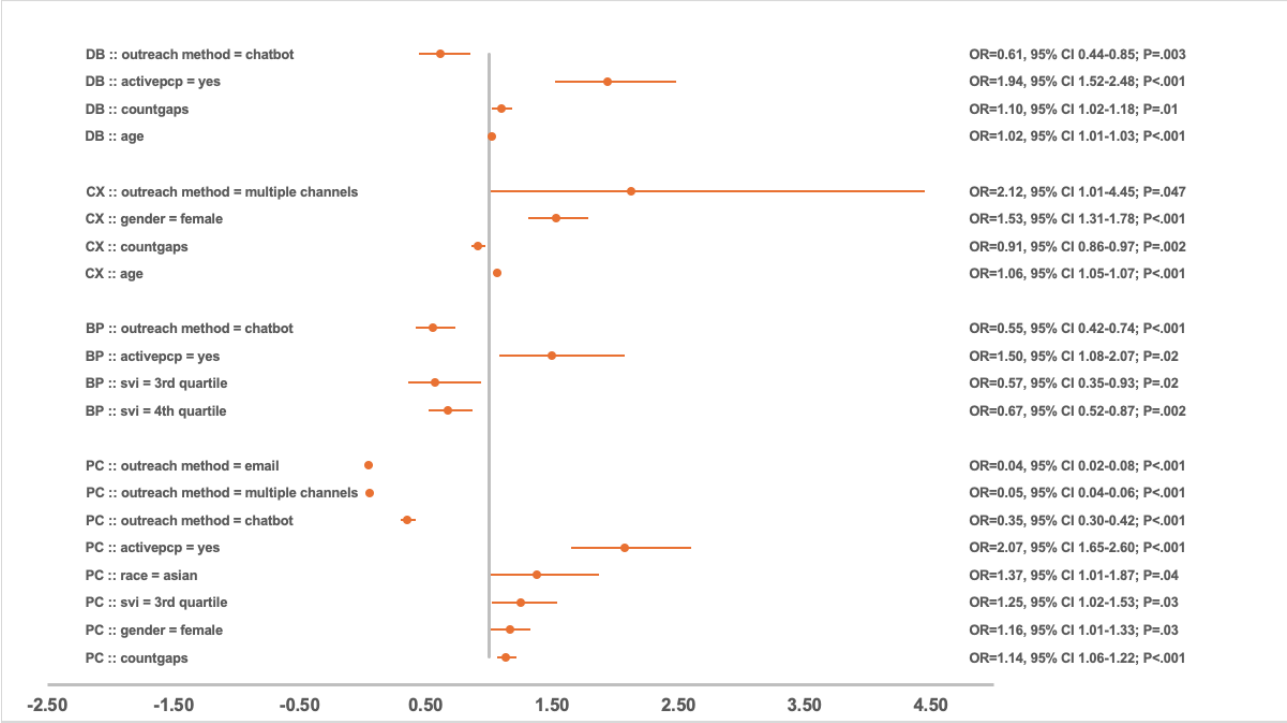
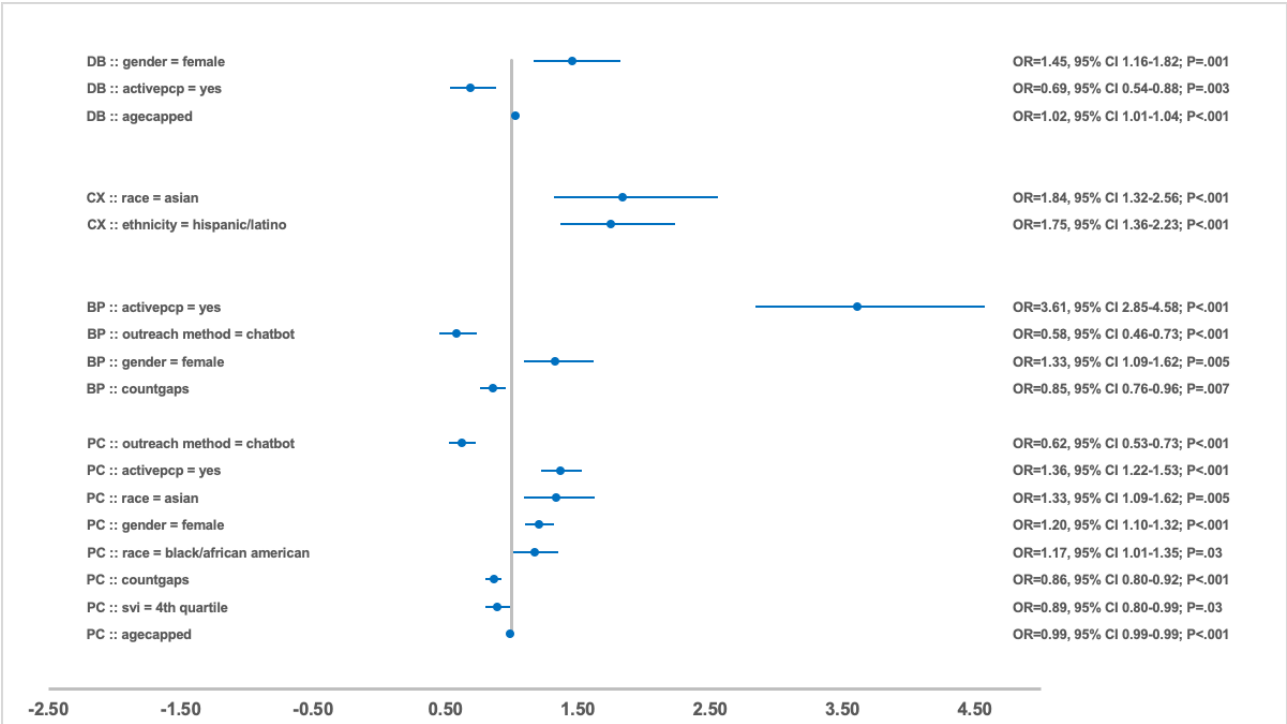


Figure 3. Odds ratio and 95% CI by significant predictors and measure group in 2022. BP: blood pressure; CX: cancer screening; DB: diabetes care; OR: odds ratio; PC: primary care; PCP: primary care provider; SVI: social vulnerability index.



Taking account of other covariates, having a PCP relationship was the strongest predictor (P<.001) for care gaps compliance in the PC measure group, with ORs 2.07 (95% CI 1.65-2.60), 1.36 (95% CI 1.22-1.53), and 2.61 (95% CI 2.31-2.94) for 2021, 2022, and 2023 respectively. BP and DB had similarly significant associations in 2021 (BP: OR 1.50, 95% CI 1.08-2.07; P=.02; DB: OR 1.94, 95% CI 1.52-2.48; P<.001) and 2022 (BP: OR 3.61, 95% CI 2.85-4.58; P<.001; DB: OR 0.69, 95% CI 0.54-0.88; P=.003), but this association was not sustained in 2023. The final model for the CX group did not surface the PCP relationship as a predictor.

The SVI quartile demonstrated a complex relationship with compliance in the PC group. Participants in the third quartile were more likely to close gaps in 2021 (OR 1.25, 95% CI 1.02-1.53; P=.03) and 2023 (OR 1.70, 95% CI 1.49-1.95; P<.001) compared with the less vulnerable first quartile. In

XSL•FO
RenderX

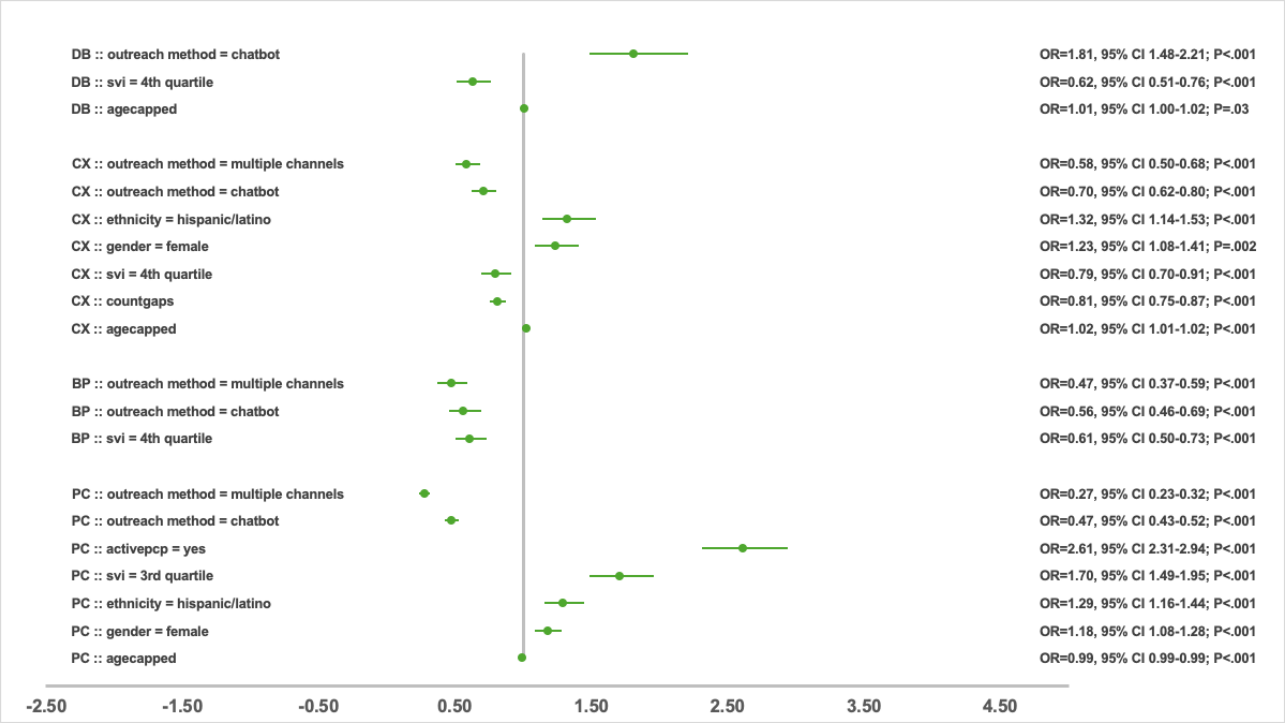
contrast, in 2022, participants in the fourth quartile were less likely to close gaps (OR 0.89, 95% CI 0.80-0.99; *P*=.03) relative to the first quartile group. For the BP group, lower odds of gap closure were observed among the third quartile (OR 0.57, 95% CI 0.35-0.93; *P*=.02) and fourth quartile (OR 0.67, 95% CI 0.52-0.87; *P*=.002) in 2021, as well as the fourth quartile in 2023 (OR 0.61, 95% CI 0.50-0.73; *P*<.001). Similar patterns were noted for the CX and DB groups in 2023, where participants in the fourth quartile had reduced odds of closing gaps (CX: OR 0.79, 95% CI 0.70-0.91; *P*<.001; DB: OR 0.62, 95% CI 0.51-0.76; *P*<.001).

Women in the PC group had higher odds of compliance than men across all years, with ORs of 1.16 (95% CI 1.01-1.33; *P*=.03) in 2021, 1.20 (95% CI 1.10-1.32; *P*<.001) in 2022, and 1.18 (95% CI 1.08-1.28; *P*<.001) in 2023. Higher odds of gap closure for women were also observed in 2022 for the BP group (OR 1.33, 95% CI 1.09-1.62; *P*=.005) and DB group (OR 1.45, 95% CI 1.16-1.82; *P*=.001).

Asians were more likely to close gaps in PC group during 2021 (OR 1.37, 95% CI 1.01-1.87; *P*=.04) and 2022 (OR 1.33, 95% CI 1.09-1.62; *P*=.005) compared with White participants. They also showed higher odds of compliance in the CX group in 2022 (OR 1.84, 95% CI 1.32-2.56; *P*<.001). Race and ethnicity were not retained in the final model for BP and DB across all 3 years. However, ethnicity had a positive influence compliance in the CX group in 2022 (OR 1.75, 95% CI 1.36-2.23; *P*<.001) and 2023 (OR 1.32, 95% CI 1.14-1.53; *P*<.001), as well as in the PC group in 2023 (OR 1.29, 95% CI 1.16-1.44; *P*<.001).

Increasing age (capped at 89 years) showed a small, diminishing association with compliance. In the DB group, each additional year of age was associated with 1%-2% higher likelihood of gap closure, while in the CX group, the increased likelihood was 6% in 2021 and 2% in 2023. In contrast, in the PC group, older age was associated with a 1% lower likelihood of closing gaps in 2022 and 2023. Age was not a significant predictor in the final BP models across all 3 years. All percentage values for likelihood are reported based on the OR values in Figures 2-4.

Figure 4. Odds ratio and 95% CI by significant predictors and measure group in 2023. BP: blood pressure; CX: cancer screening; DB: diabetes care; OR: odds ratio; PC: primary care; PCP: primary care provider; SVI: social vulnerability index.



An increase in the number of gaps was associated with a higher likelihood of gap closure in the DB group only in 2021 (OR 1.10, 95% CI 1.02-1.18; *P*=.01), with no significant effects observed in later years. Unlike the DB group, BP and CX groups showed an inverse relationship, where each additional gap decreased the likelihood of compliance (BP 2022: OR 0.85, 95% CI 0.76-0.96; *P*=.007; CX 2021: OR 0.91, 95% CI 0.86-0.97; *P*=.002; CX 2023: OR 0.81, 95% CI 0.75-0.87; *P*<.001). The PC group demonstrated inconsistency, with more gaps increasing the odds of compliance in 2021 (OR 1.14, 95% CI 1.06-1.22; *P*<.001) but decreasing them in 2022 (OR 0.86, 95% CI 0.80-0.92; *P*<.001). ORs, 95% CIs, and *P* values for all variables with significant association in the final regression models are presented in Figures 2-4 across the 3 years. Each point in these figures represents the OR for the listed cohort and variable group on the left, and the line through the point represents the CI. The actual numbers for OR, CI, and *P* value are listed on the right side. Reference values for each category include male, White, non-Hispanic/Latino, SVI first quartile, no active PCP, and phone outreach. Any subgroups excluded due to low volume (<10) for the year are not included in the regression analysis.

Model discrimination, as assessed by the area under the receiver operating characteristic curve (AUC), was highest for the PC measure group in 2021 (AUC=0.82), indicating good

XSL•FO
RenderX

discriminatory ability. Other measure groups demonstrated modest discrimination, with AUCs ranging from 0.59 to 0.68 across years. In 2021, the model also achieved the highest classification accuracy for the PC group (78.6%), while the classification accuracy for other groups varied, ranging from 70.6% (BP) to 63.5% (CX) and 67% (DB).

Across all models, chi-square statistics were significant (P<.001), suggesting that the models were statistically associated with the outcomes. Nagelkerke R^2 values, reflecting the proportion of variance explained, were highest in 2021 for the PC group (R^2=0.41) and lower for other models, ranging from 0.02 to 0.13. Calibration, as assessed by the Hosmer-Lemeshow goodness-of-fit test, showed acceptable fit for some models (eg, CX in 2022 and 2023: P=.85 and P=.98; DB in 2022 and 2023: P=.34), but poorer fit in others (eg, PC in 2023: P<.001), indicating room for improvement in predicted versus observed outcomes alignment. Complete model diagnostics are provided in Table 2.

Table 2. Final regression model metrics for each measure group, 2021-2023.

Year and measure group	Chi-square (*df*)	*P* value	Nagelkerke R^2	*P* value (Hosmer–Lemeshow)	Overall classification accuracy, %	AUC[a]
2021						
PC[b]	1942.15 (9)	<.001	0.41	.02	78.6	0.82
BP[c]	37.44 (5)	<.001	0.04	.10	70.6	0.59
CX[d]	244.08 (6)	<.001	0.08	.002	63.5	0.65
DB[e]	71.39 (4)	<.001	0.06	<.001	67.0	0.60
2022						
PC	408.45 (10)	<.001	0.06	<.001	59.3	0.63
BP	180.73 (4)	<.001	0.13	.21	67.7	0.68
CX	48.79 (8)	<.001	0.02	.85	83.2	0.59
DB	37.00 (4)	<.001	0.04	.34	69.3	0.60
2023						
PC	882.59 (9)	<.001	0.12	<.001	60.7	0.67
BP	90.76 (6)	<.001	0.05	.09	58.1	0.61
CX	145.14 (7)	<.001	0.03	.98	78.9	0.60
DB	68.64 (6)	<.001	0.04	.34	55.6	0.60

[a]AUC: area under the receiver operating characteristic curve.

[b]PC: primary care.

[c]BP: blood pressure.

[d]CX: cancer screening.

[e]DB: diabetes care.

Discussion

Key Findings

Outreach modality remains the key modifiable lever for influencing preventive care gap closure strategies. However, its impact is declining in the PC group across the 3 years, is marginal in the BP and CX group, and needs continued studies in the DB group to establish a trend and sustained evidence of association. In VBC, where outreach for gaps in care is conducted at scale, even smaller effect sizes are worth noting, as they highlight the need for ongoing research to further understand nuances between measure groups and outreach technology preferences for optimizing operational efficiencies.

Among the outreach methods, phone calls demonstrated a stronger positive impact on outcomes compared to chatbots. This may be attributed to the chatbot's deterministic, rule-based design, which lacks the interactivity and adaptability of more advanced artificial intelligence (AI)–powered conversational agents. These results are also in agreement with the recent findings from McDowell et al [20], who presented health communication preferences from their study to show that while the majority of participants preferred SMS text messaging communication in interactions with family and friends, their preference in communicating with health care providers was primarily by use of the phone call. Even though the chatbot design for this study was a web-based chat conversation, it was initiated through an SMS text message seeking consent for further communication. As noted by Fanaroff [21], SMS text messaging–based communication, though universally accessible, may face barriers in health communications due to privacy laws such as the HIPAA (Health Insurance Portability and Accountability Act)–related security constraints imposed during implementation, which can create barriers like requiring multiple clicks to get to the core of the message. This can erode patient trust in digital health technologies, especially as incidences of

XSL•FO
RenderX

SMS text message phishing attacks have increased in the last 5 years [22]. The 2023 diabetes group exception may point to a self-management culture within DB, suggesting potential for enhanced impact through digital engagement. It certainly points to a need for segmentation by measure groups when designing outreach strategies, as well as a need for further exploration of the underlying factors driving these emerging behavioral patterns.

Sustained PCP relationships more than doubled the odds of closing gaps in PC, reinforcing the understanding that it facilitates preventive measures and use in agreement with previous findings [23]. However, the effect reversed for diabetes in 2022, suggesting disease-specific pathways (eg, endocrinology follow-up may supplant the PCP role for glycemic monitoring). The lack of association between PCP relationships and BP, DB, and CX in the 2023 regression models suggests that outreach strategies should be customized by measure and specifically address the distinct impact of PCP relationships within each measure group.

Socioeconomic vulnerability appears to be an emerging barrier. The strengthening association and effect size for this covariate in the BP group over time suggest that an equity-focused resource allocation for hypertension is warranted. The association that emerged in the CX and DB groups in 2023 cannot be ignored. These results are consistent with evidence that shows neighborhood-level disadvantage suppresses screening adherence, with Montgomery et al [24] presenting one recent example of such evidence in a cancer cohort. However, the differential effect between third and fourth quartile participants in the PC group contradicts the usual linear-deprivation narrative. This highlights the need to study community-level resilience factors and changing attitudes toward PC since the 2020 pandemic.

Other demographic predictors had a relatively smaller impact. Race and ethnicity became significant variables for PC and CX outcomes in 2022 and 2023 with small effect sizes. Specifically, the positive direction of the OR for minority groups like Asians and Hispanics is surprising and suggests the need for further targeted research into culture-based practices that promote positive outcomes, so outreach strategies can be culturally tailored for better effectiveness. Gender effects in the PC group are consistently small but show higher odds of compliance among women, agreeing with prior findings [11]. In contrast, gender effects in the BP and DB groups are inconsistent across the 3 years, while interpretation within the CX group is limited due to 2 of its 3 measures being restricted to female participants.

The overall compliance rate difference between the various measure groups does indicate some weight for the burden of completion. For example, the BP measure can be completed with a single PCP visit during the year, whereas CX measures require multiple engagements with health care, which may include imaging appointments or laboratory work in addition to follow-up visits. Complex measures should be complemented with additional support or touchpoints for optimal outreach strategies.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, although the analysis draws from a large sample of individuals with preventive care gaps, the retrospective observational design limits the explanatory power of the findings. Additionally, the outreach method distribution was not uniform across measure groups; for instance, email-only and multichannel outreach cohorts did not consistently meet inclusion thresholds, restricting comparative analyses within certain subgroups. Because outreach assignments followed operational workflows rather than randomization, some selection bias may exist. Exclusions for incomplete or low-volume subgroups were applied to ensure valid statistical testing and consistent data quality. While these steps may modestly limit generalizability, the retained sample remained representative of the health system's attributed population, supporting the internal validity of findings. Generalizability is therefore most applicable to similar real-world VBC settings.

An additional limitation of this study was the presence of unmeasured variables that may account for differences in the efficiency of outreach methods. Factors such as participants' language preference, digital literacy, barriers to completion, comorbidities, and other psychosocial or contextual influences were not consistently captured across all outreach methods, likely contributing to the limited variance explained by several models. Although classification accuracy was acceptable in some cases (eg, 78.6% for PC in 2021), lower performance metrics, including the area under the receiver operating characteristic curve (eg, 0.59 for CX in 2022), indicate potential model under specification. Future research should incorporate more granular clinical, behavioral, and sociodemographic variables to enhance understanding and optimization of outreach effectiveness.

Recommendations for Practice

Even though this is not a randomized trial, the findings from this study apply to a large representative population, therefore several practical recommendations can be made to enhance the design and implementation of outreach strategies for value-based cohorts. Phone calls consistently outperformed chatbots and multichannel campaigns in driving preventive care gap closure across most measure groups and years. Health systems should continue to prioritize phone-based outreach, particularly for high-impact measure groups such as PC, while evaluating digital methods as supplementary rather than primary channels until they are more effective or better personalized. The study highlights meaningful variations in outreach effectiveness by measure group and socioeconomic status. To maximize effectiveness, outreach strategies should be customized according to the specific clinical area and complexity of the care gap, such as differentiating approaches for PC, chronic disease management, and more complex screening measures. For example, the stronger performance of chatbot outreach in the DB group in 2023 suggests that measure-specific digital readiness may guide outreach channel selection.

We also recommend enhancing chatbots for better engagement. Chatbots are an emerging technology that offers better

XSL•FO
RenderX

scalability than phone calls. The rule-based design of the chatbot used in this study likely limited its effectiveness. Transitioning to more advanced, AI-powered conversational agents that support interactive dialogue and user personalization could improve patient engagement and outcomes [25].

Socioeconomic vulnerability was a significant and increasing barrier to care gap closure, particularly in the BP group. Health systems should adopt equity-focused outreach strategies that prioritize high-SVI communities for additional support, such as community health worker involvement or tailored messaging. This is especially important in measure groups with heavier burdens of completion, like CX or chronic disease monitoring.

The relatively low explanatory power of the models underscores the need to expand the range of included variables. Incorporating clinical data (eg, comorbidities and medication adherence), behavioral factors (eg, health beliefs and prior screening history), and social determinants of health (eg, transportation access and housing stability) could significantly enhance the precision of predictive models used to guide outreach prioritization.

Conclusion

In conclusion, this study found that there is variability in the effectiveness of chatbots in outperforming traditional outreach methods for preventive screenings. The findings confirm that outreach modality continues to be a key, modifiable factor in closing preventive care gaps, with phone-based outreach consistently outperforming chatbots and multichannel strategies across most measure groups. However, the effectiveness of each outreach method varied notably by clinical context and patient subgroup, highlighting the need for a segmented outreach design that accounts for differences in patient preferences, health conditions, and burden of gap completion. The limited effectiveness of deterministic, rule-based chatbots points to the potential value of transitioning to more adaptive, AI-enabled engagement tools, particularly those capable of personalized communication and responsiveness to digital literacy. Socioeconomic vulnerability remained a key barrier, with differing trends between the third and fourth SVI quartiles indicating that vulnerability is not a linear predictor of behavior. Incorporating structured data on preferences, barriers, and comorbidities will improve model accuracy. Future research using prospective, randomized designs should further evaluate outreach modality, language, and communication preferences to guide inclusive, high-impact preventive care.

Acknowledgments

The authors would like to recognize the contributions of Mathew Sabu for data cleaning, deidentification, and analysis support during the early stages of this study.

The authors would also like to acknowledge the use of artificial intelligence tools, such as the built-in Microsoft text editor and Grammarly, to refine and clarify text that was primarily written by humans in the preparation of this manuscript.

Funding

This study did not receive external funding. The analysis was conducted as part of an operational quality improvement initiative at Northwell Health and as a component of the author’s doctoral dissertation at Rutgers University, with academic oversight provided by faculty advisors.

Data Availability

The datasets analyzed in this study were derived from health system operational data and accessed under applicable data use agreements that restrict public dissemination, even in deidentified form. Due to such contractual, ethical, and governance constraints, the data are not publicly available. Access may be considered upon reasonable request to the corresponding author and is subject to approval by the data-owning institution and execution of an appropriate data use agreement.

Authors' Contributions

JC led the conceptualization and project administration of the study and was responsible for data curation, formal analysis, visualization, and writing of the original draft. Conceptualization and methodology were jointly contributed by JC, YJ, KS, KM, DK, MD, and LC. Investigation and interpretation of findings were conducted by JC, YJ, KS, and LC. Data curation and resources were supported by KS and DK. Supervision and methodological oversight were provided by YJ, LC, KM, and MD. Writing, review, and editing were performed by all authors. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work in accordance with the International Committee of Medical Journal Editors and CRediT taxonomy guidelines.

Conflicts of Interest

Some authors are employed by Northwell Health, which supported this study as part of internal evaluation and quality improvement activities. Other authors are affiliated with Rutgers University as doctoral student or faculty and contributed to the study

XSL•FO
RenderX

independently as part of academic research. None of the authors reports any financial conflicts of interest or competing interests related to the content of this manuscript.

Multimedia Appendix 1
Measure group volume of participants and gaps per year of study. Each cell lists participant volume with measure gaps volume in brackets (n).
[PNG File , 53 KB - medinform_v14i1e81370_app1.png]

Multimedia Appendix 2
Table S1: Distribution table for final analysis of PC group 2021-2023. Table S2: Distribution table for final analysis of BP group 2021-2023. Table S3: Distribution table for final analysis of CX group 2021-2023. Table S4: Distribution table for final analysis of DB group 2021-2023.
[XLSX File (Microsoft Excel File), 17 KB - medinform_v14i1e81370_app2.xlsx]

Multimedia Appendix 3
Compliance rate by measure group 2021-2023.
[PNG File , 25 KB - medinform_v14i1e81370_app3.png]

Multimedia Appendix 4
Table A: All Associations between covariates and outcome.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v14i1e81370_app4.xlsx]

Multimedia Appendix 5
Eligibility criteria for each included measure for 2021-2023.
[DOCX File , 21 KB - medinform_v14i1e81370_app5.docx]

Multimedia Appendix 6
Distribution (%) of gaps by outreach method and measure group 2021-2023.
[PNG File , 29 KB - medinform_v14i1e81370_app6.png]

References

1. Value-based care. CMS. URL: https://www.cms.gov/priorities/innovation/key-concepts/value-based-care [accessed 2025-07-13]
2. Preventive health services. HealthCare.gov. 2025. URL: https://www.healthcare.gov/coverage/preventive-care-benefits/ [accessed 2025-07-13]
3. Figueroa JF, Duggan CE, Joynt Maddox KE. Value-based payment in Medicare: progress, challenges, and future directions. J Health Polit Policy Law 2025;50(6):1059-1079. [doi: 10.1215/03616878-11995200] [Medline: 40538335]
4. Seither R, Yusuf OB, Dramann D, Calhoun K, Mugerwa-Kasujja A, Knighton CL. Coverage with selected vaccines and exemption from school vaccine requirements among children in kindergarten - United States, 2022-23 school year. MMWR Morb Mortal Wkly Rep 2023;72(45):1217-1224 [FREE Full text] [doi: 10.15585/mmwr.mm7245a2] [Medline: 37943705]
5. Mykyta L, Cohen RA. Characteristics of adults aged 18-64 who did not take medication as prescribed to reduce costs: United States, 2021. NCHS Data Brief 2023(470):1-8 [FREE Full text] [Medline: 37314379]
6. Ma Z, Richardson LC. Cancer screening prevalence and associated factors among US adults. Prev Chronic Dis 2022;19:E22 [FREE Full text] [doi: 10.5888/pcd19.220063] [Medline: 35446757]
7. Murugesu L, Heijmans M, Rademakers J, Fransen MP. Challenges and solutions in communication with patients with low health literacy: perspectives of healthcare providers. PLoS One 2022;17(5):e0267782 [FREE Full text] [doi: 10.1371/journal.pone.0267782] [Medline: 35507632]
8. Coombs NC, Campbell DG, Caringi J. A qualitative study of rural healthcare providers' views of social, cultural, and programmatic barriers to healthcare access. BMC Health Serv Res 2022;22(1):438 [FREE Full text] [doi: 10.1186/s12913-022-07829-2] [Medline: 35366860]
9. Lieu TA, Elkin EP, Escobar PR, Finn L, Klein NP, Durojaiye C, et al. Effect of electronic and mail outreach from primary care physicians for COVID-19 vaccination of black and Latino older adults: a randomized clinical trial. JAMA Netw Open 2022;5(6):e2217004 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.17004] [Medline: 35713906]
10. Bickmore TW, Picard RW. Chatbots in health care: connecting patients to information. In: Health Technology and Informatics. Bethesda, MD: National Library of Medicine (US); 2023.
11. Chacko J, Mazza K, Stathakos K, Kim D, Carlo L. Evaluating technology-driven strategies for enhancing patient outreach for preventive care: a systematic review. Cureus 2025;17(2):e79467. [doi: 10.7759/cureus.79467] [Medline: 40130097]

XSL•FO
RenderX

12. Social Vulnerability Index. ATSDR. 2024. URL: https://www.atsdr.cdc.gov/place-health/php/svi/index.html [accessed 2026-01-22]
13. HUD USPS ZIP code crosswalk files. HUD USPS. URL: https://www.huduser.gov/portal/datasets/usps_crosswalk.html [accessed 2024-06-24]
14. Measures inventory. CMS.gov. 2025. URL: https://www.cms.gov/priorities/innovation/key-concepts/value-based-care [accessed 2025-07-13]
15. HEDIS measures. National Committee for Quality Assurance. 2025. URL: https://www.ncqa.org/hedis/measures/ [accessed 2025-07-13]
16. Microsoft Power BI desktop. Microsoft. 2025. URL: https://www.microsoft.com/en-us/power-platform/products/power-bi/desktop [accessed 2027-01-22]
17. IBM SPSS Statistics for Windows. IBM. Armonk, NY: IBM Corp URL: https://www.ibm.com/products/spss-statistics [accessed 2026-01-22]
18. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(2):143-149 [FREE Full text] [doi: 10.11613/bm.2013.018] [Medline: 23894860]
19. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49(12):1373-1379 [FREE Full text] [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
20. McDowell B, Dumais KM, Gary ST, de Gooijer I, Ward T. Preferences and attitudes towards digital communication and symptom reporting methods in clinical trials. Patient Prefer Adherence 2025;19:255-263 [FREE Full text] [doi: 10.2147/PPA.S474535] [Medline: 39901901]
21. Fanaroff J. A high-tech wreck: HIPAA roadblocks to texting patients. J Perinatol 2025;45(2):271-272. [doi: 10.1038/s41372-024-02131-x] [Medline: 39354212]
22. Lindsey D, Sinkey R, Travers C, Budhwani H, Richardson M, Quinney R, et al. When HIPAA hurts: legal barriers to texting may reinforce healthcare disparities and disenfranchise vulnerable patients. J Perinatol 2025;45(2):278-281. [doi: 10.1038/s41372-024-02080-5] [Medline: 39147791]
23. Ibrahim R, Sainbayar E, Pham HN, Shahid M, Saleh AA, Javed Z, et al. Social Vulnerability Index and cardiovascular disease care continuum: a scoping review. JACC Adv 2024;3(7):100858 [FREE Full text] [doi: 10.1016/j.jacadv.2024.100858] [Medline: 39130018]
24. Montgomery KB, McLeod MC, DePalo DK, Dugan MM, Zager JS, Elleson KM, et al. Impact of social determinants of health on melanoma nodal surveillance in a multi-institutional cohort. Ann Surg Oncol 2025;32(3):1453-1462. [doi: 10.1245/s10434-024-16498-w] [Medline: 39576454]
25. Bhimani M, Baker RH, Ausin MS, Meixiong G, Lasko R, Raglow-Defranco M, et al. Using a multilingual AI care agent to reduce disparities in colorectal cancer screening for higher fecal immunochemical test adoption among Spanish-speaking patients: retrospective analysis. J Med Internet Res 2025;27:e71211 [FREE Full text] [doi: 10.2196/71211] [Medline: 40561471]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
BP: blood pressure
CMS: Centers for Medicare & Medicaid Services
CX: cancer screening
DB: diabetes care
EMR: electronic medical record
HEDIS: Healthcare Effectiveness Data and Information Set
HIPAA: Health Insurance Portability and Accountability Act
OR: odds ratio
PC: primary care
PCP: primary care provider
SVI: social vulnerability index
VBC: value-based care

XSL•FO
RenderX

Edited by A Benis; submitted 01.Aug.2025; peer-reviewed by C Wang, PS Olson; comments to author 07.Oct.2025; revised version received 07.Jan.2026; accepted 08.Jan.2026; published 12.Feb.2026.

Please cite as:
Chacko J, Jia Y, Stathakos K, Mazza K, Kim D, Diefenbach M, Carlo L
Chatbot Outreach in Value-Based Preventive Care: Retrospective Analysis
JMIR Med Inform 2026;14:e81370
URL: https://medinform.jmir.org/2026/1/e81370
doi:10.2196/81370
PMID:

©Jincy Chacko, Yuane Jia, Kimon Stathakos, Kathleen Mazza, Doran Kim, Michael Diefenbach, Lorena Carlo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Practical Nomogram for Depression Risk Stratification in Older Adults With Hypertension and Diabetes: Retrospective Analysis of Data From the China Health and Retirement Longitudinal Study

Ting Peng, MD; Ying Zhang, MD; Rujia Miao*, MD; Jiangang Wang*, Prof Dr

Health Management Center, The Third Xiangya Hospital, Central South University, 138 Tongzipo Road, Yuelu District, Changsha, Hunan, China
*these authors contributed equally

Corresponding Author:
Jiangang Wang, Prof Dr
Health Management Center, The Third Xiangya Hospital, Central South University, 138 Tongzipo Road, Yuelu District, Changsha, Hunan, China

Abstract

Background: Depression affects over 40% of middle-aged and older Chinese adults living with both hypertension and diabetes, amplifying cardiovascular risk, functional decline, and mortality. Existing screening instruments—such as the 10-item Center for Epidemiologic Studies Depression Scale—focus narrowly on mood symptoms and are rarely feasible in busy primary care consultations. They also omit routine functional, cognitive, and social data that may jointly drive depressive states in cardiometabolic populations.

Objective: This study aimed to develop and validate a concise, clinically actionable nomogram that quantifies individual depression risk using readily available information in Chinese adults aged ≥45 years who have diagnosed hypertension and type 2 diabetes.

Methods: We analyzed anonymized wave 5 China Health and Retirement Longitudinal Study data collected between July 2020 and August 2020. Of 1504 eligible participants, 635 (42.2%) met the Center for Epidemiologic Studies Depression Scale cutoff score of >10 for probable depression. A total of 42 candidate predictors spanning demographics, laboratory values, comorbidities, functional status, and socioenvironmental factors were screened. Least absolute shrinkage and selection operator regression with 10-fold cross-validation identified the most parsimonious set. A multivariable logistic model was built on a 70% training set (n=1052) and evaluated on a 30% testing set (n=452). Performance was assessed using the area under the receiver operating characteristic curve (AUC), calibration plots, decision curve analysis, and Shapley additive explanations for interpretability. Multiple imputation was used to handle <20% missingness.

Results: Nine nonredundant predictors entered the final nomogram: activity of daily living score, memory impairment, number of pain sites, sleep duration, life satisfaction score, self-rated health score, social activity engagement score, retirement status, and memory test score. The model achieved excellent discrimination (training AUC=0.819; testing AUC=0.825) and calibration (mean absolute error ≤0.018). Decision curves demonstrated positive net clinical benefit across clinically relevant threshold probabilities. Shapley additive explanations analysis revealed a 3-fold increase in depression odds per 1-point increase in activity of daily living score, whereas retirement conferred substantial protection (prevalence of depression: 103/635, 16.2% in the retired group vs 269/869, 31.0% in the nonretired group; *P*<.001), mediated by greater social participation.

Conclusions: The 9-item nomogram enables <3-minute depression risk stratification in resource-limited primary care settings for adults with hypertension and diabetes. Functional decline, affective-cognitive burden, and socioeconomic disengagement constitute the dominant causal pathway. Prospective trials should examine whether interventions targeting postretirement social engagement and functional rehabilitation can reduce incident depression in this high-risk population.

(JMIR Med Inform 2026;14:e81529) doi:10.2196/81529

KEYWORDS

depression risk prediction; hypertension; diabetes; activities of daily living; nomogram

XSL•FO
RenderX

Introduction

Depression imposes an additional burden on people living with chronic conditions, especially among middle-aged and older adults who are trying to manage both high blood pressure and diabetes [1]. These physical illnesses often appear together with depressive symptoms, making treatment harder and speeding up the loss of daily function [2]. Studies have shown that, in this group, the rate of depression can reach more than 40% in some samples, and it is linked to poorer treatment adherence, higher heart-related risks, and greater chances of early death [3,4]. Even though depression is clearly important, it is still missed in many primary care visits [5]. The screening tools we have are reliable on paper, but they are hard to fit into busy clinics because of limited time and other practical barriers. The 10-item version of the Center for Epidemiologic Studies Depression Scale (CESD-10), widely used in Chinese-speaking populations, is a good example: when used alone, it cannot draw on the wider set of clinical, behavioral, and social clues that would help physicians spot depression earlier [6,7].

Most earlier models for predicting depression were built for people receiving psychiatric care or used general risk factors, so they do not work well for people who also have long-term illnesses [8,9]. A key weakness is that these models rarely include measures of daily function, such as the ability to bathe, dress, cook, and handle money, together with behavioral and social factors [10]. This omission matters because poor physical function can be both an early sign of depression and a result of uncontrolled blood pressure or blood sugar, making it hard to know which came first [11]. For these reasons, we set out to create a simple, practical tool to estimate depression risk in middle-aged and older adults with hypertension and diabetes. Using data from the China Health and Retirement Longitudinal Study (CHARLS), we examined 42 possible predictors covering medical, psychological, and social areas. Using a logistic regression model improved through machine learning and checked through several validation steps, we carried out 3 tasks: choosing the smallest set of predictors that still separates high-risk from low-risk individuals; turning these predictors into a clear, easy-to-use score that can be applied in busy, low-resource clinics; and showing, with numbers, how much heart- and metabolism-related functional problems act as a link between illness and depression. Rather than focusing only on symptoms, our approach places depression risk within the patient's overall health picture. The resulting model can guide clinicians and also point to potentially modifiable risk factors.

Methods

Ethical Considerations

This study used anonymized data from the fifth wave of the CHARLS. The project was approved by the Peking University Biomedical Ethics Committee in June 2008 (approval IRB00001052–11015). All personal identifiers were removed before our analysis. Patients and the public were not involved in the design, conduct, reporting, or dissemination of this research. All participants provided written informed consent, and the study followed the ethical principles set out in the Declaration of Helsinki. No financial or material compensation was offered to study participants.

Study Population

We carried out a retrospective study using data from the fifth wave of the CHARLS collected between July 2020 and August 2020. Participants had to be aged at least 45 years and have both hypertension and diabetes. We excluded anyone who was missing key details such as CESD-10 score, diagnosis of hypertension or diabetes, or age. After these exclusions, 1504 people were left for the analysis. Depression status was based on the CESD-10, a brief questionnaire that mixes positively and negatively worded statements. Each item is scored from 0 to 3, yielding total scores from 0 to 30; higher scores mean more severe depressive symptoms [12,13]. In Chinese older adults, a score above 10 is generally taken to indicate possible depression, and this cutoff has been shown to work well in previous studies and in latent profile analyses [13,14]. Using this cutoff, of the 1504 participants, we labeled 635 (42.2%) as the depression group and 869 (57.8%) as the nondepression group.

Candidate Predictor Variables

Guided by earlier studies and clinical practice, we chose 42 factors that might relate to depression risk [15-17]. These factors cover basic personal details, overall health, daily habits, blood test results, and how well people manage daily tasks. The full list is as follows: sex, age, marital status, smoking status, drinking status, cancer, lung disease, heart disease, stroke, mental illness, arthritis, high blood fat, liver disease, kidney disease, stomach disease, asthma, memory problems, hip fracture, number of painful body sites, exercise level, sleep hours, social activity engagement, whether the patient was retired, medical insurance, pension insurance, years of schooling, household size, number of living children, household income, recent physician visits, memory test score, activity of daily living (ADL) score, instrumental ADL (IADL) score, self-rated health, and satisfaction with life. Regarding "social activity engagement," it was categorized into the following eight types: (1) interacted with friends (visiting friends or socializing with friends); (2) played mahjong, played chess, played cards, or went to community clubs; (3) provided help to family members, friends, or neighbors who do not live with the participant and who did not pay them for the help; (4) went to a sports, social, or other type of club; (5) took part in community-related organization activities; (6) did voluntary or charity work; (7) attended an educational or training course; and (8) other.

Participants completed the questionnaire based on their actual situation: participation in 1 type of the above activities was recorded as "1," participation in 2 types was recorded as "2," and a higher count indicated a higher level of social activity engagement of the participant. For "life satisfaction," the CHARLS survey asked respondents about their overall satisfaction with their lives, with 5 response options: "extremely satisfied," "very satisfied," "relatively satisfied," "not very satisfied," and "not satisfied at all." For the convenience of statistical analysis, we assigned numerical values to these levels: "1" represented "not satisfied at all," "5" represented "extremely

XSL•FO
RenderX

satisfied," and so on, with each level corresponding to a specific number.

Model Construction Methods

We used least absolute shrinkage and selection operator (LASSO) regression to pick out the most useful risk factors while keeping the list short. Ten-fold cross-validation helped us find the best penalty value (λ). We then split the data into a training set (1052 people) and a testing set (452 people) in a 7:3 ratio. The training set was used to find the key factors linked to depression. These factors were entered into a standard logistic regression, and we drew a simple chart (nomogram) to show how each one affected the risk of depression. To keep the model easy to use, we removed predictors with little impact, reran the logistic regression with the remaining ones, and built the final nomogram.

Model Evaluation Methods

To check how well the model performed, we used receiver operating characteristic curves and the area under these curves (AUC) to measure overall ability to separate people with and without depression, calibration plots to check whether the predicted chances matched what really happened, and decision curve analysis to weigh the benefits of true positive predictions against the harms of false ones, giving a clear picture of both accuracy and clinical value.

Model Interpretation Methods

To make the model easier to understand, we used Shapley additive explanations (SHAP) values to show how much each factor mattered. We drew 4 simple charts (an overall importance plot, a swarm plot, a waterfall plot, and a force plot) that displayed how every factor pushed the risk of depression up or down. These pictures help readers see which factors carry the most weight and how they work together in the model.

Data Processing Methods

All statistical analyses and figure generation were performed using R (R Foundation for Statistical Computing). When less than 20% of the data were missing, we used the *mice* package to run 5 rounds of imputation and kept the set that best matched the overall pattern. We report counts and percentages for categorical variables and means with SDs for continuous variables. We compared groups using 2-tailed *t* tests, chi-square tests, or nonparametric tests as needed; the latter were chosen when the data did not follow the rules for parametric tests. A *P* value <.05 was considered statistically significant.

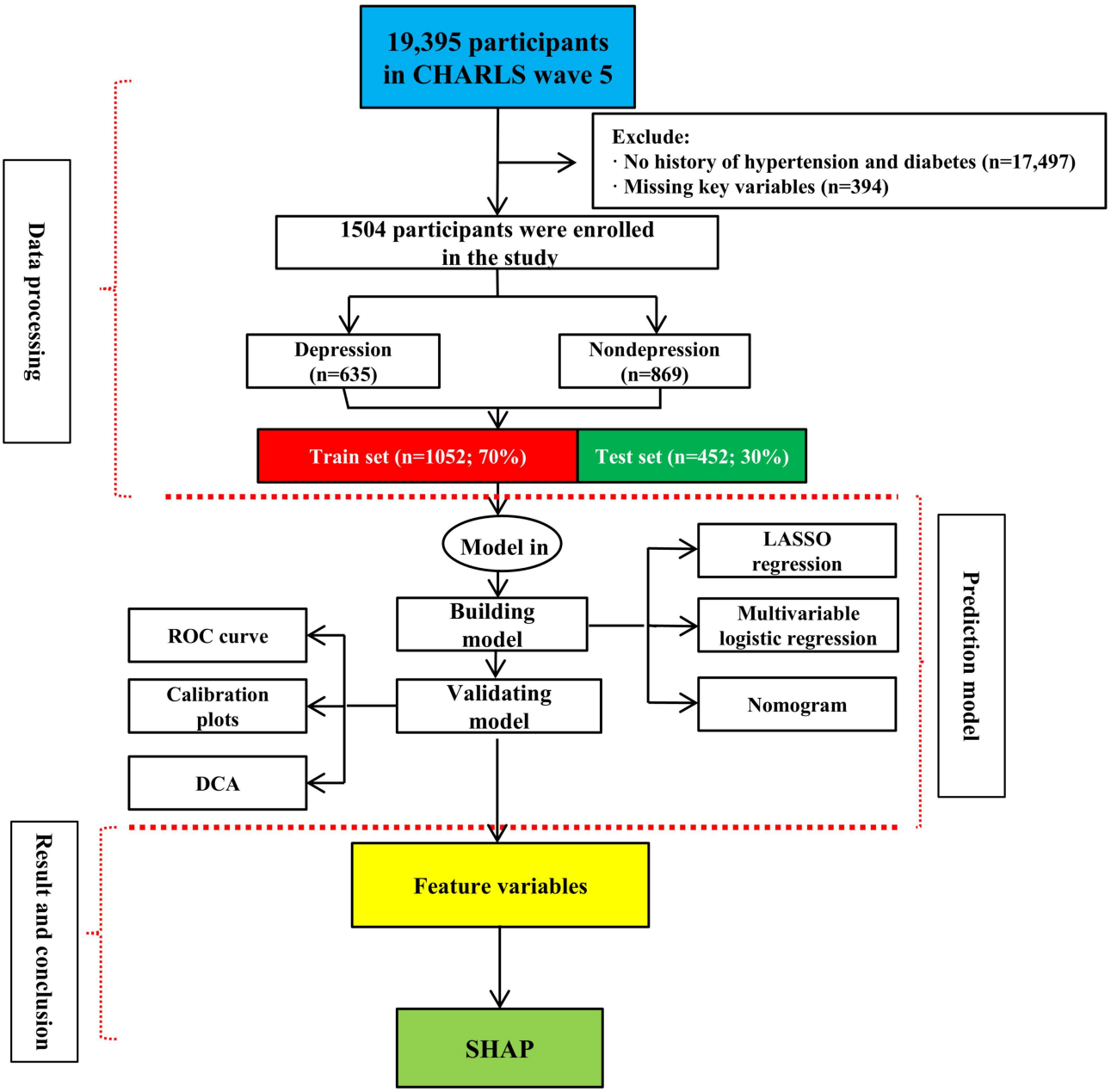
Results

Flowchart

The study flowchart is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Study flowchart. Analysis of China Health and Retirement Longitudinal Study (CHARLS) wave 5 participants with hypertension and diabetes comorbidity (aged ≥45 years). DCA: decision curve analysis; LASSO: least absolute shrinkage and selection operator; ROC: receiver operating characteristic; SHAP: Shapley additive explanations.



Baseline Characteristics

We included 1504 adults with an average age of 65.99 (SD 8.28) years and 64.93 (SD 8.72) years for those with and without depression, respectively; 678 (45.1%) were men, and 826 (54.9%) were women. Of these 1504 participants, 635 (42.2%) were classed as having possible depression, and 869 (57.8%) were classed as nondepressed. Table 1 shows that the 2 groups differed on many key points. Those with possible depression were more often women, widowed or separated, and less active and reported more illnesses, such as lung disease, heart disease, stroke, arthritis, mental illness, liver disease, kidney disease, stomach disease, asthma, hip fracture, and memory problems. They also had more pain sites; poorer sleep; lower income; fewer years of schooling; smaller households; more physician visits; fewer social activities; and lower scores on memory, daily activities, self-rated health, and life satisfaction (P<.05 in all cases).

XSL•FO
RenderX

Table . Baseline characteristics of study participants by depression status. *P* values were calculated using *t* tests, chi-square tests, or nonparametric tests as appropriate. A *P* value of <.05 was considered statistically significant.

Characteristic	Nondepression (n=869)	Depression (n=635)	*P* value
Medical insurance, n (%)	834 (96.0)	608 (95.7)	.93
Pension insurance, n (%)	765 (88.0)	543 (85.5)	.17
Exercise, n (%)	793 (91.3)	548 (86.3)	.003
Retired, n (%)	269 (31.0)	103 (16.2)	<.001
Sex (male), n (%)	463 (53.3)	215 (33.9)	<.001
Married, n (%)	746 (85.8)	504 (79.4)	.001
Hip fracture, n (%)	9 (1.0)	13 (2.0)	.16
Cancer, n (%)	30 (3.5)	24 (3.8)	.84
Lung disease, n (%)	120 (13.8)	156 (24.6)	<.001
Heart disease, n (%)	326 (37.5)	339 (53.4)	<.001
Stroke, n (%)	98 (11.3)	136 (21.4)	<.001
Psychiatric disease, n (%)	25 (2.9)	58 (9.1)	<.001
Arthritis, n (%)	324 (37.3)	389 (61.3)	<.001
Dyslipidemia, n (%)	572 (65.8)	457 (72.0)	.01
Liver disease, n (%)	90 (10.4)	107 (16.9)	<.001
Kidney disease, n (%)	144 (16.6)	174 (27.4)	<.001
Stomach disease, n (%)	261 (30.0)	265 (41.7)	<.001
Asthma, n (%)	55 (6.3)	85 (13.4)	<.001
Memory problems, n (%)	47 (5.4)	97 (15.3)	<.001
Vigorous physical activity, n (%)	237 (27.3)	193 (30.4)	.21
Moderate physical activity, n (%)	474 (54.5)	319 (50.2)	.11
Light physical activity, n (%)	727 (83.7)	482 (75.9)	<.001
Drinking, n (%)	318 (36.6)	142 (22.4)	<.001
Smoking, n (%)	203 (23.4)	95 (15.0)	<.001
Hospitalization in the previous year, n (%)	241 (27.7)	266 (41.9)	<.001
Recent physician visits, n (%)	227 (26.1)	220 (34.6)	<.001
Total income, mean (SD)	¥54,055.20 (US $7858.73; ¥79,512.32 [US $10,687.50])	¥32,470.65 (US $4720.71; ¥88,119.77 [US $12,811.20])	<.001
Family size (number of individuals), mean (SD)	2.88 (1.56)	2.58 (1.40)	<.001
Number of healthy children, mean (SD)	2.37 (1.25)	2.69 (1.21)	<.001
Financial support from children to parents, mean (SD)	¥6984.23 (US $1015.40; ¥15,871.02 [US $2307.39])	¥5448.44 (US $792.12; ¥8541.60 [US $1241.81])	.03
Financial support from parents to children, mean (SD)	¥9176.07 (US $1334.05; ¥49,285.04 [US $7165.26])	¥4270.93 (US $620.93; ¥21,830.53 [US $3173.81])	.02
Self-rated health score (1-5), mean (SD)	2.82 (0.90)	2.18 (0.89)	<.001
Life satisfaction score (1-5), mean (SD)	3.41 (0.66)	2.87 (0.87)	<.001
Number of pain sites, mean (SD)	2.11 (3.01)	4.53 (4.35)	<.001
Sleep duration (h), mean (SD)	6.21 (1.68)	5.27 (2.06)	<.001

XSL•FO
RenderX

Characteristic	Nondepression (n=869)	Depression (n=635)	*P* value
Social activity engagement (number of activities participated in), mean (SD)	0.94 (1.14)	0.65 (0.90)	<.001
Age (y), mean (SD)	64.93 (8.72)	65.99 (8.28)	.02
Educational level[a], mean (SD)	2.26 (1.11)	1.85 (1.02)	<.001
Total MET[b] (min), mean (SD)	4627.13 (4936.56)	4832.40 (5150.61)	.43
Memory score (0-11), mean (SD)	4.51 (1.74)	3.61 (1.84)	<.001
ADL[c] score (0-6), mean (SD)	0.34 (0.85)	1.20 (1.60)	<.001
IADL[d] score (0-5), mean (SD)	0.32 (0.84)	1.18 (1.50)	<.001

[a]Educational level was scored as follows: 1=illiterate; 2=below junior high school; 3=junior high school and above, below university; 4=university and above.

[b]MET: metabolic equivalent of task.

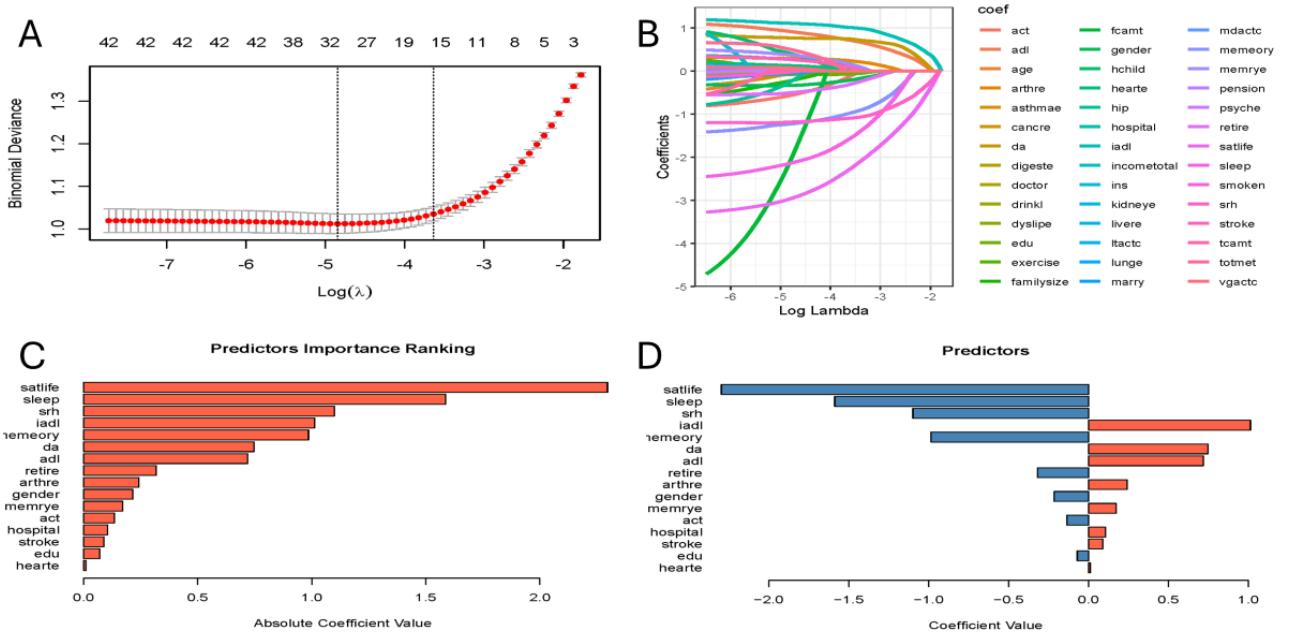
[c]ADL: activity of daily living.

[d]IADL: instrumental ADL.

Results of Model Construction

As the regularization strength (λ) increases, the model error initially decreases and then increases. We chose the simplest model within 1 SE of the minimum error. This left 16 variables whose coefficients stayed above 0 (Figure 2A). Each line in the plot shows one variable's coefficient; as λ grows (moving right), many lines drop to 0 and are removed. In the end, 16 factors remained: life satisfaction score, sleep hours, self-rated health score, IADL score, memory score, pain sites, ADL score, retirement status, arthritis, sex, memory problems, social activity engagement type, recent physician visits, stroke, years of schooling, and heart disease, listed from most to least important (Figure 2C). Figure 2D shows the direction of each factor: red bars mean higher risk, and blue bars mean lower risk.

Figure 2. Predictor selection using least absolute shrinkage and selection operator (LASSO) regression. (A) Cross-validation plot showing the relationship between the regularization parameter (λ) and model deviance. The vertical dashed line indicates the optimal λ value (minimum λ+1 SE). (B) LASSO coefficient path plot showing how variable coefficients shrink toward 0 as λ increases. (C) Ranking of the 16 selected predictors based on their standardized regression coefficients. (D) Coefficient direction plot indicating risk (red=positive coefficient) and protective (blue=negative coefficient) factors for depression. Act: social activity engagement; ADL: activity of daily living; DA: number of pain sites; IADL: instrumental ADL; memrye: memory problems; satlife: life satisfaction score; SRH: self-rated health score.

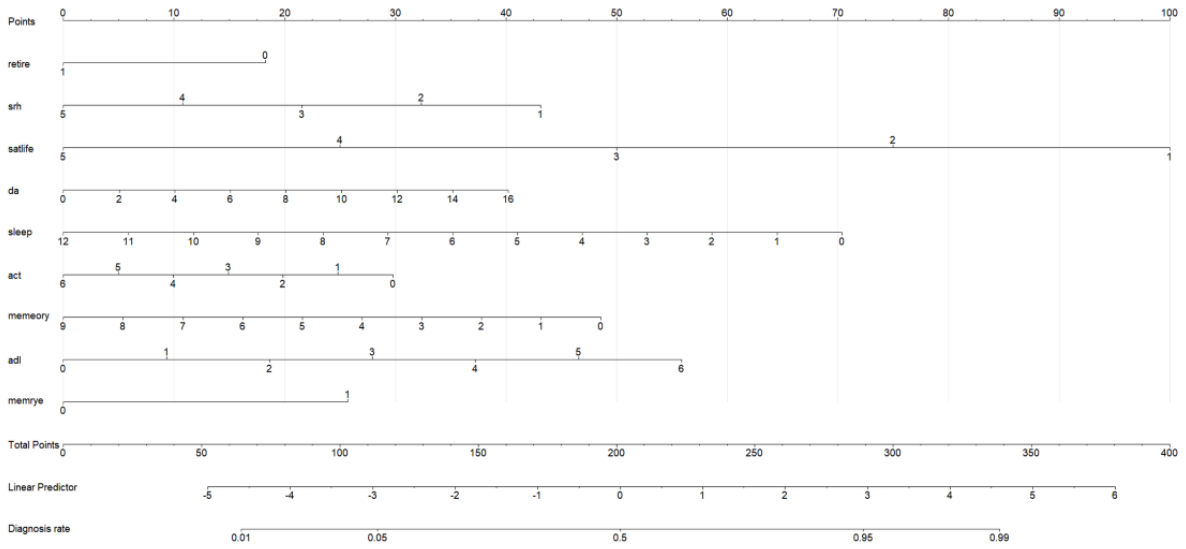


We took the 16 variables and ran them together in a new regression model and then drew a simple point chart (nomogram). The chart showed which factors mattered the most so that we could drop the weaker ones. After this step, 9 factors remained: retirement status, self-rated health score, life satisfaction score, number of pain sites, sleep hours, social activity engagement type, memory score, ADL score, and memory-related disease. We put these 9 factors into a final regression and built the last nomogram (Figure 3). On the chart, each line shows how many points a person obtains for that

XSL•FO
RenderX

factor; the bottom scale turns the total points into the chance of depression. Higher total points mean higher risk.

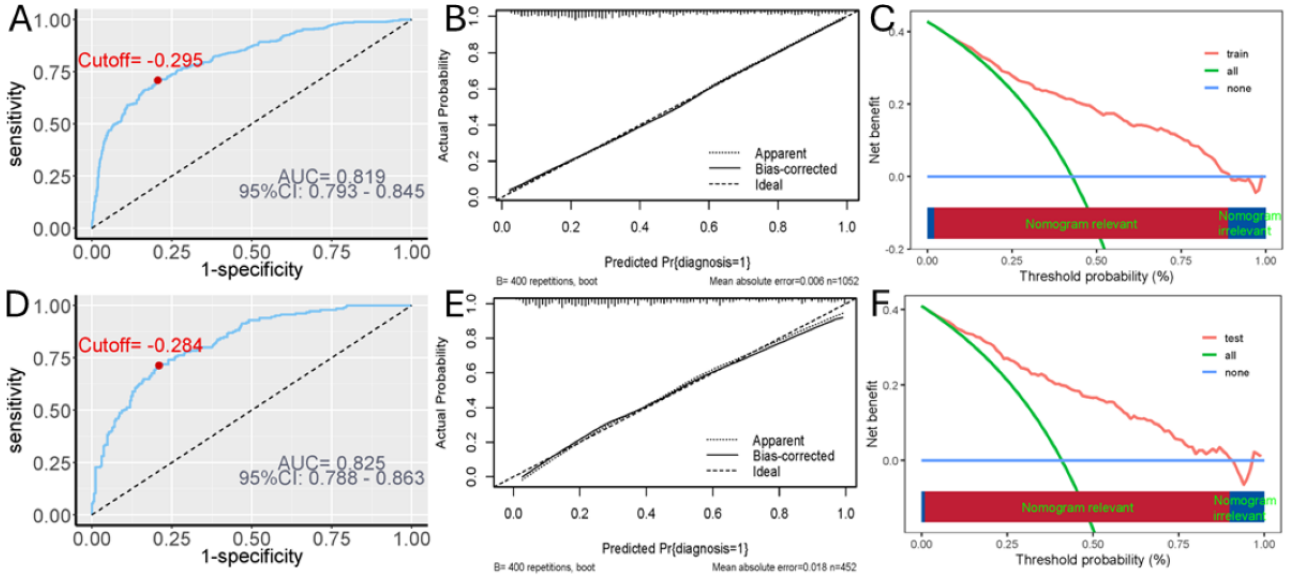
Figure 3. Nomogram for predicting depression risk. The nomogram was constructed using 9 predictors selected from multivariate logistic regression. Each predictor is assigned a point value based on its regression coefficient. The total score corresponds to the predicted probability of depression in the bottom scale. Higher total scores indicate a higher risk of depression. Act: social activity engagement; ADL: activity of daily living; DA: number of pain sites; memrye: memory problems; satlife: life satisfaction score; SRH: self-rated health score.



Results of Model Evaluation

We tested how well the model separated people with and without depression using the AUC. In the training data (Figure 4A), the AUC was 0.819, and in the testing data (Figure 4D), it was 0.825, showing good discrimination. The predicted probabilities matched the real outcomes closely in both sets; the average error was only 0.006 in training (Figure 4B) and 0.018 in testing (Figure 4E). Decision curve analysis (Figures 4C and 4F) also showed clear clinical benefit: the model yielded much larger net benefits than the "treat all" or "treat none" approaches. Overall, the nomogram was accurate and clinically useful.

Figure 4. Evaluation of the predictive model. (A) Receiver operating characteristic (ROC) curve and area under the curve (AUC) value in the training set (AUC=0.819). (B) Calibration plot in the training set (mean absolute error=0.006). (C) Decision curve analysis (DCA) in the training set. (D) ROC curve and AUC value in the testing set (AUC=0.825). (E) Calibration plot in the testing set (mean absolute error=0.018). (F) DCA in the testing set. The model demonstrates strong discriminative ability, good calibration, and high clinical utility in both datasets.



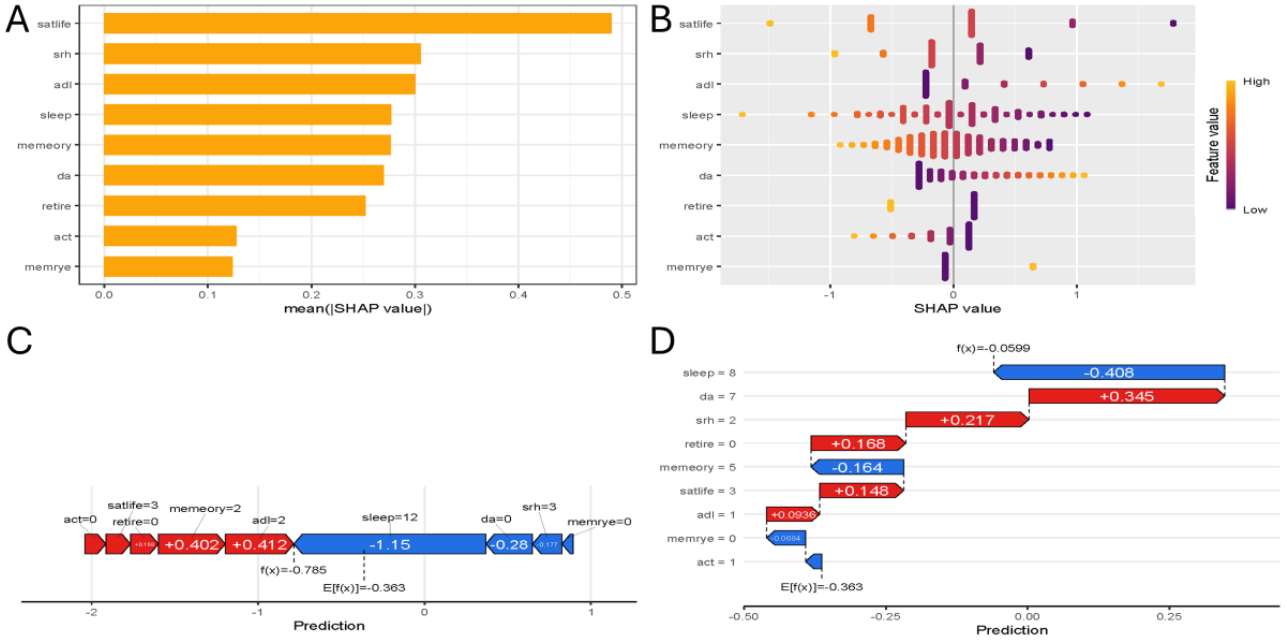
Results of Model Interpretation

We used SHAP values to look at the 9 main variables. Global and swarm plots (Figure 5A-B) show how each variable affects the entire sample. ADL score, memory problems, and number of pain sites pushed the risk upward, whereas retirement status, social activity engagement type, sleep hours, life satisfaction score, self-rated health score, and memory score pushed it

XSL•FO
RenderX

downward. To see how these factors worked for individual people, we picked 2 participants at random and used waterfall and force plots (Figure 5C-D) to show each factor's exact contribution.

Figure 5. Shapley additive explanations (SHAP) analysis of the final model. (A) Global SHAP importance plot showing the average impact of each feature on model predictions. (B) SHAP swarm plot illustrating the distribution of SHAP values for each feature across all participants (red=increased risk; blue=decreased risk). (C) Waterfall plot for a randomly selected individual showing the contribution of each feature to the final prediction. (D) Force plot for another individual, highlighting the cumulative effect of features on the predicted probability of depression. Positive SHAP values (red) increase the risk of depression, whereas negative values (blue) decrease it. Act: social activity engagement; ADL: activity of daily living; DA: number of pain sites; memrye: memory problems; satlife: life satisfaction score; SRH: self-rated health score.



Discussion

Principal Findings

This study establishes a clinically operable prediction model for depression risk among middle-aged and older patients with hypertension and diabetes, leveraging multidimensional predictors to achieve robust discrimination (AUC=0.819 - 0.825) and calibration (mean absolute error≤0.018). Nine core variables emerged as significant predictors in the final nomogram: ADL score, memory disorders, number of pain sites, sleep duration, life satisfaction score, self-rated health score, social activity engagement, retirement status, and memory score. The centrality of functional impairment as a mechanistic pathway is underscored by the ADL score's dominant predictive role [18,19], where individuals with depression exhibited markedly elevated scores (mean 1.20, SD 1.60 vs mean 0.34, SD 0.85; *P*<.001). Each unit increase in ADL score corresponded to a 3-fold increase in depression risk, particularly at critical thresholds beyond ADL score ≥2. Retirement status provided substantial protection against depression [19,20], with prevalence rates nearly halved among retirees (103/635, 16.2%) compared to nonretirees (269/869, 31.0%; *P*<.001) mediated by significantly greater social participation (mean activity score 0.94, SD 1.14 vs mean 0.65, SD 0.90; *P*<.001). SHAP value distributions (Figure 5A-B) corroborated a triadic pathogenesis framework. First, functional autonomy erosion (ADL and IADL decline interacting with multisite pain) imposes physical constraints that amplify helplessness [21,22]. Second, affective-cognitive burden manifests through memory disorders (there were memory disorders in 97/635, 15.3% vs 47/869, 5.4% of participants with and without depression, respectively; *P*<.001), and reduced life satisfaction (mean score 2.87, SD 0.87 vs mean score 3.41, SD 0.66 among participants with and without depression; *P*<.001) depletes emotional reserves [23]. Third, socioeconomic disengagement was evidenced by significantly lower household incomes among participants with depression (mean ¥32,470.65 [US $4720.71], SD ¥88,119.77 [US $12,811.20] vs mean ¥54,055.20 [US $7858.73], SD ¥79,512.32 [US $10,687.50], respectively; *P*<.001), constraining access to health-promoting resources [24,25]. The nonlinear inflection point observed in the ADL-depression relationship indicates that moderate functional decline triggers disproportionately severe psychological sequelae, offering an identifiable intercept for targeted interventions [26].

These findings advance beyond prior depression prediction paradigms in several critical dimensions. Traditional screening instruments reliant on mood-based symptomatology such as the Patient Health Questionnaire series demonstrate limited applicability in cardiometabolic populations due to diagnostic circularity: they cannot resolve whether fatigue or anhedonia stems from depression vs underlying chronic disease pathology [27,28]. Our model effectively dissociates these confounds by operationalizing ADLs as an independent construct validated through SHAP's feature dissociation capability. For instance, while arthritis prevalence was markedly higher among participants with depression (389/635, 61.3% vs 324/869, 37.3%; *P*<.001), LASSO selected number of pain sites rather than arthritis diagnosis as the principal pain predictor due to its ability to quantify functional impact severity [29]. When

XSL•FO
RenderX

benchmarked against models with similar AUC values to those of the model evaluated in this study, such as the depressive symptom predictor by Zheng et al [30,31], our algorithm demonstrates superior translational efficiency by using information already captured in routine chronic disease management ADL documentation during nursing assessments, medication review for memory issues, and social history documenting retirement status. This obviates the need for additional screening instruments, reducing assessment time to under 3 minutes. Moreover, we identify novel modifiable pathways absent in prior frameworks. The protective effect of retirement when linked to sustained social activity engagement reveals life transitions as windows for resilience-building interventions [32]. Enhancing life satisfaction emerged as another actionable leverage point given its strong negative association with depression (β=−0.63 via LASSO), suggesting that psychosocial enrichment may disrupt pathogenic cascades more effectively than purely biomedical approaches [33].

Methodologically, the integration of LASSO regression with SHAP value interpretation constitutes a primary strength. LASSO refined the predictor pool from 42 candidates to 9 nonredundant variables, efficiently eliminating noncontributory markers such as cancer or lung disease. SHAP analysis then transformed statistical coefficients—for example, the negative β value for retirement (−0.89)—into clinically interpretable narratives. At an individual level (Figure 5D), retirees maintaining active social roles exhibited an 18-fold depression risk reduction compared to their socially isolated counterparts. The CHARLS cohort's population representativeness further enhances generalizability to China's aging cardiometabolic patient demographic. Four principal limitations warrant acknowledgment. Due to the cross-sectional nature of wave 5 CHARLS data, causal precedence between functional decline and depression could not be definitively established. However, the biological gradient observed in ADL risk proportionality (higher scores predicting depression probability stepwise) lends credibility to functional deterioration as a driver rather than a secondary consequence [34,35]. Second, depression ascertainment using a CESD-10 score >10 may overlook subsyndromal presentations, potentially underestimating risk in highly functional individuals [36]. Third, unmeasured variables influencing both pathways, particularly medication adherence patterns or pharmacodynamic interactions between psychotropics and antihypertensives, were unaccounted for. Finally, sociocultural proxies unique to Chinese aging populations' children's financial support demonstrating borderline significance (*P*=.02-.03) were excluded from the final model despite their potential clinical relevance [37]. Future iterations should address this through culturally adapted predictor sets.

This study was conducted based on statistical data from the CHARLS database. Therefore, the results of this study have limitations in terms of applicability to young and middle-aged populations, as well as populations from different regions and ethnic groups. In future research, we may consider combining data from public databases in the Americas, Europe, and other regions to conduct joint modeling, thereby improving the generalization performance of the model. These findings suggest priority areas for future research and clinical application. Longitudinal analyses using multiwave CHARLS data should test the proposed retirement–social engagement–protection pathway, examining whether weekly social activity sustains resilience to depressive symptoms. A 3-tier risk stratification protocol based on nomogram scores may guide care: individuals scoring ≤150 can continue annual screening; those scoring 150 to 250 should receive functional preservation strategies, including task-oriented exercise and cognitive behavioral approaches to improve ADLs, sleep, and social engagement; and those scoring ≥250 would require multidisciplinary management due to high comorbidity. Embedding this algorithm into electronic health records within China's National Essential Public Health Service Package could enable automated alerts when thresholds are met. Pilot programs should focus on communities where the baseline ADL score is ≥1.5; in this study, 51.2% (325/635) of participants with depression were in this category. Randomized trials are needed to evaluate mediator-targeted interventions, including combined ADL and pain management, structured postretirement social role adaptation, and hemoglobin A_{1c}–stratified memory screening in diabetes. Targeting domain-specific vulnerabilities rather than symptoms alone may reduce depression incidence while optimizing primary care resources.

Conclusions

We developed and validated a depression risk model for middle-aged and older adults with hypertension and diabetes. The nomogram, integrating 9 routinely collected predictors, including ADL score, memory impairment, and pain burden, showed strong discrimination and calibration. Depression risk rose 3-fold for each unit increase in ADL score, whereas retirement was protective. Functional impairment, emotional cognitive burden, and socioeconomic disadvantage formed the dominant risk pathway. LASSO-SHAP integration streamlined predictor selection, enabling screening in under 3 minutes. Prospective trials should test interventions targeting postretirement social engagement and functional rehabilitation, alongside implementing a 3-tier risk stratification protocol in primary care. Targeting these domain-specific risks could reduce depression incidence.

Acknowledgments

The authors are grateful to all the participants and staff of the China Health and Retirement Longitudinal Study.

XSL•FO
RenderX

Funding

This study was funded by the Chinese Cardiovascular Association-ASCVD Fund (2023-CCA-ASCVD-018), the Project of State Key Clinical Department (Z2023058), and the Natural Science Foundation of Hunan Province (2025JJ90208).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

TP contributed to data curation and writing—original draft. YZ managed writing—review and editing. JW and RM managed writing—review and editing.

Conflicts of Interest

None declared.

References

1. Bi YH, Pei JJ, Hao C, Yao W, Wang HX. The relationship between chronic diseases and depression in middle-aged and older adults: a 4-year follow-up study from the China Health and Retirement Longitudinal Study. J Affect Disord 2021 Jun 15;289:160-166. [doi: 10.1016/j.jad.2021.04.032] [Medline: 33984686]
2. Hairi NN, Bulgiba A, Mudla I, Said MA. Chronic diseases, depressive symptoms and functional limitation amongst older people in rural Malaysia, a middle income developing country. Prev Med 2011 Oct;53(4-5):343-346. [doi: 10.1016/j.ypmed.2011.07.020] [Medline: 21864564]
3. Lam MI, Chen P, Xie XM, et al. Heart failure and depression: a perspective from bibliometric analysis. Front Psychiatry 2023 Mar 2;14:1086638. [doi: 10.3389/fpsyt.2023.1086638] [Medline: 36937736]
4. Hare DL, Toukhsati SR, Johansson P, Jaarsma T. Depression and cardiovascular disease: a clinical review. Eur Heart J 2014 Jun 1;35(21):1365-1372. [doi: 10.1093/eurheartj/eht462] [Medline: 24282187]
5. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 2009 Aug 22;374(9690):609-619. [doi: 10.1016/S0140-6736(09)60879-5] [Medline: 19640579]
6. Zhang Y, Ting RZW, Lam MHB, et al. Measuring depression with CES-D in Chinese patients with type 2 diabetes: the validity and its comparison to PHQ-9. BMC Psychiatry 2015 Aug 18;15:198. [doi: 10.1186/s12888-015-0580-0] [Medline: 26281832]
7. Chin WY, Choi EPH, Chan KTY, Wong CKH. The psychometric properties of the Center for Epidemiologic Studies Depression Scale in Chinese primary care patients: factor structure, construct validity, reliability, sensitivity and responsiveness. PLoS ONE 2015 Aug 7;10(8):e0135131. [doi: 10.1371/journal.pone.0135131] [Medline: 26252739]
8. Tan J, Ma C, Zhu C, et al. Prediction models for depression risk among older adults: systematic review and critical appraisal. Ageing Res Rev 2023 Jan;83:101803. [doi: 10.1016/j.arr.2022.101803] [Medline: 36410622]
9. Jin H, Wu S, Di Capua P. Development of a clinical forecasting model to predict comorbid depression among diabetes patients and an application in depression screening policy making. Prev Chronic Dis 2015 Sep 3;12:E142. [doi: 10.5888/pcd12.150047] [Medline: 26334714]
10. Karpouzian-Rogers T, Dworak EM, Bucko P, et al. Measures of decision-making ability and functional outcomes in older adults: results from a scoping review in the ARMCADA study. Front Psychol 2025 Mar 21;16:1540493. [doi: 10.3389/fpsyg.2025.1540493] [Medline: 40191575]
11. Pietrzak B, Kujawa J, Lipert A. Depressive disorders, cognitive and physical function of older people in early dementia detection. Life (Basel) 2023 Oct 4;13(10):2010. [doi: 10.3390/life13102010] [Medline: 37895392]
12. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). Am J Prev Med 1994;10(2):77-84. [Medline: 8037935]
13. Fu H, Si L, Guo R. What is the optimal cut-off point of the 10-item Center for Epidemiologic Studies Depression Scale for screening depression among Chinese individuals aged 45 and over? An exploration using latent profile analysis. Front Psychiatry 2022 Mar 14;13:820777. [doi: 10.3389/fpsyt.2022.820777] [Medline: 35360127]
14. Zhao Z, Chu Y, Chen X, Liu T. Association between outdoor and indoor air pollutant exposure and depression among middle-aged and older adults in China. Sci Rep 2025 Jul 1;15(1):22162. [doi: 10.1038/s41598-025-07185-1] [Medline: 40594657]
15. Wu Q, Feng J, Pan CW. Risk factors for depression in the elderly: an umbrella review of published meta-analyses and systematic reviews. J Affect Disord 2022 Jun 15;307:37-45. [doi: 10.1016/j.jad.2022.03.062] [Medline: 35351490]
16. Bottomley C, Nazareth I, Torres-González F, et al. Comparison of risk factors for the onset and maintenance of depression. Br J Psychiatry 2010 Jan;196(1):13-17. [doi: 10.1192/bjp.bp.109.067116] [Medline: 20044653]

XSL•FO
RenderX

17. Morelli D, Dolezalova N, Ponzo S, Colombo M, Plans D. Development of digitally obtainable 10-year risk scores for depression and anxiety in the general population. Front Psychiatry 2021 Aug 13;12:689026. [doi: 10.3389/fpsyt.2021.689026] [Medline: 34483986]
18. Zhu X, Wang Y, Luo Y, Ding R, Shi Z, He P. Bidirectional, longitudinal associations between depressive symptoms and IADL/ADL disability in older adults in China: a national cohort study. BMC Geriatr 2024 Aug 6;24(1):659. [doi: 10.1186/s12877-024-05248-y] [Medline: 39107705]
19. Arayici ME, Kose A, Dolu S, et al. Epidemiological assessment of depression, activities of daily living and associated factors in elderly individuals aged 65 years and older: evidence from a population-based study. J Clin Med 2025 Apr 21;14(8):2853. [doi: 10.3390/jcm14082853] [Medline: 40283685]
20. Zhan Y, Ning H, Zhang Y. From work to community: how social participation and retirement can benefit older adults' depression. Front Psychiatry 2025 Mar 3;16:1522222. [doi: 10.3389/fpsyt.2025.1522222] [Medline: 40099148]
21. Liu P, Chen H, Tong B, Zhu D, Cong X, Shang S. Association between multisite musculoskeletal pain and disability trajectories among community-dwelling older adults. Aging Clin Exp Res 2024 May 23;36(1):115. [doi: 10.1007/s40520-024-02764-0] [Medline: 38780859]
22. Denkinger MD, Lukas A, Nikolaus T, Peter R, Franke S, ActiFE study group. Multisite pain, pain frequency and pain severity are associated with depression in older adults: results from the ActiFE Ulm study. Age Ageing 2014 Jul;43(4):510-514. [doi: 10.1093/ageing/afu013] [Medline: 24603284]
23. Hallford DJ, Rusanov D, Yeow JJE, Barry TJ. Overgeneral and specific autobiographical memory predict the course of depression: an updated meta-analysis. Psychol Med 2021 Apr;51(6):909-926. [doi: 10.1017/S0033291721001343] [Medline: 33875023]
24. Ridley M, Rao G, Schilbach F, Patel V. Poverty, depression, and anxiety: causal evidence and mechanisms. Science 2020 Dec 11;370(6522):eaay0214. [doi: 10.1126/science.aay0214] [Medline: 33303583]
25. Campbell D, Green MJ, Davies N, et al. Effects of depression on employment and social outcomes: a Mendelian randomisation study. J Epidemiol Community Health 2022 Jun;76(6):563-571. [doi: 10.1136/jech-2021-218074] [Medline: 35318279]
26. Yan W, Wang L, Li C, Meng Y, Guo Q, Li H. Bidirectional association between ADL disability and depressive symptoms among older adults: longitudinal evidence from CHARLS. Sci Rep 2025 Feb 28;15(1):7125. [doi: 10.1038/s41598-025-91680-y] [Medline: 40021702]
27. Nouwen A, Deschênes SS, Balkhiyarova Z, Albertorio-Díaz JR, Prokopenko I, Schmitz N. Measurement invariance testing of the Patient Health Questionnaire-9 (PHQ-9) across people with and without diabetes mellitus from the NHANES, EMHS and UK Biobank datasets. J Affect Disord 2021 Sep 1;292:311-318. [doi: 10.1016/j.jad.2021.05.031] [Medline: 34139403]
28. Cheng HT, Ho MC, Hung KY. Affective and cognitive rather than somatic symptoms of depression predict 3-year mortality in patients on chronic hemodialysis. Sci Rep 2018 Apr 12;8(1):5868. [doi: 10.1038/s41598-018-24267-5] [Medline: 29651018]
29. Hawker GA, Gignac MAM, Badley E, et al. A longitudinal study to explain the pain-depression link in older adults with osteoarthritis. Arthritis Care Res (Hoboken) 2011 Oct;63(10):1382-1390. [doi: 10.1002/acr.20298] [Medline: 20662042]
30. Zheng Y, Zhang C, Liu Y. Risk prediction models of depression in older adults with chronic diseases. J Affect Disord 2024 Aug 15;359:182-188. [doi: 10.1016/j.jad.2024.05.078] [Medline: 38768825]
31. Zheng Y, Zhang T, Yang S, Wang F, Zhang L, Liu Y. Using machine learning to predict the probability of incident 2-year depression in older adults with chronic diseases: a retrospective cohort study. BMC Psychiatry 2024 Dec 2;24(1):870. [doi: 10.1186/s12888-024-06299-6] [Medline: 39623372]
32. Settels J, Böckerman P. Impact of retirement on quality of life: role of changes in social network involvement. Res Aging 2025;47(7-8):375-391. [doi: 10.1177/01640275251324871] [Medline: 40073282]
33. Koivumaa-Honkanen H, Kaprio J, Honkanen R, Viinamäki H, Koskenvuo M. Life satisfaction and depression in a 15-year follow-up of healthy adults. Soc Psychiatry Psychiatr Epidemiol 2004 Dec;39(12):994-999. [doi: 10.1007/s00127-004-0833-6] [Medline: 15583908]
34. Lin F, Liu C, Liu H. Development and validation of a predictive model for activities of daily living dysfunction in older adults: retrospective analysis of data from the China Health and Retirement Longitudinal Study. JMIR Med Inform 2025 Jun 19;13:e73030. [doi: 10.2196/73030] [Medline: 40537082]
35. Zhang F, Yang W. Interaction between activities of daily living and cognitive function on risk of depression. Front Public Health 2024 Feb 7;12:1309401. [doi: 10.3389/fpubh.2024.1309401] [Medline: 38384887]
36. Björgvinsson T, Kertz SJ, Bigda-Peyton JS, McCoy KL, Aderka IM. Psychometric properties of the CES-D-10 in a psychiatric sample. Assessment 2013 Aug;20(4):429-436. [doi: 10.1177/1073191113481998] [Medline: 23513010]
37. Wangliu Y. Does intergenerational support affect older people's social participation? An empirical study of an older Chinese population. SSM Popul Health 2023 Feb 21;22:101368. [doi: 10.1016/j.ssmph.2023.101368] [Medline: 36873267]

Abbreviations

ADL: activity of daily living
AUC: area under the receiver operating characteristic curve

XSL•FO
RenderX

CESD-10: 10-item Center for Epidemiologic Studies Depression Scale
CHARLS: China Health and Retirement Longitudinal Study
IADL: instrumental activity of daily living
LASSO: least absolute shrinkage and selection operator
SHAP: Shapley additive explanations

Edited by A Benis; submitted 30.Jul.2025; peer-reviewed by H Luo, H Liu, M Zhou; revised version received 05.Jan.2026; accepted 17.Feb.2026; published 12.Mar.2026.

Please cite as:
Peng T, Zhang Y, Miao R, Wang J
Development of a Practical Nomogram for Depression Risk Stratification in Older Adults With Hypertension and Diabetes: Retrospective Analysis of Data From the China Health and Retirement Longitudinal Study
JMIR Med Inform 2026;14:e81529
URL: https://medinform.jmir.org/2026/1/e81529
doi:10.2196/81529

© Ting Peng, Ying Zhang, Rujia Miao, Jiangang Wang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Deep Learning for Content-Based Medical Image Retrieval in Picture Archiving and Communication Systems for Brain Tumor Detection: Algorithm Development and Validation

Chin-Lin Lee[1], BSc; Tzu-Hsuan Hsu[1], BSc; Yu-Te Wu[2], PhD; Wan-Yuo Guo[3], MD, PhD; Woei-Chyn Chu[4], PhD; Chung-Yueh Lien[1], PhD

[1]Department of Information Management, National Taipei University of Nursing and Health Science, No. 365, Ming-te Rd, Beitou Dist, Taipei City, Taiwan

[2]Institute of Biophotonics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3]Department of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:
Chung-Yueh Lien, PhD
Department of Information Management, National Taipei University of Nursing and Health Science, No. 365, Ming-te Rd, Beitou Dist, Taipei City, Taiwan

Abstract

Background: Advances in medical imaging have led to massive archives, yet navigating these datasets remains challenging due to the limitations of traditional text-based search engines. While content-based medical image retrieval (CBMIR) offers a visual feature–based solution to enhance clinical workflows and research, its operational integration into picture archiving and communication systems (PACS) remains a significant bottleneck. Despite the progress in deep learning for feature extraction, CBMIR tools are rarely integrated and effectively implemented within existing radiology information systems due to complex protocol barriers.

Objective: To address these challenges, this study develops a CBMIR system meticulously designed to cater to 7 distinct types of brain tumors as seen in brain magnetic resonance images. Our system is tailored to assist radiologists and health care professionals in efficiently retrieving pertinent historical medical images, thereby providing quantitative decision support for radiologists and facilitating evidence-based case comparison, with the potential to improve retrieval efficiency and clinical workflow, rather than directly enhancing diagnostic accuracy.

Methods: The dataset used in this study was collected from a single medical center and is not publicly available. The core innovation is a state-of-the-art deep learning–based feature extraction algorithm specifically engineered for the CBMIR system. We use GoogLeNet as the primary architecture, incorporating generalized mean pooling to capture nuanced local features and an embedding layer for dimension reduction. Crucially, we address the integration gap by harmonizing 2 open-source projects to successfully embed the CBMIR system into a functional PACS environment via standard protocols.

Results: The image dataset contains 658 participants with 15,873 images collected from 2000 to 2017. The empirical findings of our research demonstrate the performance and robustness of the proposed CBMIR system. Our system achieves a remarkable mean average precision score of 89.16% and an equally impressive Precision@10 score of 94.08%. These metrics affirm the system's efficacy in retrieving relevant medical images. Furthermore, we successfully integrate the CBMIR system into a PACS by successfully harmonizing 2 open-source projects.

Conclusions: This study presents the design and implementation of a PACS-integrated CBMIR system for brain magnetic resonance imaging, and the experimental results demonstrate that the system can achieve efficient and accurate image retrieval within a clinical workflow.

(JMIR Med Inform 2026;14:e78300) doi:10.2196/78300

KEYWORDS

content-based medical image retrieval; CBMIR; deep learning; Digital Imaging and Communications in Medicine; DICOM; brain tumor; picture archiving and communication system; PACS; tumor

XSL•FO
RenderX

Introduction

Background

Advances and rapid growth in medical imaging technology have led to the emergence of massive medical image archives. Effective image retrieval is crucial for both research and clinical applications, yet navigating these large-scale archives remains a time-consuming task, often limited by traditional text-based search engines [1]. Radiologists, in particular, face the challenge of interpreting cross-sectional studies composed of thousands of images retrieved from picture archiving and communication systems (PACS) [2].

PACS is a health care IT system that manages the storage, retrieval, distribution, and presentation of medical digital images within a hospital. It works closely with the DICOM (Digital Imaging and Communications in Medicine) standard, which ensures that medical images and related data are stored and transmitted in a consistent format across different machines and systems. To integrate a third-party system, it must interface directly with the PACS via the DICOM protocol. This integration is critical as it allows health care providers to successfully access and share patient images within their workflows, enabling the content-based medical image retrieval (CBMIR) system to function effectively as a clinical decision support tool.

CBMIR offers a more efficient solution by enabling image searches based on the visual features of source images. CBMIR systems have the potential to significantly enhance clinical imaging workflows, delivering meaningful improvements across multiple areas, including medical education, research, and clinical diagnostics [3].

Röhrich et al [4] found that integrating CBMIR into medical imaging reduced image interpretation time by 31.3% for both junior and senior radiologists, while also helping them access more valuable consultative resources. When combined with an intelligent interactive visual browser, CBMIR could further enhance medical education, research, and clinical care by improving both image navigation and diagnostic outcomes [1]. Müller et al [5] highlighted the usefulness of CBMIR in academic settings, particularly for lecturers and students who can visually explore educational image repositories, facilitating more engaging and insightful learning experiences.

The retrieval principle encompasses several key aspects that collectively contribute to the effective retrieval of information. In particular, intricate processes are involved, such as analyzing and converting user requisites into queries suitable for retrieving data from indexed databases. In the process of CBMIR operation, the features of the images, relying on captions, visual words [6], preannotations [7], and image attributes [8], are first extracted and stored in a database. A distance metric summarizing information about image features measures the similarity between any given pair of images [9,10]. During CBMIR development, the key components are feature extraction and similarity determination [8], with feature extraction being the crucial step. In the past, CBMIR techniques have focused on semantic content matching of images, for example, analyzing and retrieving images based on their color, texture, and layout. Ponciano-Silva and colleagues [11] addressed the limitations of past CBMIR systems that typically used only a single perceptual parameter (ie, feature extraction and distance function), thus restricting such systems to a single search space.

In recent years, the rapid development of artificial intelligence (AI) technology and significant improvements in image processing capabilities have led to the application of AI technology to enhance CBMIR performance. The powerful algorithmic capabilities of deep learning have also been applied to CBMIR systems. Maji et al [12] compared the accuracy of various AI pretrained models, such as DenseNet [13], InceptionResNetV2 [14], InceptionV4 [14], MobileNetV2 [15], ResNet50 [16], and VGG19 (Visual Geometry Group model 19) [17]. They concluded that InceptionResNetV2 exhibited the highest accuracy and chose it as the feature extraction model. They further proposed using principal component analysis and clustering methods to increase retrieval speed by at least a factor of 2.

Rashad et al [18] adopted pretrained deep learning models to extract additional high-level features from medical images to improve the accuracy of the retrieved results in extracting compact, deep, and high-level features. Cheng and colleagues [19] tackled the limited texture representation of brain tumor images in traditional image processing methods. They used the Fisher Vector to aggregate local features into a single vector and used closed-form metric learning to effectively measure similarity between 2 given images. They achieved a mean average precision (mAP) of 94.8%. While this method achieved excellent accuracy, the concatenation of feature vectors from various subregions led to a high dimensionality of up to 24,576, potentially resulting in decreased query speed.

Qayyum et al [20] collected a diverse medical image dataset encompassing 24 anatomical regions and various imaging modalities. They used deep learning to train and build an 8-layer model whose architecture transforms 4096 features into a probability distribution over 24 categories. The average accuracy achieved in classification was 99.76%. They found that without prior classification and querying, mAP reached a mere 53%, while performing classification before querying improved mAP to 69%. Swati et al [21] focused on a contrast-enhanced magnetic resonance imaging brain tumor dataset [22] and used the pretrained VGG19 model (Visual Geometry Group model 19) for fine-tuning transfer learning (TL). Their study applied deep learning by using the last fully connected layer to extract the features, which is the input of the closed-form metric learning for contrast-enhanced magnetic resonance imaging to achieve a mAP of 96.13% [21]. Deepak and Ameer [23] evaluated various pretrained model architectures in terms of classification accuracy, training time, and feature extraction time. They selected GoogLeNet as the basis for their deep learning architecture, requiring a 2-stage training process; the achieved mAP was 97.64%.

Vision Transformer (ViT)–based models have demonstrated significant advantages when dealing with large datasets, enabling them to attend to critical image features, superior capacity for global feature modeling and context awareness,

XSL•FO
RenderX

and thereby improve medical image classification performance [24]. For instance, Anupama and Anitha [25] developed a ViT-based CBMIR system that reliably and precisely extracts pertinent medical images. The predicted results, particularly for magnetic resonance imaging data, confirm the model's capability to manage intricate imaging characteristics and provide precise outcomes [25]. Furthermore, studies by Sunday et al [26] highlight the use of the DINOv2 self-supervised framework to enhance anatomical structure visibility and reduce image noise. This demonstrates the potential of general-purpose self-supervised models in CBMIR, offering high transferability and robustness to varying image quality without requiring extensive domain adaptation [26]. In related work, Sarıateş and Özbay [27] used a TL framework to successfully classify melanoma images and enhance the efficiency of their content-based image retrieval systems.

However, due to technological bottlenecks that have yet to be overcome, CBMIR is rarely used in existing PACS and radiology information systems [1]. CBMIR could greatly benefit researchers and physicians in PACS [28-31]. Valente et al [32] used the open-source PACS software Dicoogle (University of Aveiro) as the development base for CBMIR for clinical applications. Besides general image, they focused on a mammography profile specifying the mammography features, allowing users to access CBMIR features integrated with a PACS to ensure high scalability.

Our Contributions

Our primary contributions are summarized as follows: This study proposes an innovative CBMIR approach using ArcFace Loss [33] to enhance the understanding of image feature similarity and replaces the pooling layer with generalized mean (GeM) pooling. This allows the model to focus on key regions while preserving overall image information. This strategy strengthens the feature space's discriminability, resulting in significantly improved retrieval accuracy (mAP). We developed a robust CBMIR system successfully integrated with open-source DICOM viewer named BlueLight [34], achieving real-time clinical applicability by enabling rapid consultation of structurally similar cases.

Methods

Dataset

We gathered from Taipei Veterans General Hospital (TVGH) a comprehensive collection of brain magnetic resonance (MR) images formatted in DICOM. This image dataset was retrospectively collected and features a wide array of neurological conditions. These conditions have been categorized into 7 distinct groups: metastasis, glioma, meningioma, pituitary adenoma, vestibular schwannoma, and categories for images with no tumor contour and those representing normal brain scans.

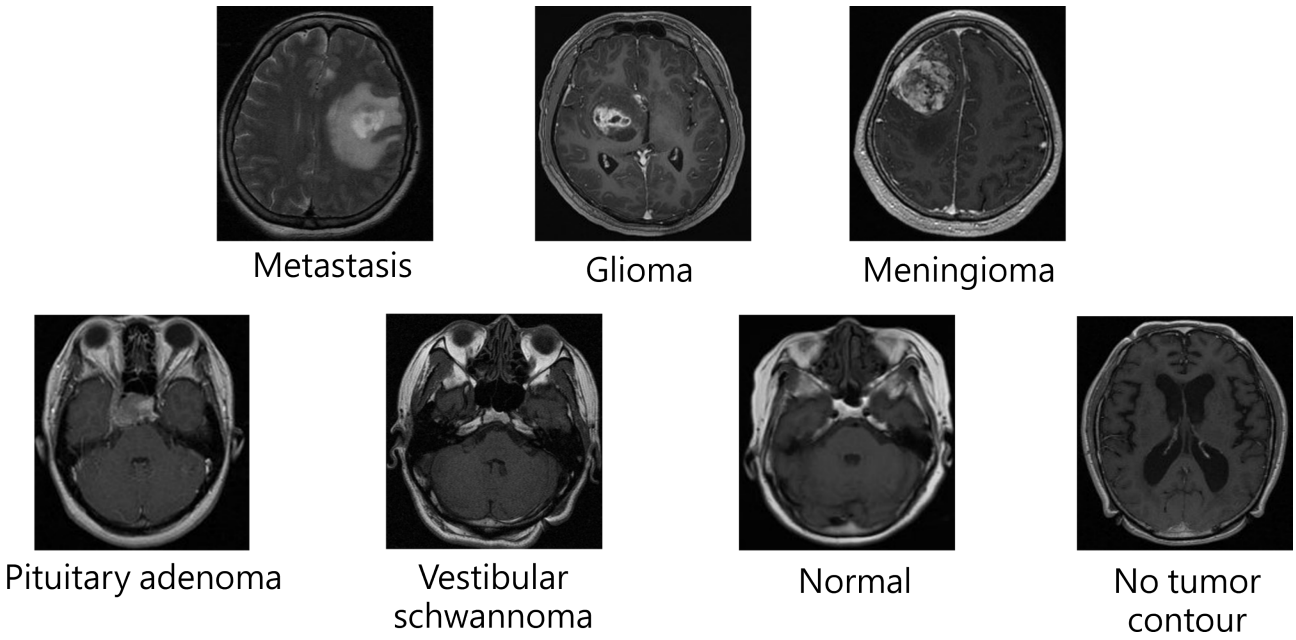
All classes of tumors were delineated by neurosurgeons and validated by neuroradiologists from the TVGH, referred to as "ground truth." Table 1 provides statistics regarding the classes of tumors in the dataset. The image dataset, comprising 658 participants, was divided into training (70%, 461 participants, 11,207 images) and test (30%, 197 participants, 4666 images) subsets collected from 2000 to 2017. The median age was 56.4 years. Gender distribution was 57% (375/658) male and 43% (283/658) female. The distribution of image laterality was 52% left and 48% right. The median tumor volume, excluding normal images, was 2.02 mL, calculated from manual segmentation of lesions in MR images. Some patients underwent Gamma Knife radiosurgery that contains both pre–radiation therapy images and follow-up images. No patient overlapped between the training and test sets. Data in the test subset were not augmented. Examples of all classes of tumors are shown in Figure 1.

Table . Classification of tumors in the dataset.

Class	Training set	Test set	Total
Metastasis	1479	651	2130
Glioma	1484	633	2117
Meningioma	1466	632	2098
Pituitary adenoma	1520	629	2149
Vestibular schwannoma	1492	645	2137
No contour	2387	884	3271
Normal brain	1379	592	1971
Overall	11,207	4666	15,873

XSL•FO
RenderX

Figure 1. Examples of all tumor classes.



A brain image is inherently composed of a series of image slices, constituting a set that encompasses images both with and without tumor contours. In real-world clinical scenarios, such a dual presence of image types is a common occurrence. Focusing solely on images displaying tumors would inadvertently result in the exclusion of those lacking tumor contours, thereby impeding the accurate extraction of the latter. To mitigate this challenge, we implemented a strategy to segregate such images into a distinct category labeled “no tumor contour.” This deliberate classification enables us to effectively handle and differentiate images without tumor contours, thereby enhancing the comprehensiveness of our dataset and ensuring a more accurate analysis.

Imaging Parameters

All images were acquired using consistent imaging acquisition protocols, detailed as follows. MR scans were obtained using 1.5T parametric MR scanners from manufacturers General Equipment and Philips. Each image had a spatial resolution of 512×512 pixels, with 16-bit depth per pixel, and a slice thickness of 3 mm. The imaging protocols included both T1-weighted (T1W) and T2-weighted (T2W) sequences. These sequences were conducted as follows: (1) T1W+C: acquired using axial 2D spin echo (repetition time = 416 ms, echo time = 9 ms, flip angle = 90°). (2) T2W: acquired using axial 2D spin echo (repetition time = 4050 ms, echo time = 109 ms, flip angle = 90°).

Image Preprocessing and Augmentation

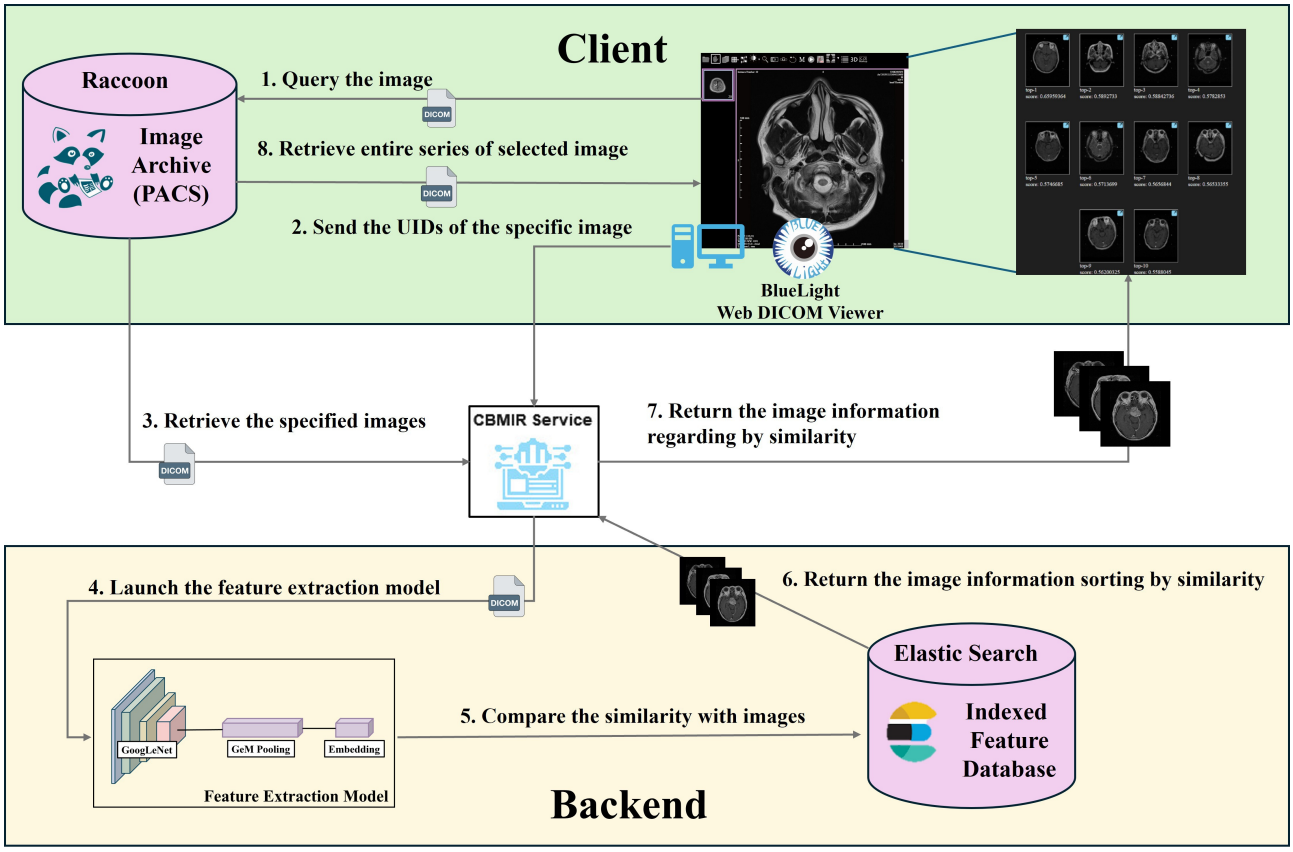
The original 512×512-pixel single-channel DICOM images were resized to 224×224 pixels using bilinear interpolation. Although the GoogLeNet backbone, integrated with GeM pooling, supports variable spatial dimensions, this specific scale was selected to align with the ImageNet pretraining distribution. Such standardization optimizes the balance between representation learning and computational efficiency, significantly reducing graphics processing unit memory footprint and inference latency. To leverage pretrained weights without altering the input layer, the resulting grayscale images were stacked into 3 identical channels (red-green-blue). Finally, while the training set underwent data augmentation (including random rotation, flipping, and scaling) to enhance robustness, the test set remained nonaugmented to ensure an evaluation reflective of clinical practice. This choice represents a practical trade-off between computational efficiency and spatial resolution, which is appropriate for large-scale similarity retrieval rather than pixel-level diagnosis.

System Architecture

The CBMIR system, including the feature extraction model, is illustrated in Figure 2. As seen therein, radiologists can select the images they wish to query for the CBMIR system within the web DICOM viewer. Upon selection, the query image is initially input into a pretrained AI model for feature extraction. Subsequently, CBMIR takes these feature vectors and matches them against the database, retrieving images with similar feature vectors from the database. For each search, the user sends a single query request containing 3 DICOM UIDs: the Study Instance UID, Series Instance UID, and SOP Instance UID. In response, the CBMIR system retrieves the specified image from the PACS and sends it to the feature extraction model. After processing, the retrieved images are then presented in the web DICOM viewer. Finally, the user can select the most closely related image and retrieve the entire study from the PACS server.

XSL•FO
RenderX

Figure 2. Feature extraction model and content-based medical image retrieval service architecture. CBMIR: content-based medical image retrieval; DICOM: Digital Imaging and Communications in Medicine; PACS: picture archiving and communication systems; UIDs: unique identifiers.



Feature Extraction Model

The foundational network of the feature extraction model architecture is based on GoogLeNet, with a modification replacing the original global average pooling with GeM pooling [35] to obtain more generalized features. Finally, an embedding layer is added to reduce the feature dimensions from 1024 to 256, compressing the data for improved computational efficiency. An explanation of each of the components of the model follows in the subsections, including (1) GoogLeNet, (2) GeM pooling, (3) embedding, and (4) the loss function.

1. *GoogLeNet layer*: The primary design core is the inception architecture, which uses 1×1 convolutional layer to reduce dimensionality, thereby lowering computational complexity. The inception architecture also splits the input into various sizes for parallel processing and subsequently aggregates them to extract features at various scales. The choice of GoogLeNet as the network backbone in this study is influenced by a previous study [23]. According to their results, GoogLeNet demonstrates several advantages over other potential network backbones, including lower computational complexity (number of operations), minimal memory consumption (parameter count), higher validation accuracy, shorter training time, and faster feature extraction. Furthermore, although ResNet50 exhibits higher accuracy, its training time is approximately 3 times that of GoogLeNet. Therefore, ResNet50 was not considered for this study due to our preference for efficient computation.

2. *GeM pooling layer*: GeM pooling was proposed by Ponciano-Silva et al [11] as a method that uses a universal formula to replace traditional global max pooling, that is,

$$(1) f(m)=f1(m)\ldots fkm\ldots fK(m)\ ,fkm=\max x\in Xkx,$$

and global average pooling, that is,

$$(2) f(a)=[f1(a)\ldots fk(a)\ldots fK(a)]\ ,fk(a)=1|Xk|\sum x\in Xkx$$

Formula equation (3) shows the process of GeM pooling, which uses a 3D tensor x with dimensions $W \times H \times K$ as input to compute the result, represented as the f vector.

$$(3) f(g)=[f1(g)\ldots fk(g)\ldots fK(g)]\ ,fk(g)=(1|Xk|\sum x\in Xkxpk)1pk$$

In this formula, pk is a parameter that determines the norm level. When $pk\to\infty$, GeM pooling becomes equivalent to max pooling; when $pk=1$, it simplifies to average pooling. Although pk can be treated as a trainable parameter, in this study, it was fixed to $pk = 3$ for all experiments, following prior image retrieval benchmarks [35]. This design choice ensures a stable and reproducible feature embedding space across all evaluations.

By maintaining pk as a constant rather than a trainable parameter, we ensure a stable feature embedding space, particularly when processing domain-specific medical images. This approach contributes to improved accuracy and consistency in image retrieval tasks.

- 3. *Embedding layer*: The embedding layer exists primarily to reduce the dimensionality of the output from the global pooling layer. In this study, we used an embedding structure [33,36] that consists of a sequence of FC (fully connected) − BN (BatchNorm) − PReLU (parametric rectified linear

XSL•FO
RenderX

unit) layers. The FC layer maps the original 1024-dimensional features to 256 dimensions while attempting to preserve the original features as much as possible. The BN layer normalizes the values to mitigate the problem of vanishing gradients, and the PReLU layer adds nonlinearity to the model.

- 4. *Loss function*: This study uses the ArcFace loss function [33], as represented in formula equation 4.

$$(4) L=-1b\log es(\cos(\theta yi+m))es(\cos(\theta yi+m))+\sum j=1,j\neq yines\cos\theta j$$

In this formula, *b* represents the batch size, *s* corresponds to the fixed feature values after L2-normalization,θyi represents the weight and angle between the correct class and feature, and *m* controls the additional angle for separating different classes. The primary objective of this loss function is to ensure that embedding features from different classes are pushed as far apart as possible while keeping embedding features from the same class close together.

Enhanced Feature Discriminability and Aggregation Strategy

In our feature extraction model, 2 key techniques were used to optimize the quality and representativeness of the feature vectors.

ArcFace loss significantly increases the quality and purity of the feature clusters. This robust separation ensures that when the system queries an image, the retrieved results are demonstrably and reliably from the most similar cluster, thus leading to higher retrieval accuracy and confidence.

GeM Pooling allows the network to intelligently decide where to focus its attention for feature extraction. It enables the model to effectively highlight the most relevant regional features (like specific pathological textures) while retaining sufficient global structural context for comprehensive retrieval, making the resulting feature vector more representative of the entire image.

Indexed Features

We use Elasticsearch as the indexing database due to its native support for high-dimensional vector storage that supports a data type for feature vectors and can perform similarity evaluations on vectors that enable retrieval of data results sorted by similarity. The features are 1024-dimensional embeddings extracted from the GoogLeNet backbone. To enhance searching performance, the indexing phase involves the following steps:

For the core similarity calculation between feature vectors (per image), we adopt the Euclidean Distance defined in formula equation 5. This metric calculates the straight-line geometric distance between 2 feature vectors in the high-dimensional space. The retrieval results are then precisely ranked based on this distance, where a smaller distance signifies higher similarity.

To significantly enhance retrieval efficiency, we used a precalculation strategy. During the indexing phase, we precalculate and store the similarity difference values between each image and all other images in the database. Each image record is uniquely identified by 3 UIDs: Study instance UID, Series instance UID, and SOP instance UID. When a new image is queried, the system calculates its distance to the query and uses this preestablished similarity difference lookup table for comparison, which avoids the traditional method of sequentially searching and calculating every single entry.

It should be noted that this precalculation strategy is primarily designed for intradatabase browsing and retrospective analysis and applies only to images that have already been indexed. For new query images not included in the database, similarity computation is performed on the fly using the Hierarchical Navigable Small World (HNSW) algorithm, which is the default vector search method provided by Elasticsearch. Specifically, when a new query image is submitted, its feature vector is extracted and compared against the indexed feature vectors using HNSW-based approximate nearest neighbor search.

Similarity Calculation

The similarity between feature vectors of different samples is typically measured using distance metrics to gauge the closeness or dissimilarity of each sample. In this study, we used the Euclidean distance, as represented in formula equation 5.

$$(5) D(p,q)=(p1-q1)2+(p2-q2)2+\ldots+(pn-qn)2 Similarity sore(p,q)=1D(p,q)$$

Here, *p* represents the feature vector of the image that the user wishes to query, and *q* represents the feature vector of the image stored in the database. The Euclidean distance measures the real distance between any 2 elements in an *n*-dimensional space. While a smaller Euclidean distance mathematically signifies higher feature-level similarity, we transformed these raw distances into "Similarity Scores" via an inverse distance mapping for display within the CBMIR interface. This ensures that higher scores directly reflect closer visual matches, thereby aligning the computational output with standard clinical assessment logic. This design choice aligns with the study's objectives by leveraging the capacity of deep learning models to encode semantic relationships as geometric distances within a high-dimensional feature space, while ensuring that the retrieval outputs are presented in a format intuitive for clinical practitioners.

Evaluation Factors

Once the CBMIR system is developed, objective evaluation metrics are essential to assess its performance. In this context, mAP and Precision@10 are commonly used evaluation measures [19,21,23]. mAP is extended from precision and average precision (AP). Precision is a fundamental metric and is defined as

$$(6) Precision=Relevant images\frown Retrieved images Retrieved images.$$

Precision is often computed as Precision@k for various values of *k*, where *k* is typically set to 10, 20, or 30. This computation helps determine the performance of the retrieval results when a certain number of images are returned. However, precision provides insight into performance only at a specific rank. AP extends precision by considering all ranks of retrieved images. AP is defined as

$$(7) AP=\sum k=1nPrecision(k)\times rel(k)Relevant images rel(k)=\{1, if image is relevant 0, otherwise$$

Here, *rel*(*k*) is a binary indicator, assuming a value of 1 if the image at rank *k* is relevant and 0 otherwise. AP provides a more comprehensive view of retrieval performance than ordinary precision. However, AP is typically used for a single query; to

XSL•FO
RenderX

assess retrieval performance across multiple queries, mAP is used. mAP is defined as

(8)mAP=∑q=1nAP(q)Q.

Here, *Q* represents the total number of queries, and *q* denotes the current query number. mAP gives an aggregate performance measure across multiple queries, providing a more comprehensive evaluation of the CBMIR system's effectiveness than ordinary precision measures could provide.

Ethical Considerations

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (1) of TVGH (protocol code 021-05-0005CC; approval date: March 14, 2023). Informed consent was waived because of the retrospective nature of the study and the use of deidentified clinical data.

Results

Overview

In this study, the image dataset contains 658 participants with 15,873 images collected from 2000 to 2017. The experimental environment for this study used the following hardware specifications: an AMD Ryzen 3 PRO 4200G CPU, 24 GB of DDR4 memory, and a dedicated Nvidia GeForce RTX 3060 Ti Graphics Card. The graphics processing unit features 4864 CUDA cores and 8 GB of GDDR6 video memory. The operating system used was Linux Ubuntu (version 22.04; Canonical). For software development, the primary Integrated Development Environment used was PyCharm (JetBrains). The programming language was Python (Python Software Foundation), and the deep learning framework used for code development was PyTorch (PyTorch Foundation).

Evaluation of AI Models for Feature Extraction Model

In evaluating the base feature extraction model for the CBMIR system, our core focus was centered on balancing model performance with clinical use and efficiency. To ensure that the system delivers fast, effective retrieval services within a real-world clinical workflow, we conducted a comprehensive comparative analysis of several state-of-the-art AI models when evaluated on the dataset. This comparison was intended to assess how modern, large-scale ViT architectures perform on the dataset relative to established CNNs (Convolutional Neural Networks).

For consistency and to leverage learned feature representations, TL was applied for all models. Each model was initialized with ImageNet pretrained weights and subsequently fine-tuned on our dataset. To explicitly examine the trade-off between model capacity and dataset size, we evaluated both base-scale ViT architectures (Swin-Base and DinoV2-Base) and their lightweight variants (Swin-Tiny and DinoV2-Small). The comparative results incorporating 95% CIs are shown in Table 2.

The results indicate that, under this TL setting and dataset scale, ViT-based models exhibit higher computational complexity and do not achieve superior retrieval performance compared with the CNN-based approach. In contrast, GoogLeNet demonstrates consistent and robust superiority across accuracy, mAP, F_1-score, and computational efficiency. As evidenced by the clearly separated CIs in Table 2, GoogLeNet (95.9% accuracy) substantially outperforms the second-best architectures (upper bound 87.1%), supporting that its performance advantage is stable and unlikely to result from sampling variability. Therefore, within the context of TL on a specific clinical magnetic resonance imaging dataset, GoogLeNet was selected as the backbone feature extractor for the proposed CBMIR system.

XSL•FO
RenderX

Table . Comparative performance and efficiency of different artificial intelligence models for feature extraction[a].

Model	Parameter count (M)	FLOPS[b] (giga)	Validation accuracy (%)	mAP[c]	F_1-score	Training time, minutes	Inference time, seconds
AlexNet	61.1	*0.71* [d]	80.5 (79.4-81.6)	82.8 (81.7-83.9)	81.4 (80.3-82.5)	151	39
VGG16	138.36	15.47	83.4 (82.3-84.5)	86.6 (85.6-87.6)	84.5 (83.5-85.5)	258	47
VGG19	143.67	19.63	86.1 (85.1-87.1)	86.1 (85.1-87.1)	85.4 (84.4-86.4)	278	45
ResNet50	25.56	4.13	83.3 (82.2-84.4)	87.2 (86.2-88.2)	84.6 (83.6-85.6)	258	59
InceptionV4	42.62	6.12	50.3 (48.9-51.7)	76.7 (75.5-77.9)	66.4 (65.0-67.8)	310	99
DinoV2-Base	85.51	21.96	73.3 (72.0-74.6)	76.3 (75.1-77.5)	73.8 (72.5-75.1)	502	60
Swin-Base	87.7	15.17	86.1 (85.1-87.1)	86.8 (85.8-87.8)	86.4 (85.4-87.4)	494	112
DinoV2-Small	21.52	5.52	70.4 (69.1-71.7)	73.9 (72.6-75.2)	71.1 (69.8-72.4)	182	59
Swin-Tiny	28.27	4.37	60.3 (58.9-61.7)	76.7 (75.5-77.9)	58.8 (57.4-60.2)	175	35
GoogLeNet	*6.8* [d]	1.51	*95.9* (*95.3-96.5*)[d]	*89.2* (*88.3-90.1*)[d]	*95.1* (*94.5-95.7*)[d]	*57* [d]	*22* [d]

[a]Performance metrics are reported as mean (95% CI).

[b]FLOPs: Floating Point Operations.

[c]mAP: mean average precision.

[d]The values in italics indicate the best performance or efficiency in the comparative results.

Performance Evaluation for Feature Extraction Model

The feature extraction model underwent training for 150,000 iterations. To evaluate its performance, feature extraction was first applied to all images in both the training and test datasets. Next, the feature vectors from the test dataset were individually compared and ranked against those from the training dataset to measure the similarity between the respective images. Notably, an image was considered relevant (with rel(k)=1 in AP) if its category matched that of the query image.

A summary of the evaluation metrics, including mAP and Precision@10, is shown in Table 3 for each image class. These metrics assess the performance and retrieval quality of the model for each image class as well as overall. The overall mAP for the study was 89.16%, with a Precision@10 of 94.08%. The best performance was observed in the normal brain category, achieving a mAP of 94.95% (95% CI 94.32 - 95.58) and a Precision@10 of 97.57% (95% CI 97.13 - 98.01), indicating high reliability in identifying normal tissues.

The results show that the overall classification performance of this model was excellent, but its mAP for CBMIR was only 89.16%. This disparity highlights the difference between classification and image retrieval in practical terms. In classification, the main training objective of the AI model is to identify the tumor class to which the current image belongs. However, in image retrieval, the training objective must consider the similarity (distance relationship) between the current image and other images to learn the relationships between the features of each image in the pair of images. These are 2 conceptually distinct tasks.

XSL•FO
RenderX

Table . Evaluation results by tumor class[a].

Tumor class	mAP[b] (%)	Precision@10 (%)
Metastasis	88.01 (87.08-88.94)	92.47 (91.71-93.23)
Glioma	88.33 (87.41-89.25)	92.70 (91.95-93.45)
Meningioma	90.94 (90.12-91.76)	96.20 (95.65-96.75)
Pituitary adenoma	87.76 (86.82-88.70)	98.01 (97.61-98.41)
Vestibular schwannoma	87.89 (86.95-88.83)	94.51 (93.86-95.16)
No tumor contour	87.39 (86.44-88.34)	89.31 (88.42-90.20)
Normal brain	94.95 (94.32-95.58)	97.57 (97.13-98.01)
Overall	89.16 (88.27-90.05)	94.08 (93.40-94.76)

[a]Performance metrics are reported as mean (95% CI).

[b]mAP: mean average precision.

Performance Evaluation for Tumor Classification

The main purpose of this study was to use an AI model to extract feature vectors for image retrieval. To realize this purpose, we chose to extend the ArcFace loss function, which is based on the softmax loss function, using a margin of 38 and a scale of 64, as the classification for a model that can theoretically be used for image classification. Therefore, this study also investigated the performance of the AI model for classification tasks. Table 4 shows the precision, recall, and F_1-score for each class, while Figure 3 provides the confusion matrix for each tumor class.

Table . Classification performance evaluation for each tumor class.

Tumor class	Precision, %	Recall, %	F_1-score, %
Metastasis	91.75	94.01	92.87
Glioma	98.17	93.36	95.71
Meningioma	99.02	96.36	97.67
Pituitary adenoma	99.68	97.93	98.80
Vestibular schwannoma	94.62	95.50	95.06
No tumor contour	88.79	89.59	89.19
Normal brain	93.72	98.31	95.96
Overall	95.11	95.01	95.04

XSL•FO
RenderX

Figure 3. Confusion matrix for classes of brain tumors.

Performance Comparison at Various Embedding Dimensions

In our study, ArcFace loss was used not merely for classification but as a feature embedding mechanism to ensure that images belonging to the same tumor class are mapped closer together in the embedding space, while those from different classes are more widely separated. This results in well-clustered and discriminative feature vectors, which form the necessary basis for our similarity-based retrieval. Hence, although these are distinct tasks, they are synergistically integrated within the same CBMIR system: classification helps structure the embedding space, thereby significantly enhancing the discriminative power and ultimately the retrieval performance of the system.

We initially drew inspiration from a study [37] in which the optimal embedding dimension for their specific objective of facial recognition was determined to be 128. However, we recognize that this dimension may not necessarily be optimal for brain tumor classification. Consequently, we undertook a comprehensive comparison of various embedding dimensions, as outlined in Table 5. The findings of this analysis unequivocally indicated that 256 dimensions yielded the highest performance for our retrieval task.

Furthermore, we note that the disparities in performance across dimensions, ranging from 64 to 512, were relatively marginal. This observation is particularly significant when considering the cost-effectiveness of deploying the model on devices with limited computational capabilities. Opting for a dimensionality of 64 can substantially curtail computational overheads while still maintaining respectable performance levels.

XSL•FO
RenderX

Table . Performance comparison of different embedding dimensions.

Embedding size	Overall mAP, %
64	88.16
128	88.76
256	89.16[a]
512	89.14

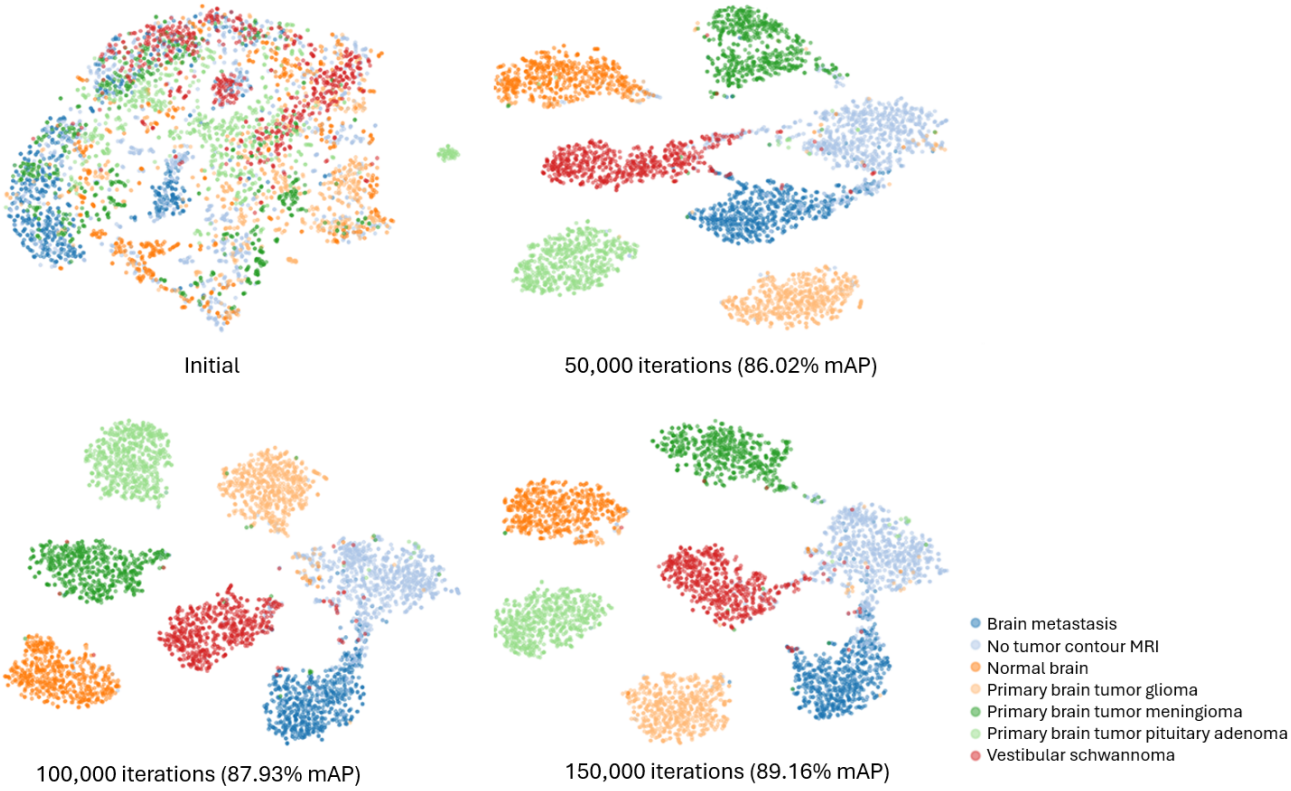
[a]Indicates the best performance in different various embedding dimensions.

Feature Vector Projection

Our methodology consolidated similar feature vectors while actively distancing dissimilar ones. While mAP serves as a useful indirect metric for assessing effectiveness, the feature vector coherence can be qualitatively assessed through projection. To qualitatively visualize the evolution of feature representations during training, we used t-distributed Stochastic Neighbor Embedding to project high-dimensional feature vectors into a lower-dimensional space. As a stochastic dimensionality reduction method, t-distributed Stochastic Neighbor Embedding is sensitive to parameter settings and is used here solely for qualitative inspection of learning trends rather than as quantitative evidence of feature space cohesion or class separability. The visual representation of feature vector projection is illustrated in Figure 4.

Figure 4 illustrates that the initially untrained feature vectors, denoted as "Initial," exhibit a scattered pattern with minimal cohesion among different classes. However, as training progresses through 50,000, 100,000, and 150,000 iterations, we observe a gradual convergence of feature vectors within each class, reflecting increasing cohesion. Concurrently, an observable qualitative visual separation implies that feature vectors from different classes become more distantly grouped, highlighting the model's progressive learning and adaptability.

Figure 4. Visual representation of feature vector projection using t-distributed Stochastic Neighbor Embedding across various training iterations. The projection was implemented with 3 components, a perplexity of 30, 1000 iterations, and the Euclidean metric. This visualization is intended to illustrate trends in feature learning rather than to provide quantitative evidence of class separation. mAP, mean average precision; MRI, magnetic resonance imaging.



System Integration With PACS

Our system operates as follows. Users can initiate image queries from the PACS and subsequently retrieve and display these images within BlueLight via the QIDO (query based on ID for DICOM Objects) and WADO (web access to DICOM objects) protocols, respectively. To use the CBMIR functionality, users provide the StudyInstanceUID, SeriesInstanceUID, and SOPInstanceUID for the specific DICOM image they require.

Once these details are specified, the CBMIR system retrieves the corresponding DICOM image from the PACS using WADO and then processes this image to calculate the CBMIR results, returning the top 10 images that bear the highest similarity to

XSL•FO
RenderX

the input image. The operational process and data workflow are depicted in Figure 5.

A demonstration video is provided in Multimedia Appendix 1, and a screenshot of the user interface of the CBMIR client prior to executing a search is shown in Figure 6A. To enhance the successful integration of our CBMIR system into clinical workflows, we integrated it with an open-source DICOM viewer known as BlueLight [34] and the DICOM server Raccoon [38]. This client application communicated with a PACS-supporting DICOMweb, significantly simplifying user access to and interpretation of brain tumor images, including the display of top retrieval results ordered by similarity scores (Figure 6B). Moreover, considering that brain images are inherently 3D, a single image may not always convey the complete information for a whole series of images. The system's workflow ensures that the entire series of images associated with a case is retrieved from the PACS, rather than just a single instance (Figure 6C).

Figure 5. Sequence diagram for the integration of content-based medical image retrieval with a picture archiving and communication system. AI: artificial intelligence; CBMIR: content-based medical image retrieval; DICOM: Digital Imaging and Communications in Medicine; PACS: picture archiving and communication system; UIDs: unique identifiers.

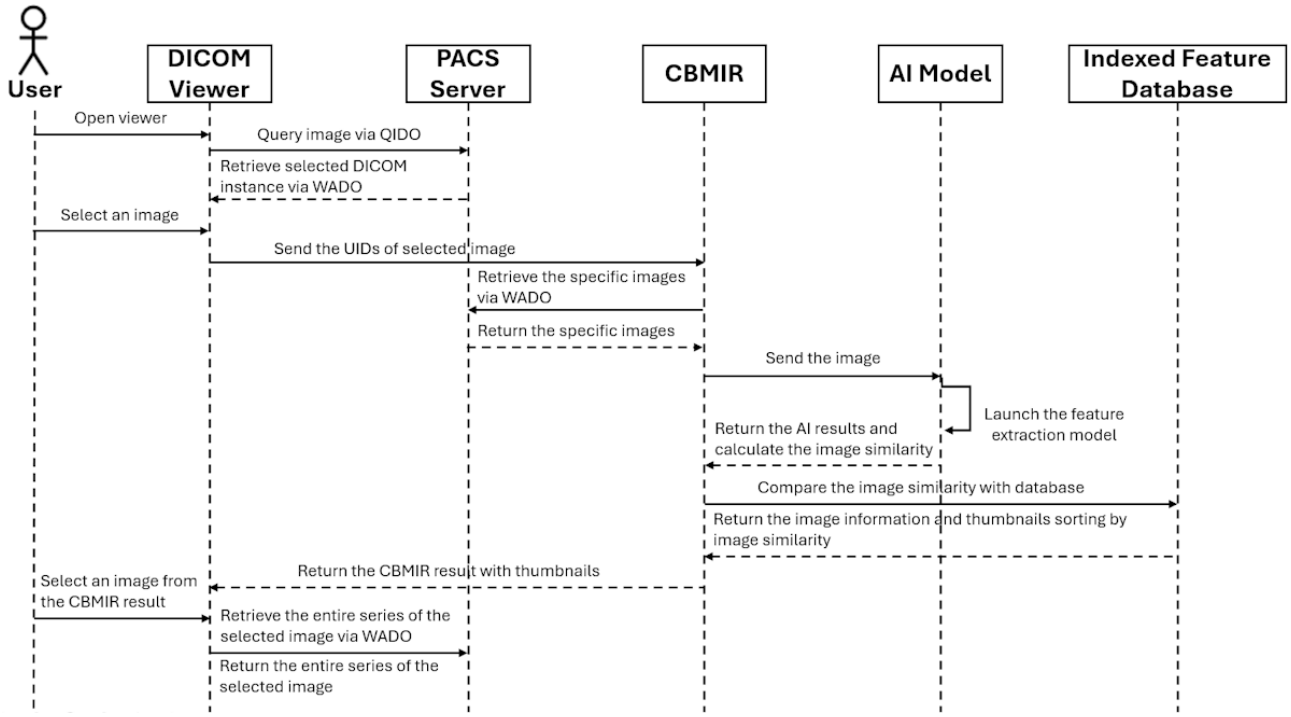
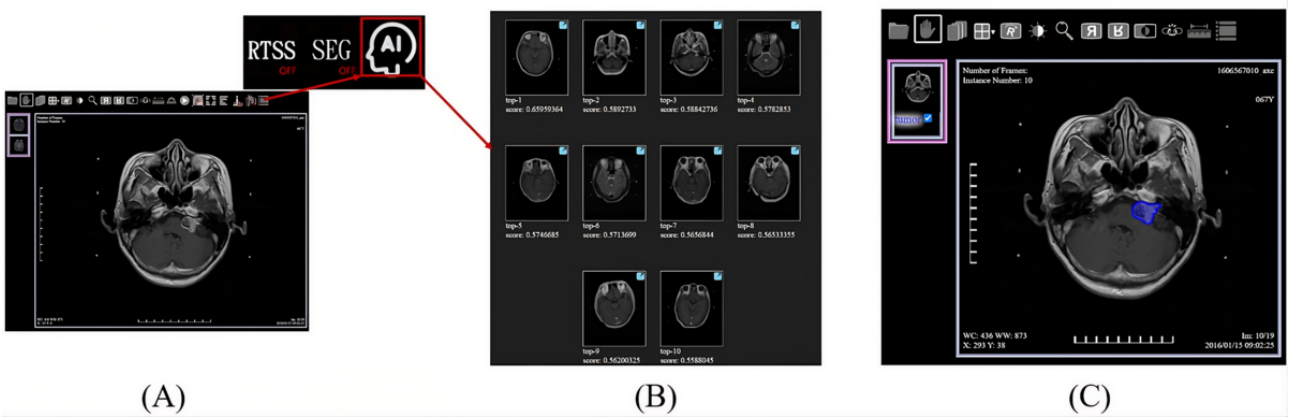


Figure 6. The user interface of the content-based medical image retrieval (CBMIR) services is integrated into a web Digital Imaging and Communications in Medicine viewer. (A) Before performing the CBMIR system; (B) Top 10 retrieval results ordered by descending similarity scores. A higher score denotes greater visual similarity between the query and the retrieved images; (C) Retrieve the selected whole series of images from picture archiving and communication system.



Performance of CBMIR System

During the usage of the CBMIR system, the "time" and "stability" of querying for similar images can significantly impact the overall operational process. Therefore, this study conducted a performance evaluation of the CBMIR system. The response time represents the end-to-end round trip for image retrieval, including WADO communication, database execution, and network transfer latency. The overall response time evaluation involved the following steps: (1) sending image to the CBMIR system, (2) retrieving images from PACS, (3) extracting image feature vectors using the CBMIR AI model,

XSL•FO
RenderX

(4) indexing image feature vectors into the database, (5) querying for similar images within the database, and (6) assessing the response time back to the user.

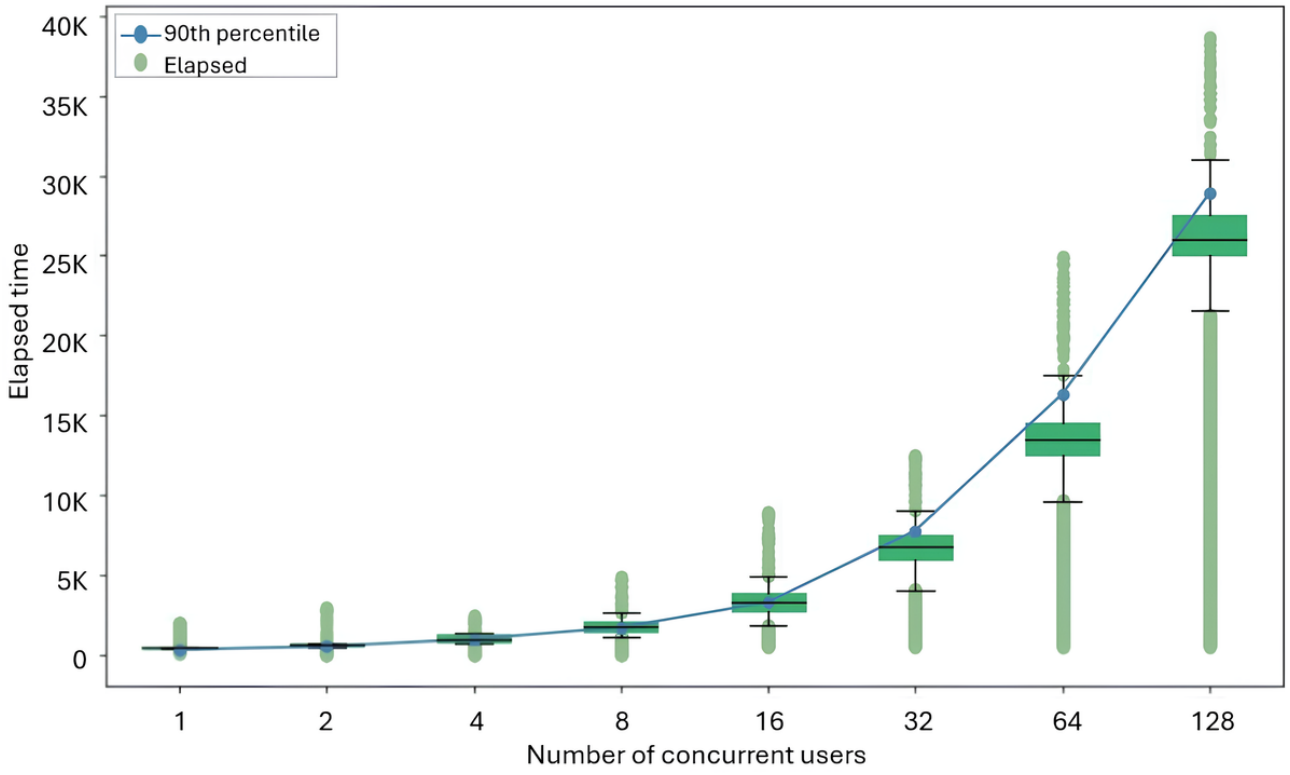
The stability assessment involved simulations with 2^*n* users, ranging up to 256 concurrent users, using the CBMIR system. The evaluation results are shown in Table 6; a box-and-whisker plot depicting the evaluation results is shown in Figure 7.

Additionally, as no system crashes or errors were observed for user counts ranging from 1 to 128, this study conducted further testing, revealing that when more than 256 users simultaneously used the system, all requests began to experience 30-second time-outs, the default value of request time-out for all web browsers, thereby all requests failed. The results demonstrate that as the number of concurrent users increases, the response time exhibits a nonlinear growth. Specifically, under extreme stress testing conditions (exceeding 64 simultaneous users), the variability in response time expands significantly. When the concurrent load surpasses 256 users, the system reaches its current architectural capacity limit, resulting in request time-outs (exceeding 30 seconds). These findings delineate the operational boundaries of the current implementation, indicating that while the system is stable for routine, low-concurrency clinical usage, high-concurrency scenarios impose a significant load that impacts real-time stability.

Table . Response time results under various user loads.

Users	Response time, ms		
	Average	Maximum	90th percentile
1	273.98	1373	333.00
2	434.48	2696	531.00
4	918.89	2233	1187.00
8	1758.38	4423	2219.00
16	3225.49	8230	3903.10
32	6432.68	12,541	7775.40
64	13629.18	24,391	16337.40
128	25672.47	39,196	28909.10

Figure 7. Relationship between the elapsed response time and the number of concurrent users in the content-based medical image retrieval system evaluation. The green box-and-whisker plots and scatter points display the statistical distribution of the response time (median, quartiles, and extreme values), while the blue line graph indicates the 90th percentile response time.



XSL•FO
RenderX

Discussion

Principal Findings

We developed and evaluated a CBMIR system tailored for brain tumor detection within a PACS environment. The results demonstrate a substantial improvement in retrieval accuracy enabled by the innovative deep learning architecture, enhancing feature discriminability. By integrating our AI model with 2 open-source tools—Raccoon for backend DICOM management and BlueLight for frontend web-based image viewing—our CBMIR system provides seamless, real-time decision support directly within the clinician's workflow. This practical integration helps bridge the gap between advanced AI retrieval techniques and routine clinical practice. This achievement represents an important step toward establishing the CBMIR system as a valuable tool for both clinical diagnosis support and medical research.

Rationale for GoogLeNet Selection

We selected GoogLeNet as the backbone feature extractor for our CBMIR system due to its demonstrated optimal balance between clinical use and overall performance.

1. *Performance versus computational cost*: As Table 2 illustrates, while ViT-based and deeper CNN architectures are state of the art in many domains, their performance metrics on our specific medical imaging task are significantly lower than those achieved by GoogLeNet. This discrepancy is likely influenced by the interaction between model complexity and the relatively homogeneous characteristics of the dataset. In contrast, GoogLeNet's moderate depth and multiscale receptive fields appear to facilitate more stable embedding learning and improved generalization in similarity-based retrieval scenarios.
2. *Overwhelming computational advantage*: GoogLeNet demonstrates a massive advantage in efficiency. Its Floating Point Operations are nearly one-tenth that of Swin (1.51 G vs 15.17 G), and its parameter count is approximately one-tenth that of the ViT models (6.8 M vs ≈85 M). Furthermore, its training time is significantly shorter (57 minutes vs ≈500 minutes).
3. *Critical clinical deployment consideration*: Given that a CBMIR system must be lightweight and support low-latency retrieval for real-time clinical use, efficiency is paramount. GoogLeNet provides a high mAP of 89.16% while maintaining minimal computational demands. Therefore, GoogLeNet was chosen as it achieves the best balance among accuracy, model compactness, and deployment efficiency, making it the superior foundation for our CBMIR architecture.

Heatmaps of Image Similarity

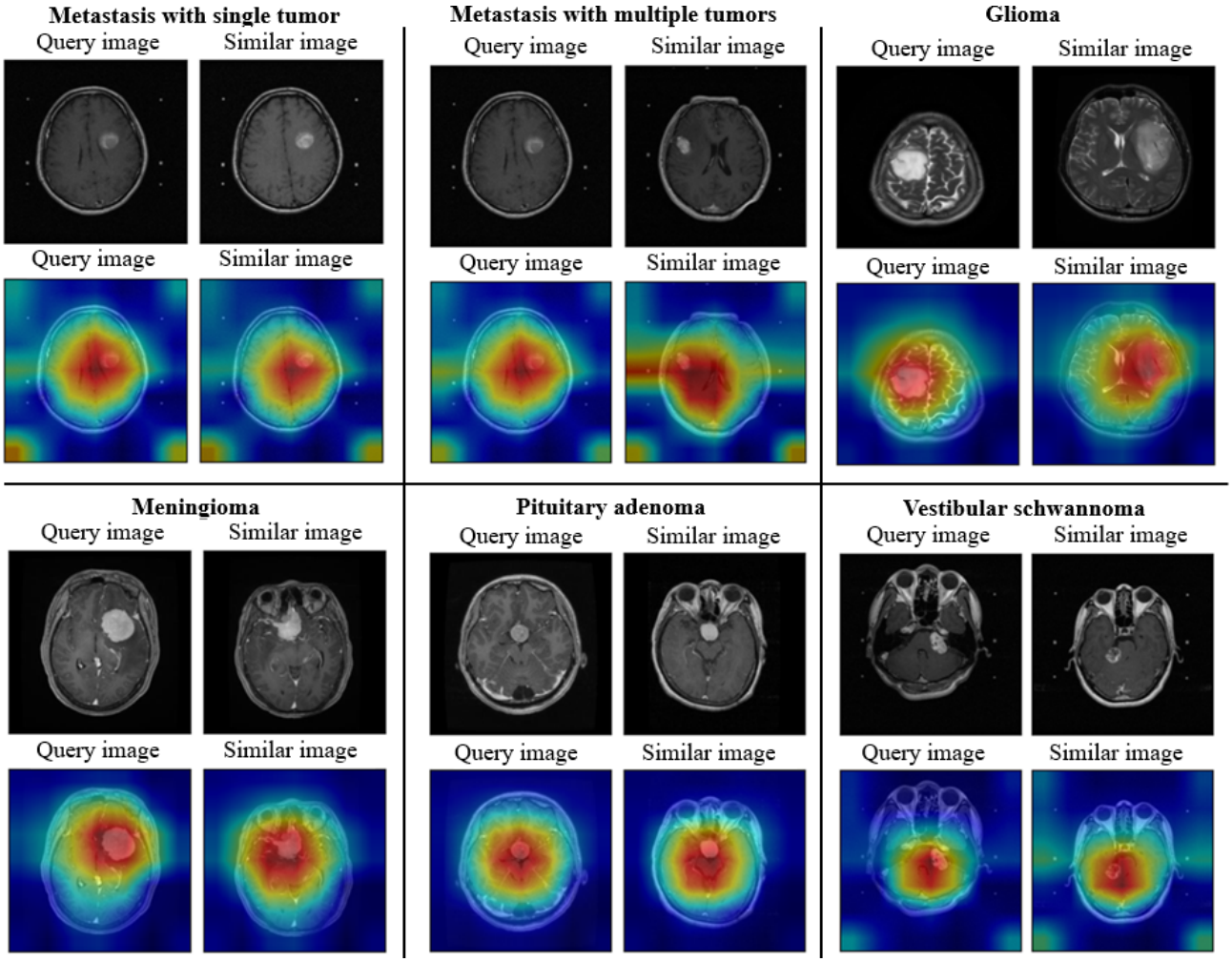
To provide interpretability for our CBMIR model's retrieval decisions, we used a visualization technique based on Gradient-weighted Class Activation Mapping, adapted for deep similarity networks following the approach introduced by Stylianou et al [39].

Our heatmaps aim to quantify the degree of feature matching between 2 images. Our methodology integrates Gradient-weighted Class Activation Mapping with the calculation of distances between the final convolutional layer and the pooling layer to gauge the level of attention or similarity across different regions within the images. In the visualization, regions exhibiting high similarity are represented in red, while those with lower similarity are depicted in blue. Figure 8 shows 5 examples of brain tumors, illustrating our primary goal of identifying images of brain tumors that share similarities. The ensuing sections showcase detailed heatmaps for each of the tumor classes.

For all tumor classes presented in Figure 8, except for the case of multiple tumors in metastasis, the examples used contain only a single tumor within the image. This choice was deliberate to ensure that the heatmaps clearly illustrate that retrieval similarity is based on matching the pathological features of the tumor (where the hot spot is located), rather than simply reflecting class probabilities as in a classification task. The high feature consistency observed between the query and retrieved images across these single-lesion cases strongly supports this interpretation. However, it is essential to note that while this analysis explains the underlying mechanism, the study does not draw definitive clinical conclusions regarding these results. It should be emphasized that these heatmaps are presented for qualitative illustration only. Determining whether such activation patterns correspond to meaningful anatomical features or to nonanatomical factors would require systematic image-by-image analysis, which is beyond the scope of this study.

XSL•FO
RenderX

Figure 8. Heatmaps of similar images for various tumor classes: query image and retrieved similar image with corresponding heatmaps.



Single Versus Multiple Tumors (Metastasis Cases)

We included metastasis cases with both single and multiple tumors within a single image. The heatmaps illustrate visually different activation patterns between these 2 scenarios. However, the primary focus of our analysis remains the observed differences in image similarity behavior between metastasis cases with multiple tumors and single-tumor cases, rather than the anatomical interpretation of activation maps.

1. *Single tumor*: In single-tumor cases, the query image heatmap shows a visually concentrated activation region (red or yellow) localized around the dominant lesion. The corresponding retrieved image exhibits a similarly localized activation pattern. This observation reflects a spatially focused activation pattern in both the query and retrieved images.
2. *Multiple tumors*: In multiple-tumor cases, the activation patterns appear more spatially distributed. While high-activation regions remain present near individual lesions, the heatmaps often extend into broader areas of the image. The similarity image heatmap exhibits horizontally distributed activation patterns or band-like structures. These patterns indicate a more dispersed activation distribution in images containing multiple lesions.

Clinical Considerations for Cross-Class Image Similarity

Our heatmap analysis differs from traditional methods, which often focus only on areas of interest within a single image for classification. Our objective is to describe and quantify the similarity between 2 images in terms of mutual focus on specific regions, an understanding pivotal for elucidating the model's retrieval effectiveness. Considering clinical evaluation, 2 radiologists (KHK and WYG) used the hospital's DICOM viewer to cross-validate brain tumor cases and assess the correctness of similarity scores between different tumor classes.

We acknowledge that highly similar images from the query results may not always belong to the same tumor class. Indeed, in some instances, the similarity score may be substantially higher for images assigned to other tumor classes. These cases indicate a significant degree of visual and structural image similarity, even if the retrieved image's classification label may not be totally accurate. The similarity score is calculated for the entire image. While a partial score calculated from the tumor lesion may result in a high score, it is worth noting that when 2 images share a similar location in an axial brain slice, the similarity of surrounding organs and structures may yield a higher final score than that of the tumor lesion alone.

This result shows a clinical potential use in identifying shared characteristics among images across tumor classes. However, it must be emphasized that definitive conclusions regarding

XSL•FO
RenderX

these results were not drawn in this study. To mitigate this potential disparity, the user interface has been thoughtfully crafted to label these images accordingly, providing a clear notification to the user regarding this outcome.

Normal Brain Versus No Tumor Contour

From a clinical perspective, the "Normal Brain" class represents images from patients without a brain tumor diagnosis, whereas the "No Tumor Contour" class consists of image slices acquired from patients with known brain tumors in which no visible lesion is present on the specific slice. Although these 2 classes may appear visually similar at the slice level, they differ fundamentally at the case level, which explains the observed confusion in similarity-based retrieval. Specifically, while the "Normal Brain" class achieves a high mAP of 94.95%, the "No Tumor Contour" class yields a lower mAP of 87.39%, which is below the overall average of 89.16%.

This confusion reflects a realistic diagnostic scenario in routine clinical practice, where individual slices without visible lesions may be indistinguishable from normal brain images. This observation highlights the system's ability to handle visually overlapping categories in routine clinical practice. The inclusion of these classes underscores the system's ability to handle visual overlaps based on deep feature embeddings learned from pixel-level representations, rather than pathology-specific descriptors. Consequently, the reported mAP reflects the global coherence of the learned visual embedding space across diverse clinical categories.

It should be noted that this behavior does not imply a false-negative diagnostic result, as the proposed CBMIR system is designed for similarity-based image retrieval rather than automated diagnosis. Instead, it underscores the necessity of contextual interpretation at the case or series level when reviewing retrieved images. To mitigate potential clinical misunderstanding, the system retrieves the complete image series rather than a single slice, allowing clinicians to examine adjacent slices where lesions may be present. The interface further indicates that tumor annotations exist elsewhere within the retrieved series.

Limitations

While this study successfully developed a high-performing CBMIR system and integrated it into PACS, there are limitations regarding dataset characteristics, system scalability, and model interpretability. First, regarding the dataset characteristics and partitioning, this study used a dataset spanning a long retrospective period (2000 - 2017) from a single medical center, acquired primarily on 1.5T scanners. Although this provided consistency in clinical protocols and image quality, it may limit the model's generalizability. In particular, robustness across different scanner manufacturers, field strengths (eg, 3T), and heterogeneous acquisition protocols remains to be systematically validated. Given the increasing prevalence of 3T scanning in modern neuro-oncology, the model's performance on higher-resolution and advanced acquisition protocols warrants further investigation, especially in multicenter clinical settings.

To ensure balanced class distributions for benchmarking, we used a random split strategy. Unlike a prospective temporal split, this approach introduces a potential for "future leakage" and may not fully simulate clinical deployment timelines. A temporal split, in which earlier data are used for training and later data for testing, may better reflect real-world clinical deployment scenarios and provide a more stringent evaluation of generalization across time. However, such a design requires strict longitudinal harmonization of acquisition protocols, and scanner variations could provide a more stringent assessment of real-world generalization and will be a potential improvement in future studies.

Second, regarding system latency and scalability, although GoogLeNet was selected for its relatively low computational cost, the processes of calculating feature vectors and similarity computation create a computational bottleneck under heavy workloads. Our stress testing demonstrates that response time increases substantially under high concurrency, exceeding 13 seconds for 64 simultaneous users and approaching 26 seconds for 128 users. Such latency is incompatible with real-time clinical workflows and therefore represents a critical limitation of the current implementation.

Third, regarding the analysis of metastasis cases with multiple tumors, we observed distinct horizontal bands of high activation in the heatmaps. While these patterns may suggest distributed feature attention, we cannot definitively rule out the possibility that they represent acquisition or reconstruction artifacts (eg, motion or noise). Consequently, this ambiguity serves as a limitation regarding the model's focus and explainability in complex, multilesion scenarios.

Fourth, to accommodate the input requirements of the pretrained GoogLeNet architecture, the original 512×512 pixels images were resized to 224×224 pixels before feature extraction. This downsampling inevitably reduces spatial resolution and may obscure fine-grained texture details that could be relevant for differentiating certain tumor subtypes. While the loss of fine-grained texture information may affect the representation of subtle local patterns, the system is designed to capture global and midlevel feature similarities that are sufficient for retrieval tasks within the intended clinical workflow. Nonetheless, this resolution reduction is a limitation of the current implementation.

Fifth, similarity in our CBMIR system does not rely on predefined morphological or pathological criteria (eg, tumor size, location, and texture). Instead, it is quantified by computing the Euclidean distance between high-dimensional deep embeddings derived directly from pixel-level representations. Consequently, similarity ranking is fundamentally driven by pixel-derived feature distances, rather than by pathological labels or manually defined structural attributes.

Finally, the retrieval performance metrics reported in this study were computed using Euclidean distance across the full test set to provide standardized and reproducible benchmarking. In contrast, the deployed CBMIR system uses an approximate nearest neighbor search strategy based on HNSW graphs to ensure low-latency, real-time responses and system scalability. The use of HNSW enables efficient similarity search for unseen queries without requiring their prior inclusion in a precalculated distance matrix. While this approximate nature may lead to

XSL•FO
RenderX

minor deviations in retrieval precision or recall compared with an exhaustive search, such a trade-off is essential for practical clinical usability, effectively balancing theoretical accuracy with the demands of real-time medical decision support.

Future Work

Building upon the current findings and addressing the identified limitations, we propose the following three main directions for future research:

1. *Expansion to multicenter dataset validation*: Future work will prioritize incorporating image data from diverse health care institutions to test and fine-tune the model's robustness. Training and validating the system on multicenter datasets will significantly enhance the model's generalizability and practical usefulness.
2. *Exploration of modern AI model*: Although preliminary experiments showed modern AI models, for example, Swin and DinoV2, to be less computationally efficient for our specific task, their inherent strength in capturing long-range dependencies warrants further investigation. Future research may explore methods to optimize or prune these AI models to harness their superior feature learning capabilities without unduly sacrificing computational efficiency.
3. *Integration of multimodal retrieval*: The current system relies solely on visual features. In clinical practice, structured and unstructured radiology reports provide critical contextual information. Future work may explore multimodal deep learning models to combine image features with textual features, enabling more comprehensive and accurate multimodal retrieval functionality.

Conclusions

This study represents the design, implementation, and evaluation of a CBMIR tailored for the classification and retrieval of brain MR images. We successfully developed a CBMIR system focusing on structured tumor classification and incorporating deep learning–based feature extraction, including GoogLeNet combined with ArcFace Loss to enhance the discrimination of image similarities and differences, thereby improving retrieval relevance. In addition, GeM pooling was used as the feature aggregation strategy and evaluated in comparison with conventional average pooling and maximum pooling methods. These results indicate that GeM pooling provides a more refined representative embedding of image features, contributing to improved retrieval accuracy and robustness in the CBMIR system. To facilitate practical deployment, the training workflow was designed as a streamlined, single-stage process, reducing complexity and enabling efficient model preparation. The system was trained using a dataset that includes both tumor-positive and tumor-negative brain MR images, reflecting the diversity of cases encountered in routine clinical practice. Finally, the proposed CBMIR system was successfully integrated with the open-source web DICOM viewer BlueLight with a PACS. The retrieved results were able to be communicated with a PACS to retrieve corresponding medical images for display in a web DICOM viewer, demonstrating the feasibility of integrating CBMIR functionality into existing clinical imaging workflows.

Acknowledgments

The authors acknowledge Dr. Kuei-Hong Kuo (Far Eastern Memorial Hospital) for the feedback of the design of the user interface, user interaction, and validation of the results while developing the software.

Generative artificial intelligence was not used in any portion of the manuscript writing.

Funding

This study was funded on November 1, 2022. This work was supported by the Taiwan's National Science and Technology Council (NSTC 114-2221-E-227-001-MY2, NSTC 114-2321-B-A49-018, 114-2634-F-A49-006, and NSTC 112-2221-E-227-001).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical considerations and restrictions imposed by the institutional review board, but the data could be accessed via data use agreements and a new application to the institutional review board (1) of Taipei Veterans General Hospital.

Authors' Contributions

CYL and WCC participated in conceptualization. CYL, CLL, THH, and YTW contributed to methodology. CLL contributed to software. CYL, CLL, YTW, WYG, and WCC participated in validation. WYG participated in data curation. CYL, CLL, and THH participated in Writing – original draft. CYL, THH, and WCC participated in writing – review and editing. CYL participated in funding acquisition. All authors revised and approved the final version of the manuscript, and all take responsibility for its accuracy.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Demonstration of the content-based medical image retrieval (CBMIR) system integration. This video illustrates the complete workflow of the CBMIR system integrated with the BlueLight DICOM viewer and picture archiving and communication system. The user interface shown in this demonstration displays the raw Euclidean distance as the “score” (where lower values indicate higher similarity).
[MP4 File, 9712 KB - medinform_v14i1e78300_app1.mp4]

References

1. Sotomayor CG, Mendoza M, Castañeda V, et al. Content-based medical image retrieval and intelligent interactive visual browser for medical education, research and care. Diagnostics (Basel) 2021 Aug 13;11(8):1470. [doi: 10.3390/diagnostics11081470] [Medline: 34441404]
2. Serieys G, Kurtz C, Fournier L, Cloppet F. Text-guided visual representation learning for medical image retrieval systems. In: Cloppet F, editor. Presented at: 2022 26th International Conference on Pattern Recognition (ICPR); Aug 21-25, 2022. [doi: 10.1109/ICPR56361.2022.9956402]
3. Welter P, Deserno TM, Fischer B, Wein BB, Ott B, Günther RW. Integration of CBIR in radiological routine in accordance with IHE. 2009 Feb 26 Presented at: SPIE Medical Imaging; Mar 13, 2009. [doi: 10.1117/12.812100]
4. Röhrich S, Heidinger BH, Prayer F, et al. Impact of a content-based image retrieval system on the interpretation of chest CTs of patients with diffuse parenchymal lung disease. Eur Radiol 2023 Jan;33(1):360-367. [doi: 10.1007/s00330-022-08973-3] [Medline: 35779087]
5. Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform 2004 Feb;73(1):1-23. [doi: 10.1016/j.ijmedinf.2003.11.024] [Medline: 15036075]
6. Shamna P, Govindan VK, Abdul Nazeer KA. Content-based medical image retrieval by spatial matching of visual words. Journal of King Saud University - Computer and Information Sciences 2022 Feb;34(2):58-71. [doi: 10.1016/j.jksuci.2018.10.002]
7. Sascha S, Marisa T, Florian S, Matthias H, Martin K, et al. Combined semantic and similarity search in medical image databases. : ProcSPIE; 2011 Presented at: Proceedings of the SPIE; Feb 25, 2011; Lake Buena Vista (Orlando), Florida. [doi: 10.1117/12.878179]
8. Pilevar AH. CBMIR: content-based image retrieval algorithm for medical image databases. J Med Signals Sens 2011 Jan;1(1):12-18. [doi: 10.4103/2228-7477.83460] [Medline: 22606654]
9. Latif A, Rasheed A, Sajid U, et al. Content ‐ based image retrieval and feature extraction: a comprehensive review. Math Probl Eng 2019 Jan;2019(1):9658350. [doi: 10.1155/2019/9658350]
10. Arakeri MP, Ram Mohana Reddy G. An intelligent content-based image retrieval system for clinical decision support in brain tumor diagnosis. Int J Multimed Info Retr 2013 Sep;2(3):175-188. [doi: 10.1007/s13735-013-0037-5]
11. Ponciano-Silva M, Bedo MVN, Pereira Dos Santos D, de Azevedo-Marques PM, Ferreira de Carvalho ADL, Traina C Jr. Endowing a content-based medical image retrieval system with perceptual similarity using ensemble strategy. J Digit Imaging 2016 Feb;29(1):22-37. [doi: 10.1007/s10278-015-9809-1] [Medline: 26259520]
12. Maji S, Bose S. CBIR Using Features Derived by Deep Learning. ACM/IMS Trans Data Sci 2021 Aug 31;2(3):1-24. [doi: 10.1145/3470568]
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) . [doi: 10.1109/CVPR.2017.243]
14. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. AAAI 2017;31(1). [doi: 10.1609/aaai.v31i1.11231]
15. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. 2018 Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-23, 2018. [doi: 10.1109/CVPR.2018.00474]
16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Sun J, editor. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016. [doi: 10.1109/CVPR.2016.90]
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015 Presented at: 3rd International Conference on Learning Representations (ICLR); May 7-9, 2015 URL: https://arxiv.org/pdf/1409.1556.pdf [accessed 2026-03-25]
18. Rashad M, Afifi I, Abdelfatah M. RbQE: an efficient method for content-based medical image retrieval based on query expansion. J Digit Imaging 2023 Jun;36(3):1248-1261. [doi: 10.1007/s10278-022-00769-7] [Medline: 36702987]
19. Cheng J, Yang W, Huang M, et al. Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. PLoS One 2016;11(6):e0157112. [doi: 10.1371/journal.pone.0157112]
20. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. Neurocomputing 2017 Nov;266:8-20. [doi: 10.1016/j.neucom.2017.05.025]
21. Swati ZNK, Zhao Q, Kabir M, et al. Content-based brain tumor retrieval for MR images using transfer learning. IEEE Access 2019;7:17809-17822. [doi: 10.1109/ACCESS.2019.2892455]

XSL•FO
RenderX

22. Cheng J. Brain tumor dataset. Figshare. 2017. URL: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427 [accessed 2026-04-06]
23. Deepak S, Ameer PM. Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings. Comput Biol Med 2020 Oct;125:103993. [doi: 10.1016/j.compbiomed.2020.103993]
24. M VRS, Gupta S, Aziz AA. Advancing brain tumor classification through pre-trained transformer and transfer learning models. Franklin Open 2026 Mar;14:100493. [doi: 10.1016/j.fraope.2026.100493]
25. Anupama N, Anitha J, editors. Revolutionizing medical diagnosis: a deep learning approach for content-based image retrieval. 2025 Presented at: 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI); Apr 4-6, 2025. [doi: 10.1109/ICCSAI64074.2025.11064005]
26. Sunday PN, Amagasaki M, Kiyama M, Diana M. Medical image similarity matching using deep learning. In: Diana M, editor. 2025 Presented at: 2025 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC); Jul 7-10, 2025. [doi: 10.1109/ITC-CSCC66376.2025.11137720]
27. Sarıateş M, Özbay E. Transfer learning-based ensemble of CNNs and vision transformers for accurate melanoma diagnosis and image retrieval. Diagnostics (Basel) 2025 Jul 31;15(15):1928. [doi: 10.3390/diagnostics15151928] [Medline: 40804891]
28. Valente F, Silva LAB, Godinho TM, Costa C. Anatomy of an extensible open source PACS. J Digit Imaging 2016 Jun;29(3):284-296. [doi: 10.1007/s10278-015-9834-0] [Medline: 26497879]
29. Welter P, Fischer B, Günther RW, Deserno né Lehmann TM. Generic integration of content-based image retrieval in computer-aided diagnosis. Comput Methods Programs Biomed 2012 Nov;108(2):589-599. [doi: 10.1016/j.cmpb.2011.08.010] [Medline: 21975083]
30. Frederico V, Carlos C, Augusto S. Content based retrieval systems in a clinical context. In: Okechukwu Felix E, editor. Medical Imaging in Clinical Practice: IntechOpen; 2013. [doi: 10.5772/53027]
31. Gravina M, Marrone S, Piantadosi G, Moscato V, Sansone C. Developing a smart PACS: CBIR system using deep learning. Presented at: Pattern Recognition ICPR International Workshops and Challenges: Virtual Event Proceedings, Part II; Jan 10-15, 2021. [doi: 10.1007/978-3-030-68790-8_24]
32. Valente F, Costa C, Silva A. Dicoogle, a PACS featuring profiled content based image retrieval. PLoS One 2013;8(5):e61888. [doi: 10.1371/journal.pone.0061888] [Medline: 23671578]
33. Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. In: Zafeiriou S, editor. Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019. [doi: 10.1109/CVPR.2019.00482]
34. Chen TT, Sun YC, Chu WC, Lien CY. BlueLight: an open source DICOM viewer using low-cost computation algorithm implemented with JavaScript using advanced medical imaging visualization. J Digit Imaging 2023 Apr;36(2):753-763. [doi: 10.1007/s10278-022-00746-0] [Medline: 36538245]
35. Radenovic F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell 2018;41(7):1655-1668. [doi: 10.1109/TPAMI.2018.2846566]
36. Henkel C. Efficient large-scale image retrieval with deep feature orthogonality and hybrid-swin-transformers. arXiv. Preprint posted online on Oct 7, 2021. [doi: 10.48550/arXiv.2110.03786]
37. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Schroff F, Kalenichenko D, Philbin J, editors. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 7-12, 2015. [doi: 10.1109/CVPR.2015.7298682]
38. Raccoon – web-based dicomweb & FHIR imaging study server. GitHub. 2020. URL: https://github.com/cylab-tw/raccoon [accessed 2026-03-06]
39. Stylianou A, Souvenir R, Pless R. Visualizing deep similarity networks. In: Stylianou A, Souvenir R, Pless R, editors. 2019 Presented at: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV); Jan 7-11, 2019. [doi: 10.1109/WACV.2019.00220]

Abbreviations

AI: artificial intelligence
AP: average precision
BN: BatchNorm
CBMIR: content-based medical image retrieval
CNN: Convolutional Neural Network
DICOM: Digital Imaging and Communications in Medicine
GeM: generalized mean
HNSW: Hierarchical Navigable Small World
mA: mean average precision
MR: magnetic resonance
PACS: picture archiving and communication system
PReLU: parametric rectified linear unit

XSL•FO
RenderX

QIDO: query based on ID for DICOM Objects
TL: transfer learning
TVGH: Taipei Veterans General Hospital
VGG19: Visual Geometry Group model 19
ViT: Vision Transformer
WADO: web access to DICOM objects

Edited by A Coristine; submitted 30.May.2025; peer-reviewed by A Dureja, FK Jush, S Mohanadas, S Aluvaka; revised version received 01.Feb.2026; accepted 02.Feb.2026; published 06.Apr.2026.

Please cite as:
Lee CL, Hsu TH, Wu YT, Guo WY, Chu WC, Lien CY
Deep Learning for Content-Based Medical Image Retrieval in Picture Archiving and Communication Systems for Brain Tumor Detection: Algorithm Development and Validation
JMIR Med Inform 2026;14:e78300
URL: https://medinform.jmir.org/2026/1/e78300
doi:10.2196/78300

© Chin-Lin Lee, Tzu-Hsuan Hsu, Yu-Te Wu, Wan-Yuo Guo, Woei-Chyn Chu, Chung-Yueh Lien. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Two-Minute Deep Learning–Powered Brain Quantitative Mapping: Accelerating Clinical Imaging With Synthetic Magnetic Resonance Imaging

Yawen Liu[1,2], PhD; Hongxia Yin[2,3], MEng; Zuofeng Zheng[4], MD; Wenjuan Liu[5], MD; Tingting Zhang[2], MEng; Linkun Cai[6], PhD; Haijun Niu[6], PhD; Han Lv[1,2], MD; Zhenghan Yang[2], MD; Zhenchang Wang[1,2], MD; Pengling Ren[1,2], PhD

[1]Precision and Intelligence Medical Imaging Lab, Beijing Friendship Hospital, Capital Medical University, No.95 Yongan Road, Xicheng District, Beijing, China

[2]Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing, China

[3]Department of Medical Engineering, Beijing Friendship Hospital, Capital Medical University, Beijing, China

[4]Department of Radiology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

[5]Department of Radiology, Aerospace Center Hospital, Beijing, China

[6]School of Biological Science and Medical Engineering, Beihang University, Beijing, China

Corresponding Author:
Pengling Ren, PhD
Precision and Intelligence Medical Imaging Lab, Beijing Friendship Hospital, Capital Medical University, No.95 Yongan Road, Xicheng District, Beijing, China

Abstract

Background: Quantitative magnetic resonance imaging (MRI) is an advanced technique that can map the physical properties (T1, T2, and proton density [PD]) of different tissues, offering crucial insights for disease diagnosis. Nonetheless, the practical application of this technology is indeed constrained by several factors, with the most notable being the protracted scanning duration.

Objective: This study aimed to explore whether deep learning (DL)–based superresolution reconstruction of ultrafast whole brain synthetic MRI can obtain quantitative T1/T2/PD maps that are closely approximated to those from routine clinical scans, while substantially shortening scan time and preserving diagnostic image quality.

Methods: A total of 151 healthy adults and 7 individuals with different pathologies were prospectively enrolled. Each individual was examined twice on a 3.0T scanner using routine and fast synthetic MRI protocols. The routine scans (acquisition matrix: 320×256) were interpolated to 512 by 512 for clinical display and served as reference images. The fast scans (acquisition matrix: 192×128) were preprocessed to 256 by 256 and used as inputs to a superresolution generative adversarial network (SRGAN), which reconstructed them to the same 512 by 512 interpolated resolution as the reference. For each quantitative chart, 120 (75.95%) healthy individuals' images were used for training, and 38 (24.05%) individuals' images (healthy individuals: n=31, 19.62%; patients: n=7, 4.43%) were used for testing. Agreement was assessed with a paired *t* test, two 1-sided tests, Bland-Altman analysis, and coefficients of variation.

Results: DL reconstructed and reference T1/T2/PD values were strongly correlated (T1: R^2=0.98; T2: R^2=0.97; and PD: R^2=0.99). The slopes of the linear regression were near 1.0 both for T1 (0.9418) and PD (0.9946), whereas T2 values were moderate, as the slope of the linear regression was 0.8057. Additionally, the average biases of T1, T2, and PD values were small (0.93%, −0.85%, and 0.31%, respectively). The intra- and intergroup coefficient of variation for most of the brain regions stayed below 5%, especially for PD values, and after DL reconstruction, it still has quantitative accuracy for lesions. Quantitative and qualitative analyses of image quality also indicate that SRGAN markedly suppressed noise and artifacts in fast acquisitions, restoring structural fidelity (structural similarity image measure) and signal fidelity (peak signal-to-noise ratio) close to the level of routine scans while substantially improving perceptual naturalness over fast scans (as measured by the naturalness image quality evaluator), although not yet matching that of routine imaging.

Conclusions: SRGAN superresolution applied to ultrafast synthetic MRI yields whole brain T1, T2, and PD maps that show strong correlation with routine synthetic MRI while halving acquisition time and maintaining diagnostic image quality. Although T1 and PD values exhibit near-ideal agreement, and T2 values demonstrate a moderate systematic underestimation, this approach represents a promising step toward accelerating clinical deployment of quantitative brain imaging.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e79389) doi:10.2196/79389

KEYWORDS

magnetic resonance imaging; qMRI; fast MRI; deep learning; brain imaging

Introduction

Quantitative magnetic resonance imaging (qMRI) can reflect the inherent properties of human tissue relaxation times and proton density (PD), providing valuable support for verifying visual assessments of structures and tissues against a normal quantitative standard, thereby holding significant clinical diagnostic value [1]. Nevertheless, the extended scan time of qMRI imposes limitations on its practical use. Synthetic MRI, based on multidynamic, multiecho sequences, can provide multiple quantitative information and multicontrast images in a single scan, enabling fast qMRI scans [2,3]. It has been proven to have comparable diagnostic value to conventional scanning methods, with the generated quantitative data being valuable for disease identification [4-7]. However, the single scan time of this imaging method still typically ranges from 4 to 7 minutes, particularly for brain MRI imaging, which can lead to individual discomfort and potentially impact image quality. Therefore, improving the imaging speed of synthetic MRI is the key to realizing ultrafast qMRI.

It is also crucial to ensure the quality of qMRI and the reliability of quantitative values while accelerating imaging. Usually, fast imaging methods such as parallel acquisition and compressed sensing may result in a loss of image quality [8]. The reconstruction technology based on deep learning (DL) has been applied in various medical image reconstruction tasks, and its development makes it possible to solve the trade-off between image quality and scan time. Generative adversarial networks, renowned for generating high-fidelity superresolution images, have been successfully used in MRI [9]. Many researchers have implemented the concept of generative adversarial networks for MRI image reconstruction, achieving high-quality reconstruction from highly undersampled data [10-14]. In addition, the generative adversarial approach can be used for tasks, such as artifact correction and denoising in MRI [15,16]. In clinical practice, in addition to focusing on image quality, it is more important to determine whether the DL-based reconstruction method can provide less bias and consistent quantitative data [17].

To shorten synthetic MRI quantitative acquisitions and increase clinical throughput, we hypothesized that DL superresolution can generate T1/T2/PD maps that are close to routine long acquisition quantitative values while maintaining image quality, thereby providing faster yet equally reliable neuroimaging and ultimately enhancing diagnostic precision and patient comfort.

Methods

Participants

This study was approved by the Medical Research Ethics Committees of the Beijing Friendship Hospital, Capital Medical University (2020-P2-122-02). Written informed consent was obtained from all participants prior to enrollment. A total of 151 healthy individuals were prospectively recruited at 1 institution for this study. The inclusion criteria were as follows: (1) the age of the individual was greater than or equal to 18 years, (2) the individual showed no structural changes or signs of disease in brain MRI, and (3) the individual showed no contraindications or adverse reactions to MRI examination. The exclusion criteria were as follows: (1) there were obvious artifacts in the images, and (2) the individual failed to complete all MRI examinations. In addition, we also collected images from 7 individuals with different pathologies (white matter [WM] hyperintensities, cerebral infarcts, and encephalomalacia) to evaluate the potential for clinical application.

Synthetic MRI Acquisition

All MRI examinations were performed using a SIGNA Pioneer 3.0 T MRI scanner (GE Healthcare). All individuals underwent two synthetic MRI scans using the multidynamic multiecho sequence (MDME sequence, named MAGiC in the GE scanner): a routine scan protocol and a fast scan protocol. The routine scan uses commonly used clinical acquisition parameters: TR=4000 ms; TE1=18.3 ms; TE2=91.4 ms; TI=28.2 ms; thickness=5 mm; FOV=220×220 mm; acquisition matrix=320×256; echo-train length=16; bandwidth=31.25 kHz; delay times=170, 670, 1840, and 3840 ms; and acceleration factors=2, 24 slices, with a pixel size of 0.7×0.8 and an acquisition time of 4 minutes 55 seconds. Fast scan modifies the acquisition matrix and acceleration factor, which affect the acquisition time, by changing the acquisition matrix to 192×128 and the acceleration factor to 3. The remaining parameters remained the same as those of the routine scan, with a pixel size of 1.1×1.7 and an acquisition time of 1 minute 52 seconds.

Data Preprocessing

The same method was used to retrieve the quantitative maps (T1 maps, T2 maps, and PD maps) as mentioned in a previous study [18]. In this study, quantitative maps from the routine scan (reconstruction matrix=512×512, 0.43×0.43 mm² pixel) served as the high-resolution (HR) reference, whereas those from the fast scan (reconstruction matrix=256×256, 0.86×0.86 mm² pixel) were used as low-resolution (LR) inputs. DICOM images were directly read and prepared for network training. It should be noted that the maximum image intensities of T1 maps, T2 maps, and PD maps are 43,000, 20,000, and 1600, respectively, which correspond to 10 times the actual values. This scaling is due to the rescaling slope value of 0.1 specified in the DICOM header file.

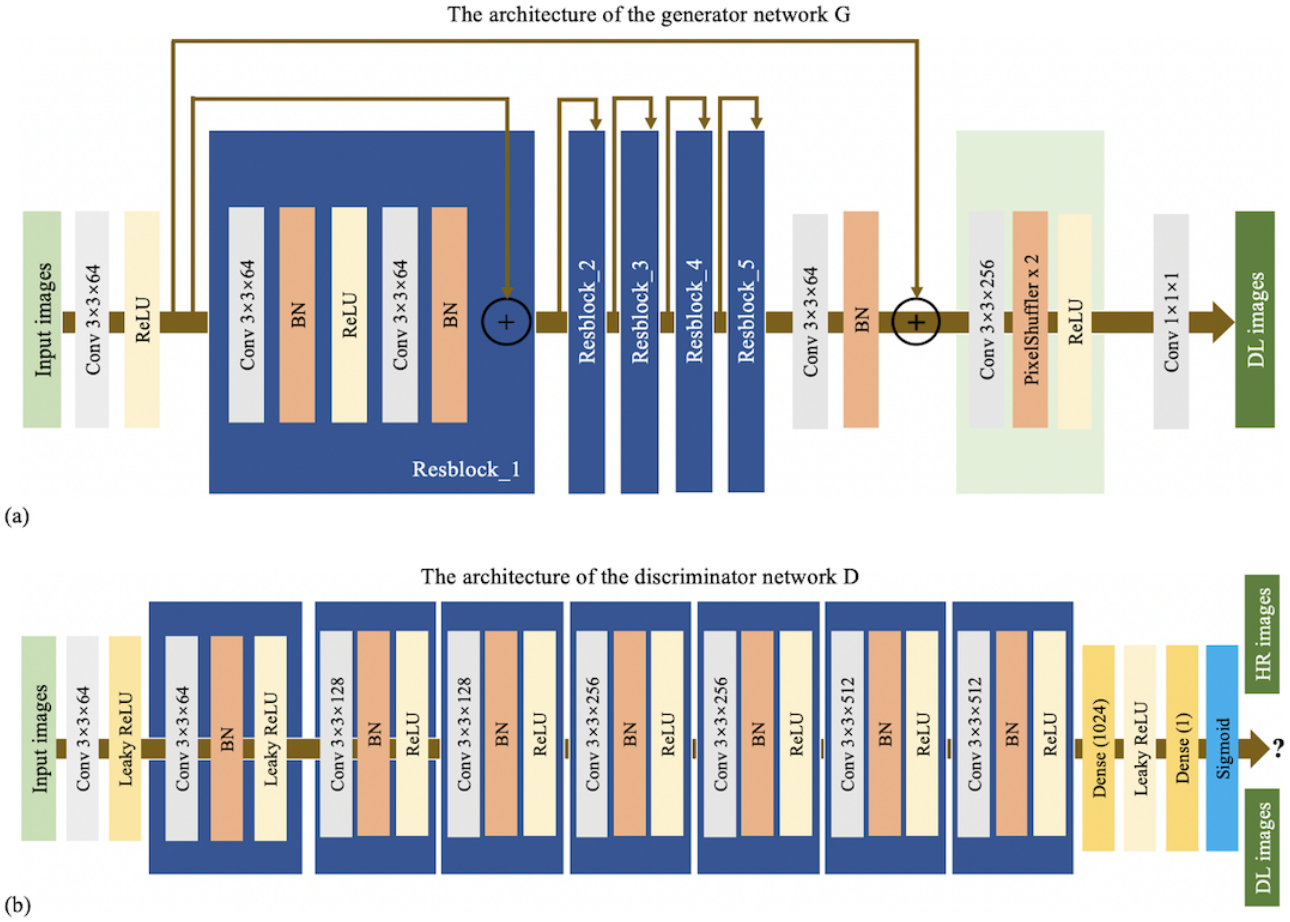
Network Architecture

The detailed generator network structure is shown in Figure 1A. It starts with a 9 by 9 convolutional filter and connects to 5 residual blocks, each of which consists of two 3 by 3 convolutional layers alternating with batch normalization layers, and activation is performed using a rectified linear unit (ReLU) function. After the residual network, a 3 by 3 convolution and a 1 by 1 convolution are connected, and a subpixel convolutional

XSL•FO
RenderX

layer is used to upsample the image. In the generator network, except for the last layer that uses tanh as the activation function, all other layers use ReLU as the activation function. The discriminator network structure is shown in Figure 1B. In the discriminator network D, our model first uses an architecture comprising eight 3×3 convolutional layers and Leaky ReLU function activation. All convolutional layers except the first have batch normalization layers. The Sigmoid activation function is applied to the last fully connected layer, which outputs the probability of discriminating whether the input HR image is a real HR image or an image generated by the generator. The network architecture includes a pretrained VGG-19 network, which is used for feature extraction and loss function calculation.

Figure 1. The architecture of the superresolution generative adversarial network. BN: batch normalization; DL: deep learning; HR: high resolution; ReLU: rectified linear unit.



Loss Function

The pixel-by-pixel error method used is the L1 loss function, also known as the mean absolute error, which is calculated using equation 1. Although SRGAN completely discards the pixel-by-pixel error, we still add this error in a certain proportion during the actual training process to amplify the difference and guide the optimization of the model. The cross-entropy loss function of the discriminator is used as the adversarial error of the network, and the calculation method is expressed using equation 2. In addition to using the adversarial error, SRGAN also uses a content error, which is defined as the Euclidean distance between the superresolved image and the feature map of the reference image, and the calculation method is expressed using equation 3. The content error is used to align the content of LR images and HR images, which plays the same role as the mean square error. In this network, the pretrained VGG-19 network is used to extract the feature parameter map. Therefore, the loss function for the generator in this network is described using equation 4.

$$(1) L_{mae} = \frac{1}{CHW}\sum_{i,j,k} |I_{i,j,k}^{DL} - I_{i,j,k}^{GT}|$$

where I^{DL} is the output image of the generator and I^{GT} is the input reference image.

$$(2) L_{gan} = -\sum_{i} \log D(G(I^{DL}))$$

where G and D are the generative network and the discriminative network, respectively.

$$(3) L_{VGG} = \frac{1}{C_j H_j W_j}\sum_{x=1}^{W_j}\sum_{y=1}^{H_j}\sum_{c=1}^{C_j}(\phi_j(G(I^{DL}))_{x,y,c} - \phi_j(G(I^{GT}))_{x,y,c})^2$$

where $(\varnothing j)$ is the feature map obtained from the jth convolutional layer of the VGG-19 network, C_j, H_j, and W_j are the number of channels and height and width of the feature map, respectively, I^{DL} is the output image of the generator, and I^{GT} is the input reference image.

$$(4) L_{total} = L_{mae} + 10^{-3} L_{gan} + 2 \times 10^{-6} L_{VGG}$$

Implementation

Approximately 80% (120 healthy individuals) of the dataset were allocated to training, and the remaining individuals (healthy individuals: 31, 19.62%; and patients: 7, 4.43%) were used for

XSL•FO
RenderX

testing. Slices from each individual were pooled and randomly shuffled within their respective sets to prevent any sequential bias and improve the performance of the model. Before training, all quantitative maps (T1, T2, and PD) underwent intensity normalization to match the output range of the generator network, that is, imagenorm=(pixelvalue/normalizepara)×2−1, where normalize_para is set to 45,000, 22,000, and 1800 for T1, T2, and PD maps, respectively.

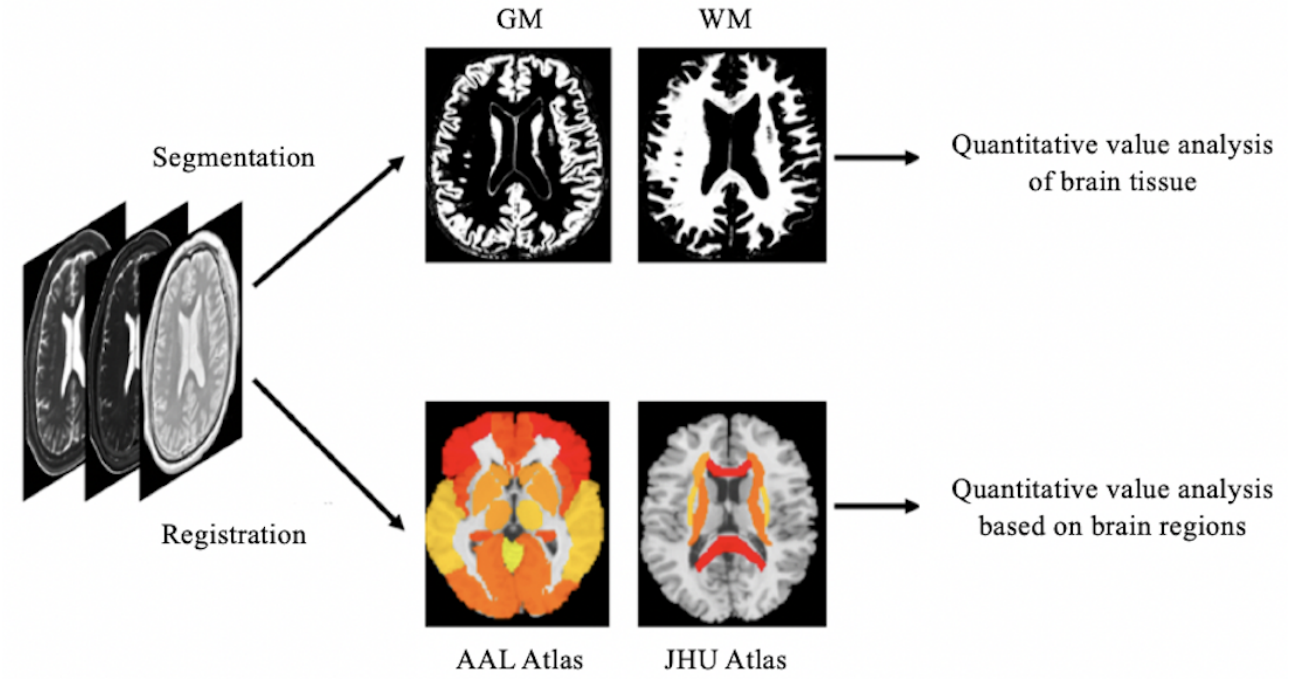
The DL models for T1, T2, and PD maps were trained separately but with identical hyperparameters. Each model was implemented using the TensorLayer framework [19] and optimized via the adaptive moment estimation (Adam) optimizer with a learning rate of 1×10^{3} and step-wise decay every 1000 iterations. Training was performed on an NVIDIA Tesla V100 GPU with a batch size of 8 and a total of 200 epochs. All other parameters were kept at their default settings.

Quantitative Value Measurements

The flowchart illustrating the process of quantitative value measurements is presented in Figure 2.

For healthy participants, quantitative value analysis of brain tissue and brain regions was conducted. First, brain tissue segmentation was performed on each individual's quantitative map using the New Segment tool within SPM [20] software. Then, quantitative value extraction of gray matter (GM) and WM was performed using self-written programs in the MATLAB (The MathWorks Inc.) platform.

Figure 2. The process of quantitative value measurements for healthy participants. AAL: anatomical automatic labeling; GM: gray matter; JHU: Johns Hopkins University; WM: white matter.



In addition to analyzing quantitative values for both GM and WM tissue, analysis was also performed by different brain regions for clinical evaluation and application. Using the anatomical automatic labeling atlas of the Montreal Neurological Institute (MNI), 10 GM regions of interest (ROIs) were created on anatomical structures, including frontal cortex, temporal cortex, parietal cortex, occipital cortex, insula, hippocampus, caudate nucleus, putamen, pallidum, and thalamus. Additionally, using the JHU-WhiteMatter-48 atlas, 9 WM ROIs were generated, including the middle cerebellar peduncle, genu of corpus callosum, splenium of corpus callosum, cerebral peduncle, internal capsule, anterior corona radiata, superior corona radiata, posterior corona radiata, and external capsule [21]. In all analyses, right and left quantitative data were averaged. The quantitative value of the corresponding brain region in the quantitative map reconstructed by DL and from routine scans was obtained for each individual for subsequent statistical analysis.

For the images of individuals with pathologies, an experienced radiologist used the ITK-SNAP software (version 3.8.0; [22]) to delineate the ROI of typical lesions manually and acquired the mean value of the signal in the ROI. To ensure paired comparison, identical ROI size and location were replicated on the DL reconstructed and reference routine scan images. A total of 12 ROIs were obtained from 7 (4.43%) patients (individual 3: 3 ROIs, individual 5: 3 ROIs, individual 7: 2 ROIs, and the remaining 4 individuals: 1 ROI each).

Image Quality Evaluation

Two full-reference evaluation indices—peak signal-to-noise ratio (PSNR) and structural similarity image measure (SSIM)—were used in this study for image quality assessment [23]. The larger the value of the 2 indices, the better the image quality. The PSNR was defined by equation 5:

$$MSE=\sum_{m=1}^{M}\sum_{n=1}^{N}[R(m,n)-I(m,n)]^{2}M\times N$$

$$(5)PSNR=20\log L^{2}MSE$$

The calculation method of SSIM is given by equation 6:

$$,$$

$$(6)SSIM(X,Y)=(2\mu_X\mu_Y+C1)(2\sigma XY+C2)(\mu X2+\mu Y2+C1)(\sigma X2+\sigma Y2+C2)$$

XSL•FO
RenderX

$C1=(K1L)2, C2=(K2L)2$

where μ and μ, σ and σ, and σ are the local means, SDs, and cross-covariance for the DL quantitative images and routine scans, respectively, and C1 and C2 are 2 quantities used to stabilize the division in the case of a weak denominator, with L=65,535, K1 = 0.01, and K2= 0.03.

In addition, a no-reference image quality assessment method, the naturalness image quality evaluator (NIQE), was used to evaluate the quality of quantitative images reconstructed by DL, with fast scan and with routine scan, respectively [24]. Unlike PSNR and SSIM, the lower the value of NIQE, the less distortion and higher image quality.

Statistical Analysis

For the quantitative analysis of brain tissue in healthy participants, Paired *t*-test was used to compare the differences in quantitative values between DL and routine scans, and equivalence was confirmed using two one-sided tests (TOST). Linear regression analysis and Bland-Altman analysis were used in the study, and we combined GM and WM to comprehensively evaluate overall bias. This method captures the overall trends and variability of different tissue types, which is crucial for evaluating the performance of DL reconstruction in clinically relevant contexts. For the region-based quantitative value analysis in healthy participants, coefficients of variation (CVs) were calculated within each imaging method (intragroup CV) and across each imaging method (intergroup CV). Intergroup CV was calculated using the average of the quantitative values from each method. For the images of individuals with pathologies, Paired *t*-test was used to compare quantitative values in different ROIs between groups. All statistical analyses were performed using SPSS 22.0 and GraphPad Prism 9. A *P* value <.05 was considered statistically significant.

Ethical Considerations

This study was approved by the Medical Research Ethics Committees of Beijing Friendship Hospital, Capital Medical University (2020-P2-122-02). All participants provided informed consent before participating. Participants' privacy and confidentiality were protected, with all data being deidentified during analysis. No compensation was provided for participation. No identifiable information of participants was included in the paper or supplementary materials in Multimedia Appendix 1. All procedures adhered to ethical guidelines for participant autonomy, safety, and confidentiality.

Results

Time Consumption

The average whole brain acquisition time was 4 minutes 55 seconds for routine scans, with an additional 1 minute for vendor postprocessing. Fast scans were completed in 1 minute 52 seconds; applying the SRGAN network trained in this study then reconstructed, the 512 by 512 quantitative T1/T2/PD maps were completed in approximately 1 second. Thus, the combined acquisition and reconstruction time was reduced from 5 minutes 55 seconds to 1 minute 53 seconds, promoting the wider clinical application of quantitative neuroimaging and synthetic MRI.

Quantitative Value Accuracy

A paired *t* test showed significant differences in GM tissue values between DL and fast scans (T1: *P*<.001, PD: *P*=.002), while there was no significant difference in GM tissue values between DL and routine scans at T1 (*P*=.66) and PD (*P*=.18). However, compared with routine scans, T2 values still showed significant differences after DL (*P*<.001). For WM, compared with routine scans, there was no significant difference in T2 after DL (*P*=.11). Although there were still significant differences in T1 and PD, the differences were reduced compared with fast scans (Table 1).

Table . The quantitative value of brain tissue obtained by fast scan, deep learning (DL), and routine scan.

	Fast, mean (SD)	DL, mean (SD)	Routine, mean (SD)	*P* value[a]	*P* value[b]
GM[c]					
T1 (ms)	1200.03 (36.01)	1230.17 (34.12)	1232.65 (34.98)	<.001	.66
T2 (ms)	97.87 (2.07)	91.36 (2.60)	93.33 (3.50)	<.001	<.001
PD (%)	79.69 (2.35)	80.29 (1.51)	80.11 (1.23)	.002	.18
WM[d]					
T1 (ms)	814.88 (35.78)	855.06 (32.12)	837.86 (35.88)	<.001	<.001
T2 (ms)	84.57 (2.17)	80.23 (1.61)	79.92 (2.05)	<.001	.11
PD (%)	67.55 (1.94)	68.49 (1.44)	68.21 (1.29)	<.001	.01

[a]*t* test results between DL and fast scans.

[b]*t* test results between DL and routine scans.

[c]GM: gray matter.

[d]WM: white matter.

Setting the clinically acceptable brain tissue quantification value limit to ±5% of the mean values from the routine scans (T1_GM epsilon: 61.63 ms, T2_GM epsilon: 4.67 ms, PD_GM epsilon: 4.01%, T1_WM epsilon: 41.89 ms, T2_GM epsilon: 4.00 ms, and PD_WM epsilon: 3.41%), the 90% TOST CI for the brain tissue in 3 quantitative maps (T1_GM: −11.88 to 6.91, T2_GM:

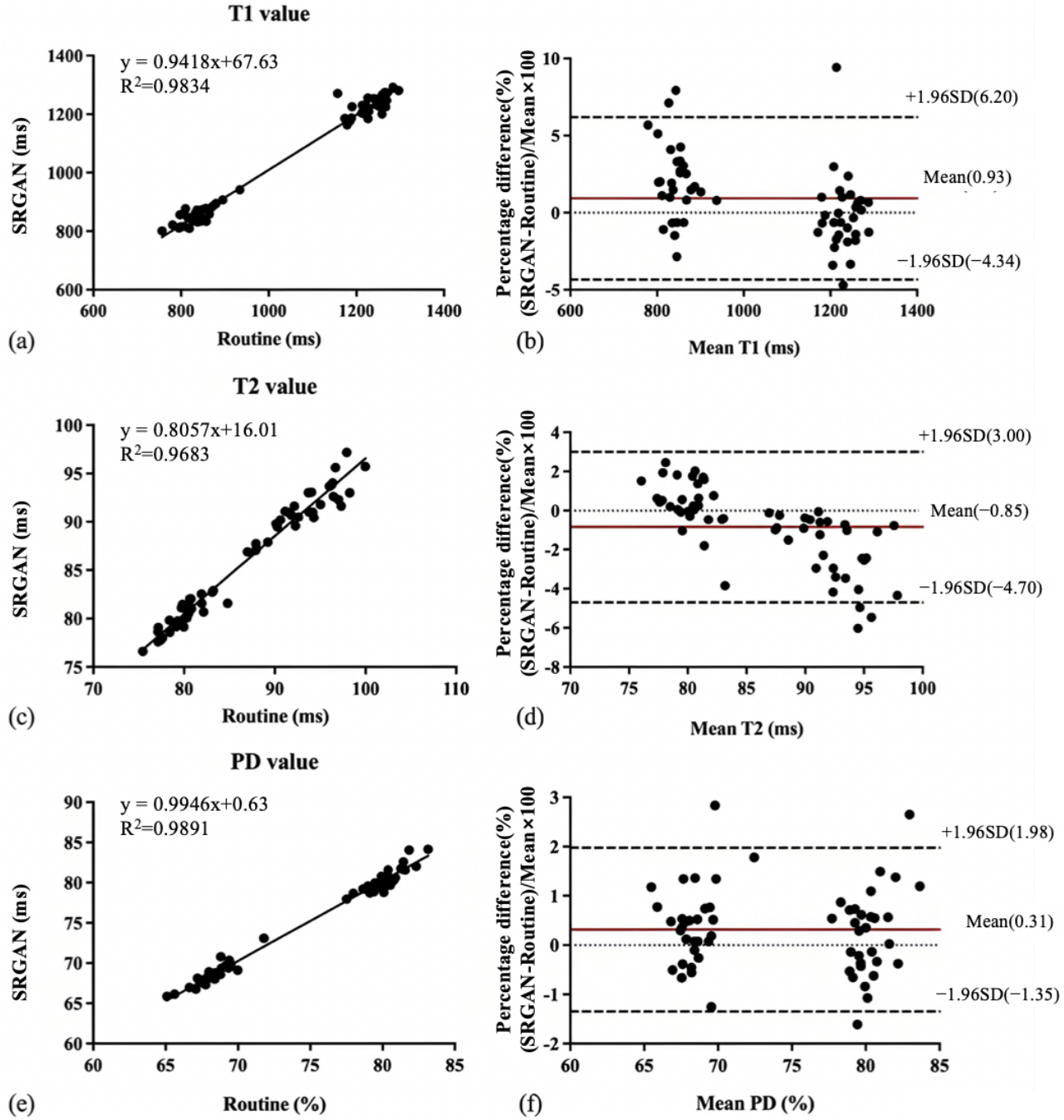
XSL•FO
RenderX

−2.48 to−1.46, PD_GM: −0.04 to 0.41, T1_WM: 10.94 to 23.46, T2_WM: −0.01 to 0.03, and PD_WM: 0.10 to 0.45) were all within the predefined margins (P<.001).

Figure 3 shows the relationship and Bland-Altman plots between the T1, T2, and PD values obtained by DL and routine scans (reference). The correlations to the reference value were excellent, as the R^2 for the T1, T2, and PD values of DL compared to the reference value was 0.98, 0.97, and 0.99, respectively. The slopes of the linear regression were near 1.0 both for T1 (0.9418) and PD (0.9946). In contrast, the T2 values were moderate, as the slope of the linear regression was 0.8057 (Figure 3A,C,E). In addition, the average percentage bias for T1, T2, and PD was 0.93%, −0.85%, and 0.31%, respectively. The 95% limits of agreement were 6.20% to −4.34% for T1, 3.00% to −4.70% for T2, and 1.98% to −1.35% for PD (Figure 3B,D,F). Similarly, Figure 4 shows the relationship and Bland-Altman plots of the T1, T2, and PD acquired by DL and fast scans. The results of linear regression analysis also showed a strong correlation between the 2 methods in T1 and PD (T1: R^2=0.99 and PD: R^2=0.98), with a slightly lower correlation in T2 (R^2=0.80). The slopes of the linear regression were near 1.0 both for T1 (0.9680) and PD (0.9381). In contrast, the T2 values were moderate, as the slope of the linear regression was 0.7755 (Figure 4A,C,E). The results of the bias trend of T1, T2, and PD between these 2 methods also showed similar results, that is, the mean percentage difference for T1 and PD was smaller (3.66% and 1.09%, respectively), whereas the mean percentage difference for T2 was slightly higher (−6.32%) (Figure 4B,D,F).

XSL•FO
RenderX

Figure 3. Scatterplots and Bland-Altman plot results for deep learning versus routine scans. The linear regression lines represent a strong linear relationship with a robust fit between the tissue value measurements of T1, T2, and proton density (PD) obtained from the 2 magnetic resonance imaging methods. Bland-Altman plots show mean percentage differences (solid red line) and the 95% CIs (dashed black line). SRGAN: superresolution generative adversarial network.



XSL•FO
RenderX

Figure 4. Scatterplots and Bland-Altman plot results for deep learning versus fast scans. The linear regression lines represent a strong linear relationship with a robust fit between the tissue value measurements of T1, T2, and proton density (PD) obtained from the 2 magnetic resonance imaging methods. Bland-Altman plots show mean percentage differences (solid red line) and the 95% CIs (dashed black line). SRGAN: superresolution generative adversarial network.

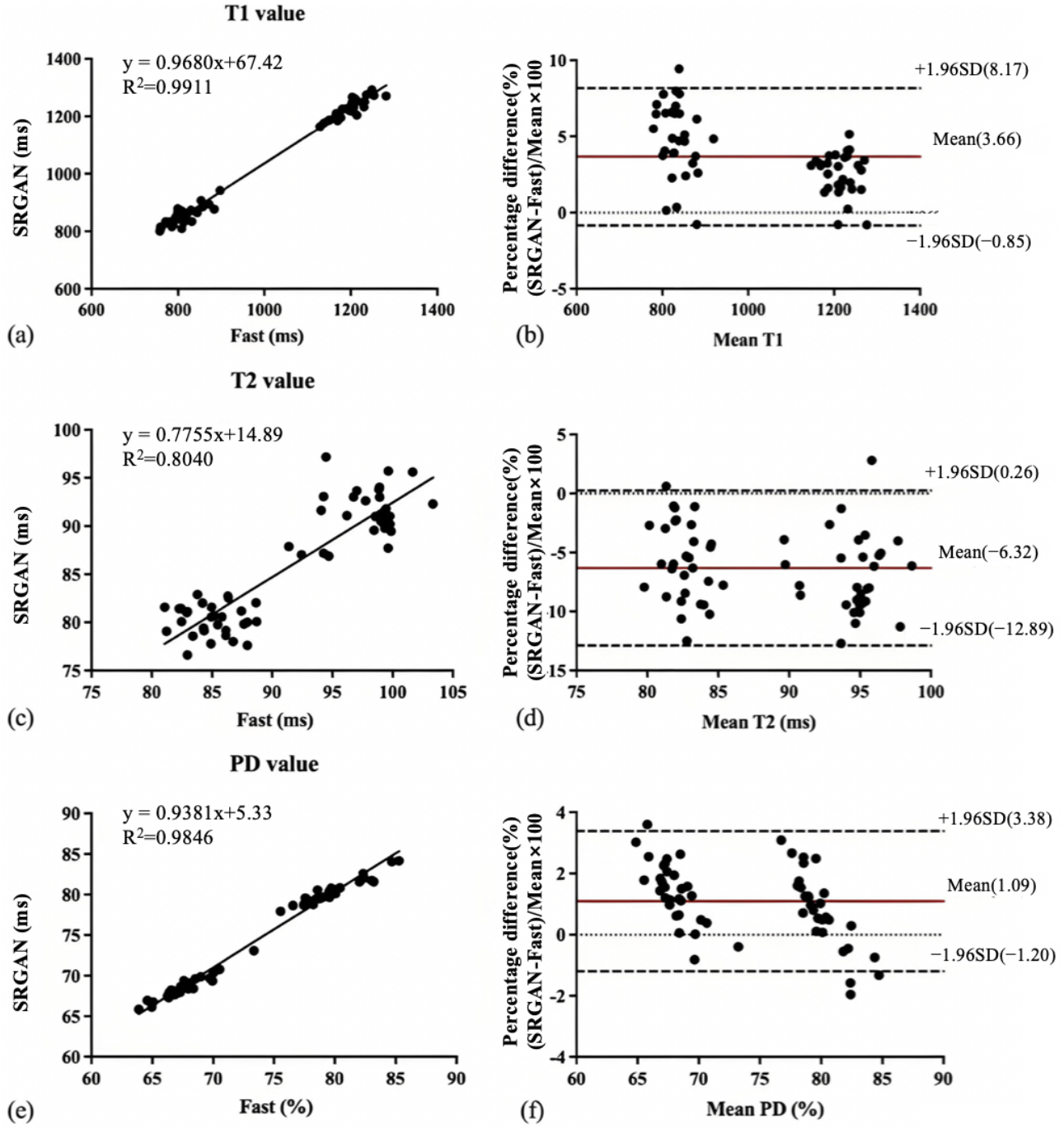


Table 2 presents the intragroup CV and intergroup CV of brain regions' quantitative values acquired by DL and routine scans. The intragroup variability of the quantitative values obtained by these 2 methods is small, as the intragroup CV for the routine scan's PD values ranges from 1.64% to 6.44%, and the PD values for DL range from 1.98% to 5.84%. The highest intragroup CV of T1 and T2 values was noted in the pallidum (T1: 18.77% vs 17.25%; and T2: 17.00% vs 16.33%). The intergroup CV was lower than the intragroup CV for all brain regions. For intergroup CV, there were no significant differences in the quantitative values of T1, T2, and PD for the WM region after DL reconstruction. However, significant differences in T2 values were observed in the GM regions, including the frontal cortex, temporal cortex, parietal cortex, occipital cortex, insula, and hippocampus.

Table 3 presents the quantitative values in ROI of typical lesions acquired by DL and routine scans. Compared with the quantitative values under routine scans, there were no significant differences in the quantitative values reconstructed by deep learning (T1, *P*=.67; T2, *P*=.73; and PD, *P*=.75).

XSL•FO
RenderX

Table . Intragroup and intergroup coefficient of variation (CV) for T1, T2, and proton density (PD) based on anatomical automatic labeling (AAL) and Johns Hopkins University (JHU) atlas in healthy individuals.

Brain region	Intragroup CV (%)		Intergroup CV (%) (T1/T2/PD)	*P* value (T1/T2/PD)
	DL[a] (T1/T2/PD)	Routine (T1/T2/PD)		
AAL atlas				
Frontal	7.41/7.28/3.62	8.20/7.14/3.44	2.07/5.17/0.61	.29/.001/.94
Temporal	5.10/8.00/3.31	5.50/8.62/3.27	1.65/4.98/0.40	.38/.003/.83
Parietal	9.32/14.40/3.02	9.08/15.90/2.97	1.30/8.11/0.45	.38/<.001/.42
Occipital	7.43/11.07/2.15	7.19/13.32/2.41	1.72/5.10/0.47	.60/.005/.55
Insula	9.03/15.23/2.39	9.86/13.44/2.53	1.80/7.12/0.72	.95/.02/.36
Hippocampus	8.95/11.43/2.53	9.02/14.30/2.71	2.62/6.08/0.72	.22/.005/.34
Caudate	13.69/16.20/5.29	14.65/17.14/5.37	2.60/5.32/0.76	.70/.35/.92
Putamen	13.48/12.37/2.62	14.48/17.33/3.47	2.04/5.44/0.73	.99/.11/.92
Pallidum	18.77/16.33/2.26	17.25/17.00/2.61	3.54/3.37/0.85	.51/.56/.50
Thalamus	13.37/12.39/2.38	13.99/13.96/3.00	1.69/4.05/0.88	.77/.27/.67
JHU atlas				
MCP[b]	7.28/7.04/1.72	5.89/8.13/1.64	1.67/2.79/0.51	.50/.08/.28
GCA[c]	15.33/9.77/5.84	15.22/9.11/6.44	1.99/2.89/0.76	.87/.52/.55
SCA[d]	6.44/3.70/1.98	6.00/4.64/1.94	1.74/1.71/0.63	.19/.05/.13
CP[e]	10.16/7.81/4.77	9.62/7.54/5.01	2.73/2.55/1.08	.41/.14/.55
IC[f]	9.55/4.29/3.95	9.63/4.52/4.20	1.30/1.74/0.68	.83/.21/.45
ACR[g]	9.01/5.62/2.94	9.78/6.91/3.86	1.97/1.34/0.91	.92/.97/.98
SCR[h]	9.95/4.68/3.73	8.35/4.95/3.07	2.16/1.31/1.03	.27/.14/.28
PCR[i]	14.19/3.72/4.25	12.20/4.59/3.69	2.60/2.49/0.96	.42/.03/.23
EC[j]	11.32/5.02/2.82	11.23/4.89/3.36	1.38/0.92/0.95	.65/.69/.21

[a]DL: deep learning.
[b]MCP: middle cerebellar peduncle.
[c]GCA: genu of corpus callosum.
[d]SCA: splenium of corpus callosum.
[e]CP: cerebral peduncle.
[f]IC: internal capsule.
[g]ACR: anterior corona radiata.
[h]SCR: superior corona radiata.
[i]PCR: posterior corona radiata.
[j]EC: external capsule.

XSL•FO
RenderX

Table . Quantitative value in the region of interest (ROI) of typical lesions acquired by deep learning (DL) and routine scan.

ROI	Description	T1[a] (ms)		T2[b] (ms)		PD[c,d] (%)	
		DL	Routine	DL	Routine	DL	Routine
1	WMH[e]	1019.68	1052.20	100.84	112.9	80.70	79.16
2	WMH	1034.41	1052.78	96.32	96.71	75.21	77.09
3	WMH	955.68	1013.20	94.78	104.52	70.77	72.97
4	CI[f]	1217.30	1398.16	126.53	149.71	80.00	81.26
5	Encephalomalacia	3597.26	3683.45	310.67	276.99	102.83	106.59
6	WMH	1088.41	1214.51	109.50	121.23	82.11	84.70
7	CI	1832.00	1679.00	136.10	133.80	90.50	90.90
8	Encephalomalacia	3719.04	3892.54	292.16	272.32	101.31	104.98
9	WMH	1734.53	1517.03	163.45	182.01	85.14	86.37
10	WMH	1171.00	1238.00	110.60	124.70	80.70	81.10
11	Encephalomalacia	3666.00	3469.00	350.20	295.80	107.80	106.30
12	WMH	1253.00	1285.00	113.20	105.20	78.70	83.00

[a]*P*=.67.

[b]*P*=.73.

[c]PD: proton density.

[d]*P*=.75.

[e]WMH: white matter hyperintensities.

[f]CI: cerebral infarcts.

Image Quality

Figure 5 and Tables 4 and 5 compare image quality among the 3 protocols. After SRGAN reconstruction, DL maps simultaneously preserved quantitative accuracy and outperformed fast acquisitions in every objective metric, demonstrating high image quality.

XSL•FO
RenderX

Figure 5. Representative results of the proposed method. Results of fast scans (left), deep learning (DL) reconstruction (middle), and routine scans (right) are displayed for T1 map, T2 map, and proton density (PD) map.

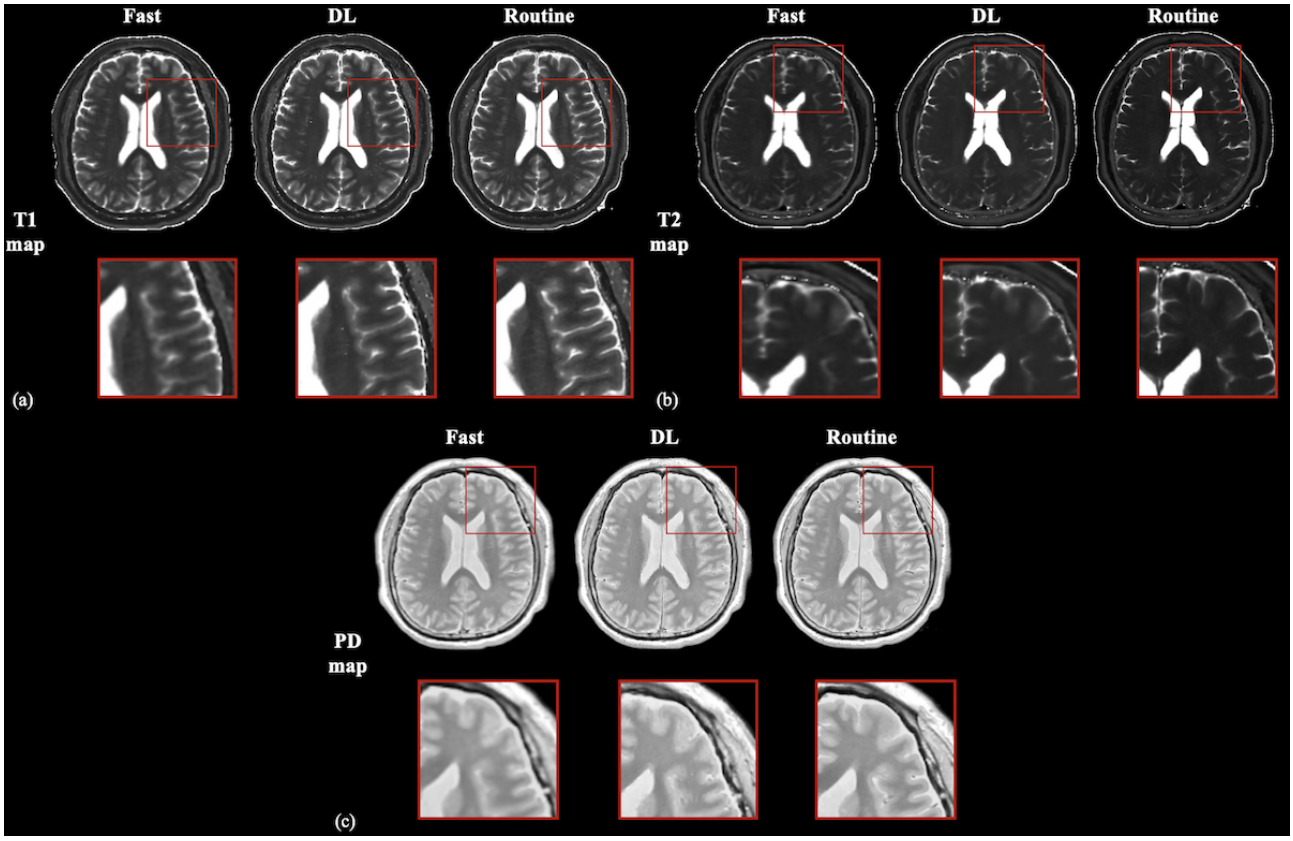


Table . Comparison of image quality between the deep learning quantitative maps and routine scans using the structural similarity image measure (SSIM) and peak signal-to-noise ratio (PSNR).

	PSNR, mean (SD)	SSIM, mean (SD)
T1 maps	21.94 (2.36)	0.82 (0.08)
T2 maps	29.14 (1.81)	0.93 (0.03)
PD[a] maps	58.01 (3.15)	0.99 (0.01)

[a]PD: proton density.

Table . The naturalness image quality evaluator of quantitative maps under fast scan, deep learning (DL), and routine scan.

	Fast scan, mean (SD)	DL, mean (SD)	Routine, mean (SD)
T1 maps	29.46 (6.42)	13.36 (1.88)	10.90 (1.84)
T2 maps	30.56 (5.57)	18.79 (1.46)	15.21 (1.78)
PD[a] maps	34.66 (2.99)	11.74 (2.17)	10.85 (2.53)

[a]PD: proton density.

Discussion

Principal Findings

In this paper, we trained an SRGAN network to reconstruct whole brain T1, T2, and PD maps from accelerated synthetic MRI data. Experimental results show that the application of DL for reconstruction can generate quantitative maps from fast scans with a 50% reduction in acquisition time. The quantitative instability caused by LR acquisition is reduced, with no significant difference between the reconstructed and routine scan values, and intra- and inter-CVs are small in most brain regions. Our study also found that the resulting maps still allow clear visual detection of common pathologies, thereby helping to shorten clinical scanning time.

Comparison With Prior Work

In previous studies, the use of commercial algorithms for fast image reconstruction resulted in a total brain acquisition time exceeding 3 minutes [25]. Our research further reduces the collection time to less than 2 minutes while maintaining accuracy. DICOM raw data were used as input and output for

XSL•FO
RenderX

the network, which conforms to the standard practice of DICOM image processing, where metadata (including rescaled slopes) are crucial for accurate image analysis and quantitative measurement. By adhering to these specifications, the study maintains the integrity of the qMRI data, facilitating a reliable evaluation of the DL reconstruction techniques' impact on the accuracy of quantitative brain tissue values.

The MDME sequence was used to acquire the quantitative maps, and several studies have investigated the repeatability and reproducibility of quantitative values using this sequence. In this study, we applied multiple quantitative metrics and statistical methods to analyze differences among DL reconstructed, fast scan, and routine scan values. The results of our study, as indicated by the paired *t* tests, provide valuable insights into the performance of DL reconstruction and demonstrate that it improves the accuracy of fast quantitative MRI scans, bringing tissue-specific T1, T2, and PD values closer to those of routine clinical acquisitions. In GM, the method achieved no statistically significant difference compared to routine scans. However, a small but statistically significant underestimation in T2 remained. A previous study examining the influence of parameter modifications on MDME-derived quantitative values reported that alterations in acquisition matrix or acceleration factor induced negligible variance in measured constants, with the exception of CSF T2 estimation [26,27]. Notably, this investigation was performed on an isolated, homogeneous phantom; consequently, partial volume effects and other anatomical confounders were absent. The present fast scan protocol, in contrast, simultaneously varies both matrix size and acceleration factor without a commensurate increase in the number of echo times. This divergence in experimental design may compromise the accuracy of T2 mapping—particularly through reduced sampling density in echo space—thereby introducing systematic bias into the fitted relaxation constants. However, TOST analysis further indicates that the 90% CIs of all brain tissue parameters in the 3 quantitative maps are within clinically acceptable ranges, providing the possibility for the clinical applications of DL quantitative reconstruction.

In our quantitative analysis, the linear regression slope for T2 values (DL vs routine) was 0.8057, which deviates from the ideal value of 1.0. Clinically, this may affect the diagnostic accuracy of high T2 lesions, leading to misdiagnosis or delayed intervention. Most of the samples in this study were healthy volunteers, and the accuracy of DL for T2 values under specific pathological manifestations was not validated. Such bias may arise primarily from the L1 loss component, which enforces pixel-wise fidelity at the expense of high-frequency details and can lead to oversmoothing—particularly in regions with high tissue contrast or rapid T2 decay. Consequently, although the average T2 error across healthy tissue may satisfy TOST criteria, it may reduce the accuracy of this method in extreme cases of dynamic range that often occur in pathology. Despite this limitation, the SRGAN model demonstrated good performance in overall image quality metrics, indicating its practicality in routine clinical workflows, although it requires appropriate attention.

Linear regression and Bland-Altman plots also demonstrated that the vast majority of quantitative parameters exhibited high correlation and agreement between the DL and routine scans, with 95% of data points lying within the limits of agreement, which is in line with previous research [28,29]. In contrast, T2 estimates derived from the DL exhibited a marked bias relative to the reference (Figure 3 vs Figure 4). These findings indicate that the DL model successfully compensates for systematic errors introduced by the fast scan, yielding quantitative values that closely approximate those obtained under routine scans. Such error mitigation directly supports the primary objective of this study: to establish the feasibility of using DL for accurate, ultrafast quantitative MR reconstruction.

In addition to measuring the quantitative values of gray and WM, the study also applied classic atlases from brain science analysis: the anatomical automatic labeling and the Johns Hopkins University atlases [30]. These atlases divide the brain into regions based on structure or function, which allows for a more comprehensive characterization of the accuracy and reliability of the quantitative values. Generally, the intragroup CV should ideally be below 10%, and the intergroup CV should remain under 15%, yet a significant intragroup CV was observed in deep GM nucleus regions, such as pallidum. This may be due to the inhomogeneity of B1 and differences in data acquisition parameters or variations in the age and gender of the individuals, which have also been shown and analyzed in previous studies [31-33]. Additionally, these numerical disparities may stem from spatial registration errors between individual quantitative maps and brain region templates, potentially affecting the extraction of quantitative values. Our study also observed significant differences in the T2 CVs between DL and routine scans in some GM brain regions. This difference indicates that although the SRGAN model has good consistency at the overall organizational level, it may introduce variability at the regional scale, which may be due to the model's sensitivity to local texture or noise, potentially limiting its practicality in precise quantitative neuroimaging applications. However, the intra- and inter-group CV for the remaining brain regions remained below 5%, especially for PD values, which show minimal intragroup and intergroup variation across all brain regions. This implies that T2 is relatively sensitive in quantitative analysis and should be a key focus in disease analysis.

The outlined lesion areas in the study are predominantly distributed in the WM. The average T1, T2, and PD quantitative values within these lesion areas exceed 1000 milliseconds, 100 milliseconds, and 80 milliseconds, respectively, which are higher than the ranges observed in the healthy participants' quantitative values. Furthermore, different types of lesions exhibit distinct quantitative values. WM hyperintensities and acute-phase cerebral infarction have similar quantitative values. In contrast, the quantitative values in encephalomalacia are significantly higher than those in the other 2 types of lesions, particularly the T1 values within encephalomalacia (approaching 4000 ms), which are similar to the T1 value of cerebrospinal fluid because the components of encephalomalacia are similar to cerebrospinal fluid. These findings align with clinical knowledge, indicating that lesion areas generally have longer T1 and T2 relaxation

XSL•FO
RenderX

times. Importantly, the quantitative values obtained through DL still maintain diagnostic efficacy. Through careful examination of quantitative values, it was found that with a small amount of lesion data, although the difference in quantitative values between DL and routine scans did not reach statistical significance, 5 reconstructed T2 values were lower than those of the routine scan in 7 WMH and infarct ROIs. Given that these lesions have already exhibited high T2 relaxation times (usually >100 ms), any systematic underestimation may weaken their abnormal manifestations. This may be because superresolution networks are only trained on quantitative images of healthy adults and do not include any pathological cases. Therefore, the model has not yet learned the characteristic diastolic curves of common brain injuries, such as significant prolongation of WM hyperintensities on T2. This lack of exposure may lead to the network interpreting high T2 values in lesions as outliers or reconstruction artifacts, thereby regressing these values to the healthy tissue distribution. Therefore, the sample size will continue to be expanded in the future to verify the quantitative accuracy of DL reconstruction in various neurological diseases.

While the DL maps have basically met the predefined quantitative equivalence margins, the no-reference NIQE metric reveals a residual perceptual gap: compared with fast scanning, the DL image quality has significantly improved (as indicated by a decreased NIQE score), but compared with routine scans, the NIQE scores of the 3-parameter images are slightly higher. This suggests that the current model ensures the accuracy of clinical quantification and significantly improves accuracy on the basis of fast scans, but further improvement is needed to achieve complete perceptual equality with routine HR imaging.

Limitations

This study still has some limitations. First, the study adopted the widely validated SRGAN framework and focused on quantitative agreement and image quality comparison; however, emerging architectures such as ESRGAN or SwinIR were not compared, nor were traditional non-DL upsampling methods used as baselines. Second, the network was trained exclusively on data acquired with a GE 3.0 T scanner and entirely on normal-appearing brain tissue (only 7 patients with 3 common lesion types were used for testing), so the performance of the method on pathological tissue is a preliminary proof of concept rather than evidence of clinical equivalence. The small number of pathology cases limits statistical power and generalizability to diverse or high-contrast lesions, which may pose a barrier to routine clinical deployment. Third, based on the current amount of data, using the training set and testing set with a single subject-level split of 8:2 may prevent data leakage from related slices but may not fully capture the variability between multiple random partitions. Due to limited training data, the model runs on 2D slices instead of 3D volumes, which may result in a loss of spatial continuity between slices. Fourth, acceleration was limited to an R of 3, reconstruction was performed on postprocessed quantitative maps, raw k-space data were not used, and weighted images generated from the quantitative maps were not evaluated. Additionally, the ground truth images have been reconstructed from a matrix of 320 by 256 to a matrix of 512 by 512, and thus, the model learns to reproduce the interpolated image content rather than recovering genuine high-frequency details that would be present in a natively HR acquisition. Finally, evaluation was performed based on quantitative metrics without introducing a qualitative visual assessment by radiologists.

Future Directions

Future research should focus on the following improvements: systematically compare the current SRGAN implementation with newer superresolution networks and quantify the individual contributions of perceptual and adversarial losses to quantitative accuracy, and compare them with non-DL baseline reconstruction methods to better illustrate the benefits of DL. A prospective cohort including demyelinating disease, tumor edema, and microhemorrhages will be enrolled, and the model will be fine-tuned or retrained if necessary before wider clinical deployment. To improve robustness, we will collect both k-space and image-domain data and develop a dual-domain joint reconstruction framework. Future work will explore 3D or slice-aware architectures based on richer training data and computing resources and attempt to use different data partitioning methods to validate generalization further. Finally, the weighted images derived from the superresolved quantitative maps will be generated and assessed for diagnostic quality.

Conclusions

In conclusion, accelerating synthetic MRI is pivotal for its broader clinical adoption in neuroimaging. In this study, we undertook the optimization and enhancement of images acquired through fast synthetic MRI via DL-based reconstruction techniques. Our efforts may result in a remarkable 50% reduction in clinical scan time, all the while achieving better image quality. Importantly, reconstructed T1 and PD maps show excellent agreement with reference scans, supporting their potential for reliable quantification, although T2 values exhibit a consistent underestimation at higher intensities. Despite this limitation, the overall strong correlations and low average biases suggest that using DL methods to mitigate biases in quantitative values is feasible. Furthermore, the shortened scan time not only enhances patient comfort but also reduces the likelihood of motion artifacts. This optimization streamlines the scanning workflow, minimizes redundancy, and maximizes the efficient use of MRI machines.

Acknowledgments

The authors are grateful to Dr Yi Zhu and Dr Rong Wei for their assistance with debugging the deep learning models. The authors also thank Dr Qi Sun for his help with the statistical methods.

XSL•FO
RenderX

Funding

This study was supported by the Space Medical Experiment Project of China Manned Space Program (HYZHXMH01005), Beijing Scholar 2015 (Zhenchang Wang), Beijing Friendship Hospital, Capital Medical University (YYZZ202351), and State Key Laboratory of Digestive Health (25ZX-QN05).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YL (lead), HY (equal), PR (equal)
Data curation: YL
Formal analysis: YL (lead), ZZ (supporting)
Funding acquisition: ZW (lead), PR (equal)
Investigation: YL (lead), TZ (supporting)
Methodology: YL, ZZ (supporting), WL (supporting), TZ (supporting), LC (supporting), HL (supporting)
Project administration: PR
Resources: YL (lead), HL (supporting)
Supervision: HY (lead), ZW (equal), ZY (supporting)
Validation: YL
Visualization: YL
Writing – original draft: YL (lead), HY (supporting), HN (supporting)
Writing – review & editing: YL (lead), ZY (supporting), ZW (supporting), PR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
The details of network structure and loss function.
[DOCX File, 15 KB - medinform_v14i1e79389_app1.docx]

References

1. Cashmore MT, McCann AJ, Wastling SJ, McGrath C, Thornton J, Hall MG. Clinical quantitative MRI and the need for metrology. Br J Radiol 2021 Apr 1;94(1120):20201215. [doi: 10.1259/bjr.20201215] [Medline: 33710907]
2. Warntjes JBM, Dahlqvist O, Lundberg P. Novel method for rapid, simultaneous T1, T2*, and proton density quantification. Magn Reson Med 2007 Mar;57(3):528-537. [doi: 10.1002/mrm.21165] [Medline: 17326183]
3. Blystad I, Warntjes JBM, Smedby O, Landtblom AM, Lundberg P, Larsson EM. Synthetic MRI of the brain in a clinical setting. Acta Radiol 2012 Dec 1;53(10):1158-1163. [doi: 10.1258/ar.2012.120195] [Medline: 23024181]
4. Gao W, Yang Q, Li X, et al. Synthetic MRI with quantitative mappings for identifying receptor status, proliferation rate, and molecular subtypes of breast cancer. Eur J Radiol 2022 Mar;148(157):110168. [doi: 10.1016/j.ejrad.2022.110168] [Medline: 35078137]
5. Zhang Z, Li S, Wang W, et al. Synthetic MRI for the quantitative and morphologic assessment of head and neck tumors: a preliminary study. Dentomaxillofac Radiol 2023 Sep;52(6):20230103. [doi: 10.1259/dmfr.20230103] [Medline: 37427697]
6. Ma L, Lian S, Liu H, et al. Diagnostic performance of synthetic magnetic resonance imaging in the prognostic evaluation of rectal cancer. Quant Imaging Med Surg 2022 Jul;12(7):3580-3591. [doi: 10.21037/qims-22-24] [Medline: 35782274]
7. Li M, Fu W, Ouyang L, et al. Potential clinical feasibility of synthetic MRI in bladder tumors: a comparative study with conventional MRI. Quant Imaging Med Surg 2023 Aug 1;13(8):5109-5118. [doi: 10.21037/qims-22-1419] [Medline: 37581035]
8. Bauer S, Markl M, Honal M, Jung BA. The effect of reconstruction and acquisition parameters for GRAPPA-based parallel imaging on the image quality. Magn Reson Med 2011 Aug;66(2):402-409. [doi: 10.1002/mrm.22803] [Medline: 21381100]
9. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]
10. Mardani M, Gong E, Cheng JY, et al. Deep generative adversarial neural networks for compressive sensing MRI. IEEE Trans Med Imaging 2019 Jan;38(1):167-179. [doi: 10.1109/TMI.2018.2858752] [Medline: 30040634]
11. Quan TM, Nguyen-Duc T, Jeong WK. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. IEEE Trans Med Imaging 2018 Jun;37(6):1488-1497. [doi: 10.1109/TMI.2018.2820120] [Medline: 29870376]

XSL•FO
RenderX

12. Yang G, Yu S, Dong H, et al. DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans Med Imaging 2018 Jun;37(6):1310-1321. [doi: 10.1109/TMI.2017.2785879] [Medline: 29870361]
13. Liu Y, Liu Y, Vanguri R, et al. 3D isotropic super-resolution prostate MRI using generative adversarial networks and unpaired multiplane slices. J Digit Imaging 2021 Oct;34(5):1199-1208. [doi: 10.1007/s10278-021-00510-w] [Medline: 34519954]
14. Su PY, Shih HJ, Xu JL. Rapid liver fibrosis evaluation using the UNet-ResNet50-32 × 4d model in magnetic resonance elastography: retrospective study. JMIR Med Inform 2025 Oct 20;13:e80351. [doi: 10.2196/80351] [Medline: 41115237]
15. Ran M, Hu J, Chen Y, et al. Denoising of 3D magnetic resonance images using a residual encoder-decoder Wasserstein generative adversarial network. Med Image Anal 2019 Jul;55:165-180. [doi: 10.1016/j.media.2019.05.001] [Medline: 31085444]
16. Latif S, Asim M, Usman M, Qadir J, Rana R. Automating motion correction in multishot MRI using generative adversarial networks. arXiv. Preprint posted online on Nov 24, 2018. [doi: 10.48550/arXiv.1811.09750]
17. Qiu S, Chen Y, Ma S, et al. Multiparametric mapping in the brain from conventional contrast-weighted images using deep learning. Magn Reson Med 2022 Jan;87(1):488-495. [doi: 10.1002/mrm.28962] [Medline: 34374468]
18. Liu Y, Niu H, Ren P, et al. Generation of quantification maps and weighted images from synthetic magnetic resonance imaging using deep learning network. Phys Med Biol 2022 Jan 17;67(2). [doi: 10.1088/1361-6560/ac46dd] [Medline: 34965516]
19. Don H, Supratak A, Mai L, et al. Tensorlayer: a versatile library for efficient deep learning development. Presented at: MM '17: Proceedings of the 25th ACM international conference on Multimedia; Oct 23-27, 2017; Mountain View, CA p. 1201-1204. [doi: 10.1145/3123266.3129391]
20. SPM8. UCL: Department of Imaging Neuroscience. URL: https://www.fil.ion.ucl.ac.uk/spm/software/spm8/ [accessed 2026-01-14]
21. Zheng Z, Liu Y, Yin H, et al. Evaluating T1, T2 relaxation, and proton density in normal brain using synthetic MRI with fast imaging protocol. Magn Reson Med Sci 2024 Oct 1;23(4):514-524. [doi: 10.2463/mrms.tn.2022-0161] [Medline: 37690836]
22. itk-SNAP. URL: http://www.itksnap.org/pmwiki/pmwiki.php [accessed 2026-01-14]
23. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004 Apr;13(4):600-612. [doi: 10.1109/tip.2003.819861] [Medline: 15376593]
24. Saad MA, Bovik AC, Charrier C. Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans on Image Process 2012 Aug;21(8):3339-3352. [doi: 10.1109/TIP.2012.2191563]
25. Kim E, Cho HH, Cho SH, et al. Accelerated synthetic MRI with deep learning-based reconstruction for pediatric neuroimaging. AJNR Am J Neuroradiol 2022 Nov;43(11):1653-1659. [doi: 10.3174/ajnr.A7664] [Medline: 36175085]
26. Zheng Z, Yang J, Zhang D, et al. The effect of scan parameters on T1, T2 relaxation times measured with multi-dynamic multi-echo sequence: a phantom study. Phys Eng Sci Med 2022 Jun;45(2):657-664. [doi: 10.1007/s13246-022-01128-0] [Medline: 35553390]
27. Hirano Y, Ishizaka K, Sugimori H, et al. Assessment of accuracy and repeatability of quantitative parameter mapping in MRI. Radiol Phys Technol 2024 Dec;17(4):918-928. [doi: 10.1007/s12194-024-00836-4] [Medline: 39198352]
28. McAllister A, Leach J, West H, Jones B, Zhang B, Serai S. Quantitative synthetic MRI in children: normative intracranial tissue segmentation values during development. AJNR Am J Neuroradiol 2017 Dec;38(12):2364-2372. [doi: 10.3174/ajnr.A5398] [Medline: 28982788]
29. Saccenti L, Andica C, Hagiwara A, et al. Brain tissue and myelin volumetric analysis in multiple sclerosis at 3T MRI with various in-plane resolutions using synthetic MRI. Neuroradiology 2019 Nov;61(11):1219-1227. [doi: 10.1007/s00234-019-02241-w] [Medline: 31209528]
30. Rolls ET, Huang CC, Lin CP, Feng J, Joliot M. Automated anatomical labelling atlas 3. Neuroimage 2020 Feb 1;206:116189. [doi: 10.1016/j.neuroimage.2019.116189] [Medline: 31521825]
31. Hagiwara A, Fujimoto K, Kamagata K, et al. Age-related changes in relaxation times, proton density, myelin, and tissue volumes in adult brain analyzed by 2-dimensional quantitative synthetic magnetic resonance imaging. Invest Radiol 2021 Mar 1;56(3):163-172. [doi: 10.1097/RLI.0000000000000720] [Medline: 32858581]
32. Hagiwara A, Hori M, Cohen-Adad J, et al. Linearity, bias, intrascanner repeatability, and interscanner reproducibility of quantitative multidynamic multiecho sequence for rapid simultaneous relaxometry at 3 T: a validation study with a standardized phantom and healthy controls. Invest Radiol 2019 Jan;54(1):39-47. [doi: 10.1097/RLI.0000000000000510] [Medline: 30300164]
33. Zheng Z, Liu Y, Wang Z, Yin H, Zhang D, Yang J. Evaluating age-and gender-related changes in brain volumes in normal adult using synthetic magnetic resonance imaging. Brain Behav 2024 Jul;14(7):e3619. [doi: 10.1002/brb3.3619] [Medline: 38970221]

Abbreviations

AAL: anatomical automatic labeling

XSL•FO
RenderX

CV: coefficient of variation
DL: deep learning
GM: gray matter
HR: high resolution
LR: low resolution
MRI: magnetic resonance imaging
NIQE: naturalness image quality evaluator
PD: proton density
PSNR: peak signal-to-noise ratio
ReLU: rectified linear unit
ROI: region of interest
SRGAN: superresolution generative adversarial network
SSIM: structural similarity image measure
TOST: two 1-sided tests
WM: white matter

Edited by A Coristine; submitted 20.Jun.2025; peer-reviewed by B Feige, J Bao; accepted 23.Dec.2025; published 23.Jan.2026.

Please cite as:
Liu Y, Yin H, Zheng Z, Liu W, Zhang T, Cai L, Niu H, Lv H, Yang Z, Wang Z, Ren P
Two-Minute Deep Learning–Powered Brain Quantitative Mapping: Accelerating Clinical Imaging With Synthetic Magnetic Resonance Imaging
JMIR Med Inform 2026;14:e79389
URL: https://medinform.jmir.org/2026/1/e79389
doi:10.2196/79389

© Yawen Liu, Hongxia Yin, Zuofeng Zheng, Wenjuan Liu, Tingting Zhang, Linkun Cai, Haijun Niu, Han Lv, Zhenghan Yang, Zhenchang Wang, Pengling Ren. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Nomograms Based on X-Ray Radiomics for Predicting Pain Progression in Knee Osteoarthritis Using Data From the Foundation for the National Institutes of Health: Development and Validation Study

Yingwei Sun[1*], MD; Jing Liu[2*], MSc; Chunbo Deng[3*], MD; Chengbao Peng[2], MSc; Shinong Pan[4], MD; Xueyong Liu[5], MD

[1]Department of Radiology, Affiliated Hospital of Liaoning University of Traditional Chinese Medicine, Shenyang, China

[2]Platform Engineering Research Center, Neusoft Research Institute of Healthcare Technology, Shenyang, China

[3]Department of Sports and Trauma, Central Hospital Affiliated to Shenyang Medical College, Shenyang, China

[4]Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, China

[5]Department of Rehabilitation, Shengjing Hospital of China Medical University, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Xueyong Liu, MD
Department of Rehabilitation
Shengjing Hospital of China Medical University
No.16 Puhe Road, Shenbei New District
Shenyang, 110134
China
Phone: 86 2418940112266
Email: liusjh@sj-hospital.org

Abstract

Background: Knee osteoarthritis (KOA) is one of the most prevalent chronic musculoskeletal disorders among the older adult population. Screening populations at risk of rapid progression of osteoarthritis and implementing appropriate early intervention strategies is advantageous for the treatment and prognosis of affected patients.

Objective: This study aimed to construct and validate a nomogram model based on x-ray radiomics to effectively identify individuals experiencing progression of KOA pain.

Methods: The Foundation for the National Institutes of Health Biomarkers Consortium included a total of 600 participants who were classified as pain progressors (n=297, 49.5%) and non–pain progressors (n=303, 50.5%) according to an increase in the Western Ontario and McMaster Universities Osteoarthritis Index pain score of ≥9 points (on a scale from 0 to 100) during the follow-up period of 24 to 48 months. X-rays that lacked defined spacing in the DICOM image were excluded. Fully automatic selection of subchondral bone regions on the inner and outer edges of the tibia and femur as regions of interest and extraction of radiomics features for different combinations of regions of interest were conducted. Least absolute shrinkage and selection operator regression was used to select features and generate a radiomics score using Shapley additive explanations for interpretability. The radiomics score, along with clinical indicators, was incorporated into nomograms using a multivariable logistic regression model. The subgroup analysis focused solely on the progression of pain and cases with no progression at all. The receiver operating characteristic curve, along with calibration and decision curves, was used to assess the discriminative performance.

Results: A total of 450 participants were included in the study. Shapley additive explanations analysis identified Wavelet-HH_gldm_HighGrayLevelEmphasis as the primary radiomics feature. Nomogram 1 and nomogram 2 for predicting KOA pain progression achieved area under the curve values of 0.766 and 0.753, respectively, with mean absolute errors of 0.012 and 0.008, respectively, in the calibration curves. Decision curve analysis showed a positive net benefit across a range of threshold probabilities. In subgroup analyses, nomogram 3 and nomogram 4 yielded areas under the curve of 0.795 and 0.740, respectively.

Conclusions: The nomograms based on x-ray radiomics demonstrated excellent predictive capability and accuracy in forecasting the progression of KOA pain.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e78338) doi:10.2196/78338

KEYWORDS

radiomics; osteoarthritis; knee; nomogram; artificial intelligence

Introduction

Osteoarthritis (OA) is one of the most common joint disorders affecting the older adult population. With the aging of the population, its incidence has been increasing year by year, imposing a significant burden on both individuals and society [1]. Knee OA (KOA) has the highest incidence among all types of OA. The etiology of KOA is multifactorial, and its pathogenesis remains unclear. KOA can be categorized into various clinical phenotypes and molecular subtypes, including metabolic, inflammatory, mechanical, and genetic [2]. The current treatment guidelines are not suitable for the early diagnosis of KOA and do not distinguish patients who may experience rapid progression [3]. Identifying patients at high risk of rapid progression of KOA will assist in patient education and in the prioritization and allocation of health care resources [4].

The subchondral bone is essential for knee joint structure. Its degeneration disrupts the homeostasis of the intra-articular environment, potentially triggering cartilage degeneration and injury. Furthermore, it is associated with the incidence and progression of OA [5,6]. X-ray imaging continues to be the most prevalent, convenient, and cost-effective modality for evaluating KOA. However, it can only provide a rough estimate of disease severity without quantitative analysis or predicting disease progression [7]. With the application of artificial intelligence in KOA, x-rays can now predict the progression of KOA [3]. Leung et al [8] developed a deep learning predictive model to assess the risk of progression in KOA using knee x-ray images. The findings of their study indicated that this deep learning model surpassed traditional binary outcome models, which are based on standard scoring systems, in predicting the likelihood of total knee arthroplasty among patients with OA [8]. Although deep learning technology is regarded as the most sophisticated image classification technique currently available, its inherent lack of interpretability poses challenges for users in comprehending the model's decision-making outcomes [9].

Radiomics is a novel approach that harnesses advanced image analysis tools and the rapid development of medical imaging data. By integrating artificial intelligence, it enables high-throughput extraction of information from standard medical images and applies the extracted data to clinical decision support systems to enhance the accuracy of diagnosis, prognosis, and prediction [10]. Radiomics has been gradually applied in the field of KOA. Current research primarily focuses on using radiomics to diagnose KOA and predict its radiological progression [11-14]. However, studies on radiomics for predicting the pain progression of KOA are still relatively scarce. A recent study that used magnetic resonance imaging (MRI) radiomics to predict pain progression in KOA demonstrated that the constructed radiomics model achieved an area under the receiver operating characteristic curve (AUC) of 0.79 to 0.86 for KOA pain progression prediction [15]. Nevertheless, the study incorporated 200 radiomics parameters and included omics data from 12- and 24-month follow-ups, resulting in high model complexity and limited practical applicability.

Therefore, this study aimed to develop a nomogram model integrating x-ray subchondral radiomics features and clinical characteristics to predict pain progression in KOA. The Shapley additive explanations (SHAP) method was incorporated to offer a comprehensive explanation and intuitive visualization of the model's predictive mechanisms. This approach allows for the identification of high-risk individuals with rapid pain progression among patients with KOA using cost-effective detection methods, thereby providing clinicians with evidence-based guidance for early intervention strategies.

Methods

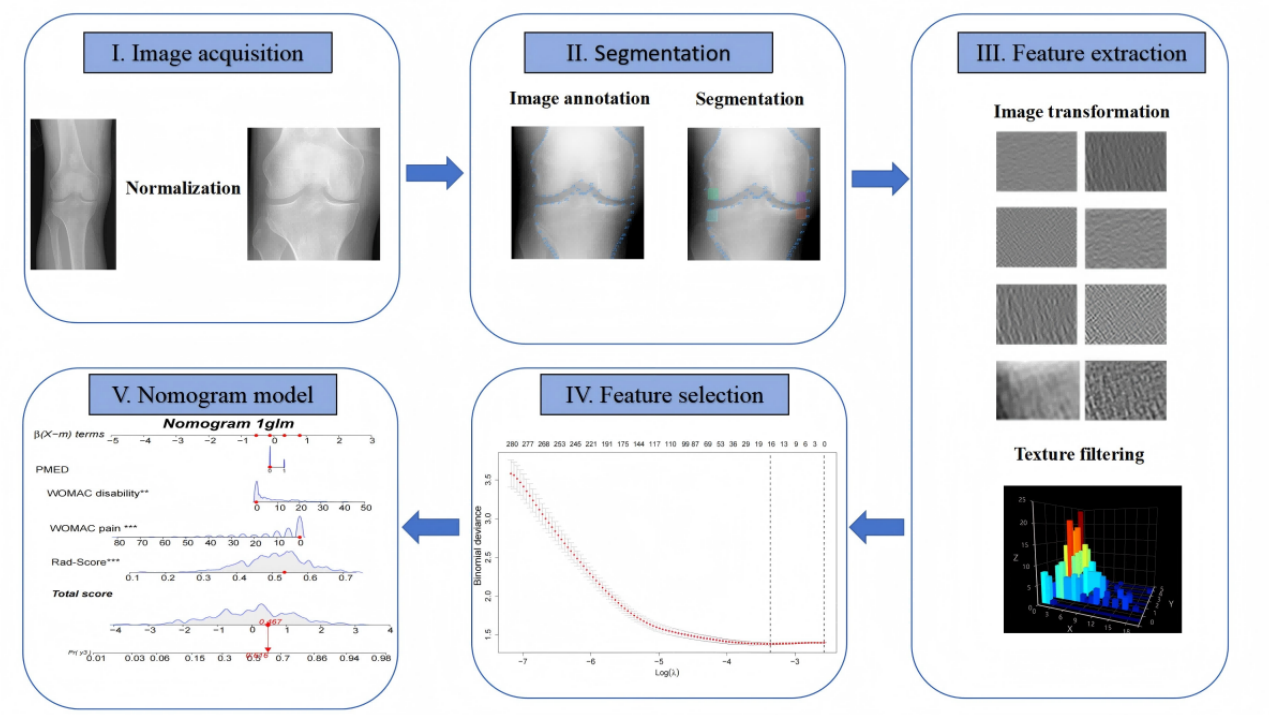
Participants

This study involved participants from the Foundation for the National Institutes of Health (FNIH) Osteoarthritis Biomarkers Project. The design of this research is detailed in previously published literature [16,17]. In brief, the FNIH Osteoarthritis Biomarkers Project includes 600 participants with KOA of Kellgren-Lawrence grades 1 to 3. According to the results of the follow-up, the 600 participants at baseline were categorized into four distinct subgroups: (1) participants exhibiting both radiological and pain progression (n=194, 32.3%), (2) participants with only radiological progression (n=103, 17.2%), (3) participants experiencing only pain progression (n=103, 17.2%), and (4) participants who exhibited neither radiological nor pain progression (n=200, 33.3%). Radiological progression was defined as a loss of minimum joint space width of ≥0.7 mm in the medial femorotibial compartment between baseline and follow-up at 24, 36, or 48 months. Pain progression was characterized by a sustained increase in the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain subscale (≥9 on a scale from 0 to 100) based on the minimum clinically important difference during the same time frame. In the Digital Imaging and Communications in Medicine (DICOM) standard, pixel spacing is a critical factor for converting image pixels into physical dimensions. Spacing ensures the reliability and reproducibility of medical images in diagnosis, research, and treatment planning [18]. However, some datasets lack or contain erroneous spacing information. They use default values (1.0, 1.0), which result in calculated knee dimensions significantly exceeding normal human anatomical ranges. Such data can be defined as "DICOM data with invalid spacing information." In the standardization process for the Osteoarthritis Initiative (OAI) knee x-ray images, cases with unspecified pixel spacing were excluded. As the pixel spacing of 0.2 mm was predominant in the dataset, all retained images were calibrated to this value to ensure consistent spatial scaling [19]. The specific process of normalization is described in Multimedia Appendix 1 [20-22]. Groups 1 and 3, which

XSL•FO
RenderX

showed pain progression, were combined as the case group, whereas groups 2 and 4, with no pain progression, served as the control group for analysis. Additionally, we conducted subgroup analyses on group 3 (with only pain progression) and group 4 (without both pain and radiological progression). The study flow diagram is shown in Figure 1.

Figure 1. Flowchart of the study. The x-ray images of the knee joint were normalized, and the regions of interest were automatically identified and selected. Subsequently, imaging omics parameters were extracted, followed by a screening process for feature parameters. Imaging omics labels were then constructed to evaluate clinical features, ultimately leading to the generation of a nomogram and an assessment of its performance.



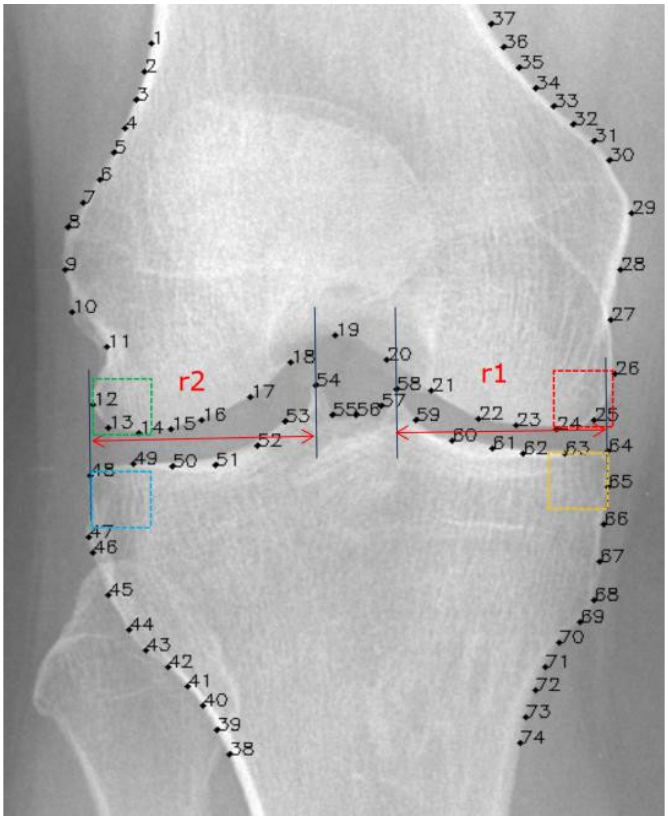
Segmentation and Region of Interest Annotation

BoneFinder (The University of Manchester) is a fully automated software tool designed to outline and segment skeletal structures from 2D radiographs by placing a series of points along bone contours or at key anatomical landmarks. In brief, BoneFinder uses a random forest voting mechanism to precisely determine the optimal placement of model feature points, enabling robust and accurate shape model matching. The system can accurately identify and segment bones not previously present in new images. This method demonstrates exceptional robustness and precision across various skeletal applications, with its generated point data being widely applicable to morphometric analyses, including the construction of statistical morphological models and automatic derivation of standard geometric measurement parameters [23], resulting in 74 markers per image (Figure 2). The fully automatic BoneFinder knee module achieves a mean point-to-curve error of less than 1.0 mm for 99% of 500 images (ie, in 99% of images, the error is less than 1.0 mm) without considering knee size [24].

Four vertical lines are constructed at the marked points 64, 58, 54, and 48 on the knee joint x-ray image, as identified by BoneFinder. These lines are sequentially designated as line 1, line 2, line 3, and line 4. The distance between line 1 and line 2 is denoted as R1, whereas the distance between line 3 and line 4 is referred to as R2. With point 64 designated as the vertex and a side length of R1/4, a square encompassing the medial tibia was defined as the region of interest (ROI). With marker point 64 designated as the vertex and a quarter of R1 as the side length, a square encompassing the medial tibia was defined as the ROI. First, using marker point 48 as the vertex and a quarter of R2 as the side length, a square area on the lateral tibia was selected to define ROI. Second, the selection process for the femur ROI differs from that of the tibia ROI. ROI 3 is defined with the line intersecting marker 24 as its lower boundary, while the line intersecting marker point 26 serves as the side boundary of this ROI. ROI 4 of the lateral femur uses the line indicated at point 14 as the lower boundary of the ROI and the column line marked at point 12 as its vertical boundary. Then, squares with side lengths of a quarter of R1 and a quarter of R2 are constructed. To assess the repeatability of ROI delineation, a total of 40 cases were randomly selected from the collected dataset. Two senior musculoskeletal radiologists (YWS and SNP) conducted manual segmentation of ROIs. Dice similarity coefficients were used to assess manual and automated segmentation agreement. Image annotation was performed using Neusoft Darts [25].

XSL•FO
RenderX

Figure 2. Schematic diagram of the automated region of interest segmentation process.



Radiomics Feature Extraction

The extraction of radiomics features was based on Neusoft Discovery [25]. The features that were extracted can be categorized into 3 distinct groups. First-order features were derived from the gray values of images, reflecting the distribution and heterogeneity of intensity within the ROI, typically including maximum, minimum, mean, median, range, variance, kurtosis, skewness, and entropy. The second-order features refer to the extracted texture characteristics, which include texture feature values derived from the gray-level co-occurrence matrix, gray-level run-length matrix, gray-level connected region size matrix, neighborhood difference matrix, and gray-level correlation matrix. The texture features can quantify the distribution of textures within the ROI that are difficult to perceive visually. Third, features were extracted after log transformation, first-level wavelet decomposition, and gradient transformation were applied to the ROI. All features were normalized using the following formula: $Z=(x–\mu)/\sigma$.

Radiomics Feature Selection and Radiomics Signature Construction

The Spearman correlation was used to remove the redundant features. The feature subset was optimized using the least absolute shrinkage and selection operator (LASSO) method in logistic regression. A 10-fold cross-validation approach was used to identify the optimal model that minimizes the mean squared error, and the correlation coefficients of the features selected by the best LASSO model were calculated. According to the coefficients derived from LASSO regression, a weighted calculation was performed to obtain the radiomics score (rad-score). Radiomics parameters from 4 ROIs (ROIs 1-4) were extracted individually, and rad-score 1 was developed following LASSO screening of the nonzero parameters. The radiomics parameters were applied to ROI 1 and ROI 3, leading to the development of rad-score 2. In subgroup analysis, rad-score 3, obtained from extracted and screened nonzero parameters from ROIs 1 to 4, and rad-score 4, obtained from extracted and screened nonzero parameters from ROI 1 and ROI 3, predicted pain progression only and no progression at all, respectively. SHAP analysis was used to mitigate the “black-box” effect of machine learning, enabling feature prioritization and interpretability of individual radiomics features on the rad-score. Feature maps were used to visualize the results of the rad-score.

Development and Evaluation of Nomograms

The analysis used stepwise backward multivariable logistic regression to identify clinical risk factors at baseline and construct nomograms. The Akaike information criterion likelihood ratio test was used as the stopping rule for the backward stepwise logistic regression analysis. Clinical information from patients at baseline included age, sex, history of injury, BMI, the minimum joint space width of the medial tibial-femoral compartment, Kellgren-Lawrence grade, WOMAC pain score and WOMAC disability score, medicine for pain in the previous 12 months, and race. The nomograms were developed based on the prediction probability value of the optimal radiomics models and clinical independent risk factors. The discriminative ability of the nomogram was evaluated using the AUC and accuracy. Calibration of the nomogram was performed through calibration curves, with optimal calibration indicated by alignment with the diagonal line. Internal validation was conducted using a bootstrap method involving 1000

XSL•FO
RenderX

resampling iterations, and the mean absolute error was used to assess model calibration. Additionally, decision curve analysis (DCA) quantified net benefits at various threshold probabilities to evaluate the clinical effectiveness of the nomograms.

Statistical Analysis

The comparison of categorical variables was conducted using the Pearson chi-square test. Continuous variables were expressed as mean and SD and analyzed using either the 2-tailed *t* test or Mann-Whitney *U* test. The normality of continuous data was assessed using the Kolmogorov-Smirnov test; measurements that followed a normal distribution were subjected to the *t* test, whereas those that did not conform to a normal distribution underwent the Mann-Whitney *U* test. The DeLong test was used for comparing 2 models. *P* values were calculated using a 2-tailed *z* test, with $P<.05$ considered statistically significant. All statistical analyses were performed using the EmpowerStats software (X&Y Solutions, Inc) and the R software (version 4.2.2; R Foundation for Statistical Computing).

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Institutional Review Board of the University of California, San Francisco (approval 10-00532), and the 4 clinical centers of the OAI project recognized the study as HIPAA (Health Insurance Portability and Accountability Act) compliant. All participants provided written informed consent. Patient-identifiable information (including names and hospital ID numbers) was carefully removed from the x-ray images and reports before analysis. Due to the retrospective design of this study, participants received no compensation.

Results

Baseline Clinical Characteristics

Due to the lack of spacing in DICOM images, a total of 450 participants were ultimately included in the study, with 227 (50.4%) and 223 (49.6%) in the control and case groups, respectively. The baseline characteristics of participants in both the case and control groups were frequency matched for factors including age, BMI, WOMAC functional scores, Kellgren-Lawrence grade, pain medication history, and race. Notably, a statistically significant difference was observed in WOMAC pain scores ($P=.04$) and medication history ($P=.02$) between the 2 groups (Table 1). In subgroup analysis, 18% (n=81) of the individuals demonstrated only pain progression during the follow-up period, whereas 32.9% (n=148) of the individuals exhibited neither pain progression nor radiographic deterioration. The baseline characteristics of participants across both groups were well matched, with no statistically significant differences observed (Table S1 in Multimedia Appendix 2).

Table 1. Baseline characteristics of the study population.

Characteristic	Control group (n=223)	Case group (n=227)	*P* value
Demographics			
Age (years), mean (SD)	62.14 (8.91)	61.48 (8.79)	.38
Female sex, n (%)	131 (58.7)	136 (59.9)	.80
Race, n (%)			.44
Asian	3 (1.3)	2 (0.9)	
Black or African American	21 (9.4)	32 (14.1)	
White	196 (87.9)	191 (84.1)	
Other non-White race	3 (1.3)	2 (0.9)	
Clinical measures			
BMI (kg/m^2), mean (SD)	30.36 (4.85)	30.50 (4.71)	.75
WOMAC[a] pain score (0-20), mean (SD)	2.70 (3.40)	1.87 (2.47)	.04
WOMAC function score (0-68), mean (SD)	8.15 (10.98)	8.13 (9.18)	.15
History of knee injury, n (%)	77 (34.5)	83 (36.6)	.65
Use of pain medication, n (%)	49 (22.0)	73 (32.2)	.02
Radiographic measures			
Kellgren-Lawrence grade, n (%)			.70
1	29 (13.0)	29 (12.8)	
2	112 (50.2)	106 (46.7)	
3	82 (36.8)	92 (40.5)	
Minimum medial joint space width (mm), mean (SD)	3.77 (1.10)	3.80 (1.32)	.87

[a]WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

XSL•FO
RenderX

Segmentation and Feature Extraction

The automated segmentation dataset was used for radiomics feature extraction as the Dice similarity coefficients between automatic and manual segmentation were excellent (0.800; Table S2 in Multimedia Appendix 2). We used 10-fold cross-validation to identify the optimal tuning parameter (alpha). In total, 16 nonzero coefficient features were derived from the optimal tuning parameters in rad-score 1. Rad-score 2 comprised 12 nonzero coefficient features. In subgroup analysis, rad-score 3 included 11 nonzero feature parameters, whereas rad-score 4 consisted of 3 nonzero features (Tables S3-S6 in Multimedia Appendix 2 and Figures S1-S4 in Multimedia Appendix 2). The correlation among features was analyzed using Spearman correlation analysis, and heat maps are presented in Figures S1 to S4 in Multimedia Appendix 2.

Development and Evaluation of Radiomics Signatures

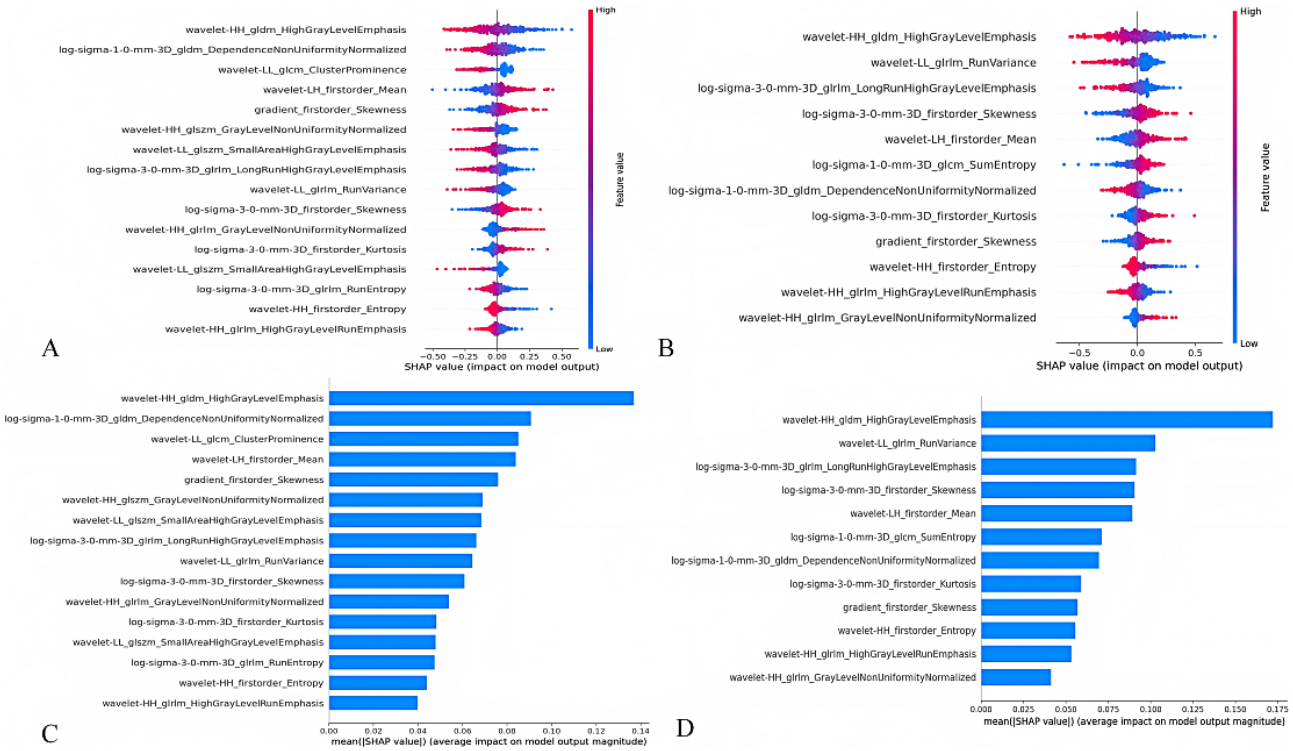
Table 2 shows the diagnostic proficiency of 4 radiomics signatures. The AUC value for rad-score 1 was 0.728 (95% CI 0.681-0.774), the AUC value for rad-score 2 was 0.716 (95% CI 0.635-0.736), the AUC value for rad-score 3 was 0.775 (95% CI 0.714-0.837), and the AUC value for rad-score 4 was 0.679 (95% CI 0.604-0.755; Table 2). The SHAP plot in Figure 3 visually illustrates how radiomics features influenced model predictions in terms of magnitude and direction. Features are sorted vertically by global importance. Each point represents the SHAP value of each feature for a specific patient plotted horizontally and stacked vertically to show density distribution. The color of the point reflects the feature value, from low (blue) to high (red; Figures 3A and C). The bar chart illustrates the impact of various imaging features on the rad-score. Wavelet-HH_gldm_HighGrayLevelEmphasis showed significant effects on both rad-score 1 and rad-score 2 (Figures 3B and D). The SHAP values of each feature in rad-score 3 and rad-score 4 are shown in Figure S5 in Multimedia Appendix 2. Figure 4 shows representative participants with x-ray subchondral bone feature maps from 4 ROIs. The case group participant demonstrated higher ROI heterogeneity than the control group participant, with a higher score (0.58 vs 0.44).

Table 2. Diagnostic efficacy of the 4 radiomics scores.

Radiomics score	AUC[a] (95% CI)	Accuracy	Sensitivity	Specificity
1	0.728 (0.681-0.774)	0.684	0.797	0.570
2	0.716 (0.635-0.736)	0.668	0.683	0.677
3	0.775 (0.714-0.837)	0.742	0.753	0.716
4	0.679 (0.604-0.755)	0.694	0.580	0.757

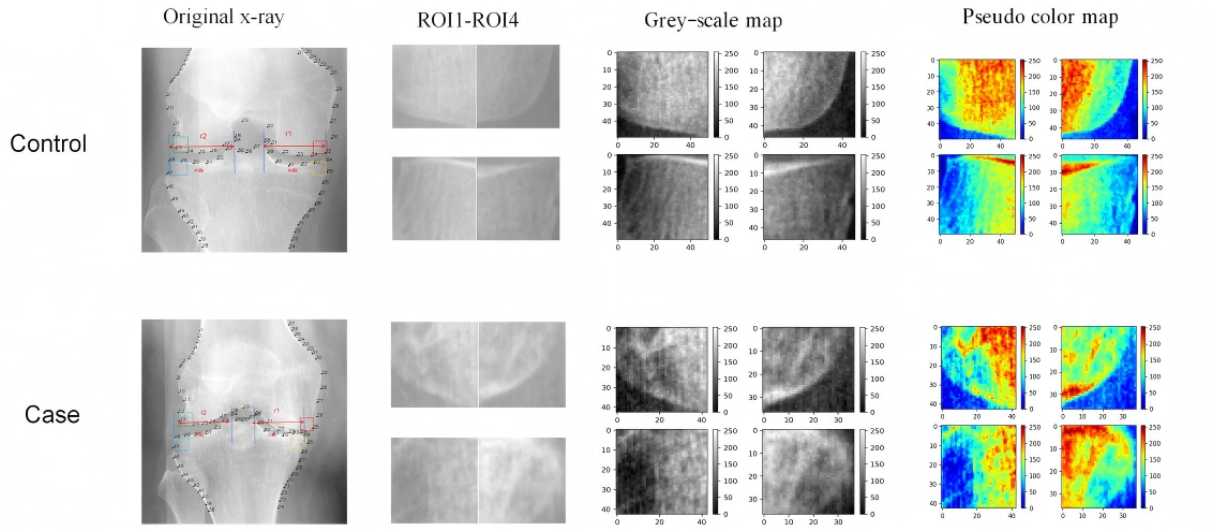
[a]AUC: area under the receiver operating characteristic curve.

Figure 3. Shapley additive explanations (SHAP) summary plots of the radiomics score. The plot illustrates the feature relevance and attributions of the model's predictive performance: (A) rad-score 1 and (B) rad-score 2. The bar chart illustrates the impact of various imaging features on the rad-score: (C) rad-score 1 and (D) rad-score 2.



XSL•FO
RenderX

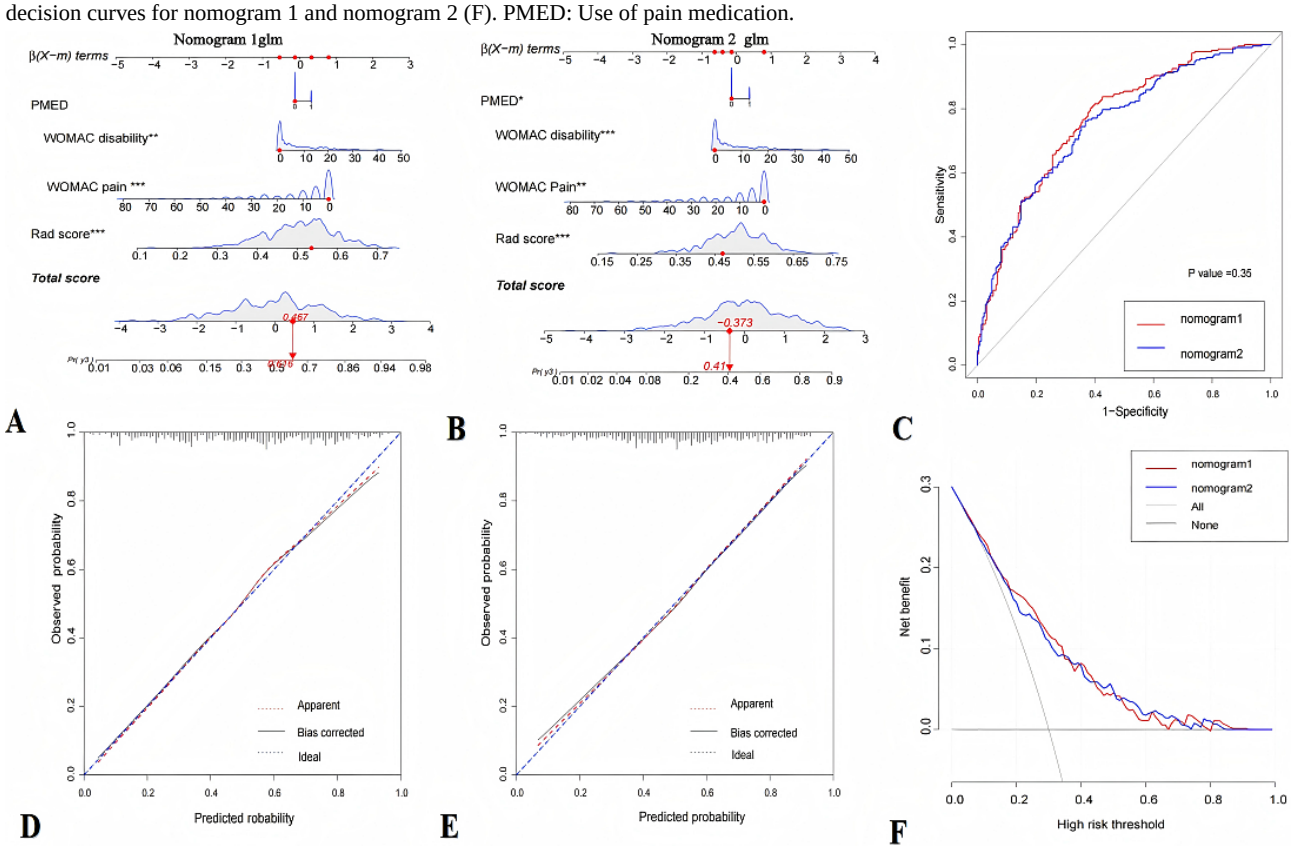
Figure 4. Visualization of radiomics signatures. The case group participant demonstrated higher region of interest heterogeneity than the control group participant, with a higher score.



Construction and Testing of Nomograms

Through multiple regression analysis, the WOMAC disability score, pain medication history, and WOMAC pain score were found to be risk factors for KOA pain progression. We constructed visual nomograms by integrating the score with selected clinical risk factors. The AUC of nomogram 1 was 0.766 (95% CI 0.722-0.809), and the AUC of nomogram 2 was 0.753 (95% CI 0.708-0.797). The comparison of AUC between the 2 models showed no statistically significant difference (DeLong test; *P*>.05; Figure 3). The calibration curve demonstrated that all cohorts were accurately aligned with the actual observations. The internal validation was conducted using the bootstrap method, with a resampling count set to 1000. The mean absolute errors of the calibration curves for the 2 models were found to be 0.012 and 0.008, indicating that the model demonstrated a high level of calibration accuracy (Figure 5 and Table 3). The DCA demonstrated that, if the threshold probability in the clinical decision was in the range of 10% to 70%, both nomogram 1 and nomogram 2 could yield more net benefits compared with either the treat-all-patients strategy or the treat-none strategy.

Figure 5. Development and validation of the nomogram to predict pain progression in knee osteoarthritis (A: nomogram 1; B: nomogram 2), the receiver operating characteristic curves of nomogram 1 and nomogram 2(C), the calibration curves of nomogram 1 and nomogram 2 (D and E, respectively), and decision curves for nomogram 1 and nomogram 2 (F). PMED: Use of pain medication.



XSL•FO
RenderX

Table 3. Diagnostic efficacy of the 4 nomogram models.

Model	AUC[a] (95% CI)	Accuracy	Sensitivity	Specificity	MAE[b]
Nomogram 1	0.766 (0.722-0.809)	0.709	0.692	0.726	0.012
Nomogram 2	0.753 (0.708-0.797)	0.693	0.648	0.740	0.008
Nomogram 3	0.795 (0.734-0.865)	0.760	0.691	0.797	0.011
Nomogram 4	0.740 (0.673-0.808)	0.686	0.740	0.695	0.013

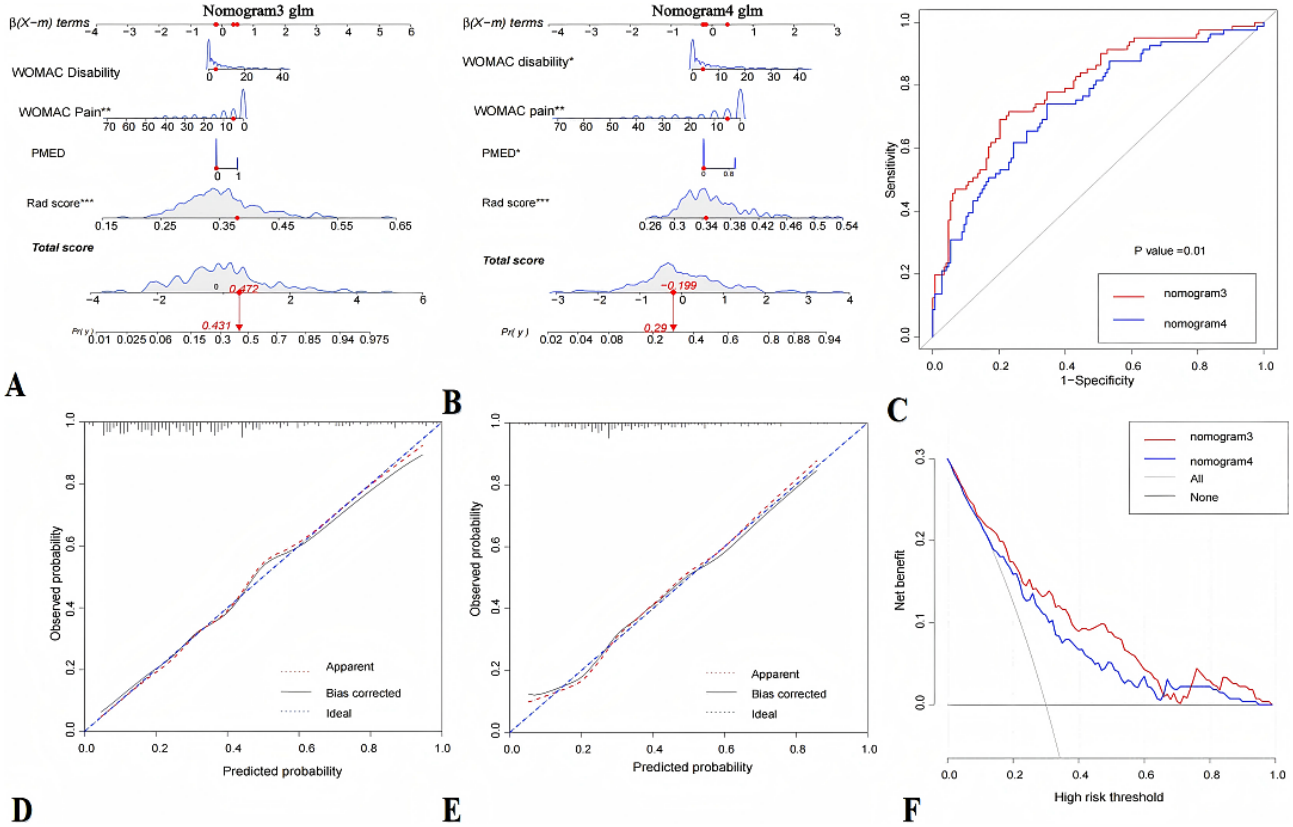
[a]AUC: area under the receiver operating characteristic curve.

[b]MAE: mean absolute error.

Subgroup Analyses

The AUC of nomogram 3 was 0.795 (95% CI 0.734-0.865), and the AUC of nomogram 4 was 0.740 (95% CI 0.673-0.808). The comparison of the AUC between the 2 models showed a statistically significant difference (DeLong test; $P<.05$). The calibration curve demonstrated a robust concordance between the probabilities predicted and the actual observed outcomes. The DCAs demonstrated that, if the threshold probability in the clinical decision was within the range of 10% to 65%, the curve corresponding to nomogram 3 was above nomogram 4, and the AUC that it formed with the "treat none" and "treat all" lines was relatively larger. The net benefit of nomogram 3 was superior to that of nomogram 4, making it the better model (Table 3 and Figure 6).

Figure 6. Development and validation of the nomogram to predict pain progression in knee osteoarthritis (A: nomogram 3; B: nomogram 4), the receiver operating characteristic curves of nomogram 3 and nomogram 4 (C), the calibration curves of nomogram 3 and nomogram 4 (D and E, respectively), and decision curves for nomogram 3 and nomogram 4 (F). PMED: Use of pain medication.



Discussion

Principal Findings

Our study developed a nomogram integrating x-ray radiomics features and clinical variables to predict pain progression in KOA. The resulting baseline-based, cost-effective nomogram demonstrated commendable accuracy and reproducibility in forecasting KOA pain progression, offering significant potential for guiding early intervention strategies in at-risk populations. The application of advanced artificial intelligence to widely available and cost-effective x-ray imaging represents a feasible and promising approach for predicting KOA pain progression. Radiomics models can automatically classify KOA [26], and texture analysis of the subchondral bone, particularly in the tibia, has shown predictive value for radiographic KOA incidence [27,28]. Given the complex etiology of KOA and the frequent discrepancy between clinical symptoms and radiological severity, we focused on subchondral bone remodeling—a process linked to pain pathogenesis through

XSL•FO
RenderX

molecules such as netrin-1 and PEG2, which promote sensory nerve innervation [29,30]. This rationale supported our use of subchondral bone radiomics for predicting symptomatic progression. Radiomics enables high-throughput extraction of quantitative features from medical images, providing data for clinical decision support systems [10]. The selection of key ROIs is critical as different image areas contain varying levels of information relevant to disease progression [31]. While the medial subchondral bone often shows a stronger correlation with OA [32] and the medial tibial margin is recognized as highly informative due to its role in absorbing uneven mechanical loads [33], the optimal ROI selection remains debated. Some deep learning studies for KOA pain progression have also concentrated on the medial joint space and osteophytes [8], whereas other radiomics research suggests that including the entire tibial subchondral bone may yield better outcomes than focusing solely on specific compartments [17].

In this study, we systematically evaluated the impact of ROI selection on model accuracy by constructing radiomics signatures both from 4 ROIs encompassing the entire knee joint and from the 2 medial ROIs alone. Our findings revealed that incorporating lateral ROIs did not improve the overall model AUC compared to using medial ROIs only. However, subgroup analysis indicated that models combining medial and lateral ROIs achieved significantly higher AUC values in specific contexts. The primary cause of this discrepancy may stem from disease heterogeneity. Notably, in the main analysis, both the pain progression group and the nonprogression group included participants with radiological progression. Pathological changes in the lateral subchondral bone may accelerate the radiological progression of the knee joint [34-36]. Therefore, the role of radiomics features in lateral ROIs during pain progression was likely diluted or masked by features from radiologically progressive participants in the control group.

When compared to prior studies, our model demonstrated competitive predictive performance. For instance, Kraus et al [16] applied 3 texture parameters to predict KOA pain progression, achieving AUCs ranging from 0.603 to 0.649. Our rad-score yielded AUCs between 0.679 and 0.775, representing a substantial improvement. This aligns with the understanding that multivariable models integrating multiple risk factors generally provide more reliable estimates than those based on single parameters [37]. The superior performance of our baseline-based model is particularly notable compared to previously reported predictive capabilities relying solely on baseline biomarkers [38]. A recent study using MRI-based radiomics to predict KOA pain progression reported high predictive accuracy, with AUC values ranging from 0.79 to 0.86 [15]. This performance is commendable. However, it is important to note that their model incorporated a substantially larger set of 200 radiomics parameters and, crucially, relied on omics data extracted from 12- and 24-month follow-up MRI scans. While this approach yields high accuracy, the model's complexity and its dependence on longitudinal imaging data significantly limit its practicality and scalability in routine clinical settings. In contrast, our study deliberately pursued a more streamlined and clinically translatable strategy. By using only baseline x-ray images and constructing a parsimonious model using key radiomics features and clinical variables, our baseline-based model eliminates the need for costly follow-up MRI scans and complex feature sets, making it more feasible for immediate clinical application and potential widespread adoption.

Research using deep learning methods to predict KOA pain progression remains scarce. A study using OAI data demonstrated that deep learning models achieved an AUC of 0.77 for predicting KOA pain progression, significantly higher than the 0.692 of traditional models [39]. By integrating demographic, clinical, and imaging risk factors with a deep learning model incorporating baseline knee x-rays, the highest diagnostic efficacy (AUC=0.807) was achieved in predicting pain progression. However, the original study failed to provide specific feature information. In contrast, our study visually presented each feature parameter's impact on the rad-score using Shapley summaries. Wavelet-HH_gldm_HighGrayLevelEmphasis, a gray-level correlation matrix feature derived from high-frequency (HH) subbands of wavelet-transformed images, reflects the concentration of high–gray value regions and emphasizes texture variations in these areas. By quantifying texture changes in knee images, this feature effectively captures tissue degeneration differences caused by pain progression, indirectly reflecting pathological changes and potentially enabling prediction of knee pain progression.

The nomogram was selected for its ability to simplify complex statistical models into user-friendly graphical interfaces that generate individualized risk estimates [40]. While nomograms have been applied in OA contexts such as predicting outcomes after total knee replacement [41,42], their application specifically for predicting KOA pain progression represents a novel contribution. Our nomogram uniquely integrates radiomics signatures with key clinical variables: baseline WOMAC pain scores, WOMAC disability scores, and analgesic use. Interestingly, the baseline WOMAC pain score served as a negative predictor in our model, potentially reflecting the development of tolerance to chronic pain in patients with KOA [43]. The inclusion of WOMAC disability scores and analgesic use is supported by their established roles as risk factors in KOA phenotypic and prognostic models [44,45].

Limitations

This study has several limitations. First, the relatively small sample size may affect statistical power and generalizability. Second, the use of data from previous nested case-control studies could introduce selection bias. Third, our analysis relied solely on internal validation because it was difficult to find an equivalent external cohort. Although the data were sourced from the well-established FNIH Osteoarthritis Biomarkers Project, external validation using independent datasets, such as those from the PROGRESS OA study (phase 2 of the FNIH Biomarkers consortium), was not feasible as these data have not been publicly released [46]. To our knowledge, no other studies have explored the specific pain progression patterns defined by the FNIH [15], and other studies using FNIH data have similarly been unable to perform external validation.

XSL•FO
RenderX

Conclusions

A nomogram based on baseline x-ray radiomics signatures can effectively predict the progression of pain in KOA. This has significant implications for supporting health care professionals in tailoring treatment plans for individuals experiencing rapid pain progression, as well as in reducing health care costs, and these findings need further validation in future trials.

Acknowledgments

The data for this study were obtained and analyzed from the controlled-access datasets distributed from the Osteoarthritis Initiative, a data repository housed within the National Institute of Mental Health Data Archive. The Osteoarthritis Initiative represents a public-private partnership consisting of 5 contracts funded by the National Institutes of Health.

Funding

This study received funding from Liaoning Province Science and Technology Plan Joint Program (2024012243-JH4/4800) and Liaoning Province Science and Technology Plan Joint Technical Breakthrough Plan Project (2024JH2/10 2600325).

Data Availability

The datasets generated or analyzed during this study are available in the Osteoarthritis Initiative repository [47].

Authors' Contributions

Conceptualization: YS, XL
Data curation: YS
Funding acquisition: YS, XL
Investigation: SP
Methodology: JL, CD, XL
Project administration: JL, XL
Resources: JL
Software: CP
Supervision: CP, SP, XL
Validation: YS, XL
Visualization: YS, XL
Writing—original draft: YS, CD
Writing—review and editing: YS, CD, XL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Size normalization and pixel normalization of the x-ray.
[DOCX File , 198 KB - medinform_v14i1e78338_app1.docx]

Multimedia Appendix 2
Baseline characteristics of the subgroup population, radiomic characteristics, and Dice and Shapley additive explanations summary plots of the radiomic score.
[DOCX File , 1921 KB - medinform_v14i1e78338_app2.docx]

References

1. Safiri S, Kolahi AA, Smith E, Hill C, Bettampadi D, Mansournia MA, et al. Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the Global Burden of Disease Study 2017. Ann Rheum Dis 2020 Jun;79(6):819-828. [doi: 10.1136/annrheumdis-2019-216515] [Medline: 32398285]
2. Fan X, Sun AR, Young RS, Afara IO, Hamilton BR, Ong LJ, et al. Spatial analysis of the osteoarthritis microenvironment: techniques, insights, and applications. Bone Res 2024 Feb 04;12(1):7 [FREE Full text] [doi: 10.1038/s41413-023-00304-6] [Medline: 38311627]
3. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nat Rev Rheumatol 2019 Jan;15(1):49-60. [doi: 10.1038/s41584-018-0130-5] [Medline: 30523334]
4. Atukorala I, Pathmeswaran A, Makovey J, Metcalf B, Bennell KL, March L, et al. Can pain flares in knee osteoarthritis be predicted? Scand J Rheumatol 2021 May;50(3):198-205. [doi: 10.1080/03009742.2020.1829035] [Medline: 33467963]

XSL•FO
RenderX

5. Hu W, Chen Y, Dou C, Dong S. Microenvironment in subchondral bone: predominant regulator for the treatment of osteoarthritis. Ann Rheum Dis 2021 Apr;80(4):413-422 [FREE Full text] [doi: 10.1136/annrheumdis-2020-218089] [Medline: 33158879]
6. Hu Y, Chen X, Wang S, Jing Y, Su J. Subchondral bone microenvironment in osteoarthritis and pain. Bone Res 2021 Mar 17;9(1):20 [FREE Full text] [doi: 10.1038/s41413-021-00147-z] [Medline: 33731688]
7. Roemer FW, Demehri S, Omoumi P, Link TM, Kijowski R, Saarakkala S, et al. State of the art: imaging of osteoarthritis-revisited 2020. Radiology 2020 Jul;296(1):5-21. [doi: 10.1148/radiol.2020192498] [Medline: 32427556]
8. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. Radiology 2020 Sep;296(3):584-593 [FREE Full text] [doi: 10.1148/radiol.2020192091] [Medline: 32573386]
9. Guan B, Liu F, Haj-Mirzaian A, Demehri S, Samsonov A, Neogi T, et al. Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-MONTH follow-up period. Osteoarthritis Cartilage 2020 Apr;28(4):428-437 [FREE Full text] [doi: 10.1016/j.joca.2020.01.010] [Medline: 32035934]
10. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017 Dec;14(12):749-762. [doi: 10.1038/nrclinonc.2017.141] [Medline: 28975929]
11. Jiang H, Peng Y, Qin SY, Chen C, Pu Y, Liang R, et al. MRI-based radiomics and delta-radiomics models of the patella predict the radiographic progression of osteoarthritis: data from the FNIH OA biomarkers consortium. Acad Radiol 2024 Apr;31(4):1508-1517. [doi: 10.1016/j.acra.2023.10.003] [Medline: 37923575]
12. Lin T, Peng S, Lu S, Fu S, Zeng D, Li J, et al. Prediction of knee pain improvement over two years for knee osteoarthritis using a dynamic nomogram based on MRI-derived radiomics: a proof-of-concept study. Osteoarthritis Cartilage 2023 Feb;31(2):267-278 [FREE Full text] [doi: 10.1016/j.joca.2022.10.014] [Medline: 36334697]
13. Almhdie-Imjabbar A, Podsiadlo P, Ljuhar R, Jennane R, Nguyen KL, Toumi H, et al. Trabecular bone texture analysis of conventional radiographs in the assessment of knee osteoarthritis: review and viewpoint. Arthritis Res Ther 2021 Aug 06;23(1):208 [FREE Full text] [doi: 10.1186/s13075-021-02594-9] [Medline: 34362427]
14. Li W, Feng J, Zhu D, Xiao Z, Liu J, Fang Y, et al. Nomogram model based on radiomics signatures and age to assist in the diagnosis of knee osteoarthritis. Exp Gerontol 2023 Jan;171:112031 [FREE Full text] [doi: 10.1016/j.exger.2022.112031] [Medline: 36402414]
15. Wang T, Liu H, Zhao W, Cao P, Li J, Chen T, et al. Predicting knee osteoarthritis progression using neural network with longitudinal MRI radiomics, and biochemical biomarkers: a modeling study. PLoS Med 2025 Aug 21;22(8):e1004665 [FREE Full text] [doi: 10.1371/journal.pmed.1004665] [Medline: 40839548]
16. Kraus VB, Collins JE, Charles HC, Pieper CF, Whitley L, Losina E, et al. Predictive validity of radiographic trabecular bone texture in knee osteoarthritis: The Osteoarthritis Research Society International/Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. Arthritis Rheumatol 2018 Jan;70(1):80-87 [FREE Full text] [doi: 10.1002/art.40348] [Medline: 29024470]
17. Almhdie-Imjabbar A, Toumi H, Lespessailles E. Performance of radiological and biochemical biomarkers in predicting radio-symptomatic knee osteoarthritis progression. Biomedicines 2024 Mar 16;12(3):666 [FREE Full text] [doi: 10.3390/biomedicines12030666] [Medline: 38540279]
18. Chen B, Duan X, Yu Z, Leng S, Yu L, McCollough C. Technical Note: development and validation of an open data format for CT projection data. Med Phys 2015 Dec;42(12):6964-6972 [FREE Full text] [doi: 10.1118/1.4935406] [Medline: 26632052]
19. Swiecicki A, Li N, O'Donnell J, Said N, Yang J, Mather RC, et al. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. Comput Biol Med 2021 Jun;133:104334. [doi: 10.1016/j.compbiomed.2021.104334] [Medline: 33823398]
20. Singh A, Horng H, Roshkovan L, Weeks JK, Hershman M, Noël P, et al. Development of a robust radiomic biomarker of progression-free survival in advanced non-small cell lung cancer patients treated with first-line immunotherapy. Sci Rep 2022 Jun 15;12(1):9993 [FREE Full text] [doi: 10.1038/s41598-022-14160-7] [Medline: 35705618]
21. Almhdie-Imjabbar A, Nguyen KL, Toumi H, Jennane R, Lespessailles E. Prediction of knee osteoarthritis progression using radiological descriptors obtained from bone texture analysis and Siamese neural networks: data from OAI and MOST cohorts. Arthritis Res Ther 2022 Mar 08;24(1):66 [FREE Full text] [doi: 10.1186/s13075-022-02743-8] [Medline: 35260192]
22. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 2018 Jan 29;8(1):1727 [FREE Full text] [doi: 10.1038/s41598-018-20132-7] [Medline: 29379060]
23. Lindner C, Thiagarajah S, Wilkinson JM, arcOGEN Consortium, Wallis GA, Cootes TF. Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE Trans Med Imaging 2013 Aug;32(8):1462-1472. [doi: 10.1109/TMI.2013.2258030] [Medline: 23591481]
24. Lindner C, Thiagarajah S, Wilkinson JM, Wallis GA, Cootes TF, arcOGEN Consortium. Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting. Med Image Comput Comput Assist Interv 2013;16(Pt 2):181-189. [doi: 10.1007/978-3-642-40763-5_23] [Medline: 24579139]

XSL•FO
RenderX

25. Li Y, Chen D, Liu S, Lin J, Wang W, Huang J, et al. Supervised training models with or without manual lesion delineation outperform clinicians in distinguishing pulmonary cryptococcosis from lung adenocarcinoma on chest CT. Mycoses 2024 Jan;67(1):e13692. [doi: 10.1111/myc.13692] [Medline: 38214431]
26. Li W, Liu J, Xiao Z, Zhu D, Liao J, Yu W, et al. Automatic grading of knee osteoarthritis with a plain radiograph radiomics model: combining anteroposterior and lateral images. Insights Imaging 2024 Jun 13;15(1):143 [FREE Full text] [doi: 10.1186/s13244-024-01719-3] [Medline: 38867121]
27. Janvier T, Jennane R, Toumi H, Lespessailles E. Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2017 Dec;25(12):2047-2054 [FREE Full text] [doi: 10.1016/j.joca.2017.09.004] [Medline: 28935435]
28. Pishgar F, Guermazi A, Roemer FW, Link TM, Demehri S. Conventional MRI-based subchondral trabecular biomarkers as predictors of knee osteoarthritis progression: data from the Osteoarthritis Initiative. Eur Radiol 2021 Jun;31(6):3564-3573 [FREE Full text] [doi: 10.1007/s00330-020-07512-2] [Medline: 33241511]
29. Jiang W, Jin Y, Zhang S, Ding Y, Huo K, Yang J, et al. PGE2 activates EP4 in subchondral bone osteoclasts to regulate osteoarthritis. Bone Res 2022 Mar 09;10(1):27 [FREE Full text] [doi: 10.1038/s41413-022-00201-4] [Medline: 35260562]
30. Zhu S, Zhu J, Zhen G, Hu Y, An S, Li Y, et al. Subchondral bone osteoclasts induce sensory innervation and osteoarthritis pain. J Clin Invest 2019 Mar 01;129(3):1076-1093 [FREE Full text] [doi: 10.1172/JCI121561] [Medline: 30530994]
31. Jiang T, Lau SH, Zhang J, Chan LC, Wang W, Chan PK, et al. Radiomics signature of osteoarthritis: current status and perspective. J Orthop Translat 2024 Mar 16;45:100-106 [FREE Full text] [doi: 10.1016/j.jot.2023.10.003] [Medline: 38524869]
32. Hirvasniemi J, Klein S, Bierma-Zeinstra S, Vernooij MW, Schiphof D, Oei EH. A machine learning approach to distinguish between knees without and with osteoarthritis using MRI-based radiomic features from tibial bone. Eur Radiol 2021 Nov;31(11):8513-8521 [FREE Full text] [doi: 10.1007/s00330-021-07951-5] [Medline: 33884470]
33. Egloff C, Hügle T, Valderrabano V. Biomechanics and pathomechanisms of osteoarthritis. Swiss Med Wkly 2012 Jul 19;142:w13583 [FREE Full text] [doi: 10.4414/smw.2012.13583] [Medline: 22815119]
34. Lim YZ, Wang Y, Wluka AE, Davies-Tuck ML, Teichtahl A, Urquhart DM, et al. Are biomechanical factors, meniscal pathology, and physical activity risk factors for bone marrow lesions at the knee? A systematic review. Semin Arthritis Rheum 2013 Oct;43(2):187-194. [doi: 10.1016/j.semarthrit.2013.03.002] [Medline: 24157092]
35. Hunter DJ, Niu J, Felson DT, Harvey WF, Gross KD, McCree P, et al. Knee alignment does not predict incident osteoarthritis: the Framingham osteoarthritis study. Arthritis Rheum 2007 Apr;56(4):1212-1218. [doi: 10.1002/art.22508] [Medline: 17393450]
36. Sharma L, Song J, Felson DT, Cahue S, Shamiyeh E, Dunlop DD. The role of knee alignment in disease progression and functional decline in knee osteoarthritis. JAMA 2001 Jul 11;286(2):188-195. [doi: 10.1001/jama.286.2.188] [Medline: 11448282]
37. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. Radiology 2016 Dec;281(3):947-957. [doi: 10.1148/radiol.2016152234] [Medline: 27347764]
38. Hunter DJ, Deveza LA, Collins JE, Losina E, Katz JN, Nevitt MC, et al. Multivariable modeling of biomarker data from the phase I Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. Arthritis Care Res (Hoboken) 2022 Jul;74(7):1142-1153 [FREE Full text] [doi: 10.1002/acr.24557] [Medline: 33421361]
39. Guan B, Liu F, Mizaian AH, Demehri S, Samsonov A, Guermazi A, et al. Deep learning approach to predict pain progression in knee osteoarthritis. Skeletal Radiol 2022 Feb;51(2):363-373 [FREE Full text] [doi: 10.1007/s00256-021-03773-0] [Medline: 33835240]
40. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008 Mar 10;26(8):1364-1370. [doi: 10.1200/JCO.2007.12.9791] [Medline: 18323559]
41. Paz-González R, Balboa-Barreiro V, Lourido L, Calamia V, Fernandez-Puente P, Oreiro N, et al. Prognostic model to predict the incidence of radiographic knee osteoarthritis. Ann Rheum Dis 2024 Apr 11;83(5):661-668 [FREE Full text] [doi: 10.1136/ard-2023-225090] [Medline: 38182405]
42. Sun H, You Y, Jiang Q, Ma Y, Huang C, Liu X, et al. Radiomics-based nomogram for predicting total knee replacement in knee osteoarthritis patients. Eur J Radiol 2025 Jan;182:111854 [FREE Full text] [doi: 10.1016/j.ejrad.2024.111854] [Medline: 39626336]
43. Li X, Li C, Zhang P. Predictive models of radiographic progression and pain progression in patients with knee osteoarthritis: data from the FNIH OA biomarkers consortium project. Arthritis Res Ther 2024 May 30;26(1):112 [FREE Full text] [doi: 10.1186/s13075-024-03346-1] [Medline: 38816759]
44. Huang Z, Bucklin MA, Guo W, Martin JT. Disease progression and clinical outcomes in latent osteoarthritis phenotypes: data from the Osteoarthritis Initiative. Research Square. Preprint posted online on January 26, 2024 2024 Jan 26:rs.3.rs-3855831 [FREE Full text] [doi: 10.21203/rs.3.rs-3855831/v1] [Medline: 38343849]
45. Li HH, Chan LC, Chan PK, Wen C. An interpretable knee replacement risk assessment system for osteoarthritis patients. Osteoarthr Cartil Open 2024 Feb 16;6(2):100440 [FREE Full text] [doi: 10.1016/j.ocarto.2024.100440] [Medline: 38385105]

XSL•FO
RenderX

46. Collins JE, Robinson D, Arden N, Bay-Jensen AC, Deveza LA, Karsdal M, et al. Biochemical biomarkers of knee osteoarthritis progression: results from the FNIH biomarkers consortium progress OA study. Osteoarthr Cartil Open 2025 Sep 05;7(4):100677 [FREE Full text] [doi: 10.1016/j.ocarto.2025.100677] [Medline: 41050370]
47. The Osteoarthritis Initiative. National Institute of Mental Health Data Archive. URL: https://nda.nih.gov/oai [accessed 2026-01-05]

Abbreviations

AUC: area under the receiver operating characteristic curve
DCA: decision curve analysis
DICOM: Digital Imaging and Communications in Medicine
FNIH: Foundation for the National Institutes of Health
HIPAA: Health Insurance Portability and Accountability Act
KOA: knee osteoarthritis
LASSO: least absolute shrinkage and selection operator
MRI: magnetic resonance imaging
OA: osteoarthritis
OAI: Osteoarthritis Initiative
rad-score: radiomics score
ROI: region of interest
SHAP: Shapley additive explanations
WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

Edited by A Benis; submitted 03.Jun.2025; peer-reviewed by HD Zuo, G Tang; comments to author 11.Nov.2025; accepted 17.Dec.2025; published 14.Jan.2026.

Please cite as:
Sun Y, Liu J, Deng C, Peng C, Pan S, Liu X
Nomograms Based on X-Ray Radiomics for Predicting Pain Progression in Knee Osteoarthritis Using Data From the Foundation for the National Institutes of Health: Development and Validation Study
JMIR Med Inform 2026;14:e78338
URL: https://medinform.jmir.org/2026/1/e78338
doi:10.2196/78338
PMID:

©Yingwei Sun, Jing Liu, Chunbo Deng, Chengbao Peng, Shinong Pan, Xueyong Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Linking Electronic Health Records for Multiple Sclerosis Research: Comparative Study of Deterministic, Probabilistic, and Machine Learning Linkage Methods

Ohoud Almadani[1*], MSHI; Yasser Albogami[2*], PhD; Adel Alrwisan[1*], PhD

[1]Saudi Food and Drug Authority, 4904 northern ring branch rd, 13513, Hittin Dist, Unit number : 1, Riyadh, Saudi Arabia

[2]Department of Clinical Pharmacy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia

[*]all authors contributed equally

Corresponding Author:
Ohoud Almadani, MSHI
Saudi Food and Drug Authority, 4904 northern ring branch rd, 13513, Hittin Dist, Unit number : 1, Riyadh, Saudi Arabia

Abstract

Background: Data linkage in pharmacoepidemiological research is commonly employed to ascertain exposures and outcomes or to obtain additional information on confounding variables. However, to protect patient confidentiality, unique patient identifiers are not provided, which makes data linkage across multiple sources challenging. The Saudi Real-World Evidence Network (SRWEN) aggregates electronic health records from various hospitals, which may require robust linkage techniques.

Objective: We aimed to evaluate and compare the performance of deterministic, probabilistic, and machine learning (ML) approaches for linking deidentified data of patients with multiple sclerosis (MS) from the SRWEN and Ministry of National Guard Health Affairs electronic health record systems.

Methods: A simulation-based validation framework was applied before linking real-world data sources. Deterministic linkage was based on predefined rules, whereas probabilistic linkage was based on a similarity score–based matching. For ML, both similarity score–based and classification approaches were applied using neural networks, logistic regression, and random forest models. The performance of each approach was assessed using confusion matrices, focusing on sensitivity, positive predictive value, F_1 score, and computational efficiency.

Results: The study included linked data of 2247 patients with MS from 2016 to 2023. The deterministic approach resulted in an average F_1 score of 97.2% in the simulation and demonstrated varying match rates in real-world linkage: 1046/2247 (46.6%) to 1946/2247 (86.6%). This linkage was computationally efficient, with run times of <1 second per rule. The probabilistic approach provided an average F_1 score of 93.9% in the simulation, with real-world match rates ranging from 1472/2247 (65.5%) to 2144/2247 (95.4%) and processing times ranging from approximately 0.1 to 5 seconds per rule. ML approaches achieved high performance (F_1 score reached 99.8%) but were computationally expensive. Processing times ranged from approximately 13 to 16,936 seconds for the classification-based approaches and from approximately 13 to 7467 seconds for the similarity score–based approaches. Real-world match rates from ML models were highly variable depending on the method used; the similarity score–based approach identified 789/2247 (35.1%) matched pairs, whereas the classification-based approach identified 2014/2247 (89.6%).

Conclusions: Probabilistic linkage offers high linkage capacity by recovering matches missed by deterministic methods and proved to be both flexible and efficient, particularly in real-world scenarios where unique identifiers are lacking. This method achieved a great balance between recall and precision, enabling better integration of various data sources that could be useful in MS research.

(JMIR Med Inform 2026;14:e79869) doi:10.2196/79869

KEYWORDS

EHR; MS; probabilistic; deterministic; ML; multiple sclerosis; electronic health record; machine learning

Introduction

Real-world data (RWD) derived from routine clinical care provide remarkable potential for generating pharmacoepidemiological studies related to medication safety, effectiveness, and utilization [1-4]. The recent emergence of electronic health records (EHRs) in the Kingdom of Saudi Arabia led to the foundation of the Saudi Real-world Evidence Network (SRWEN) by the Saudi Food and Drug Authority [5-8]. This central network plans to integrate and harmonize EHRs from various hospitals, primarily to maximize the efficacy

XSL•FO
RenderX

of RWD in contributing to drug regulatory policymaking and understanding disease dynamics in the Saudi population [6,7].

Common data models (CDMs), such as the observational medical outcomes partnership model, incorporate consolidated formats necessary for data standardization across multiple clinical data sources to perform large-scale observational studies [9-11]. CDMs enable researchers to pool multiple EHR datasets and perform federated data analyses through standardized data architectures, formats, and terminologies [10,11]. The observational medical outcomes partnership CDM, developed as part of the observational health data sciences and informatics, offers a systematic framework of clinical data across diverse subdomains such as demographics, diagnoses, medications, procedures, and laboratory results. It uses standardized terminologies, mitigating inherent variability, customized formats, or inconsistent coding in raw EHR data [9,11].

While CDMs are in place to standardize structures in collecting and evaluating clinical data, many are built on deidentified datasets that may not hold vital contextual information or complete patient histories [9-11]. Thus, restricting the ability to solve such limitations of CDM as incomplete data mapping, the lack of direct mappings for some clinical or medication concepts, laboratory or narrative data gaps, which necessitate going back to the data to retrieve context-specific information that may not be contained in the standardized model [9-11]. Therefore, EHR linkage offers a powerful solution addressing this limitation by enabling researchers to link patient records across various databases and reconstruct comprehensive longitudinal health trajectories [12-14]. Common linkage approaches are deterministic, probabilistic, and machine learning (ML) linkage, each with advantages and challenges [14-16]. For instance, the deterministic linkage is based on an exact matching across predefined variables, achieving high precision; however, sensitivity may be compromised in this case, particularly when matching data that contain errors in a given variable. Probabilistic linkage overcomes this weakness by using weighted matching, enhancing sensitivity at the cost of reduced precision in specific applications [12,13]. ML techniques, such as gradient boosting and neural networks (NNs), can achieve high levels of accuracy [12,13]. These models are particularly successful when integrating a dataset with large volumes of missing and errant data, identifying more links than deterministic and probabilistic approaches [16].

With a growing number of multiple sclerosis (MS) cases among the Saudi population—from 25 per 100,000 in 1998 to 40.4 per 100,000 in 2020—the linkage of data pertaining to MS is of priority, especially given the complexity of MS treatment management, both within Saudi Arabia and globally [17]. SRWEN contains a complex variety of data sources, reference standards, and case definitions, highlighting the intrinsic complexity associated with the reliability of RWD in health care settings [7]. Therefore, accurate linkage of EHRs is paramount for addressing the complexities associated with defining patient outcomes with MS, such as disease progression, relapse rate, and treatment response. This analysis requires frequent validation by clinical expertise and enhancement of the validity and robustness of the SRWEN[17]. To achieve this, we compared different EHR linkage methods, including deterministic, probabilistic, and ML approaches. The findings from this study may have implications for the value and design of future longitudinal studies using linked MS data.

Methods

Ethical Considerations

This study was approved by the institutional review board of Saudi Food and Drug Authority with approval number 2024_17 and by the King Abdullah International Medical Research Center with approval number ROMJ/WE/WS/059/2023. Because the data were extracted from the EHR system as part of routine clinical care, patients are required to provide consent to receive treatment within the institution. Separate study-specific consent was considered waived in accordance with institutional policies. Patient privacy and confidentiality were granted through de-identification of all records prior to data exchange, such that no direct identifiers were available to the research team.

Study Population and Data Sources

In this study, we used 2 datasets from 2 different sources. First, we included SRWEN data of patients with MS between 2016 and 2023. The SRWEN functions as a repository of longitudinal health care information and incorporates diverse EHR data, such as demographics, medical history, medication use, laboratory outcomes, hospitalizations, and clinical outcomes. Information is gathered from diverse EHRs from multiple hospitals in Saudi Arabia, including the Ministry of National Guard Health Affairs (MNGHA). Subsequently, these collected data are standardized through a CDM to ensure consistency and streamline analysis [6]. Furthermore, an anonymization process is implemented prior to exchanging data to ensure patient privacy [6]. For the purposes of this study, we focused exclusively on extracting and analyzing MNGHA data within SRWEN.

Similarly, we collected MNGHA EHR data from King Abdullah International Medical Research Center, identifying patients with MS using the International Classification of Diseases 10th version (*ICD10*: G35) for the same period of (2016 - 2023). MNGHA EHRs contain routinely collected data, including patient characteristics, diagnostic history, medications history, and visit records (ie, admissions, outpatient clinical visits, or emergency visits; Table S1 in Multimedia Appendix 1).

While SRWEN collects data already from the MNGHA, the benefit of the direct link between these 2 entities is advantageous. Data within SRWEN are organized and standardized within a CDM, providing for multi-institutional data analysis at the risk of some loss in source detail. By linking the primary MNGHA EHRs to their corresponding RERN representations, we were able to recover additional clinical granularity from narrative diagnosis notes and reports. This linkage also allowed validation of data consistency between SRWEN and its original source, resulting in a larger and more accurate longitudinal profile of patients with MS.

XSL•FO
RenderX

Simulation-Based Validation

Overview

We first cloned and deidentified the original dataset to address the challenge of linking deidentified records from the targeted database (ie, SRWEN) with original records (ie, MNGHA EHRs). This process ensured a realistic framework testing and that the dataset resembled the deidentified target dataset. Simulating the conditions of the intended linkage objective, we linked records to themselves using these deidentified cloned records, employing deterministic, probabilistic, and ML approaches. Following this self-linkage phase, we assessed each approach's performance using computational efficiency metrics. Consequently, the best-performing linkage method was used to link the original records to the target dataset, thereby maximizing the accuracy and reliability of the final linkage. This methodology enabled a rigorous, performance-based selection of the optimal linkage approach for secure and accurate record linkage of deidentified datasets [13].

Deterministic Approach

Deterministic EHR linkage is a rule-based approach that relies on exact matches between specific patient identifiers across different EHR systems [12,14,15,18]. Given the lack of patient identifiers in SRWEN, we implemented multiple matching criteria (ie, rules) based on a range of available variables, each considered both independently and in unique combinations (Table S2 in Multimedia Appendix 1). We explored different variable groupings in some detail to determine how various linkage methods perform under conditions with differing degrees of completeness and complexity of data encountered in real-world settings. This approach generates pairs of records classified as match or nonmatches. For each pair, we included multiple linkage variables, including demographic characteristics (eg, gender, date of birth, nationality, eligibility), and health care–related fields (eg, *ICD-10* codes, diagnosis description and date, medication order details, visit type and date, and facility information) [12,14,15,18]. Additional derived variables were generated from these data elements, including coverage, which was derived from eligibility and indicated a patient's treatment eligibility, the first MS diagnosis date, derived from the diagnosis data, and inpatient admission status, derived from visit records.

Probabilistic Approach

In this approach, patient records were grouped into blocks based on shared variables (eg, date of birth, initial MS diagnosis date, received medications, or admission events) to reduce the number of pairwise comparisons required for linkage [12,14,19]. Subsequently, each record pair within a block was compared across diverse fields, including gender, coverage (ie, eligibility), region, nationality, facility name, or diagnostic codes other than MS [12,14,19]. A similarity score was assigned to each record pair corresponding to the level of agreement on predefined variables. Variable weights were estimated as match probabilities (ie, m-probability)—the probability that two field values match, given that the records refer to the same individual—and unmatch probability (ie, u-probability)—the probability that 2 field values match given that the records refer to different individuals. These probabilities were computed using the expectation-maximization algorithm [12-14,19]. An overall similarity score for each pair was computed by summing the weighted scores across all variables [12,14,19].

Pairs were categorized as matches or nonmatches using a one-to-one linkage approach. This method was implemented using the "select_n_to_m" function from the rclin package, which maximizes total similarity score under the constraint that each record is linked only once [20]. This probabilistic approach incorporates imperfect matches caused by inaccuracies or missing data, thereby enhancing the quality and reliability of linked EHR datasets in clinical studies [19].

ML Approach

In the ML approach, record linkage was performed in 2 methods. First, we generated pairs of records using common variables in the cloned data to limit the number of comparisons, creating a grouped set of pairs [20] Similarity scores were computed for each pair across different selected variables 20[]. Pairs were labeled as matches or nonmatches for training using the known true matches. The pairs were then fed into different models for training, and we fitted the model on variables that had the higher impact based on each variable similarity score, allowing the model to learn patterns that accurately identify true matches [20].

Second, a classification-based ML approach was implemented by training models on a labeled training dataset comprising 75% of all possible record pairs derived from the original dataset. Match and nonmatch labels were derived from the reference dataset. The remaining 25% of the record pairs were used for the testing dataset to evaluate the classifier performance [16].

Four ML algorithms—K-nearest neighbor, logistic regression (LR), NN, and random forest (RF)—were used in both methods to evaluate their linkage accuracy [16]. Internal validation was reinforced using a 100-replication bootstrap approach in each method for more robust performance estimates and to ensure stability across repeated sampling iterations [16]

Real-World Record Linkage

In the real-world matching scenario, we linked data from SRWEN and MNGHA to identify patients diagnosed with MS between 2016 and 2023. Because the SRWEN and MNGHA datasets lack a unique common identifier, no reference standard was available to evaluate the performance of each matching algorithm [13]. Therefore, model performance was compared based on the number of matched record pairs identified.

Statistical Analysis

Descriptive analyses of both datasets, including diseases, demographics, and facilities, were tabulated and presented by counts and percentages. Using original identifiers as the reference standard, accurate matches and true nonmatches were defined for the simulation [13]. True matches were record pairs agreeing on identifiers, whereas true nonmatches were record pairs conflicting on identifiers [13]. Sensitivity and positive predictive value (PPV) were calculated along with their corresponding 95% CI. Additionally, the F_1 score was calculated as a measure of balanced accuracy between sensitivity and PPV

XSL•FO
RenderX

(equation 1) [19]. Computational time required by the ML model to complete the record linkage process.

$$F1score=2\times\frac{Sensitivity\times PPV}{Sensitivity+PPV} \quad (1)$$

Results

Descriptive Analysis

The MNGHA dataset contained 2642 patients with MS, whereas 2247 were identified from SRWEN. Because only MNGHA-related dataset within SRWEN was used, 395/2642 (14.9%) patients with MS from the original MNGHA dataset were not captured in the SRWEN data. The distribution of demographic data between the 2 data sources was comparable. For instance, the mean age of patients in the MNGHA dataset was 34.4 (SD 11) years, which was similar to the mean age of 34.6 (SD 11.3) years in SRWEN.

Deterministic Approach

The highest number of pairwise comparisons (3,854,740) was observed in the "sex" variable, compared with a significant precise pairwise comparisons (6212) using the "date of birth (DOB)" or "nationality" variables. Sensitivity was consistent with a perfect score across the variables. However, the specificity rate varied, with the highest (99.9%) observed using "DOB" and the lowest (5%) observed using "nationality." Conversely, precision was considerably low across different variables, with the highest (42.5%) observed using "DOB." Moreover, both "DOB" and "MS Date" had relatively high F_1 scores of 59.7% and 48.8%, respectively (Table 1).

Table . Results of deterministic matching in % based on an individual variable.

Metric	Sex	DOB[a]	Nationality	Coverage	Facility	Region	MS[b] date[c]	Admissions	Other diagnoses	Medications
Accuracy	44.45	99.95	5.09	69	62.92	58.06	99.92	88.15	64.37	99.73
Sensitivity (recall)	100	100	100	100	100	100	100	100	100	100
Specificity	44.43	99.95	5.05	68.99	62.91	58.04	99.92	88.14	64.36	99.73
PPV[d] (precision)	0.07	42.53	0.04	0.12	0.10	0.09	32.28	0.32	0.11	12.45
F_1 score	0.14	59.68	0.08	0.25	0.21	0.18	48.81	0.64	0.21	22.14

[a]DOB: date of birth.

[b]MS: multiple sclerosis.

[c]MS date: first multiple sclerosis diagnosis.

[d]PPV: positive predictive value.

A sequence of deterministic linkage rules was applied to evaluate data matching using a set of variables instead of a single variable (Table 1). The number of correctly matched pairs remained consistent across most rules, with 2642 pairs correctly matched for each rule. Likewise, almost all of the linkage rules resulted in a nearly perfect score across the metrics. The average F_1 score for the deterministic approach was 97.2% (Table S3 in Multimedia Appendix 1). In terms of computing time, deterministic linkage was completed in less than one second for each criterion (Table S3 in Multimedia Appendix 1).

Probabilistic Approach

To ensure a manageable dataset for analysis, we first reduced the number of comparisons by blocking records based on exact matches on DOB (6212 pairs). Precision remained consistently perfect regardless of the variables used across diverse combinations of matching criteria in the probabilistic approach. Both specificity and sensitivity increased with more variables, as the combination of sex, nationality, and coverage was 76%. Further, adding the facility name and the MS diagnosis date led to a perfect score. A similar observation was found for the F_1 score, which began at 69.9% with sex only, increased to 94.6% after adding facility name and region, and reached a perfect score of 100% with the inclusion of MS diagnosis date.

By using a combination of the best-performing variables derived from the deterministic approach to improve match accuracy further, in addition to examining different blocking criteria, precision was constantly at 100%. Sensitivity and specificity rates were high, considerably assisted by combining specific clinical variables. Likewise, the F_1 score ranged from 82% to 99.9% (Table S3 in Multimedia Appendix 1). In terms of computing time, probabilistic linkage required 0.1 to 5 seconds for each criterion (Table S3 in Multimedia Appendix 1).

Using this probabilistic approach with the one-to-one linkage (ie, select_n_to_m from the reclin package) matching criterion, the linkage models achieved perfect performance for all the metrics evaluated. In addition, m-probabilities and u-probabilities were compared across the variables under different matching criteria (Table S4 in Multimedia Appendix 1).

ML Approach

In the similarity-score method, models trained on combinations of DOB, MS diagnosis date, and medications, after being blocked (ie, exact match) on admission, performed best overall. NN, RF, and LR models demonstrated comparable performance, with slightly higher accuracy using NN (99.37% vs 99.36%). However, computational time analysis showed that LR presented a combination of high performance and acceptable processing time (15 - 766 s). In contrast, both RF and NN, while having

XSL•FO
RenderX

comparable results to LR, were less practical for large-scale linkage tasks due to their longer training times (15 - 7467 s) associated with their complexity (Table S5 in Multimedia Appendix 1).

In the classification approach, despite the blocking variable or combination of matching criteria, performance was constantly high among the three models, ranging from 91.7% to 100%, while computational efficiency varied depending on the blocking strategy and applied comparison criteria. Moreover, the classification is usually more computationally expensive (13 - 16,936 s) compared with the similarity-score approach, affecting scalability for larger datasets. However, K-nearest neighbor demonstrated an excellent trade-off between computational efficiency and linkage performance (Table S5 in Multimedia Appendix 1).

Record Linkage

The deterministic approach produced the fewest matched pairs. Increasing the number of variables (eg, adding admissions or medications) made the linkage criteria stricter, resulting in gradually fewer matched pairs, ranging from 1046/2247 (46.6%) to 1946/2247 (86.6%). The probabilistic approach achieved more pairs ranging from 1472/2247 (65.5%) to 2144/2247 (95.4%) based on the combination of variables. The ML approach, using similarity scoring or classification with an NN model, generated 789/2247 (35.1%) pairs and 2014/2247 (89.6%) pairs, respectively (Table S5 in Multimedia Appendix 1).

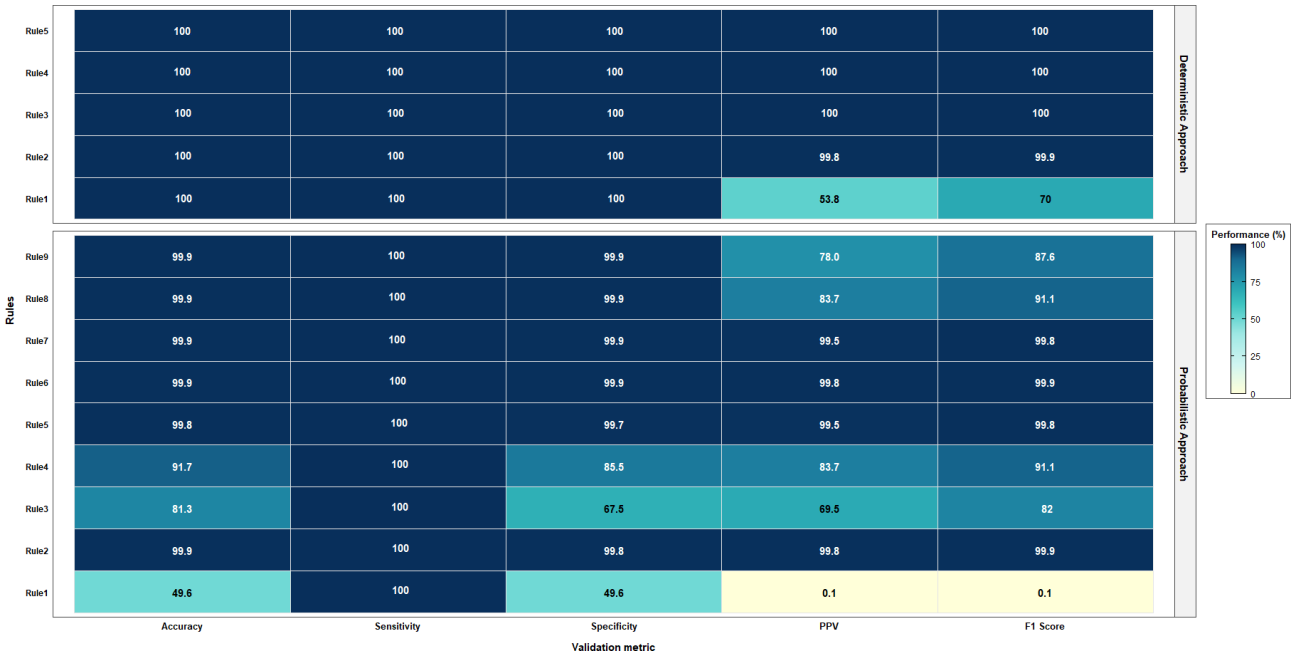
Discussion

Principal Results

Through the simulated scenario and after using a mixture of variables without identifiers, we determined that the deterministic method was the most efficient data linkage algorithm. Although this approach had limited generality, it accomplished 100% scores across all metrics without incurring high computational times. Therefore, the deterministic approach is practical when we need absolute certainty in matched pairs. In contrast, the probabilistic method also demonstrated efficient performance with a similar mixture of variables. Initially, the recall and specificity metrics exhibited low performances. However, when we applied one-to-one linkages, the performance of the probabilistic method improved significantly, yielding a 100% score. Furthermore, the core benefit of this approach is retrieving a higher number of matched pairs, signifying its advantage when maximum generalization and thorough data retrieval are required. In the case of ML methods like NN models, we attained perfect performance scores at the expense of higher computational costs, limiting their feasibility in real-world scenarios. Among all methods, the probabilistic approach provided the highest number of patients, signifying its flexibility in real-world healthcare scenarios. These outcomes give a reproducible framework for appropriate EHR record linkages, underscoring its efficacy in practical patient care situations.

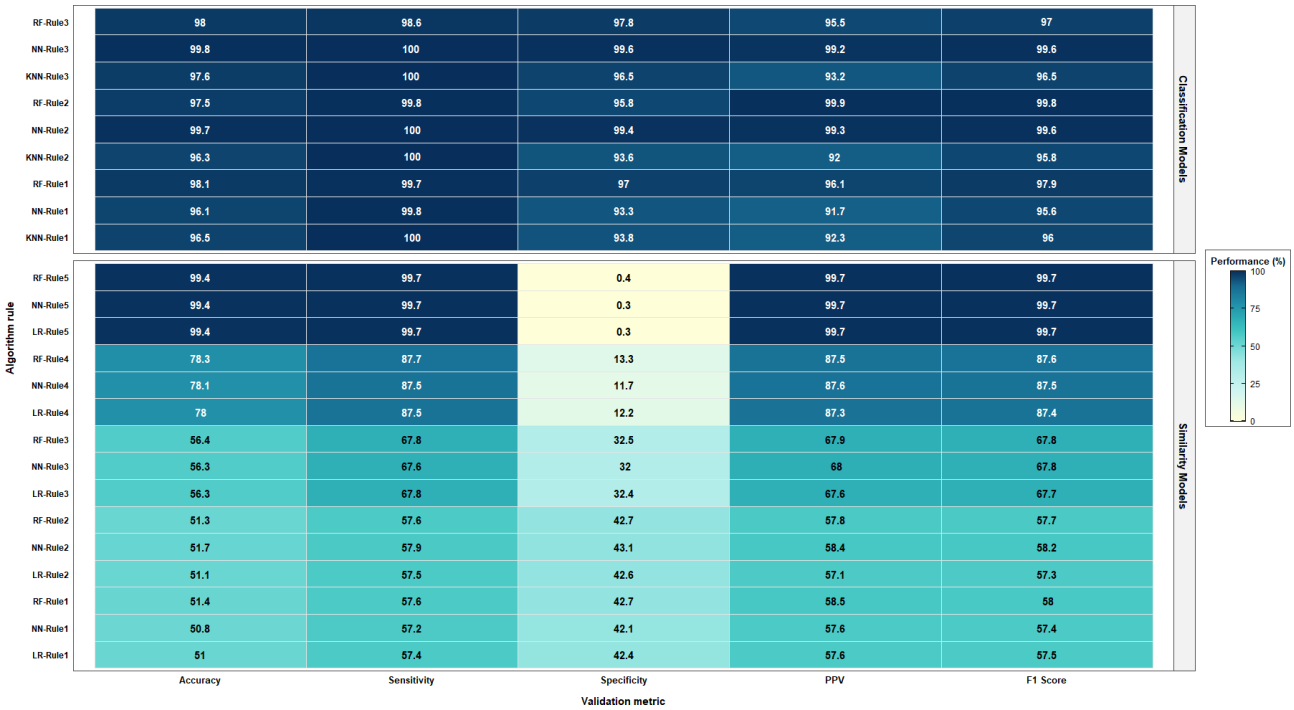
Considering the average F_1 score for the different models in each approach, the deterministic approach achieved the highest average (97.2%), followed by the probabilistic approach (93.9%). Noteworthy, the probabilistic approach using one-to-one linkage achieved an average of 100%, and in both probabilistic and ML approaches, an initial blocking step based on exact matching was performed. Examining the criteria without applying any blocking led to poor performance in the probabilistic method and was infeasible for ML. Figure 1 shows a comparison between deterministic and probabilistic approaches, while Figure 2 shows a comparison between similarity score and classification ML approaches.

Figure 1. Performance comparison of deterministic and probabilistic approaches. PPV: positive predictive value.



XSL•FO
RenderX

Figure 2. Performance comparison of machine learning models. KNN: K-nearest neighbor; LR: logistic regression; NN: neural network; PPV: positive predictive value; RF: random forest.



Comparison with Prior Work

Despite the application of a straightforward deterministic algorithm requiring exact matching among all variables of linkage rule, our deterministic solution outperformed the probabilistic method, in contrast to findings from previous studies [21,22]. For instance, Gomatam et al [21] used a forward stepwise deterministic method to evaluate variables independently before constructing matching rules. We used a similar approach and observed high specificity and recall, achieving a perfect score, using this method [21]. The probabilistic linkage was implemented similarly to prior research, and we only allowed for exact matches in the deterministic approach. However, unlike previous studies' findings where probabilistic linkage outperformed deterministic linkage in most scenarios, our results are comparable, with slightly better performance of the deterministic approach when linkage rules (rather than individual variables) were employed [19,21-23]. Implementing a one-to-one linkage method proposed by the reclin package further improved the probabilistic approach to exceed the performance of the deterministic approach when using similar matching rules [20]. Compared with Röchner et al [16], our ML approach exhibited a lower performance than the deterministic approach using a similar combination of rules. Furthermore, even though NN in the classification method achieved a near-perfect score, it required significantly more computational time than the deterministic approach. Nonetheless, we reached a similar finding to Röchner et al [16] in identifying the NN as the highest-performing algorithm in record linkage.

Computational benchmarking was performed on a standard workstation featuring an Intel Core i7 processor, 16 GB of RAM, and 2 GB of GPU (64-bit Windows OS). From a computational perspective, the most efficient method of linkage was the deterministic approach: each run finished in under one second, F_1 scores were between 70% and 100%, and memory requirements were low (<2 GB). The probabilistic approach required 0.1 to 4.88 seconds (F_1 score 82% - 99.9%), while maintaining moderate memory requirements (<8 GB) and achieving excellent accuracy. In contrast, machine-learning-based methodologies were very computationally intensive requiring: 12 to 7467 seconds (~0.2 - 124 min; F_1 score 57.3% - 99.7%, similarity-based models), and 13 to 16,936 seconds (~0.2 - 282 min; F_1 score 95% - 99.8%, classification-based models), both at a maximum memory requirement of 12 to 14 GB. These data demonstrate that the deterministic and probabilistic approaches are ideal for large scale and routine linkage tasks where the computational efficiency of the method is crucial, while ML-based methods, though highly accurate, demand greater resources and are better suited for smaller or high-value analytic linkages where accuracy outweighs runtime considerations.

Clinical and Regulatory Implications

Beyond demonstrating methodological aspects, this study supports the assertion that EHR linkage of original sources with their standardized mapping has clinical importance of linking the fragmented data of MS into a multi-institutional timeline that supports relapse reconstruction, mapping of treatment trajectories, and assessment of long-term outcomes. Importantly, this linkage addresses the challenges of RWD in MS by enabling the validation of deidentified records and matching the structured components (diagnoses, labs, prescriptions, utilization) with unstructured data (notes by physicians and radiology reports) to improve phenotype definition when ICD codes are ineffectively applied. The linked dataset supports earlier case identification (using utilization surrogates as proxies when standard scales are not available), risk stratification and

XSL•FO
RenderX

compliance monitoring across care sites, and provides more frequent and expandable data for pharmacoepidemiologic and safety surveillance that will be more robust than values obtained from trial groups. At the same time, addressing challenges related to the lack of data sufficiency, burdens related to the gaining of patient-reported outcomes data, evolving criteria for diagnosis, and the need for consistent harmonization and resourcing across systems. This linkage produces an analytically reproducible and available dataset for the practical and generalizable studies related to MS and in support of decisions concerning it, changing otherwise unstructured clinical data into actionable knowledge connecting to clinicians, regulators, and health-system leaders.

With the growing availability of large-scale clinical datasets, opportunities to create more comprehensive patient care pathways through data linkage are expanding. The findings of this study suggest that probabilistic linkage of anonymized datasets using indirect identifiers can substantially improve linkage capacity, minimize computational cost, and maintain patient privacy. Probabilistic linkage has demonstrated a greater ability to recover more pairs that may be missed by deterministic methods (1472-2144 vs 1046-1946) due to its flexibility, highlighting its potential as a complementary method to deterministic linkage. Furthermore, to ensure transparency, we examined the contributions of each linkage variable before developing the matching criteria. This approach allowed us to identify the most influential variables for effective linkage. More particularly, the date of birth and first MS diagnosis date achieved F_1 scores of 59.7% and 48.8%, respectively, suggesting strong linkage capabilities with balanced precision and recall. Moreover, incorporating both variables in the matching rules of the deterministic approach resulted in a perfect record linkage across all performance metrics. This improved linkage capability is beneficial for patients with MS by enabling data integration across different providers and supports an even more unified and comprehensive record of each patient's journey. Such a broad linkage helps researchers and clinicians track the progression of diseases over time, evaluate treatment effects on symptoms, and track long-term health outcomes.

The study findings are significant for regulatory requirements, especially when conducting policymaking and postmarketing surveillance. The low computational costs of the deterministic approach enable the regulators to evaluate its treatment outcomes and specify any safety issues, whereas the flexibility and open-ended retrieval of the probabilistic approach is remarkable, especially in the case of comprehensive data retrieval from anonymized records. Therefore, using both deterministic and probabilistic approaches is significant for creating a comprehensive clinical data source for policymaking, postmarketing assessments, and monitoring of the healthcare industry. Ultimately, a combination of these approaches can lead to excellent patient outcomes regarding safety and treatment accuracy.

Limitations

Several limitations must be acknowledged in this study. First, our selection of methods was affected by computational constraints, which highlighted the need for advanced optimization and more powerful hardware for ML models. Discrepancies in patient records count between datasets could also affect generalizability, highlighting the importance of complete data capture in future studies. In our linkage of RWD, we standardized both datasets in terms of variable names and value structure, perhaps influencing the performance of algorithms when linking across different data quality in a real-world application. Furthermore, one patient in SRWEN had an inaccurate date of birth, while 4 patients in both data sets had an unknown nationality. Also, 44 patients in MNGHA and 11 patients in SRWEN had an unknown coverage status. Variables such as other diagnoses, admissions (visits), and medications contain null values (no value), but we believe that these nulls refer to the absence of a clinical event rather than missing entries. Lastly, the lack of an identifier limited our ability to validate the results against a reference standard for the real-world linkage section. Consequently, we could not accurately evaluate the retrieval of matched pairs. Therefore, we used the number of matches solely to describe the linkage performance in real-world linkage.

Conclusion

The findings of this study demonstrate how probabilistic linkage with indirect identifiers can enhance data integration, particularly for MS research and patient care. In real-world situations where we lack unique identifiers, probabilistic linkage proved to be both flexible and efficient as a complement to traditional approaches. These results provide a more holistic overview of patients with MS care and afford researchers access to an inclusive dataset for studying disease progression, treatment outcomes, and monitoring medication safety. In conclusion, applying probabilistic linkage techniques can be expected to facilitate more informed, data-driven clinical and research practices in managing MS.

Acknowledgments

The views expressed in this paper are those of the author(s) and do not necessarily reflect those of the Saudi Food and Drug Authority or its stakeholders. Guaranteeing the accuracy and the validity of the data is a sole responsibility of the research team.

Data Availability

The data that support the study findings are not publicly available due to confidentiality restrictions.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: OA, YA, AA
Methodology: OA, YA, AA
Formal analysis: OA
Data curation: YA, AA
Investigation: OA, YA, AA
Project administration: OA
Validation: OA, YA, AA
Visualization: OA
Writing – original draft: OA
Writing – review & editing: YA, AA

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 40 KB - medinform_v14i1e79869_app1.docx]

References

1. Hernán MA. Methods of public health research — strengthening causal inference from observational data. N Engl J Med 2021 Oct 7;385(15):1345-1348. [doi: 10.1056/NEJMp2113319] [Medline: 34596980]
2. Li M, Chen S, Lai Y, et al. Integrating real-world evidence in the regulatory decision-making process: a systematic analysis of experiences in the US, EU, and China using a logic model. Front Med (Lausanne) 2021;8:669509. [doi: 10.3389/fmed.2021.669509] [Medline: 34136505]
3. Beaulieu-Jones BK, Finlayson SG, Yuan W, et al. Examining the use of real-world evidence in the regulatory process. Clin Pharmacol Ther 2020 Apr;107(4):843-852. [doi: 10.1002/cpt.1658] [Medline: 31562770]
4. Yoshida S, Takeuchi M, Tanaka-Mizuno S, et al. Clinical epidemiology and pharmacoepidemiology studies with real-world databases. Proc Jpn Acad Ser B Phys Biol Sci 2022;98(10):517-528. [doi: 10.2183/pjab.98.026] [Medline: 36504194]
5. Samra H, Li A, Soh B, Zain MA. Utilisation of hospital information systems for medical research in Saudi Arabia: a mixed-method exploration of the views of healthcare and IT professionals involved in hospital database management systems. Health Inf Manag 2020;49(2-3):117-126. [doi: 10.1177/1833358319847120] [Medline: 31046465]
6. Alnofal FA, Alrwisan AA, Alshammari TM. Real-world data in Saudi Arabia: current situation and challenges for regulatory decision-making. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1303-1306. [doi: 10.1002/pds.5025] [Medline: 32458499]
7. Althunian TA, Alrasheed MM, Alnofal FA, et al. Recording type 2 diabetes mellitus in a standardised central Saudi database: a retrospective validation study. BMJ Open 2023 Mar 21;13(3):e065468. [doi: 10.1136/bmjopen-2022-065468] [Medline: 36944455]
8. Alharbi RA. Adoption of electronic health records in Saudi Arabia hospitals: knowledge and usage. J King Saud Univ Sci 2023 Feb;35(2):102470. [doi: 10.1016/j.jksus.2022.102470]
9. Hripcsak G, Shang N, Peissig PL, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019 Aug;96:103253. [doi: 10.1016/j.jbi.2019.103253] [Medline: 31325501]
10. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341. [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
11. Papez V, Moinat M, Payralbe S, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open 2021 Feb;4(3):ooab001. [doi: 10.1093/jamiaopen/ooab001] [Medline: 34514354]
12. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets. J Clin Epidemiol 2021 Aug;136:136-145. [doi: 10.1016/j.jclinepi.2021.04.015] [Medline: 33932483]
13. Avoundjian T, Dombrowski JC, Golden MR, et al. Comparing methods for record linkage for public health action: matching algorithm validation study. JMIR Public Health Surveill 2020 Apr 30;6(2):e15917. [doi: 10.2196/15917] [Medline: 32352389]
14. Ruppert LP, He J, Martin J, et al. Linkage of Indiana State Cancer Registry and Indiana Network for patient care data. J Registry Manag 2016;43(4):174-178. [Medline: 29595920]
15. Conway RBN, Armistead MG, Denney MJ, Smith GS. Validating the matching of patients in the linkage of a large hospital system's EHR with state and national death databases. Appl Clin Inform 2021 Jan;12(1):82-89. [doi: 10.1055/s-0040-1722220] [Medline: 33567463]

XSL•FO
RenderX

16. Röchner P, Rothlauf F. Using machine learning to link electronic health records in cancer registries: on the tradeoff between linkage quality and manual effort. Int J Med Inform 2024 May;185:105387. [doi: 10.1016/j.ijmedinf.2024.105387] [Medline: 38428200]
17. Al-Saeed OM, Albogami MF, Hatem M. Multiple sclerosis in Saudi Arabia: a literature review. Saudi J Health Sci 2021;10(3):160-164. [doi: 10.4103/sjhs.sjhs_163_21]
18. Grannis SJ, Williams JL, Kasthuri S, Murray M, Xu H. Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. J Am Med Inform Assoc 2022 Jul 12;29(8):1409-1415. [doi: 10.1093/jamia/ocac068] [Medline: 35568993]
19. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform 2015 Aug;56:80-86. [doi: 10.1016/j.jbi.2015.05.012] [Medline: 26004791]
20. van der Laan J. Record linkage using machine learning. 2024. URL: https://htmlpreview.github.io/?https://github.com/djvanderlaan/reclin2/blob/master/inst/doc/record_linkage_using_machine_learning.html [accessed 2026-01-21]
21. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med 2002 May 30;21(10):1485-1496. [doi: 10.1002/sim.1147] [Medline: 12185898]
22. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011 May;64(5):565-572. [doi: 10.1016/j.jclinepi.2010.05.008] [Medline: 20952162]
23. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014;21(1):97-104. [doi: 10.1136/amiajnl-2013-001744] [Medline: 23703827]

Abbreviations

CDM: common data model
EHR: electronic health record
ICD10: International Classification of Diseases 10th version
LR: logistic regression
ML: machine learning
MNGHA: Ministry of National Guard Health Affairs
MS: multiple sclerosis
NN: neural network
RERN: Real-world Evidence Network
RF: random forest
RWD: Real-world data

Edited by A Benis; submitted 01.Jul.2025; peer-reviewed by D Chrimes, J Li; revised version received 29.Dec.2025; accepted 29.Dec.2025; published 04.Feb.2026.

Please cite as:
Almadani O, Albogami Y, Alrwisan A
Linking Electronic Health Records for Multiple Sclerosis Research: Comparative Study of Deterministic, Probabilistic, and Machine Learning Linkage Methods
JMIR Med Inform 2026;14:e79869
URL: https://medinform.jmir.org/2026/1/e79869
doi:10.2196/79869

© Ohoud Almadani, Yasser Albogami, Adel Alrwisan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Stress Hyperglycemia Ratio and the Risk of Sepsis in Patients With Heart Failure: Retrospective Cohort Study From Medical Information Mart for Intensive Care-IV

Xianzhen Cai[1*], MM; Baoxin Yan[1*], MM; Jinhao Chen[2], MD; Junjun Ye[1], MM; Jiating Su[1], MM; Shuangshuang Tong[1], MM; Bin Xie[1], MD; Bingna Zhang[1], BS; Jilin Li[1], MD

[1]Department of Cardiology, Second Affiliated Hospital of Shantou University Medical College, 69 Dongxia North Road, Shantou, China

[2]Department of Cardiology, The Sixth People's Hospital of Longgang District, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Jilin Li, MD
Department of Cardiology, Second Affiliated Hospital of Shantou University Medical College, 69 Dongxia North Road, Shantou, China

Abstract

Background: The occurrence of sepsis in patients with heart failure (HF) has received less attention in research; yet, it poses a significant clinical challenge due to the complex interplay between chronic cardiac dysfunction and acute systemic inflammation. The stress hyperglycemia ratio (SHR) has emerged as an independent risk factor in various cardiovascular diseases and patients with sepsis, but its role in predicting sepsis risk in patients with HF remains underexplored.

Objective: This study investigates the association between SHR and sepsis occurrence in patients with HF and explores the potential mediating role of inflammatory indicators.

Methods: This retrospective cohort study used data from the Medical Information Mart for Intensive Care-IV (version 3.0) database, encompassing patients with HF from critical care units. SHR was calculated based on initial blood glucose and glycated hemoglobin A_{1c} levels. The analysis population was divided into 4 groups based on the quartiles of the SHR. The primary end point was 7-day sepsis incidence, which was diagnosed following Sepsis-3 criteria.

Results: Within the 1205-patient cohort (male: 764/1205, 63.4%; median 71.51, IQR 62.45-79.47), a total of 162 (13.4%) patients with HF experienced sepsis within 7 days. In the fully adjusted model, a per-unit SHR increase was linked to an 18% higher sepsis risk (hazard ratio 1.18, 95% CI 1.01 - 1.38; *P*=.04). Restricted cubic splines analysis showed a nonlinear saturation effect association (*P* for nonlinearity=.02), which was consistent in the diabetic subgroup (*P* for nonlinearity=.01). After adjusting for 7-day mortality as a competing event using the Fine-Gray model, SHR was independently associated with an increased risk of sepsis (*P*=.01). The association between SHR and sepsis was significantly modified by diabetes mellitus, BMI, and insulin use (all *P* for interaction<.05). Furthermore, mediation analysis indicated that several inflammatory indices, including the systemic immune-inflammation index, neutrophil-to-lymphocyte ratio, platelet-to-neutrophil ratio, systemic inflammation response index, and monocyte-neutrophil-to-lymphocyte ratio, significantly mediated the association in critically ill patients with HF.

Conclusions: In critically ill patients with HF, an elevated SHR was associated with heightened 7-day sepsis risk, especially for those combined with diabetes. Furthermore, systemic inflammatory indices partially mediated this association in the overall population, implicating inflammation as a potential mechanistic link between SHR and sepsis.

(JMIR Med Inform 2026;14:e81229) doi:10.2196/81229

KEYWORDS

stress hyperglycemia ratio; heart failure; sepsis; diabetes; mediation analysis

Introduction

Heart failure (HF), a prevalent but serious disease, has seen a 29% global prevalence increase from 2010 to 2019 [1], with a 30% - 50% risk of mortality or rehospitalization within the acute phase [2].

The critically ill patients with HF are particularly vulnerable to sepsis, which may be related to immune dysfunction [3], cardiac remodeling [4], or bacterial translocation [5]. It was reported that the mortality rate reach 90% in those combined cardiac dysfunction and sepsis [6], and the recurrence rate of sepsis in patients with HF is 3-fold higher than those without HF [7].

XSL•FO
RenderX

Thus, identifying modifiable risk factors for sepsis occurrence is crucial for global public health and prevention strategies. The pathogenesis of sepsis in critical illness is multifactorial, with emerging evidence suggesting that hyperglycemia, especially under stress, may contribute to the development and prognosis of sepsis involving immune dysregulation, inflammatory activation, and organ impairment [8-10]. Stress hyperglycemia ratio (SHR) has emerged as a novel and potentially valuable metric in the assessment of acute glycemic levels, which is calculated by comparing admission blood glucose levels to glycated hemoglobin (HbA_{1c}), distinguishing an acute stress-induced hyperglycemia from preexisting poor glycemic control [11]. Multiple studies have demonstrated that higher SHR levels were associated with poorer prognosis in patients diagnosed with acute coronary syndrome [12], acute HF [13], and sepsis [14]. However, the relationship between SHR and the occurrence of sepsis among patients with HF remains unclear. Inflammatory markers such as neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio, monocyte-to-lymphocyte ratio, neutrophil-to-platelet ratio (NPR), neutrophil-monocyte-to-lymphocyte ratio (NMLR), systemic immune-inflammation index (SII), and systemic inflammation response index (SIRI) have been investigated in patients with sepsis and show promise as prognostic indicators [15-20]. Given the intricate interplay between hyperglycemia and inflammation [21], and the well-documented role of inflammatory processes in the pathogenesis of sepsis, we hypothesize that inflammation may serve as a mediating factor in the association between SHR and sepsis occurrence. Therefore, this study aims to fill this critical gap by investigating the association of SHR and the risk of sepsis among critically ill patients with HF, based on data from the Medical Information Mart for Intensive Care-IV (MIMIC-IV) database. Furthermore, we explored whether inflammation acts as a potential mediator in this relationship.

Methods

Ethical Considerations

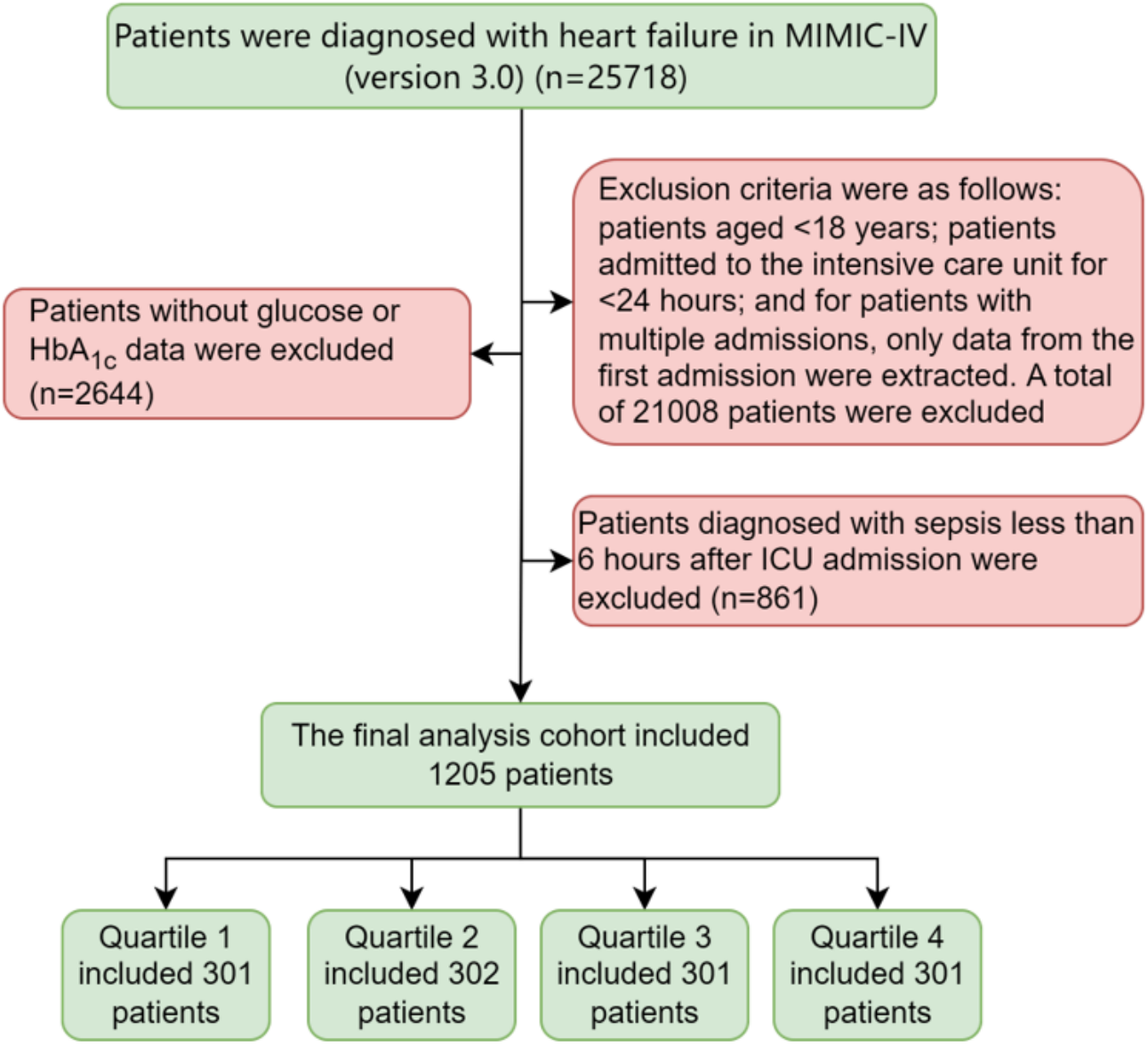
This study was a retrospective observational design with publicly available data from the MIMIC-IV database, which includes information on critically ill patients admitted to the intensive care units (ICUs) at Beth Israel Deaconess Medical Center from 2008 to 2022 [22]. The data acquisition process adhered to all relevant regulations. Author BY obtained a Collaborative Institutional Training Initiative license (Record ID 12861338) and the necessary permissions to extract data from the MIMIC-IV database. The project was approved by the institutional review boards of both the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center and followed the Strengthening the Reporting of Cohort Studies in Surgery guidelines [23]. Our research was conducted in accordance with the Declaration of Helsinki. Since the MIMIC-IV database includes fully anonymous or deidentified patient records and all protected health information has been removed in accordance with Health Insurance Portability and Accountability Act (HIPAA) privacy standards, participants or the participants' legal guardians or next of kin consent for publication is not applicable to this study. This study did not involve any ethical conflict issues.

Study Design and Participants

This is a retrospective observational study based on a large-scale critical care database, which included the ICU patients who were diagnosed with HF based on the *International Classification of Diseases, Ninth Revision* (ICD-9) and *International Classification of Diseases, Tenth Revision* (ICD-10) codes (Table S1 in Multimedia Appendix 1). We excluded (1) patients younger than 18 years of age; (2) patients with an ICU stay duration less than 24 hours; (3) for patients with multiple ICU admissions, only data from the first hospitalization were included; (4) patients lacked data of blood glucose or HbA_{1c}; and (5) patients diagnosed with sepsis [24] less than 6 hours after ICU admission. Finally, a total of 1205 patients were enrolled in the study cohort and divided into 4 groups based on SHR quartile (Figure 1). Quartile 4 representing the highest SHR was consistent with previously published studies [25].

XSL•FO
RenderX

Figure 1. Patient selection flowchart. Flowchart illustrating the inclusion and exclusion criteria for patient selection in a retrospective cohort study investigating the association between the stress hyperglycemia ratio and 7-day sepsis risk among critically ill patients with heart failure. Data were derived from the MIMIC-IV database, Beth Israel Deaconess Medical Center (Boston, MA, United States), encompassing admissions from 2008 to 2022. A total of 1205 patients were included in the final analysis. Inclusion and exclusion criteria are detailed. HbA_{1c}: glycated hemoglobin; ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care-IV.



Definition of Sepsis

Sepsis was defined according to Sepsis-3 criteria [3]. The time of sepsis onset (T_{sepsis}) was determined as the earlier of the first clinical suspicion of infection ($T_{suspicion}$) and the first time the Sequential Organ Failure Assessment (SOFA) score increased by ≥2 points (T_{SOFA}), provided that T_{SOFA} occurred no more than 24 hours before or 12 hours after $T_{suspicion}$ according to published literature [24]. To specifically capture new-onset sepsis developing during the ICU stay and to minimize the inclusion of infections likely present at admission, we excluded all patients for whom T_{sepsis} occurred within the first 6 hours after ICU admission in our primary analysis.

Inflammation Biomarkers

Inflammatory biomarkers included NLR, platelet-to-lymphocyte ratio, monocyte-to-lymphocyte ratio, NPR, NMLR, SII, and SIRI. Detailed calculation methods are provided in Table S2 in Multimedia Appendix 1.

Data Collection

The SHR was calculated using the formula: SHR=admission blood glucose (mg/dL)/ (28.7×HbA_{1c} (%)−46.7) [11]. Blood glucose and HbA_{1c} values were sourced from the initial records after ICU admission. The baseline variables were chosen on the bases of their possible influence on sepsis risk and cardiovascular risk of the individuals and followed six categories: (1) demographics (sex, age, race, and BMI), (2) vital signs (heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature, and 6-hour fluid intake), (3) laboratory parameters (hemoglobin, white blood cell, platelet, blood urea nitrogen, creatinine, glucose, HbA_{1c}, etc), (4) comorbidities (diabetes, hypertension, shock, etc), (5) disease severity scores (Charlson comorbidity

XSL•FO
RenderX

index, Acute Physiology Score III, systemic inflammatory response syndrome score, SOFA score, etc), and (6) medication (angiotensin-converting enzyme inhibitors/angiotensin receptor blockers, diuretics, antibiotics drugs, insulin, glucocorticoid, oral hypoglycemic drugs, etc). For all baseline variables, the first recorded value within the initial 6 hours after ICU admission was extracted using PostgreSQL (version 13.7.1; PostgreSQL Global Development Group) and Navicat Premium (version 17; PremiumSoft CyberTech Ltd) and used in the analysis. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational studies [26]. Variables with a missing rate ≤25% were imputed using the Multivariate Imputation by Chained Equations method, generating 20 imputed datasets, applying predictive mean matching for continuous variables and logistic regression for categorical variables. The imputation model included all baseline variables used in the analysis. Variables with a missing rate >25% were excluded to avoid bias. Further details are provided in Table S3 in Multimedia Appendix 1.

Definition of Outcomes

The primary outcome was the occurrence of sepsis within 7 days after ICU admission, which was defined based on the acute disease distribution patterns observed in our dataset (Figure S1 in Multimedia Appendix 1) and was consistent with the early prediction window commonly adopted in prior studies focusing on acute sepsis in ICU settings [27]. Secondary outcomes included 7-day mortality, a composite outcome of sepsis occurrence or mortality within 7 days, 28-day mortality, and in-hospital mortality.

Statistical Analysis

Continuous variables were described using median (IQR) or mean (SD), and comparisons between groups were performed using the Kruskal-Wallis test or ANOVA, as appropriate. Whereas categorical variables were expressed as frequency (%) and were analyzed using the Fisher exact test or the Pearson chi-square test. The Kaplan-Meier survival analysis was used to estimate the occurrence of primary outcomes among groups according to the log-rank test. To assess the impact of SHR on sepsis occurrence, the multivariate proportional hazard regression models (Cox regression models) were used to assess the hazard ratio (HR) and 95% CI for event occurrence. Specifically, model 1 was unadjusted for covariates, and model 2 further adjusted sex, age, and BMI. Considering the impacts of patient general condition, vital signs, SOFA scores, and steroid use on outcomes, based on model 2, model 3 incorporated acute HF, diabetes, shock, SOFA score, antibiotic, insulin, and glucocorticoid. To explore the potential dose-response relationship between SHR and the risk of sepsis occurrence, restricted cubic splines (RCS) analysis was performed and adjusted by the same models in model 3. Furthermore, subgroup analyses were conducted to delve deeper into the data, stratifying outcomes based on sex, age, BMI, diabetes, acute HF, and insulin use. These subgroup analyses were performed using comprehensive regression models adjusted for potential confounding factors. To test whether diabetes status modifies the relationship between SHR and sepsis risk, we introduced an interaction term between SHR (as a continuous variable) and diabetes status (yes or no) within the Cox regression model, which was also adjusted for the same covariates as in the fully adjusted model (model 3). Moreover, to rigorously evaluate the robustness of our main findings, sensitivity analyses were performed by logistic regression, Fine-Gray, and competing risk Cox models. In addition, we extended the exclusion window for incident sepsis from the original 6 hours to 12, 24, 36, and 48 hours after ICU admission, respectively, and repeated the Cox proportional hazards regression models.

To determine whether inflammatory indices (such as NLR and SII) mediate the association between SHR and sepsis occurrence, we used the Cox model in the *CMAverse* package in R (R Foundation for Statistical Computing). We performed 1000 bootstrap simulations to estimate each mediator's effect and calculate the mediation proportion. The average direct effect represents the impact of SHR on sepsis without mediation, while the average causal mediation effect indicates the effect of SHR on sepsis occurrence via mediators. The mediation proportion was calculated by dividing the average causal mediation effect by the total effect. All statistical analyses were executed using R software (version 4.4.1), with statistical significance set as a 2-sided *P* value of less than .05.

Results

Baseline Characteristics

A total of 25,718 patients with HF from the ICU were consecutively recruited. Following certain exclusion criteria, 1205 patients were ultimately included, of whom 63.4% (n=764) were male, and had a median age of 71.51 (IQR 62.45 - 79.47) years. Table 1 presents the baseline characteristics. There were 297 (24.6%) patients with hypertension, 564 (46.8%) patients with diabetes, 855 (71%) patients with acute HF, and 595 (49.4%) patients with pneumonia.

XSL•FO
RenderX

Table . Baseline characteristics of study participants stratified by stress hyperglycemia ratio (SHR) quartiles[a].

	Overall (N=1205)	Groups of SHR[b]				*P* value
		Q1 (n=301)	Q2 (n=302)	Q3 (n=301)	Q4 (n=301)	
Male, n (%)	764 (63.4)	206 (68.4)	194 (64.2)	188 (62.5)	176 (58.5)	.08
Age (years), median (IQR)	71.51 (62.45 - 79.47)	71.64 (60.73 - 79.40)	71.69 (62.35 - 79.97)	71.61 (62.25 - 79.25)	71.15 (63.13 - 79.61)	.92
Race, n (%)						.38
Black	108 (9)	26 (8.6)	30 (9.9)	21 (7)	31 (10.3)	
White	800 (66.4)	202 (67.1)	206 (68.2)	207 (68.8)	185 (61.5)	
Others	297 (24.6)	73 (24.3)	66 (21.9)	73 (24.3)	85 (28.2)	
BMI (kg/m^2), median (IQR)	28.54 (24.42 - 32.92)	28.73 (25.10 - 33.59)	27.94 (24.42 - 31.99)	28.64 (24.54 - 32.87)	28.20 (23.93 - 33.20)	.65
HbA_{1c}[c] (%), median (IQR)	6.00 (5.60 - 7.10)	6.50 (5.90 - 8.00)	5.80 (5.50 - 6.57)	5.90 (5.50 - 6.70)	6.00 (5.50 - 7.40)	<.001
Glucose (mmol/L), median (IQR)	133.00 (106.00 - 182.00)	99.00 (89.00 - 115.00)	115.00 (104.00 - 134.75)	140.00 (123.00 - 171.00)	219.00 (175.00 - 275.00)	<.001
SHR, median (IQR)	1.03 (0.85 - 1.32)	0.74 (0.62 - 0.80)	0.94 (0.90 - 0.99)	1.13 (1.08 - 1.21)	1.59 (1.42 - 1.86)	<.001
Vital signs, median (IQR)						
Heart rate (bpm)	81.28 (73.59 - 91.48)	81.04 (73.41 - 89.58)	80.85 (74.38 - 89.13)	81.54 (74.40 - 93.11)	82.31 (72.71 - 92.97)	.43
SBP[d] (mm Hg)	111.48 (104.13 - 122.04)	109.90 (104.39 - 119.23)	111.00 (105.26 - 119.29)	111.70 (103.58 - 122.68)	113.79 (104.18 - 126.50)	.08
DBP[e] (mm Hg)	60.85 (53.94 - 69.63)	58.71 (52.62 - 65.57)	59.06 (52.74 - 67.47)	62.10 (55.45 - 70.69)	64.44 (56.58 - 72.86)	<.001
MBP[f] (mm Hg)	75.91 (70.45 - 83.49)	74.24 (69.61 - 81.09)	74.66 (70.02 - 81.41)	77.53 (70.95 - 84.12)	78.42 (71.95 - 86.63)	<.001
Respiratory rate (breaths per minute)	19.08 (17.22 - 21.06)	18.65 (16.91 - 20.48)	18.87 (17.01 - 20.92)	19.36 (17.29 - 21.31)	19.50 (17.67 - 21.61)	<.001
Temperature (°C)	37.06 (36.83 - 37.44)	37.06 (36.83 - 37.40)	37.10 (36.89 - 37.50)	37.11 (36.83 - 37.44)	37.06 (36.83 - 37.50)	.54
Intake (mL per 6 hours)	499.86 (227.77 - 1221.86)	579.12 (240.00 - 1275.58)	718.86 (254.32 - 1654.32)	477.99 (244.53 - 1111.45)	382.35 (165.92 - 800.00)	<.001
Laboratory test, median (IQR)						
Hemoglobin (g/dL)	12.00 (10.40 - 13.70)	11.60 (9.80 - 13.40)	12.00 (10.70 - 13.47)	12.30 (10.50 - 13.90)	12.20 (10.40 - 13.80)	.03
White blood cell (10^9/L)	12.10 (8.80 - 16.60)	11.20 (8.00 - 16.50)	12.60 (8.70 - 17.78)	11.90 (9.20 - 15.50)	12.30 (9.40 - 16.60)	.05
Platelets (10^9/L)	217.00 (176.00 - 274.00)	207.00 (171.00 - 262.00)	214.50 (171.50 - 267.75)	210.00 (174.00 - 265.00)	234.00 (189.00 - 287.00)	<.001
HCT[g] (%)	36.80 (31.90 - 41.20)	35.60 (30.90 - 41.00)	36.75 (32.40 - 41.18)	37.70 (32.10 - 41.80)	37.00 (32.00 - 41.40)	.22
BUN[h] (mg/dL)	24.00 (18.00 - 36.00)	25.00 (18.00 - 38.00)	22.00 (17.00 - 33.00)	23.00 (17.00 - 33.00)	25.00 (19.00 - 43.00)	.001
Creatinine (μmol/L)	1.20 (0.90 - 1.60)	1.20 (0.90 - 1.70)	1.20 (0.90 - 1.58)	1.10 (0.90 - 1.50)	1.20 (1.00 - 1.80)	.005
Chloride (mmol/L)	104.00 (100.00 - 107.00)	104.00 (101.00 - 107.00)	105.00 (101.00 - 108.00)	103.00 (99.00 - 106.00)	102.00 (98.00 - 105.00)	<.001

XSL•FO
RenderX

	Overall (N=1205)	Groups of SHR[b]				*P* value
		Q1 (n=301)	Q2 (n=302)	Q3 (n=301)	Q4 (n=301)	
Sodium (mmol/L)	139.00 (136.00 - 141.00)	139.00 (137.00 - 141.00)	139.00 (137.00 - 141.00)	138.00 (136.00 - 141.00)	138.00 (135.00 - 141.00)	<.001
Potassium (mmol/L)	4.40 (4.00 - 4.80)	4.40 (4.00 - 4.70)	4.30 (4.00 - 4.70)	4.40 (4.00 - 4.70)	4.50 (4.10 - 5.00)	.004
Bicarbonate (mmol/L)	25.00 (22.00 - 27.00)	25.00 (23.00 - 28.00)	25.00 (23.00 - 27.00)	24.00 (22.00 - 27.00)	23.00 (21.00 - 26.00)	<.001
PT[i] (seconds)	14.70 (12.90 - 17.40)	15.30 (13.50 - 17.90)	15.70 (13.60 - 18.00)	14.30 (12.70 - 17.40)	13.70 (12.40 - 16.40)	<.001
APTT[j] (seconds)	41.50 (30.90 - 73.80)	40.50 (31.60 - 72.40)	41.45 (31.63 - 72.57)	41.30 (30.20 - 74.00)	43.80 (29.10 - 79.90)	.94
Comorbidities, n (%)						
Acute HF[k]	855 (71)	205 (68.1)	210 (69.5)	209 (69.4)	231 (76.7)	.08
Hypertension	297 (24.6)	73 (24.3)	83 (27.5)	73 (24.3)	68 (22.6)	.56
Coronary heart disease	934 (77.5)	229 (76.1)	220 (72.8)	234 (77.7)	251 (83.4)	.02
AMI[l]	551 (45.7)	142 (47.2)	144 (47.7)	131 (43.5)	134 (44.5)	.69
Valve disorder	540 (44.8)	166 (55.1)	146 (48.3)	122 (40.5)	106 (35.2)	<.001
Atrial fibrillation	604 (50.1)	163 (54.2)	151 (50)	157 (52.2)	133 (44.2)	.08
COPD[m]	329 (27.3)	99 (32.9)	84 (27.8)	73 (24.3)	73 (24.3)	.06
Chronic kidney disease	421 (34.9)	118 (39.2)	103 (34.1)	95 (31.6)	105 (34.9)	.26
Liver disease	74 (6.1)	23 (7.6)	22 (7.3)	19 (6.3)	10 (3.3)	.11
Cancer	40 (3.3)	5 (1.7)	8 (2.6)	16 (5.3)	11 (3.7)	.08
Diabetes	564 (46.8)	166 (55.1)	100 (33.1)	138 (45.8)	160 (53.2)	<.001
Lipoprotein metabolism disorders	800 (66.4)	216 (71.8)	203 (67.2)	186 (61.8)	195 (64.8)	.07
Shock	284 (23.6)	81 (26.9)	58 (19.2)	60 (19.9)	85 (28.2)	.01
Pneumonia	104 (8.6)	28 (9.3)	25 (8.3)	26 (8.6)	25 (8.3)	.97
Disease severity score, median (IQR)						
Charlson comorbidity index	6.00 (4.00 - 8.00)	6.00 (5.00 - 8.00)	6.00 (4.00 - 8.00)	6.00 (4.00 - 8.00)	7.00 (5.00 - 8.00)	.05
APSIII[n]	40.00 (31.00 - 50.00)	41.00 (32.00 - 53.00)	39.00 (30.00 - 48.75)	38.00 (31.00 - 49.00)	41.00 (32.00 - 52.00)	.15
SAPSII[o]	35.00 (28.00 - 43.00)	35.00 (29.00 - 43.00)	37.00 (31.00 - 44.00)	34.00 (26.00 - 43.00)	35.00 (28.00 - 42.00)	.05
SIRS[p]	3.00 (2.00 - 3.00)	3.00 (2.00 - 3.00)	3.00 (2.00 - 3.00)	3.00 (2.00 - 3.00)	3.00 (2.00 - 3.00)	.49
SOFA[q]	3.00 (1.00 - 5.00)	4.00 (2.00 - 6.00)	3.00 (1.00 - 6.00)	2.00 (1.00 - 4.00)	2.00 (1.00 - 4.00)	<.001
Medication at baseline, n (%)						
ACEI/ARB[r]	398 (33)	108 (35.9)	106 (35.1)	95 (31.6)	89 (29.6)	.31
Amiodarone	154 (12.8)	37 (12.3)	38 (12.6)	33 (11)	46 (15.3)	.45
Anticoagulants	139 (11.5)	48 (15.9)	29 (9.6)	32 (10.6)	30 (10)	.05
NSAID[s]	770 (63.9)	210 (69.8)	202 (66.9)	195 (64.8)	163 (54.2)	<.001
β-Blocker	737 (61.2)	203 (67.4)	209 (69.2)	174 (57.8)	151 (50.2)	<.001

XSL•FO
RenderX

	Overall (N=1205)	Groups of SHR[b]				*P* value
		Q1 (n=301)	Q2 (n=302)	Q3 (n=301)	Q4 (n=301)	
Diuretic drug	751 (62.3)	197 (65.4)	204 (67.5)	177 (58.8)	173 (57.5)	.03
Hypolipidemic drug	751 (62.3)	206 (68.4)	194 (64.2)	187 (62.1)	164 (54.5)	.004
Antibiotics	595 (49.4)	184 (61.1)	193 (63.9)	130 (43.2)	88 (29.2)	<.001
Insulin	747 (62)	221 (73.4)	192 (63.6)	163 (54.2)	171 (56.8)	<.001
Glucocorticoid	167 (13.9)	49 (16.3)	50 (16.6)	34 (11.3)	34 (11.3)	.09
Oral hypoglycemic drugs	19 (1.6)	8 (2.7)	5 (1.7)	1 (0.3)	5 (1.7)	.15

[a]Baseline demographic, clinical, laboratory, and treatment characteristics of 1205 critically ill patients with HF, stratified by quartiles of the stress hyperglycemia ratio. Data were derived from the Medical Information Mart for Intensive Care-IV database (Boston, MA, United States; 2008 - 2022). *P* values were calculated using the Kruskal-Wallis test for continuous variables and the Pearson chi-square or Fisher exact test for categorical variables.

[b]SHR quartiles: Q1: <0.8490; Q2: 0.8490-1.0331; Q3: 1.0331-1.3177; Q4: >1.3177.

[c]HbA_{1c}: glycated hemoglobin.

[d]SBP: systolic blood pressure.

[e]DBP: diastolic blood pressure.

[f]MBP: mean blood pressure.

[g]HCT: hematocrit.

[h]BUN: blood urea nitrogen.

[i]PT: prothrombin time.

[j]APTT: activated partial thromboplastin time.

[k]HF: heart failure.

[l]AMI: acute myocardial infarction.

[m]COPD: chronic obstructive pulmonary disease.

[n]APSIII: Acute Physiology Score III.

[o]SAPSII: Simplified Acute Physiology Score II.

[p]SIRS: systemic inflammatory response syndrome.

[q]SOFA: Sequential Organ Failure Assessment.

[r]ACEI/ARB: angiotensin-converting enzyme inhibitor/angiotensin receptor blocker.

[s]NSAID: nonsteroidal anti-inflammatory drug.

All participants were stratified into 4 groups (Q1-4) based on SHR quartiles: Q1: <0.8490 (n=301); Q2: 0.8490 - 1.0331 (n=302); Q3: 1.0331 - 1.3177 (n=301); and Q4: >1.3177 (n=301). The highest SHR group (Q4) exhibited increased vital signs, including respiratory rate and diastolic blood pressure, alongside elevated laboratory values such as platelet count and serum potassium levels. They had a higher burden of comorbidities (eg, coronary heart disease) and disease severity scores (Charlson index and Simplified Acute Physiology Score II), but lower antibiotic use. In contrast, the lowest SHR group (Q1) had higher HbA_{1c}, diabetes prevalence, and insulin use. However, parameters such as age, BMI, and race levels showed no significant variation across SHR groups.

Furthermore, for clinical outcome, Table S4 in Multimedia Appendix 1 provides a comparison of baseline characteristics between patients with sepsis and those without sepsis within 7 days after admitted to the ICU. Compared to patients without sepsis, those with sepsis exhibited significantly elevated levels of SHR (median 1.13, IQR 0.92 - 1.46 vs median 1.02, IQR 0.84 - 1.29; *P*=.001) and higher inflammation infection.

Considering SHR be associated with diabetes status, compared to patients without diabetes, those with diabetes had higher BMI (28.96 vs 27.90 kg/m²) and no significant difference in SHR (*P*=.79; Table S5 in Multimedia Appendix 1).

Relationship Between SHR and Parameters Related to Sepsis Diagnosis

Figures S2 and S3 in Multimedia Appendix 1 illustrate the relationship between SHR and markers of sepsis. The association between SHR and antibiotic use (defined as administration during the entire ICU stay) was assessed using the point-biserial correlation coefficient, yielding a weak negative correlation ($r=-0.196$). The Mann-Whitney *U* test further confirmed a difference (*P*<.001). In contrast, the linear relationship between SHR and SOFA score was evaluated using Spearman correlation and was not statistically significant ($r=0.056$; *P*=.05).

Clinical Outcomes According to SHR Levels

In this study, sepsis occurred in 162 (13.4%) patients with HF, among whom, the Q4 (SHR >1.3177) exhibited the highest sepsis occurrence at 17.3%, thus being chosen as the reference group, and Q1 to Q4 showed progressively higher risks (9.6% vs 11.3% vs 15.6% vs 17.3%). Notably, the Q4 group exhibited a significantly higher mortality rate of 17.3% (52/301) and a

XSL•FO
RenderX

composite outcome rate of 34.2% (103/301) compared with all other groups (both *P*<.001; S6 in Multimedia Appendix 1).

The Kaplan-Meier curves illustrated that patients with the highest SHR (Q4) exhibited the highest 7-day all-cause mortality across different SHR levels (*P*=.02; log-rank *P*<.001; Figure S4 in Multimedia Appendix 1). Specifically, although the association was nonsignificant in the diabetes subgroup (*P*=.17) and the nondiabetes subgroup (*P*=.07), the Q4 had relatively the highest 7-day all-cause mortality in the diabetes subgroup, while in the nondiabetes subgroup, patients in Q3 and Q4 groups had higher 7-day all-cause mortality than other groups (Figures S5 and S6 in Multimedia Appendix 1).

Association of SHR Index With Sepsis Occurrence

To explore the independent relationship of SHR index with the occurrence of sepsis, we constructed 3 Cox regression models (Table 2), and significant values were all found in models 1 - 3. When SHR emerged as a continuous variable, in the fully adjusted model 3, per 1-unit higher SHR index was associated with an 18% higher risk of sepsis occurrence (HR 1.18, 95% CI 1.01 - 1.38; *P*=.04). Additionally, similar trends were observed when patients were grouped according to quartiles of the SHR index. Taking the highest quartile of the SHR index (Q4) as the reference, the occurrence risk of sepsis in other groups showed a gradual upward trend of Q1 <Q2 <Q4 (HR 0.52, 95% CI 0.32 - 0.82; *P*=.005 vs HR 0.62, 95% CI 0.40 - 0.97; *P*=.04), while compared to Q4, Q3 showed no significant difference (HR 0.96, 95% CI 0.65 - 1.43; *P*=.85), indicating a dose-response relationship between SHR levels and sepsis risk. However, the trend test indicated a significant upward trend in sepsis risk with increasing quartiles (*P* for trends=.001).

Table . Cox regression models for the association between stress hyperglycemia ratio (SHR) and 7-day sepsis occurrence[a].

	Events, n	Cases, n	Model 1[b]		Model 2[c]		Model 3[d]	
			HR[e] (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
SHR	—[f]	—	1.31 (1.11 - 1.54)	.002	1.34 (1.13 - 1.59)	.001	1.18 (1.01 - 1.38)	.04
Q1	29	301	0.54 (0.34 - 0.84)	.007	0.54 (0.34 - 0.85)	.008	0.52 (0.32 - 0.82)	.005
Q2	34	302	0.62 (0.40 - 0.96)	.03	0.62 (0.40 - 0.96)	.03	0.62 (0.40 - 0.97)	.04
Q3	47	301	0.90 (0.61 - 1.33)	.59	0.90 (0.60 - 1.33)	.59	0.96 (0.65 - 1.43)	.85
Q4	52	301	Reference	—	Reference	—	Reference	—
P for trends	—	—	1.25 (1.09 - 1.44)	.002	1.25 (1.08 - 1.44)	.002	1.26 (1.09 - 1.46)	.001

[a]Multivariable Cox proportional hazards regression model evaluating the association between the SHR (both as a continuous variable and by quartiles) and the risk of sepsis within 7 days of intensive care unit admission among 1205 critically ill patients with heart failure. Data were derived from the Medical Information Mart for Intensive Care-IV database (Boston, MA, United States; 2008 - 2022).

[b]Model 1: unadjusted.

[c]Model 2: adjusted for sex, age, and BMI.

[d]Model 3: adjusted for sex, age, BMI, acute heart failure, diabetes, shock, Sequential Organ Failure Assessment score, antibiotic use, insulin use, and glucocorticoid use.

[e]HR: hazard ratio.

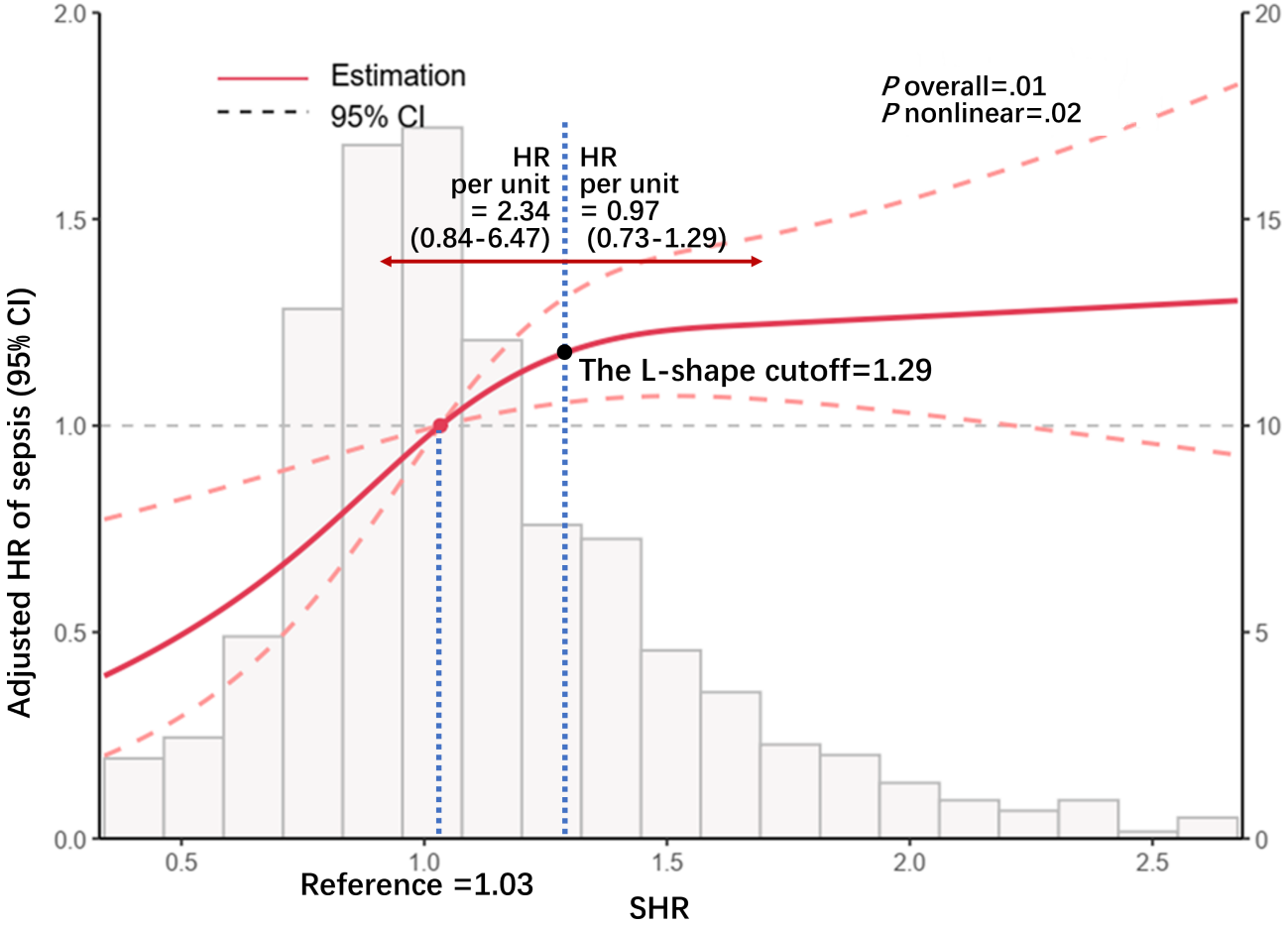
[f]Not available.

The interaction analysis between SHR and diabetes was statistically significant (HR 1.96, 95% CI 1.23 - 3.14; *P*=.005), indicating that the association between SHR and sepsis was stronger in patients with diabetes. Similarly, a significant interaction was also observed with insulin use (HR 0.48, 95% CI 0.30 - 0.77; *P*=.002), whereas interactions with BMI (both continuous and categorized) were not statistically significant (*P*=.17 and *P*=.07, respectively; Table S7 in Multimedia Appendix 1).

Additionally, multivariable adjusted RCS analysis was used to provide a comprehensive examination of the continuous relationship between SHR and the occurrence of sepsis. Relationship showed a nonlinear saturation effect as SHR increased, with the specific turning point of 1.29 (*P* for nonlinearity=.02; Figure 2). Specifically, sepsis risk increased steeply with SHR up to 1.29, after which the risk curve plateaued.

XSL•FO
RenderX

Figure 2. Restricted cubic spline analysis of the association between SHR and 7-day sepsis risk. Restricted cubic spline curve depicting the dose-response relationship between the SHR (as a continuous variable) and the adjusted HR for 7-day sepsis occurrence among 1205 critically ill patients with heart failure. Data were derived from the MIMIC-IV database (Boston, MA, United States; 2008 - 2022). The model was adjusted for sex, age, BMI, acute heart failure, diabetes, shock, Sequential Organ Failure Assessment score, antibiotic use, insulin use, and glucocorticoid use. The curve demonstrates a nonlinear saturation effect, with an inflection point at SHR=1.29 (*P* for nonlinearity=.02). The horizontal dashed line represents an HR of 1.0. The shaded area represents the 95% CI. HR: hazard ratio; MIMIC-IV: Medical Information Mart for Intensive Care-IV; SHR: stress hyperglycemia ratio.

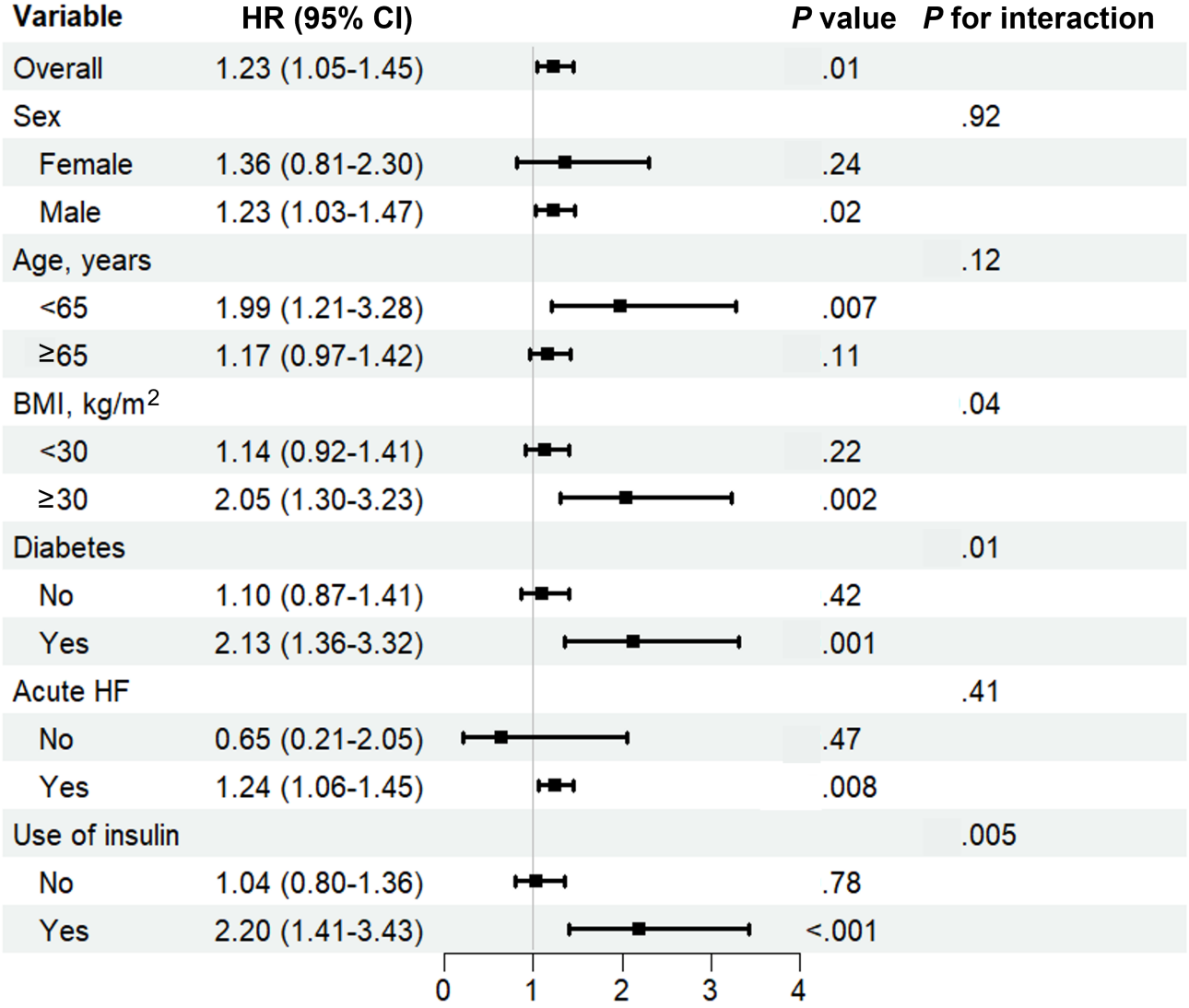


Subgroup Analysis

Subsequently, the subgroup analyses were stratified by sex (male and female), age (<65 and ≥65 years), BMI (<30 and ≥30 kg/m^2), diabetes (yes and no), acute HF (yes and no), and insulin use (yes and no). No significant interactions were observed for sex, age, or acute HF subgroups (*P* for interaction>.05). Notably, significant interactions were detected between SHR and BMI (*P*=.04), diabetes (*P*=.01), and insulin use (*P*=.005; Figure 3), thus prompting further RCS analyses within these subgroups. Results indicated that in the subgroups with diabetes, insulin use, and BMI ≥30 kg/m^2, a linear association was observed between SHR and sepsis incidence (all *P* overall<.05; *P* nonlinear>.05). In contrast, a U-shaped relationship was identified in the subgroup with BMI <30 kg/m^2, with the inflection point at 1.03 (*P* overall=.01; *P* nonlinear=.09). No significant associations were found in the remaining subgroups (*P* overall>.05; Figures S7-S9 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. Subgroup analysis forest plot. Forest plot summarizing subgroup analyses of the association between the stress hyperglycemia ratio (as a continuous variable) and 7-day sepsis risk among 1205 critically ill patients with HF. Data were derived from the MIMIC-IV database (Boston, MA, United States; 2008 - 2022). Subgroups were stratified by sex, age (<65 vs ≥65 years), BMI (<30 vs ≥30 kg/m^2), diabetes status, acute HF, and insulin use. HRs and 95% CIs were adjusted for sex, age, BMI, acute HF, diabetes, and insulin use. *P* values for interaction are displayed to assess effect modification. HF: heart failure; HR: hazard ratio; MIMIC-IV: Medical Information Mart for Intensive Care-IV.

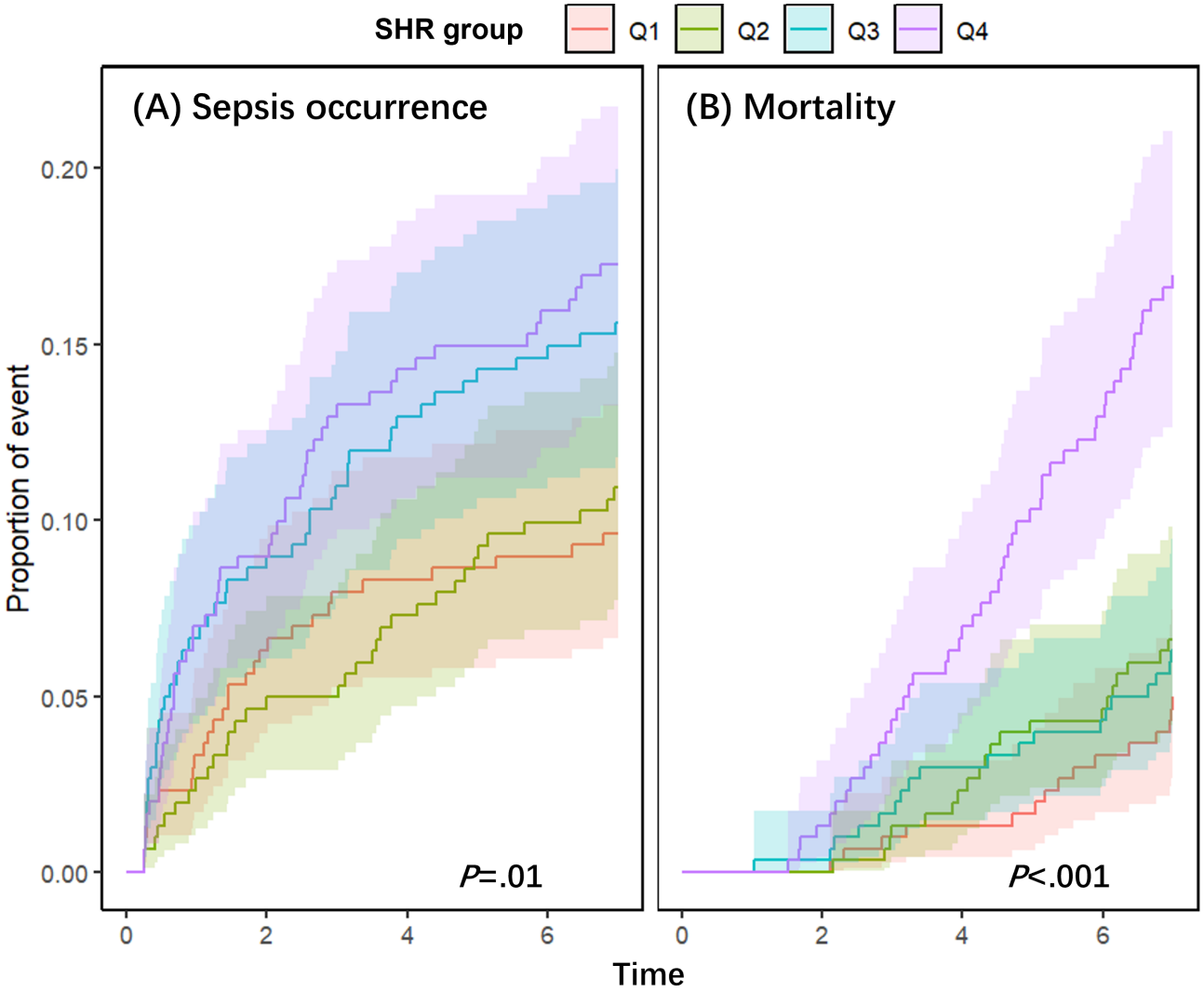


Sensitivity Analyses

We performed different sensitivity tests to validate the robustness of our results. First, considering 7-day all-cause mortality as a competing event, the Fine-Gray competing risk model was used. In the univariable-adjusted Fine-Gray model, the cumulative occurrence of sepsis increased from Q1 to Q4 (*P*=.01), maintaining significance even when accounting for the competing risk of all-cause mortality (*P*<.001; Figure 4).

XSL•FO
RenderX

Figure 4. Cumulative incidence curves from Fine-Gray competing risk models, illustrating the impact of SHR quartiles on (A) the cumulative incidence of sepsis within 7 days and (B) the cumulative incidence of 7-day all-cause mortality, with each considered a competing event for the other. The analysis is based on a retrospective cohort of 1205 critically ill patients with heart failure from the MIMIC-IV database (Boston, MA, United States; 2008-2022). SHR quartiles: Q1 (<0.849); Q2 (0.849 - 1.033); Q3 (1.033 - 1.318); Q4 (>1.318). MIMIC-IV: Medical Information Mart for Intensive Care-IV; SHR: stress hyperglycemia ratio.



In the multivariable-adjusted Fine-Gray model, SHR, as a categorical variable (quartiles), revealed a significant overall difference in sepsis incidence across groups (*P*=.01). When modeled as a continuous variable in the fully adjusted model, the association was positive but not statistically significant (HR 1.18, 95% CI 0.96 - 1.44; *P*=.11). There was a significant positive trend in sepsis risk across increasing SHR quartiles, indicating a dose-response relationship (*P* for trend<.001 in the fully adjusted model; Table S8 in Multimedia Appendix 1).

Second, the logistic regression was also made to show a significant positive association between SHR and sepsis risk (model 3: odds ratio 1.03, 95% CI 1.01 - 1.05; *P*<.001), with sepsis risk increasing significantly across SHR quartiles (*P* for trend <.001; Table S9 in Multimedia Appendix 1).

Third, when the sepsis exclusion window was extended to 12, 24, 36, and 48 hours, the overall association remained consistent. However, the association was not significant when the exposure was modeled continuously in the fully adjusted model (all *P*>.05; Table S10 in Multimedia Appendix 1).

Mediation Analysis

Table 3 shows the mediation value of inflammatory indices between SHR and sepsis occurrence.

XSL•FO
RenderX

Table . Mediation effects of inflammatory biomarkers on the stress hyperglycemia ratio-sepsis association[a].

	Case, n	ACME[b]		ADE[c]		PE[d]		TE[e]	
		Estimate (95% CI)	*P* value	Estimate (95% CI)	*P* value	Estimate (95% CI)	*P* value	Estimate (95% CI)	*P* value
SII[f]	807	1.05 (1.01 to 1.13)	.01	1.90 (1.23 to 3.02)	.01	0.10 (0.01 to 0.30)	.01	2.00 (1.28 to 3.15)	.004
NLR[g]	807	1.03 (1.00 to 1.09)	.04	1.93 (1.23 to 2.97)	.004	0.06 (0.00 to 0.23)	.04	1.99 (1.29 to 3.04)	<.001
PLR[h]	807	1.02 (0.99 to 1.06)	.23	1.92 (1.26 to 2.93)	.004	0.04 (−0.02 to 0.14)	.23	1.96 (1.27 to 2.97)	<.001
MLR[i]	807	1.01 (0.97 to 1.06)	.53	1.97 (1.26 to 3.01)	<.001	0.03 (−0.06 to 0.13)	.53	1.99 (1.27 to 3.06)	<.001
NPR[j]	808	1.05 (1.00 to 1.12)	.04	1.88 (1.22 to 3.02)	.004	0.10 (0.00 to 0.30)	.04	1.97 (1.31 to 3.14)	.004
SIRI[k]	807	1.07 (1.00 to 1.15)	.04	1.78 (1.15 to 2.75)	.01	0.13 (0.00 to 0.39)	.05	1.90 (1.23 to 2.93)	.004
NMLR[l]	807	1.03 (1.00 to 1.09)	.04	1.93 (1.23 to 2.98)	.004	0.06 (0.00 to 0.23)	.04	1.99 (1.28 to 3.04)	<.001

[a]Mediation analysis estimating the ACMEs, ADEs, PM, and TEs of systemic inflammatory indices in the association between the stress hyperglycemia ratio and 7-day sepsis occurrence. Analysis was performed among 807 patients with complete inflammatory biomarker data from the Medical Information Mart for Intensive Care-IV database (Boston, MA, United States; 2008 - 2022). Estimates were derived from a fully adjusted Cox regression model using 1000 bootstrap resamplings. The mediation effects presented in the table are estimated based on the fully adjusted model, which accounts for adjustments including sex, age, BMI, acute heart failure, diabetes, shock, Sequential Organ Failure Assessment score, antibiotic use, insulin use, and glucocorticoid use.

[b]ACME: average causal mediation effect (indirect effect).

[c]ADE: average direct effect.

[d]PE: proportion mediated.

[e]TE: total effect.

[f]SII: systemic immune-inflammation index.

[g]NLR: neutrophil-to-lymphocyte ratio.

[h]PLR: platelet-to-lymphocyte ratio.

[i]MLR: monocyte-to-lymphocyte ratio.

[j]NPR: neutrophil-to-platelet ratio.

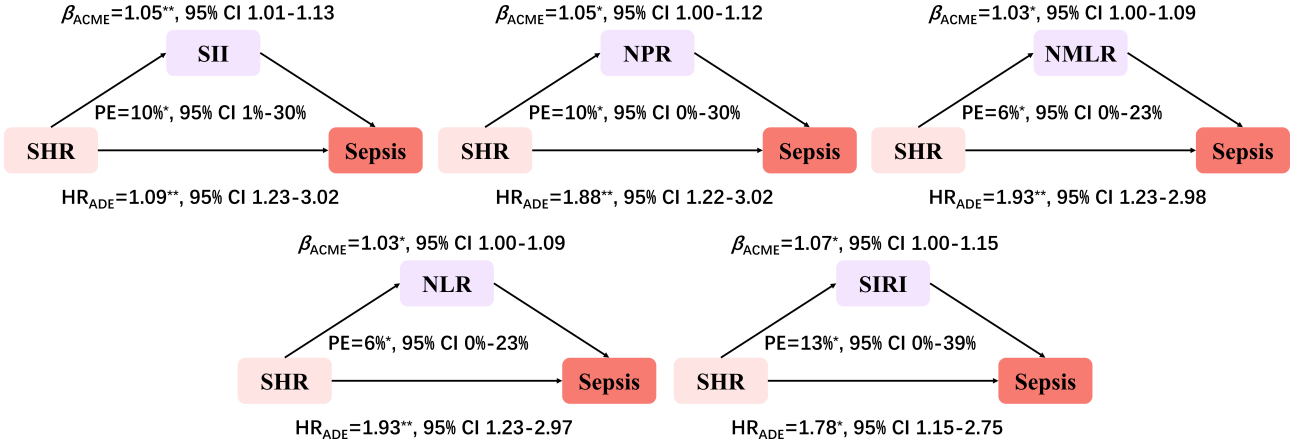
[k]SIRI: systemic inflammation response index.

[l]NMLR: neutrophil-monocyte-to-lymphocyte ratio.

The mediation analysis indicates that, for the general population, the estimated proportions of the association mediated by NLR, NPR, NMLR, SIRI, and SII were 6% (95% CI 0.00% - 23%), 10% (95% CI 0.00% - 30%), 6% (95% CI 0.00% - 23%), 13% (95% CI 0.00% - 39%), and 10% (95% CI 1% - 30%), respectively (Figure 5).

XSL•FO
RenderX

Figure 5. Mediation analysis (adjusted for sex, age, BMI, acute heart failure, diabetes, shock, Sequential Organ Failure Assessment score, antibiotic use, insulin use, and glucocorticoid use). ACME: average causal mediation effect; ADE: average direct effect; HR: hazard ratio; NLR: neutrophil-to-lymphocyte ratio; NMLR: neutrophil-monocyte-to-lymphocyte ratio; NPR: neutrophil-to-platelet ratio; PE: proportion mediated; SHR: stress hyperglycemia ratio; SII: systemic immune-inflammation index; SIRI: systemic inflammation response index. *P<.05; **P≤.01.



For the diabetes subpopulation, no significant exposure mediator interaction effects were observed (all P>.05; Table S11 in Multimedia Appendix 1). Additionally, mediation analysis in the nondiabetic subgroup was underpowered due to limited sample size, resulting in unstable estimates with wide CIs (Table S12 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we first demonstrated that a higher SHR was significantly associated with a higher risk of sepsis occurrence after fully adjusting for potential confounders in critically ill patients with HF based on the MIMIC-IV database, with an evidently positive linear relationship in the diabetes and BMI ≥30 kg/m^2 subgroup. The association between SHR and sepsis occurrence in the RCS analysis after adjusting confounders further showed a nonlinear pattern, where risk increased up to an SHR threshold (1.29) and then plateaued. Furthermore, systemic inflammatory indices partially mediated this relationship, implicating inflammation as a potential mechanistic link.

Comparison to Prior Work

The concept of SHR highlights a relative acute rise in glycemia compared to the individual's previous glycemic status in response to a stress reaction or critical illness [11], which was triggered by the release of stress hormones (eg, adrenaline and cortisol), immune responses, neural activity, and insulin resistance [28].

Extensive research has shown that SHR was associated with adverse outcomes across various diseases. Consistent with our findings, elevated SHR has been linked to poorer prognosis in patients with acute coronary syndrome [12], acute HF [13], and sepsis [14,29,30]. However, while several studies have demonstrated a U-shaped association between SHR and mortality in various cardiovascular and metabolic diseases, including HF [13], acute myocardial infarction [31], and cardiovascular-kidney-metabolic syndrome [32], the relationship between SHR and the occurrence of sepsis in critically ill patients with HF remains underexplored. Our study revealed a positive, nonlinear saturation effect between SHR and sepsis occurrence in critically ill patients with HF. However, a U-shaped association between SHR and the occurrence of acute kidney injury in patients with congestive heart failure was observed in a study by Li et al [33], which contrasts with our finding of a plateauing effect. This discrepancy may be attributable to differences in the outcomes studied (sepsis vs acute kidney injury) and the competing risk of early mortality in patients with extreme SHR values, which was addressed in our sensitivity analyses using the Fine-Gray model.

The steep increase in risk below the SHR threshold of 1.29 in our study underscores a dose-dependent hazard, which likely reflects the cumulative, detrimental effects of acute hyperglycemia on immune dysfunction [8,9] and endothelial activation [34], making patients progressively more susceptible to infection. Several biological mechanisms may explain this association. First, critical illness activates the hypothalamic-pituitary-adrenal axis, leading to stress hyperglycemia [28], which impairs immune responses such as adhesion, chemotaxis, phagocytosis, and bacterial killing by immune cells [8], thus increasing susceptibility to infection. Second, stress hyperglycemia may cause endothelial dysfunction by increasing mitochondrial reactive oxygen species production [9], leading to upregulation of adhesion molecules and enhanced leukocyte recruitment [34], thereby amplifying the inflammatory cascade. Finally, hyperglycemia-driven reactive oxygen species generation can cause myocardial inflammation and dysfunction [35], worsening HF and indirectly elevating sepsis risk.

Our subgroup analyses further revealed that the association between SHR and sepsis was significantly modified by diabetes mellitus, BMI, and insulin use. RCS analysis showed a nonlinear saturation effect in the overall population, which was consistent in the diabetic subgroup. Specifically, in the diabetic subgroup, as SHR increased, the risk of sepsis increased linearly, a finding that aligns with previous studies, demonstrating that the long-term hyperglycemic state of patients with diabetes impairs immune cell function, including adhesion, chemotaxis, phagocytosis, and bacterial killing, thereby increasing susceptibility to sepsis [10,36]. The resulting chronic low-grade

XSL•FO
RenderX

inflammation and vascular endothelial activation, typical in diabetes, also overlap with sepsis pathology [14,29], explaining why patients with HF and diabetes exhibited higher sepsis occurrence in our study.

While there is no definitive research to explain the risk plateau after SHR exceeds 1.29, we hypothesize the following potential factors. First, there is a significant competing risk of early mortality, given that patients with the highest SHR had the greatest 7-day mortality, potentially precluding the diagnosis of sepsis. Second, these patients likely receive more aggressive clinical management (eg, insulin therapy and broader-spectrum antibiotics), which may attenuate the observable sepsis risk. Therefore, an SHR >1.29 might be a stronger indicator of overall critical illness severity and metabolic derangement.

After adjusting for 7-day mortality as a competing event using the Fine-Gray model, SHR remained independently associated with an increased risk of sepsis. The subsequent sensitivity analysis further corroborated this finding, confirming the robustness of our results. Specifically, although the association was nonsignificant in the diabetes subgroup and nondiabetes subgroup when examined separately, the Q4 had the relatively highest 7-day all-cause mortality in the diabetes subgroup, while in the nondiabetes subgroup, patients in Q3 and Q4 groups had higher 7-day all-cause mortality than other groups.

Additionally, our study is the first to explore the mediating role of systemic inflammation in the SHR-sepsis association. We found that NLR, NPR, NMLR, SIRI, and SII significantly mediated this relationship, with mediation proportions ranging from 6% to 13%. This provides a new prospective between the proinflammatory effects of hyperglycemia [21,37-39] and the pathogenesis of sepsis [40], offering a mechanistic insight that has not been previously reported in this specific population. The potential mechanisms might be as follows: SHR underscores an individual's acute glycemic response to stress, and the increased blood glucose further promotes monocyte and macrophage aggregation, leading to the release of additional proinflammatory cytokines [37-39], which contributes to inflammation, tissue damage, and impaired immune function.

Sepsis is a systemic inflammatory response syndrome caused by infection [3], wherein monocytes, neutrophils, and other innate immune cells release proinflammatory cytokines to induce infection [40]. Therefore, comprehensive inflammation indices might have certain mediation value for sepsis onset and prognosis, which is consistent with our findings. Chronic hyperglycemia in patients with diabetes may lead to adaptive changes in insulin signaling and glucose transporter expression, which can attenuate the acute inflammatory response [41]. Moreover, hospitalized individuals with diabetes typically receive stricter glucose management, potentially mitigating the detrimental impact of stress-induced hyperglycemia on sepsis risk and obscuring the role of inflammation mediation [42].

From a clinical perspective, these findings highlight that even moderate elevations in SHR (well below 1.29) might carry increased risk for sepsis, thus suggesting that it could serve as a dynamic and sensitive marker for early risk stratification. The identified threshold of 1.29 could serve as a practical tool for sepsis risk stratification, enabling earlier identification of high-risk patients for intensified monitoring and early treatment intervention based on this cutoff. From a public health perspective, SHR offers a low-cost, accessible biomarker for risk stratification, particularly in resource-limited settings. The finding that inflammatory indices mediate this relationship suggests that anti-inflammatory strategies, alongside strict glycemic control, might be beneficial in mitigating sepsis risk in this vulnerable population.

Limitations

However, limitations exist. First, our retrospective cohort design, based on data from a public database, cannot establish a causal relationship between SHR and sepsis risk. Second, as HbA_{1c} is not routinely measured in all ICU admissions, restricting the cohort to patients with available HbA_{1c} may have introduced selection bias by overrepresenting individuals with known or suspected dysglycemia, thus limiting the generalizability of our findings to the broader population with HF. Third, the relatively small sample size of the diabetes subpopulation may contribute to the nonsignificant mediation effects observed in this subgroup. Fourth, the mediation analysis for the nondiabetic subgroup was underpowered due to the sample size, resulting in unstable estimates with wide CIs. Fifth, given the simultaneous measurement of SHR and inflammatory markers at baseline, the temporal sequence between exposure and mediators cannot be established, which was a significant constraint of the cross-sectional design. Sixth, the unavailability of data on inflammatory biomarkers, including C-reactive protein, procalcitonin, and interleukin-6, limited our ability to perform a comprehensive inflammation mediation analysis. Seventh, as antibiotic use is binary, the reported point-biserial correlation (r=−0.196) between SHR and antibiotic use is methodologically limited and should be interpreted descriptively. Finally, we were unable to account for HF phenotype (HF with reduced ejection fraction vs HF with preserved ejection fraction) due to data limitations.

Future Directions

First, to translate our findings into clinical practice, future studies can commence with validating the prognostic utility of SHR in large, prospective studies. Second, researchers can delve into the underlying mechanisms connecting this metabolic ratio to sepsis susceptibility. Future studies should integrate SHR into a multivariate clinical prediction model to test whether it can effectively guide preventive strategies and improve outcomes in high-risk patients with HF.

Conclusions

Elevated SHR was independently associated with increased 7-day sepsis risk in critically ill patients with HF, and this association was significantly modified by diabetes status. Furthermore, systemic inflammatory indices partially mediated the SHR and sepsis association in the overall cohort, suggesting inflammation as a potential mechanism. SHR may serve as a practical clinical marker for early identification of individuals at high risk of sepsis.

XSL•FO
RenderX

Acknowledgments

The authors are grateful to the participants of the Medical Information Mart for Intensive Care-IV database and to all the research personnel involved. No generative artificial intelligence was used in the preparation of this manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets used during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XC and BY contributed to the manuscript equally. They were responsible for studying conceptualization, data curation, data analysis, and writing the original draft. JC and JY were responsible for formal analysis, accessing and verifying the data, and writing and editing. JS and ST were responsible for the investigation, software, and data curation. BX and BZ were responsible for data curation and methodology. JL was responsible for studying conceptualization and writing reviews or editing. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables and figures. This appendix contains all supplementary material for the retrospective cohort study of 1,205 critically ill patients with heart failure (MIMIC-IV database, 2008–2022), including: diagnostic codes for heart failure (Table S1); formulas for systemic inflammatory indices (Table S2); missing data and imputation methods (Table S3); baseline characteristics stratified by sepsis (Table S4) and diabetes status (Table S5); clinical outcomes by stress hyperglycemia ratio (SHR) quartiles (Table S6); interaction analysis (Table S7); Fine-Gray competing risk models (Table S8); logistic regression sensitivity analysis (Table S9); competing risk analysis after excluding early sepsis diagnoses (Table S10); mediation analyses in diabetic (Table S11) and non-diabetic (Table S12) subgroups; distribution of time-to-sepsis onset (Figure S1); scatter plots of SHR with antibiotic use (Figure S2) and SOFA score (Figure S3); Kaplan-Meier curves for 7-day mortality overall (Figure S4), in diabetic (Figure S5) and non-diabetic (Figure S6) patients; and restricted cubic spline curves for SHR and sepsis risk stratified by diabetes (Figure S7), insulin use (Figure S8), and body mass index (Figure S9).
[DOCX File, 2161 KB - medinform_v14i1e81229_app1.docx]

References

1. Khan MS, Shahid I, Bennis A, Rakisheva A, Metra M, Butler J. Global epidemiology of heart failure. Nat Rev Cardiol 2024 Oct;21(10):717-734. [doi: 10.1038/s41569-024-01046-6] [Medline: 38926611]
2. Chen J, Normand SLT, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998-2008. JAMA 2011 Oct 19;306(15):1669-1678. [doi: 10.1001/jama.2011.1474] [Medline: 22009099]
3. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
4. Walker AMN, Drozd M, Hall M, et al. Prevalence and predictors of sepsis death in patients with chronic heart failure and reduced left ventricular ejection fraction. J Am Heart Assoc 2018 Oct 16;7(20):e009684. [doi: 10.1161/JAHA.118.009684] [Medline: 30371261]
5. Yuzefpolskaya M, Bohn B, Nasiri M, et al. Gut microbiota, endotoxemia, inflammation, and oxidative stress in patients with heart failure, left ventricular assist device, and transplant. J Heart Lung Transplant 2020 Sep;39(9):880-890. [doi: 10.1016/j.healun.2020.02.004] [Medline: 32139154]
6. Arfaras-Melainis A, Polyzogopoulou E, Triposkiadis F, et al. Heart failure and sepsis: practical recommendations for the optimal management. Heart Fail Rev 2020 Mar;25(2):183-194. [doi: 10.1007/s10741-019-09816-y] [Medline: 31227942]
7. Zhu JW, Hu XP, Jin J, et al. Association of pre-exiting heart failure with long-term mortality and the recurrence of sepsis. Sci Rep 2025 Jan 2;15(1):343. [doi: 10.1038/s41598-024-83443-y] [Medline: 39747927]
8. Koh GCKW, Peacock SJ, van der Poll T, Wiersinga WJ. The impact of diabetes on the pathogenesis of sepsis. Eur J Clin Microbiol Infect Dis 2012 Apr;31(4):379-388. [doi: 10.1007/s10096-011-1337-4] [Medline: 21805196]
9. An Y, Xu BT, Wan SR, et al. The role of oxidative stress in diabetes mellitus-induced vascular endothelial dysfunction. Cardiovasc Diabetol 2023 Sep 2;22(1):237. [doi: 10.1186/s12933-023-01965-7] [Medline: 37660030]
10. Shah BR, Hux JE. Quantifying the risk of infectious diseases for people with diabetes. Diabetes Care 2003 Feb;26(2):510-513. [doi: 10.2337/diacare.26.2.510] [Medline: 12547890]

XSL•FO
RenderX

11. Roberts GW, Quinn SJ, Valentine N, et al. Relative hyperglycemia, a marker of critical illness: introducing the stress hyperglycemia ratio. J Clin Endocrinol Metab 2015 Dec;100(12):4490-4497. [doi: 10.1210/jc.2015-2660] [Medline: 26485219]
12. Xu W, Song Q, Wang X, et al. Association of stress hyperglycemia ratio and in-hospital mortality in patients with coronary artery disease: insights from a large cohort study. Cardiovasc Diabetol 2022 Oct 19;21(1):217. [doi: 10.1186/s12933-022-01645-y] [Medline: 36261839]
13. Zhou Q, Yang J, Wang W, Shao C, Hua X, Tang YD. The impact of the stress hyperglycemia ratio on mortality and rehospitalization rate in patients with acute decompensated heart failure and diabetes. Cardiovasc Diabetol 2023 Jul 26;22(1):189. [doi: 10.1186/s12933-023-01908-2] [Medline: 37495967]
14. Yan F, Chen X, Quan X, Wang L, Wei X, Zhu J. Association between the stress hyperglycemia ratio and 28-day all-cause mortality in critically ill patients with sepsis: a retrospective cohort study and predictive model establishment based on machine learning. Cardiovasc Diabetol 2024 May 9;23(1):163. [doi: 10.1186/s12933-024-02265-4] [Medline: 38725059]
15. Li X, Chen Y, Yuan Q, Zhou H, Lu L, Guo R. Neutrophil-to-lymphocyte ratio, monocyte-to-lymphocyte ratio, platelet-to-lymphocyte ratio associated with 28-day all-cause mortality in septic patients with coronary artery disease: a retrospective analysis of MIMIC-IV database. BMC Infect Dis 2024 Jul 29;24(1). [doi: 10.1186/s12879-024-09516-5]
16. Salciccioli JD, Marshall DC, Pimentel MAF, et al. The association between the neutrophil-to-lymphocyte ratio and mortality in critical illness: an observational cohort study. Crit Care 2015 Jan 19;19(1):13. [doi: 10.1186/s13054-014-0731-6] [Medline: 25598149]
17. Zhang Y, Peng W, Zheng X. The prognostic value of the combined neutrophil-to-lymphocyte ratio (NLR) and neutrophil-to-platelet ratio (NPR) in sepsis. Sci Rep 2024 Jul 2;14(1):15075. [doi: 10.1038/s41598-024-64469-8] [Medline: 38956445]
18. Xia Y, Xu H, Xie J, et al. Prognostic value of neutrophil-to-monocyte/lymphocyte ratio for 28-day mortality in ICU sepsis patients: a retrospective cohort study. Front Med (Lausanne) 2024;11:1434922. [doi: 10.3389/fmed.2024.1434922] [Medline: 39211344]
19. Ma K, Zhang Y, Hao J, Zhao J, Qi Y, Liu C. Correlation analysis of systemic immune inflammatory index, Serum IL-35 and HMGB-1 with the severity and prognosis of sepsis. Pak J Med Sci 2023;39(2):497-501. [doi: 10.12669/pjms.39.2.6651] [Medline: 36950402]
20. Zhang Y, Xing Z, Zhou K, Jiang S. The predictive role of systemic inflammation response index (SIRI) in the prognosis of stroke patients. Clin Interv Aging 2021;16:1997-2007. [doi: 10.2147/CIA.S339221] [Medline: 34880606]
21. Berbudi A, Rahmadika N, Tjahjadi AI, Ruslami R. Type 2 diabetes and its impact on the immune system. Curr Diabetes Rev 2020;16(5):442-449. [doi: 10.2174/1573399815666191024085838] [Medline: 31657690]
22. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
23. Rashid R, Sohrabi C, Kerwan A, et al. The STROCSS 2024 guideline: strengthening the reporting of cohort, cross-sectional, and case–control studies in surgery. Int J 2024;110(6):3151-3165. [doi: 10.1097/JS9.0000000000001268]
24. Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2020 Feb;48(2):210-217. [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
25. Mi D, Li Z, Gu H, et al. Stress hyperglycemia is associated with in-hospital mortality in patients with diabetes and acute ischemic stroke. CNS Neurosci Ther 2022 Mar;28(3):372-381. [doi: 10.1111/cns.13764] [Medline: 35084107]
26. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
27. Yang H, Li J, Zhang C, Sierra AP, Shen B. Predictive model for daily risk alerts in sepsis patients in the ICU: visualization and clinical analysis of risk indicators. Precis Clin Med 2025 Mar;8(1):pbaf003. [doi: 10.1093/pcmedi/pbaf003] [Medline: 40041421]
28. Dungan KM, Braithwaite SS, Preiser JC. Stress hyperglycaemia. Lancet 2009 May 23;373(9677):1798-1807. [doi: 10.1016/S0140-6736(09)60553-5] [Medline: 19465235]
29. Li L, Zhou L, Peng X, et al. Association of stress hyperglycemia ratio and mortality in patients with sepsis: results from 13,199 patients. Infection 2024 Oct;52(5):1973-1982. [doi: 10.1007/s15010-024-02264-3] [Medline: 38679664]
30. Zhang S, Shen H, Wang Y, et al. Association between stress hyperglycemia ratio and all-cause mortality in critically ill patients with sepsis: results from the MIMIC-IV database. Eur J Med Res 2025 Jan 21;30(1):42. [doi: 10.1186/s40001-025-02281-4]
31. Liu J, Zhou Y, Huang H, et al. Impact of stress hyperglycemia ratio on mortality in patients with critical acute myocardial infarction: insight from american MIMIC-IV and the chinese CIN-II study. Cardiovasc Diabetol 2023 Oct 21;22(1):281. [doi: 10.1186/s12933-023-02012-1] [Medline: 37865764]
32. Tan MY, Zhang YJ, Zhu SX, Wu S, Zhang P, Gao M. The prognostic significance of stress hyperglycemia ratio in evaluating all-cause and cardiovascular mortality risk among individuals across stages 0-3 of cardiovascular-kidney-metabolic syndrome: evidence from two cohort studies. Cardiovasc Diabetol 2025 Mar 24;24(1):137. [doi: 10.1186/s12933-025-02689-6] [Medline: 40128747]

XSL•FO
RenderX

33. Li L, Ding L, Zheng L, et al. Relationship between stress hyperglycemia ratio and acute kidney injury in patients with congestive heart failure. Cardiovasc Diabetol 2024 Jan 13;23(1):29. [doi: 10.1186/s12933-023-02105-x] [Medline: 38218835]
34. Aird WC. The role of the endothelium in severe sepsis and multiple organ dysfunction syndrome. Blood 2003 May 15;101(10):3765-3777. [doi: 10.1182/blood-2002-06-1887] [Medline: 12543869]
35. Costantino S, Paneni F, Mitchell K, et al. Hyperglycaemia-induced epigenetic changes drive persistent cardiac dysfunction via the adaptor p66Shc. Int J Cardiol 2018 Oct 1;268:179-186. [doi: 10.1016/j.ijcard.2018.04.082] [Medline: 30047409]
36. Trevelin SC, Carlos D, Beretta M, da Silva JS, Cunha FQ. Diabetes mellitus and sepsis: a challenging association. Shock 2017 Mar;47(3):276-287. [doi: 10.1097/SHK.0000000000000778] [Medline: 27787406]
37. Shomali N, Mahmoudi J, Mahmoodpoor A, et al. Harmful effects of high amounts of glucose on the immune system: an updated review. Biotechnol Appl Biochem 2021 Apr;68(2):404-410. [doi: 10.1002/bab.1938] [Medline: 32395846]
38. Mossel DM, Moganti K, Riabov V, et al. Epigenetic regulation of S100A9 and S100A12 expression in monocyte-macrophage system in hyperglycemic conditions. Front Immunol 2020;11:1071. [doi: 10.3389/fimmu.2020.01071] [Medline: 32582175]
39. Song J, Park DW, Moon S, et al. Diagnostic and prognostic value of interleukin-6, pentraxin 3, and procalcitonin levels among sepsis and septic shock patients: a prospective controlled study according to the Sepsis-3 definitions. BMC Infect Dis 2019 Nov 12;19(1):968. [doi: 10.1186/s12879-019-4618-7] [Medline: 31718563]
40. Deng C, Zhao L, Yang Z, et al. Targeting HMGB1 for the treatment of sepsis and sepsis-induced organ injury. Acta Pharmacol Sin 2022 Mar;43(3):520-528. [doi: 10.1038/s41401-021-00676-7] [Medline: 34040166]
41. Flynn MC, Kraakman MJ, Tikellis C, et al. Transient intermittent hyperglycemia accelerates atherosclerosis by promoting myelopoiesis. Circ Res 2020 Sep 11;127(7):877-892. [doi: 10.1161/CIRCRESAHA.120.316653] [Medline: 32564710]
42. Friol A, Devautour C, Semenov A, et al. Do diabetes and poor control of acute stress-related hyperglycemia increase the risk of ICU-acquired infections? A retrospective assessment in patients with septic shock. Ann Intensive Care 2025 Oct 22;15(1):168. [doi: 10.1186/s13613-025-01596-z] [Medline: 41123877]

Abbreviations

HbA_{1c}: glycated hemoglobin
HF: heart failure
HIPAA: Health Insurance Portability and Accountability Act
HR: hazard ratio
ICD-10: International Classification of Diseases, Tenth Revision
ICD-9: International Classification of Diseases, Ninth Revision
ICU: intensive care unit
MIMIC-IV: Medical Information Mart for Intensive Care-IV
NLR: neutrophil-to-lymphocyte ratio
NMLR: neutrophil-monocyte-to-lymphocyte ratio
NPR: neutrophil-to-platelet ratio
RCS: restricted cubic splines
SHR: stress hyperglycemia ratio
SII: systemic immune-inflammation index
SIRI: systemic inflammation response index
SOFA: Sequential Organ Failure Assessment
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Coristine; submitted 06.Aug.2025; peer-reviewed by K Pang, T Wan; revised version received 28.Feb.2026; accepted 28.Feb.2026; published 17.Apr.2026.

Please cite as:
Cai X, Yan B, Chen J, Ye J, Su J, Tong S, Xie B, Zhang B, Li J
Stress Hyperglycemia Ratio and the Risk of Sepsis in Patients With Heart Failure: Retrospective Cohort Study From Medical Information Mart for Intensive Care-IV
JMIR Med Inform 2026;14:e81229
URL: https://medinform.jmir.org/2026/1/e81229
doi:10.2196/81229

© Xianzhen Cai, Baoxin Yan, Jinhao Chen, Junjun Ye, Jiating Su, Shuangshuang Tong, Bin Xie, Bingna Zhang, Jilin Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits

XSL•FO
RenderX

unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Knowledge-Guided Explainable Recommendation Tool for Cancer Risk Prediction Models Using Retrieval-Augmented Large Language Models: Development and Validation Study

Shumin Ren[1,2*], PhD; Xin Zheng[1*], MSc; Jing Zhao[1], MSc; Jiale Du[1], MSc; Yuxin Zhang[1], MSc; Cheng Bi[1], MSc; Jie Song[1], MSc; Jinyi Zhang[2], MSc; Hongmei Lang[2], MD; Fan Zhang[3], MD; Bairong Shen[1], Prof Dr

[1]Institutes for Systems Genetics, West China Hospital of Sichuan University, Frontiers Science Center for Disease-related Molecular Network, Chengdu, Sichuan, China

[2]West China School of Medicine, Sichuan University, Sichuan University affiliated Chengdu Second People's Hospital, Chengdu Second People's Hospital, Chengdu, Sichuan, China

[3]Health Management Center, General Practice Medical Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Bairong Shen, Prof Dr
Institutes for Systems Genetics, West China Hospital of Sichuan University, Frontiers Science Center for Disease-related Molecular Network, Chengdu, Sichuan, China

Abstract

Background: Cancer risk prediction models are vital for precision prevention, enabling individualized assessment of cancer susceptibility based on genetic, clinical, environmental, and lifestyle factors. However, the practical use of these models is hindered by fragmented resources, heterogeneous reporting, and the absence of transparent, structured systems for systematic discovery and comparison.

Objective: This study aimed to develop a retrieval-augmented, knowledge-guided system that provides accurate recommendations for cancer risk prediction models.

Methods: We developed CanRisk-RAG, a recommendation platform underpinned by a precisely constructed knowledge base comprising more than 800 peer-reviewed cancer risk prediction models spanning diverse cancer types, modeling approaches, and predictive variables. The system integrates (1) large language model (LLM)–based semantic tag extraction, (2) embedding vectorization of structured metadata and abstracts, (3) a multifactor ranking algorithm combining semantic similarity with multiple quality indicators, and (4) LLM-generated literature summarization to support rapid user interpretation. Performance was evaluated across 4 types of representative queries. Eight domain experts independently assessed retrieval quality. CanRisk-RAG was benchmarked against PubMed, ChatGPT-4o, ScholarAI, and Gemini 1.5 Flash.

Results: On the independent validation set, CanRisk-RAG consistently outperformed all 4 baseline applications, achieving the highest overall relevance (8.30 [SD 0.59]) and reliability (7.62 [SD 0.76]) scores on a 10-point scale (P<.05). It also demonstrated high authenticity, data completeness, and consistency. Baseline applications frequently returned incomplete, inconsistent, or fabricated results, especially for complex, multifactorial queries, whereas CanRisk-RAG delivered accurate and structured recommendations grounded in validated evidence.

Conclusions: CanRisk-RAG presents a transparent, domain-specific, and semantically enriched framework for discovering cancer risk prediction models, addressing several limitations of existing keyword-based search tools and general-purpose LLMs. By integrating structured knowledge, multifactor ranking, and LLM-based reasoning, the system aims to improve the precision, reproducibility, and usability of model selection in cancer risk prediction. While our evaluation demonstrates encouraging performance compared with baseline systems, further validation in broader clinical contexts and real-world applications is warranted. The framework's general design may also be adaptable to other clinical model domains, providing a potential foundation for advancing evidence-based model discovery in precision medicine.

(JMIR Med Inform 2026;14:e78519) doi:10.2196/78519

KEYWORDS

cancer risk prediction models; large language models; LLMs; retrieval-augmented generation; RAG; personalized medicine; biomedical information retrieval

XSL•FO
RenderX

Introduction

Cancer risk prediction models are increasingly recognized as essential components of precision medicine, offering clinicians and researchers the ability to estimate an individual's likelihood of developing cancer based on diverse risk factors. These models integrate genetic markers, clinical data, environmental exposures, and lifestyle behaviors, thereby enabling more personalized approaches to prevention and screening [1-4]. These heterogeneous factors are not independent but often interact in complex and nonlinear ways. For example, genetic susceptibility may amplify the carcinogenic effects of smoking or obesity, while lifestyle interventions can partially mitigate inherited risk [5,6]. Understanding and modeling these interdependencies is therefore crucial for accurate cancer risk prediction.

Statistical approaches such as logistic regression, Cox proportional hazards models, and Bayesian networks have traditionally been used to analyze correlations and interactions among these predictors. More recently, advanced machine learning techniques, including random forests, gradient boosting, and deep neural networks, have enabled the modeling of higher-order, nonlinear, and cross-domain relationships between genetic, clinical, and environmental factors [7-10]. These methods can capture subtle patterns that traditional regression frameworks may overlook, thus improving predictive accuracy and interpretability when properly validated.

By identifying individuals at elevated risk, cancer risk prediction models facilitate earlier interventions, targeted monitoring, and ultimately reductions in cancer incidence and mortality [11,12]. Recent advances have demonstrated that combining heterogeneous risk determinants can substantially improve predictive accuracy. For instance, multivariate models that combine germline mutations (eg, BRCA1/2), hormonal and reproductive history, and lifestyle exposures have achieved strong discriminative power in breast cancer risk prediction [13,14], while integrative models linking smoking intensity, air pollution exposure, and genomic markers have improved lung cancer risk stratification [15,16]. Together, these developments highlight both the promise and the complexity of synthesizing genetic, clinical, environmental, and behavioral information into actionable tools for precision prevention.

Despite the promise of cancer risk prediction models, their practical impact is limited by barriers related to accessibility and adoption. A fundamental problem lies in the fragmentation of available models, which vary widely in methodological approach, target population, and predictive variables. Clinicians and researchers often struggle to identify the most appropriate model for a given use case, a challenge compounded by the lack of centralized, structured resources for systematic discovery. General medical retrieval systems are ill-suited to this task, as they are designed primarily to locate literature rather than extract or recommend specialized prediction tools. As a result, existing models are underutilized, research efforts are frequently duplicated, and the translational pathway from model development to clinical implementation remains inefficient. Indeed, it has been observed that most published prediction models are never adopted in clinical practice, highlighting a persistent gap between academic research and implementation [17]. Reviews further emphasize that although many models have been developed, only a small fraction have been successfully translated into clinical or public health practice, largely due to limitations in validation, usability, and acceptability [18,19].

To date, most users seeking cancer risk prediction models have relied on mainstream biomedical databases such as PubMed [20]. While PubMed provides broad access to scientific publications, its keyword-based search mechanisms often generate an overwhelming number of results, many of which are irrelevant for model identification. Moreover, critical details regarding model structure, predictive variables, validation cohorts, or performance metrics are typically buried in unstructured text, making systematic retrieval difficult.

In parallel, institutional platforms have been developed to provide more direct access to specific risk prediction tools. The US National Cancer Institute's Division of Cancer Epidemiology and Genetics offers a set of statistical macros and online calculators covering selected cancer types [21], while the Memorial Sloan Kettering Cancer Center provides nomograms tailored to particular clinical scenarios [22]. Although these resources represent valuable contributions, their scope remains restricted to only several (typically fewer than a dozen) tools, employ relatively narrow modeling methodologies, and lack detailed metadata regarding model characteristics or generalizability.

Large language models (LLMs) have recently been explored as potential solutions to improve medical information access. Their ability to integrate knowledge and process complex clinical text holds significant promise for bridging gaps between scattered information and practical decision-making [23-26]. However, empirical studies highlight critical weaknesses in LLMs, including limited precision, the generation of fabricated or irrelevant outputs, and the opacity of reasoning processes [23,27,28]. Retrieval-augmented generation (RAG) has been proposed as a mechanism to mitigate these shortcomings by grounding model outputs in external knowledge sources [29,30]. Several initiatives illustrate the potential of this approach. RefAI [31], for instance, combines PubMed retrieval with a GPT-4–based summarization system, enabling more effective access to recent biomedical publications. Similarly, WeiseEule [32] integrates keyword-driven retrieval from PubMed with LLM-based responses to improve query precision. However, these systems largely depend on user-defined search terms and ad hoc knowledge construction, leading to inconsistent coverage and limited reproducibility. Moreover, they are not designed to handle structured, model-specific metadata, such as predictive variables, validation cohorts, or performance metrics, that are critical for evidence-based model selection.

To address these challenges, the present study introduces CanRisk-RAG, a knowledge-guided and LLM-enhanced recommendation system designed to systematically retrieve, rank, and summarize cancer risk prediction models. Built on a knowledge base that we systematically constructed to consolidate validated cancer risk prediction models across

XSL•FO
RenderX

multiple cancer types, predictive variables, and methodological frameworks, CanRisk-RAG provides a structured foundation for model discovery and reuse. By combining domain-curated knowledge with LLM-driven semantic reasoning, the system bridges the gap between unstructured user queries and structured model metadata. Furthermore, a multi-factor ranking algorithm integrates semantic similarity with evidence-based indicators, such as publication quality, model performance, and timeliness, ensuring that recommended models are both relevant and scientifically robust. To enhance interpretability and usability, CanRisk-RAG also generates concise LLM-based summaries that highlight key model characteristics.

In essence, CanRisk-RAG aims to provide a scalable, transparent, and domain-aware retrieval framework that supports both research and clinical decision-making in cancer risk assessment. By facilitating the efficient reuse of existing models and improving accessibility for diverse user needs, including population-level screening and high-risk cohorts defined by genetic predisposition or family history, it represents a step toward operationalizing precision prevention in oncology.

This paper is organized as follows: We first introduce the data sources and system development workflow. Next, we describe the design of the evaluation tasks and the specific evaluation metrics used. We then present the performance results of CanRisk-RAG in terms of model recommendation and tool usability, along with comparative analyses against four baseline applications. Finally, we discuss the implications of the findings and explore potential directions for extending the system to other cancer research or medical domains.

Methods

Overview

The Methods section outlines the construction of CanRisk-RAG, covering data preparation, semantic vectorization, tag extraction via LLM, hybrid ranking with multiple quality indicators, literature summarization, and comparative evaluation. For more implementation details and technical specifications, please refer to Supplemental Information 1.1 in Multimedia Appendix 1.

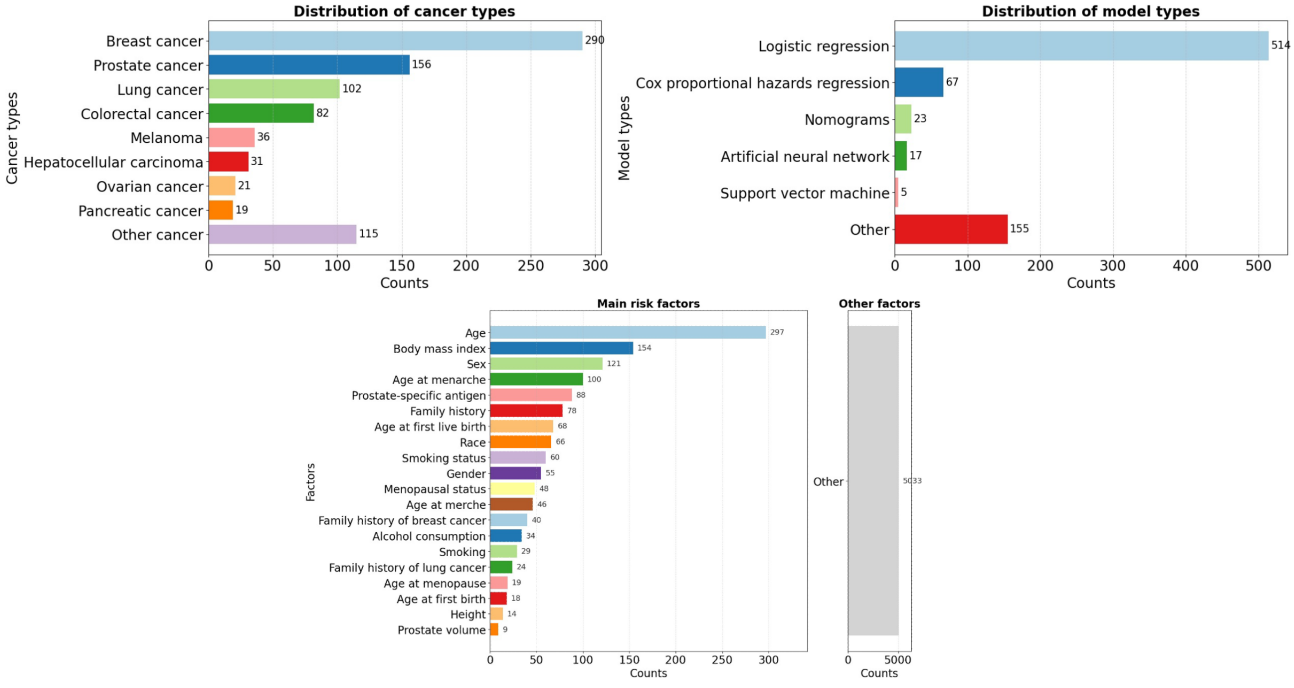
Dataset

Data Preparation

We constructed a knowledge base containing peer-reviewed cancer risk prediction models, consolidating clinically developed and validated cancer risk prediction models extracted from PubMed. The detailed search criteria and the inclusion and exclusion standards used for model collection are provided in Supplemental Information 1.2 and 1.3 in Multimedia Appendix 1.

Following this procedure, a total of 808 cancer risk prediction models were included. Each model is uniquely identified by a model ID (MID). Each model was comprehensively described using metadata covering 3 major categories, namely basic information, sample information, and model information, across a total of 36 fields. To facilitate structured storage and interpretation, an Entity–Relationship (ER) diagram was constructed (see Supplemental Information 1.4 in Multimedia Appendix 1). In addition, as illustrated in Figure 1, the knowledge base organizes data across more than 10 cancer types (see Supplemental Information 2.1 in Multimedia Appendix 2 for the full distribution), over 13 methodological frameworks (eg, logistic regression, Cox regression, nomograms), and 3095 predictive factors, including age, body mass index, sex, and family history.

Figure 1. Distribution of cancer types, model types, and predictive factors in the CanRisk-RAG.



XSL•FO
RenderX

Data Segmentation and Vectorization

To enable semantic retrieval, data from the knowledge base were transformed into structured vector representations. Key attributes of each model, including cancer type, methodological approach, research stage, predictive factors, and geographic context were integrated to form composite descriptors that reflect the essential semantic characteristics of the models. These descriptors were then encoded using advanced sentence embedding techniques to generate dense vector representations, ensuring that key semantic meaning was preserved.

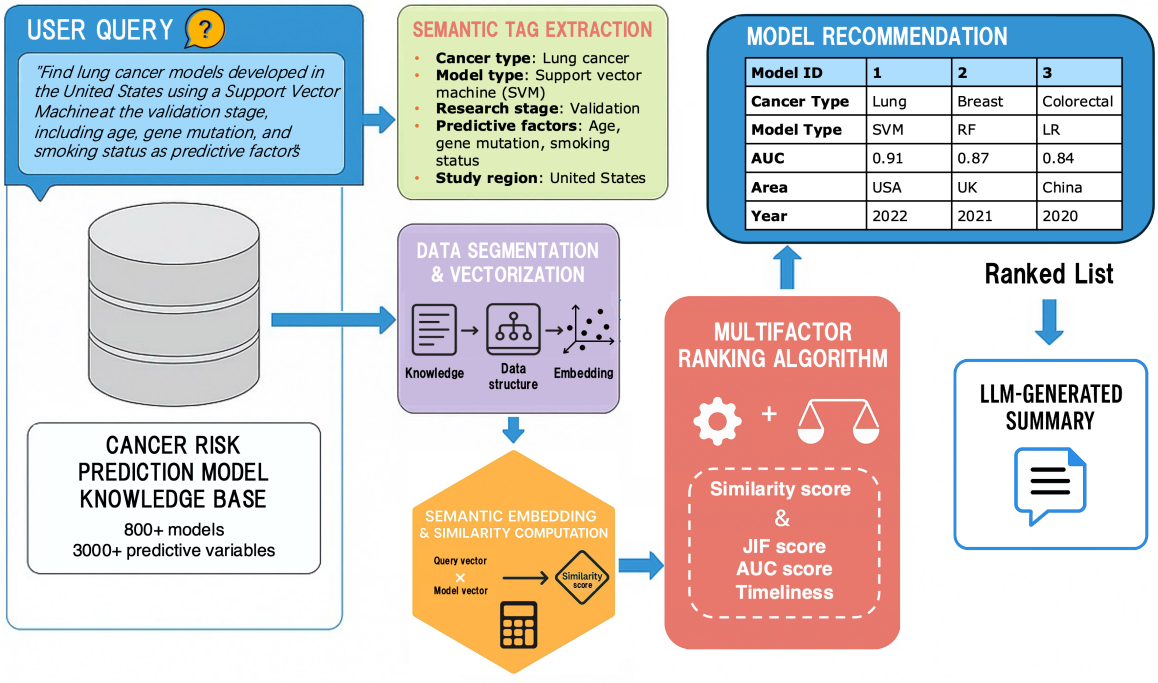
In parallel, the textual abstracts of the model research were independently embedded to capture complementary contextual details not fully represented by structured metadata. The combination of structured semantic vectors and unstructured textual embeddings provided a comprehensive representation of each model (see Supplemental Information 1.1 in Multimedia Appendix 1 for details). This dual encoding strategy established the basis for subsequent similarity search and model ranking, allowing the system to align user queries with the most relevant prediction models in a context-aware and interpretable manner.

Workflow of CanRisk-RAG

System Architecture

The overall workflow of CanRisk-RAG comprises a multistage pipeline that transforms unstructured model resources into structured model recommendations (Figure 2). The system integrates several key components: (1) semantic tag extraction, in which user queries are parsed by a LLM into predefined biomedical fields; (2) semantic embedding and vectorization, which encode both structured metadata and unstructured abstracts into dense representations for similarity computation; (3) multifactor ranking, combining semantic similarity with quality indicators such as the journal impact factor (JIF), the area under the curve (AUC), and publication recency to ensure reliable prioritization; and (4) LLM-based summarization, which synthesizes retrieved content into concise and readable outputs. Together, these stages enable CanRisk-RAG to perform context-aware retrieval, interpret complex user intents, and deliver transparent, evidence-based model recommendations supported by structured domain knowledge.

Figure 2. Pipeline of CanRisk-RAG construction. This figure presents the full pipeline of CanRisk-RAG, a retrieval-augmented framework designed to identify and summarize validated cancer risk prediction models from peer-reviewed resources. The workflow proceeds from user query interpretation to model recommendation: (1) users enter a natural language query describing desired model features; (2) a large language model (LLM)–based semantic tag extractor converts the query into structured fields (eg, cancer type, model type, predictive factors, study region); (3) the query is matched against a knowledge base constructed from PubMed containing over 800 cancer risk models and 3000 predictive variables; (4) structured metadata and text are vectorized using optimized sentence embeddings; (5) semantic similarity between query and models is computed; (6) a multifactor ranking algorithm integrates similarity, journal impact factor (JIF), model area under the curve (AUC), and publication recency into a composite score; and (7) top-ranked models are presented with metadata and an LLM-generated summary.



Tag Extraction for Enhanced Semantic Matching

To support efficient retrieval, user queries were first transformed from free text into structured semantic representations. This process was facilitated by a LLM (DeepSeek-V3) [33], chosen for its open-source nature and ability to support scalable deployment. Rather than directly matching raw text against the database, the model extracted predefined semantic fields, including cancer type, model type, research stage, predictive factors, and study region, thereby aligning user intent with structured knowledge (see Supplemental Information 1.5 in Multimedia Appendix 1 for the prompt contents).

XSL•FO
RenderX

An evaluation set of 40 queries (Supplemental Information 1.6 in Multimedia Appendix 1) was constructed to evaluate the tag extraction process. The queries were designed to cover diverse cancer types, model types, predictive factor combinations, geographic contexts, and noisy inputs (eg, typos and incomplete descriptions). Two independent annotators created gold labels following the predefined schema (Cancer_Type, Model_Type, Research_Type, Involved_Factors, Country), and disagreements were adjudicated by a third expert. Precision, recall, and F_1-score were calculated per category and additionally provided uncertainty estimates using Wilson 95% CIs for precision (Supplemental Information 1.7 in Multimedia Appendix 1).

Embedding Model Selection

To prioritize relevant models, CanRisk-RAG employs a ranking strategy primarily driven by semantic similarity. User queries are semantically aligned with database entries through embedding-based similarity, which captures both structured tags and contextual meaning. The similarity score serves as the main determinant, reflecting the degree of correspondence between user intent and model characteristics.

Before performing multifactor ranking, to determine which embedding model achieved the best semantic representation performance, we evaluated four widely used sentence-embedding architectures [34-36], including all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2, and all-distilroberta-v1. Each model's output was assessed by 2 independent evaluators across 12 predefined recommendation queries derived from 4 thematic categories (Textbox 1). For each model, average relevance and reliability scores were calculated. All-distilroberta-v1 achieved the highest overall performance, with statistically significant advantages in both relevance and reliability metrics (see Supplemental Information 2.2 in Multimedia Appendix 2). Therefore, all-distilroberta-v1 was selected as the optimal embedding model for subsequent tool development and deployment.

Textbox 1. Evaluation themes and representative queries used for CanRisk-RAG development.

Model recommendation tasks based on cancer type

- Breast cancer
- Lung cancer
- Prostate cancer

Model recommendation tasks based on model type

- Logistic regression
- Support vector machine
- Neural network

Model recommendation tasks based on predictive factors

- Age, gene mutation, and family history
- PSA level, digital rectal examination, and prior biopsy
- Sex, smoking status, emphysema, and pulmonary nodule

Comprehensive model recommendation tasks

- Gastric cancer; logistic regression; predictive factors: age, family history, and BMI
- Colorectal cancer; Cox proportional hazards regression; predictive factors: prior biopsy, alcohol intake, and height
- Bladder cancer; nomogram models; predictive factors: race, gene mutation, smoking status, and education

Multifactor Ranking Algorithm

To enhance credibility and practical utility, 3 additional factors were integrated into the ranking: the JIF as an indicator of publication quality [31], the AUC as a measure of model performance, and publication recency to emphasize timely contributions. The AUC was treated as a continuous metric using a smoothed compression function, while JIF values were normalized through a log-smoothing transformation to mitigate right-skewed distributions [37]. The publication year was linearly normalized to yield a continuous timeliness score. These components were then combined into a unified weighted scoring function (Equation 1).

(1)Total score=α×similarity score+β×journal impact factor score+γ×AUC score+δ×timeliness score

LLM-Based Literature Summarization

To support user interpretation, CanRisk-RAG incorporates an automated summarization component that synthesizes the content of retrieved studies into concise, coherent overviews. By integrating information across multiple abstracts, the system highlights key findings and methodological features, enabling users to rapidly grasp the essential contributions of relevant models.

XSL•FO
RenderX

Assessment Experiments

Benchmarking and Evaluation Design

To optimize the performance of CanRisk-RAG, we conducted a comprehensive comparative evaluation using 12 representative queries spanning 4 thematic areas, including cancer types, modeling approaches, predictive factors, and integrated multicomponent queries. Each theme was further divided into subthemes to reflect major research directions in cancer risk prediction [38-47] (see Textbox 1). Eight domain experts, including 3 clinical cancer researchers and 5 biomedical informatics specialists, independently assessed the outputs for all themes, with each query evaluated by 2 experts. In parallel, multiple ablation experiments on the training dataset (Table 1) examining alternative weighting schemes enabled the optimization of the final ranking configuration, ensuring balanced and robust system performance.

Table . Weight configurations for multifactor ranking algorithm used in ablation experiments.

Group	α (Similarity)	β (JIF)[a]	γ (AUC)[b]	δ (Timeliness)
G1 (Proposed)	0.85	0.05	0.05	0.05
G2 (Similarity only)	1	0	0	0
G3 (AUC+)	0.7	0.05	0.2	0.05
G4 (Recency+)	0.7	0.05	0.05	0.2
G5 (JIF+)	0.7	0.2	0.05	0.05
G6 (Similarity + Recency)	0.85	0	0	0.15
G7 (Similarity + AUC)	0.85	0	0.15	0
G8 (Similarity + JIF)	0.85	0.15	0	0
G9 (Similarity + Recency + AUC)	0.8	0	0.1	0.1
G10 (Similarity + Recency + JIF)	0.8	0.1	0	0.1
G11 (Similarity + AUC + JIF)	0.8	0.1	0.1	0

[a]JIF: journal impact factor.

[b]AUC: area under the curve.

At the application level, CanRisk-RAG was benchmarked against PubMed, ChatGPT-4o [48], ScholarAI [49], and Gemini 1.5 Flash [50], each accessed through their respective application programming interfaces under standardized query conditions. Prompts were standardized across systems to maintain fairness in comparison (see Supplemental Information 1.8 in Multimedia Appendix 1).

This stage was evaluated by domain experts. Performance was scored on 5 key dimensions: relevance, reliability, authenticity, data completeness, and consistency (0 - 10 scale). Table 2 summarizes the evaluation types, metrics, and definitions used to assess the 5 applications, including CanRisk-RAG (see Supplemental Information 1.9 in Multimedia Appendix 1 for the scoring criteria of the 5 evaluation metrics).

XSL•FO
RenderX

Table . Evaluation metrics and definitions used for model recommendation and tool usability assessment.

Evaluation metrics type and metric names		Definition
Model recommendation performance evaluation		
	Relevance	The purpose of the relevance assessment is to determine the extent to which the recommended models match the user's specified criteria.
	Reliability	The reliability assessment examines the authority of the sources where the models were published and evaluates the performance of the models themselves.
	Authenticity	Authenticity checks whether the sources of the models are genuine and can be verified in various journal databases.
	Data completeness	Data completeness evaluates the thoroughness of the model information provided by each application, including metadata, cancer type, modeling type, model performance, predictive factors, and more.
	Consistency	Consistency evaluation measures how consistently the model recommendation results align when the same search conditions or prompts are used in 2 separate instances.
Tool usability evaluation		
	User goal achievement	This metric evaluates whether the tool meets users' specific needs and expectations under open-ended, self-defined query scenarios. Participants independently formulated queries based on their own research or clinical information needs. The score reflects the extent to which users were able to successfully locate relevant cancer risk prediction models that satisfied their personalized model recommendation requirements.
	Operative accessibility	Operative accessibility refers to the overall ease of using the tool and encompasses aspects such as deployment complexity, interface design, navigation intuitiveness, and loading speed. A user-friendly tool should minimize operational complexity and cognitive load while offering clearly structured content, intuitive workflows, and comprehensive help pages.

Validation and Usability

To prevent data leakage between hyperparameter tuning and final performance assessment, we strictly separated the datasets used for weight optimization and for comparative evaluation. Weight optimization and ablation analyses were conducted solely on the development dataset, whereas the final benchmarking against baseline applications (PubMed, ChatGPT-4o, ScholarAI, and Gemini 1.5 Flash) was performed on an independent, held-out evaluation set. This evaluation set consisted of 20 carefully designed queries (Supplemental Information 1.10 in Multimedia Appendix 1) covering previously unseen cancer types, modeling approaches, and multifactorial scenarios that jointly specified cancer type, model type, and predictive factors. Recommendation outputs were independently assessed by eight domain experts using the predefined five evaluation dimensions.

In addition, usability was assessed by 10 potential end users, including clinicians and biomedical informatics researchers, based on 2 criteria: goal achievement and operative accessibility (see Supplemental Information 1.11 in Multimedia Appendix 1 for the scoring criteria). Inclusion criteria of potential end users required participants to be able to independently complete retrieval tasks. To mitigate potential conflict-of-interest bias, individuals involved in system development or dataset curation were excluded from usability scoring. None of the evaluators were members of the development team.

This multiperspective evaluation enabled a systematic assessment of CanRisk-RAG's retrieval quality and practical applicability in comparison to existing systems. Table 2 summarizes the evaluation types, metrics, and definitions used for the assessment of CanRisk-RAG.

XSL•FO
RenderX

Statistical Analysis

The evaluation was conducted in 2 phases: first, the embedding models were assessed, then the 5 applications were evaluated. Descriptive statistics on the scores were given by the evaluators, including the mean and SD. To determine the consistency among evaluators for every metric assessed, we conducted a 2-way intraclass correlation coefficient analysis.

For the evaluation of different applications, the Shapiro-Wilk test was employed to determine whether the scores assigned by different evaluators for a specific metric followed a normal distribution. Due to the departure of the scores from normality, the Mann-Whitney *U* test was selected to compare the overall performance scores. To explore variations in performance across the five applications and various subthemes, the Kruskal-Wallis rank-sum test was employed. When significant differences were identified by the Kruskal-Wallis test, Dunn post hoc multiple comparisons test was used to evaluate the performance of the different baselines in relation to CanRisk-RAG. Cliff delta was reported for all pairwise comparisons. Multiplicity control across post hoc tests was handled using Benjamini–Hochberg adjustments (α=.05). A significance level of P<.05 was applied to all statistical tests. All statistical tests were performed using Python (version 3.12).

Ethical Considerations

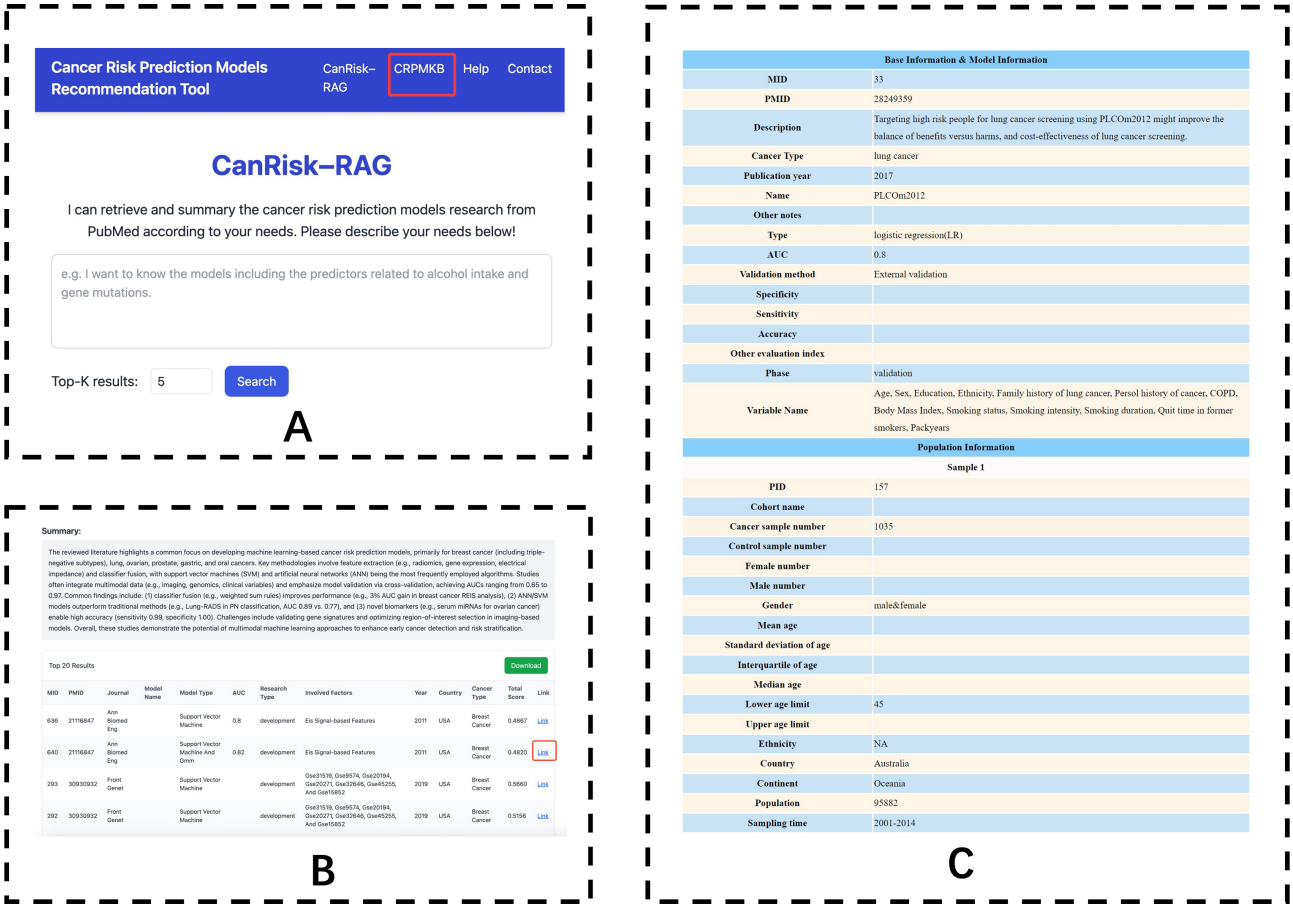
This study did not involve human or animal subjects; therefore, an ethics review board assessment was not required in accordance with institutional and national guidelines of Sichuan University. All analyses were conducted on publicly available, nonidentifiable data, consistent with regional research ethics policies [51].

Results

Tool Interface

As shown in Figure 3, the homepage of CanRisk-RAG features an interactive input form (Figure 3A) that allows users to specify any retrieval criteria related to cancer risk prediction models. Users may define the desired number of recommended results through a customizable selection option. Additionally, users can directly access the data source associated with the homepage by clicking the “CRPMKB” link in the main navigation bar. After entering the desired model criteria in the input field and clicking the “Search” button, the system displays the results page (Figure 3B), which presents a list of recommended models retrieved based on the user’s specified conditions. In the table result, each recommended model is presented with detailed information, including attributes such as MID, PMID, Journal, Model Name, Model Type, AUC, Research Type, Involved Factors, Year, Country, Cancer Type, Total Score. Each model entry is hyperlinked to its corresponding detailed record (Figure 3C). Furthermore, for all recommended models, the tool automatically generates a summary of the corresponding research abstract (Figure 3B), providing a concise overview of these studies to support rapid understanding and comparison.

Figure 3. CanRisk-RAG tool interfaces. (A) Homepage query form. (B) Recommendation results page. (C) Detailed model information page.



XSL•FO
RenderX

Model Recommendation Evaluation Results

Weight Configuration

We systematically evaluated eleven alternative weighting configurations (G1-G11) that varied the relative contributions of similarity (α), JIF (β), model AUC (γ), and publication timeliness (δ) (Table 3). Each configuration was tested on the same dataset of representative queries, and the average scores for relevance, reliability, and total performance were calculated.

A nonparametric Kruskal–Wallis *H* test revealed a statistically significant overall difference among the eleven configurations (H=47.916, P<6.430 × 10 , see Supplemental Information 2.3 in Multimedia Appendix 2).

The proposed configuration (G1: α=.85, β=.05, γ=0.05, δ=0.05) achieved the highest overall score, significantly outperforming variants such as SimOnly and Recency+ (P_{Holm}<.01). G1 consistently yielded higher scores across all configurations, demonstrating a robust performance on relevance and reliability.

Table . Performance comparison of different weighting configurations (mean [SD] for relevance, reliability, and total scores).

Group	Avg_Relevance, mean (SD)	Avg_Reliability, mean (D)	Avg_Total, mean (SD)
G1 (Proposed)	9.50 (0.78)	8.88 (0.85)	18.38 (1.44)
G8 (Similarity + JIF[a])	8.50 (0.66)	9.21 (0.78)	17.71 (1.16)
G7 (Similarity + AUC[b])	8.92 (0.97)	8.75 (0.85)	17.67 (1.55)
G6 (Similarity + Recency)	9.25 (0.74)	8.04 (0.81)	17.29 (1.12)
G3 (AUC+)	7.96 (0.96)	9.25 (0.90)	17.21 (1.59)
G5 (JIF+)	8.17 (0.92)	8.67 (0.96)	16.83 (1.37)
G10 (Similarity + Recency + JIF)	8.08 (0.78)	8.25 (0.79)	16.33 (0.87)
G9 (Similarity + Recency + AUC)	8.17 (1.05)	8.08 (0.93)	16.25 (1.51)
G11 (Similarity + AUC + JIF)	7.83 (0.64)	7.92 (0.72)	15.75 (0.79)
G2 (Similarity only)	9.42 (0.65)	5.46 (0.72)	14.88 (1.08)
G4 (Recency+)	7.33 (0.96)	7.08 (0.72)	14.42 (1.06)

[a]JIF: journal impact factor.

[b]AUC: area under the curve.

Performance Comparison of Applications

In the comparative evaluation across 5 applications, CanRisk-RAG consistently achieved the highest relevance and reliability across all 4 themes of the validation set (Table 4, see Supplemental Information 2.4 and 2.5 in Multimedia Appendix 2 for detailed statistics), delivering markedly superior retrieval precision compared with existing applications. For relevance, the average score was 8.30 (SD 0.59; 95% CI 8.02-8.58), indicating that the retrieved models closely matched the intended cancer types, modeling approaches, and predictive factors specified in the queries. In terms of reliability, the average score reached 7.62 (SD 0.76; 95% CI 7.27-7.98), reflecting strong consistency with authoritative and well-validated literature sources. These results confirm that, when applied to independent validation themes, CanRisk-RAG maintains robust generalization and stable interpretability, effectively extending to unseen data while preserving retrieval accuracy and quality consistency, thereby underscoring its scalability and real-world applicability. PubMed showed reasonable accuracy for straightforward cancer-type queries but struggled when tasks involved modeling approaches or specific predictive factors, often returning incomplete or no usable results. Scholar AI performed moderately in simpler recommendation tasks, but its accuracy dropped sharply for more complex, integrative queries. ChatGPT-4o occasionally retrieved relevant cancer-type models but frequently missed methodological or variable-level details. Gemini 1.5 Flash repeatedly failed to interpret user intent or generate sufficient recommendations, underscoring its limited capability for structured retrieval.

XSL•FO
RenderX

Table . Comparative evaluation of relevance and reliability on the test set across 5 applications.

Applications	Relevance, mean (SD)	P_{bh} [a]	Reliability, mean (SD)	P_{bh}
Topic 1				
CanRisk-RAG	8.70 (0.45)	—[b]	8.10 (0.55)	—
ChatGPT-4o	5.90 (1.24)	<.05	3.50 (0.35)	<.05
Gemini 1.5 Flash	2.40 (0.42)	<.05	1.80 (0.57)	<.05
PubMed	5.70 (0.97)	<.05	7.20 (1.04)	ns[c]
ScholarAI	7.90 (1.02)	ns	6.60 (0.96)	ns
Topic 2				
CanRisk-RAG	8.50 (0.50)	—	7.80 (0.76)	—
ChatGPT-4o	1.30 (0.84)	<.05	1.00 (0.35)	<.05
Gemini 1.5 Flash	1.20 (0.45)	<.05	1.80 (0.27)	<.05
PubMed	2.40 (1.14)	<.05	6.50 (1.00)	ns
ScholarAI	6.10 (0.74)	<.05	5.20 (1.30)	<.05
Topic 3				
CanRisk-RAG	8.40 (0.42)	—	7.20 (0.76)	—
ChatGPT-4o	1.80 (0.27)	<.05	1.10 (0.42)	<.05
Gemini 1.5 Flash	1.10 (0.42)	<.05	0.90 (0.55)	<.05
PubMed	1.20 (0.57)	<.05	5.60 (2.90)	ns
ScholarAI	5.10 (0.89)	<.05	4.70 (1.82)	ns
Topic 4				
CanRisk-RAG	7.60 (0.42)	—	7.40 (0.82)	—
ChatGPT-4o	2.20 (0.57)	<.05	1.10 (0.55)	<.05
Gemini 1.5 Flash	1.90 (0.22)	<.05	1.70 (0.57)	<.05
PubMed	0.10 (0.22)	<.05	0.40 (0.42)	<.05
ScholarAI	1.80 (1.15)	<.05	1.20 (1.04)	<.05
Average score				
CanRisk-RAG	8.30 (0.59)	—	7.62 (0.76)	—
ChatGPT-4o	2.80 (2.01)	<.05	1.68 (1.15)	<.05
Gemini 1.5 Flash	1.65 (0.65)	<.05	1.55 (0.60)	<.05
PubMed	2.35 (2.28)	<.05	4.92 (3.13)	<.05
ScholarAI	5.22 (2.44)	<.05	4.42 (2.37)	<.05

[a]P_{bh}: Benjamini–Hochberg adjusted significance levels, representing statistical comparisons against the corresponding CanRisk-RAG scores within the same topic or in the overall average.

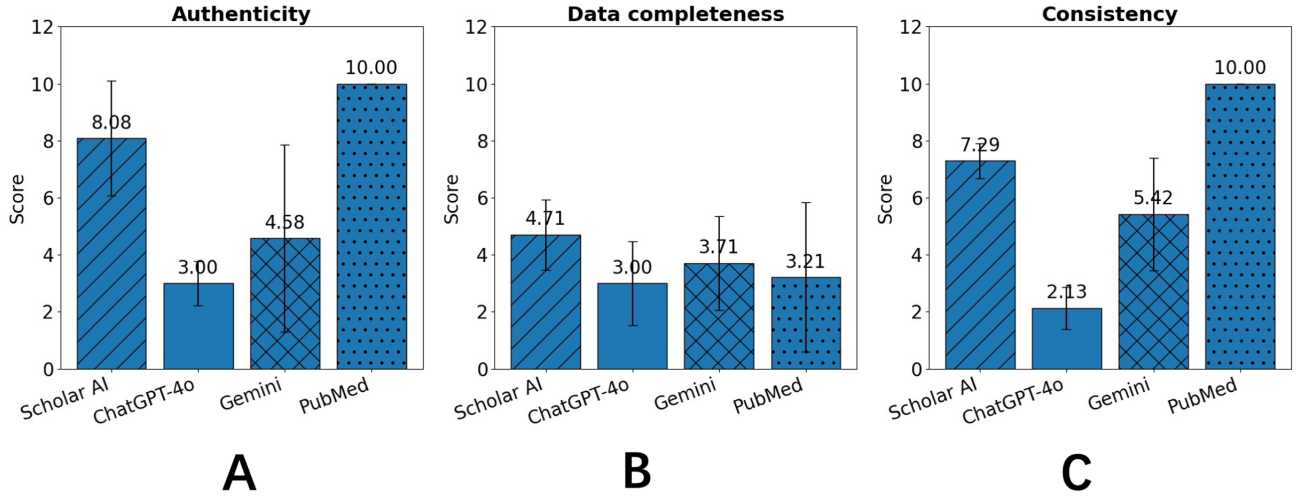
[b]Not applicable.

[c]Not significant.

As a structured and knowledge-grounded recommendation system, CanRisk-RAG achieved high performance in authenticity, data completeness, and consistency, and was therefore used as the benchmark for evaluating other applications (Figure 4). PubMed, while fully reliable in authenticity and consistency due to its literature-based retrieval, showed only moderate data completeness, particularly failing in the comprehensive recommendation task. The 3 LLM-based tools, Scholar AI, ChatGPT-4o, and Gemini 1.5 Flash, performed less reliably overall and often generated incomplete or inconsistent metadata (see Supplemental Information 2.6 in Multimedia Appendix 2). Scholar AI displayed partial strength across themes but still suffered from missing model details, while Gemini 1.5 Flash showed some consistency in recommending models based on cancer types and modeling approaches yet frequently fabricated or cited nonscientific sources such as blogs or online news (see relevant examples in Supplemental Informations 1.12 and 1.13 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 4. Comparative evaluation of authenticity, data completeness, and consistency among 4 baseline applications in cancer risk prediction model recommendation tasks.



Overall, these findings demonstrate that CanRisk-RAG effectively bridges structured domain knowledge with LLM-driven semantic reasoning, achieving both high retrieval precision and reproducibility. Across all evaluation metrics, CanRisk-RAG consistently outperformed PubMed, Scholar AI, ChatGPT-4o, and Gemini 1.5 Flash, attaining markedly higher relevance and reliability alongside near-perfect scores in authenticity, data completeness, and consistency. In contrast, the baseline systems frequently struggled with complex, multifactorial queries, often returning incomplete, inconsistent, or fabricated information.

Usability Evaluation Results

The usability scores are summarized in Supplemental Information 14 in Multimedia Appendix 1. On a 3-point scale, CanRisk-RAG achieved a mean score of 2.7 (SD 0.48) for user goal achievement, indicating that the system was generally effective in meeting users' personalized retrieval needs. In addition, the tool demonstrated good operative accessibility, with an average score of 2.5 (SD 0.53), suggesting that users found the interface and overall operation intuitive and easy to use. Taken together, these results indicate that CanRisk-RAG is user-friendly, enabling most users to accomplish their intended tasks efficiently with minimal operational burden.

Discussion

Principal Findings

The results highlight CanRisk-RAG's clear advantages in recommending cancer risk prediction models, outperforming general-purpose LLM tools and traditional systems like PubMed. By combining an organized knowledge base, which provides rich metadata including cancer types, modeling methods, and performance metrics with a RAG framework, CanRisk-RAG enables high-quality personalized recommendations and advances beyond keyword-based retrieval toward expert-level semantic recommendation. Unlike ChatGPT-4o, Scholar AI, and Gemini 1.5 Flash, which often hallucinate content or misinterpret domain-specific queries, CanRisk-RAG delivers accurate recommendations with structured information grounded in validated literature. It should be noted that as a peer-reviewed, specialized biomedical literature database, PubMed achieved near-perfect scores in authenticity and consistency, which should be interpreted as a reference baseline rather than an algorithmic achievement. While PubMed inherently ensures authenticity through direct literature indexing, it lacks the semantic parsing capabilities required for complex, multifactorial queries, thereby underscoring the value of domain-specific systems that integrate structured knowledge with semantic understanding. It should also be acknowledged that a degree of circularity exists between the system design of CanRisk-RAG and the reliability evaluation. Specifically, JIF is incorporated as a low-weight auxiliary signal in the ranking algorithm, while the reliability metric used in expert evaluation partially reflects publication in authoritative journals. Consequently, the high reliability scores achieved by CanRisk-RAG should be partly interpreted as evidence that the system adheres to its predefined design principles, namely prioritizing models supported by established, higher-reputation peer-reviewed sources, rather than being overinterpreted as direct indicators of clinical effectiveness.

The system adopts a dual-embedding strategy that fuses tag-level semantic representations with embeddings of the original user query, allowing it to jointly capture structured domain constraints and broader user intent. This design enables effective semantic matching even in the absence of exact keyword overlap. Building on this representation, a multi-factor ranking algorithm integrates semantic similarity with complementary quality indicators—including JIF, model AUC, and publication recency—to enhance the credibility and practical usefulness of the recommended models. This is particularly beneficial for model comparison and systematic review scenarios, where both relevance and evidential strength are critical. In addition, our tag-extraction evaluation demonstrates that DeepSeek-V3 achieves high overall accuracy (micro-F_1=0.994), with only two omission cases observed in the 40-query test set. These rare errors were primarily attributable to implicit semantic descriptors or typographical issues rather than systematic extraction failures. Importantly, the fused embedding framework provides partial robustness to such upstream imperfections, as the original-query embedding channel can still encode missing concepts even when a structured tag is not completely extracted.

XSL•FO
RenderX

With respect to the comparative evaluation strategy, baseline LLM systems, namely ChatGPT-4o, ScholarAI, and Gemini 1.5 Flash, were evaluated using static zero-shot prompts to ensure standardized and reproducible conditions across tools. Although this design choice reduces prompt engineering bias and enhances fairness, it may underestimate the upper bound performance of general purpose LLMs, which can benefit from chain of thought prompting, multiround clarification, or interactive refinement. However, incorporating adaptive and tool-specific prompt optimization would compromise reproducibility and comparability, as performance would then depend on prompt engineering expertise rather than intrinsic system capability. Importantly, CanRisk-RAG does not rely on iterative user clarification. Instead, it embeds domain knowledge and semantic constraints directly into the retrieval pipeline, reflecting a distinct design philosophy that shifts cognitive and procedural burden from the user to the system. This property is particularly important in clinical and translational research settings. Moreover, unlike interactive LLM-based tools, CanRisk-RAG produces deterministic and reproducible outputs for identical inputs, which is essential for evidence-based model discovery and systematic research workflows.

Although the superior performance of CanRisk-RAG highlights the advantage of integrating structured domain knowledge with retrieval-augmented generation, its effectiveness largely depends on the built-in knowledge base, which supplies high-quality, expert-verified model information. Consequently, the "extensibility" of the framework depends on the sustainable expansion of such constructed resources. To address this scalability challenge, we developed a semiautomated update pipeline that combines automated literature retrieval, rule-based metadata extraction, and LLM-assisted knowledge extraction in a workflow [52]. This tool substantially improves data update efficiency while reducing manual workload. Nevertheless, essential human review and expert supplementation are retained to ensure the accuracy of curated content. The knowledge base is planned to be updated every six months to maintain both recency and reliability.

In addition, although currently focused on cancer risk prediction models, the CanRisk-RAG framework is adaptable to other disease domains and scenarios, including prognosis and treatment response models, underscoring its potential as a universal platform for the discovery and application of clinical prediction models.

Limitations

Despite the encouraging performance of CanRisk-RAG, several limitations should be acknowledged.

First, although CanRisk-RAG offers enhanced flexibility and effective retrieval, its performance is inevitably shaped by multiple structured layers of uncertainty spanning internal (algorithmic) and external (evidence-base) factors. Internally, reliance on semantic similarity may reduce field-level specificity; LLM-based tag extraction can over-generalize factors (eg, interpreting "family history of breast cancer" as "family history" or generalizing a specific genetic factor such as BRCA1/2 mutation into a broad category like "genetic risks") and changes in multi-factor weights (eg, raising "Recency" weight from 0.05 to 0.15) can alter rankings. These factors introduce algorithmic uncertainty, which we partially mitigated through ablation experiments and iterative weighting optimization. Externally, heterogeneity in study design, sample size, population, and reporting limits comparability and generalizability. For instance, lung cancer survival models from small regional cohorts may not generalize to multi-institutional datasets, and models trained on East Asian populations may underperform in Western populations due to demographic and healthcare-system differences. Even established models (eg, Gail) exhibit cohort-dependent AUC and calibration, highlighting cross-dataset variability as an important uncertainty layer. Future work will incorporate domain adaptation and uncertainty quantification methods to reduce these effects.

Second, the current ranking strategy relies on a limited set of quality-related indicators. JIF is included as a low-weight auxiliary signal and is normalized to mitigate its influence; however, it remains a coarse, venue-level metric that does not directly reflect study-level methodological rigor. Important indicators such as sample size, calibration reporting, and the extent of external validation are not yet uniformly available across the curated literature and therefore could not be systematically incorporated as first-class ranking features in the present version. As a result, residual bias may be introduced when finer-grained quality signals are missing or inconsistently reported, even though JIF is used only to refine the ordering among semantically similar candidates. In addition, the ranking strategy incorporates reported AUC values as a quantitative indicator of model performance without explicitly differentiating or weighting AUCs derived from model development versus external validation stages. This design choice may inadvertently favor development studies that report optimistic performance over rigorously validated models with stronger generalizability. Future versions of CanRisk-RAG will aim to integrate validation-aware performance stratification and explicit study-level quality indicators, such as cohort size, validation design, and calibration metrics, to better balance apparent accuracy with real-world applicability and progressively reduce reliance on venue-level signals.

Third, the present study relied on a single LLM (DeepSeek-V3) for semantic tag extraction and literature summarization. While this design choice ensured internal consistency and reproducibility across experiments, it inevitably limited robustness to model-specific variability and potential biases inherent to a single LLM. In addition, baseline comparisons with general-purpose LLMs were conducted under a static zero-shot prompting setting without access to external retrieval, web browsing, or tool invocation capabilities. As a result, these baselines were constrained to parametric-only knowledge, which may underestimate their performance in more interactive or retrieval-augmented configurations. In this context, the evaluation primarily highlights the hallucination risks and structural limitations of ungrounded LLM responses when external knowledge is unavailable, rather than serving as an exhaustive benchmark of the full capabilities of modern LLM ecosystems. Future work will extend this evaluation framework by incorporating multi-model ensemble strategies for tag extraction and summarization to enhance robustness and

XSL•FO
RenderX

stability. In parallel, more comprehensive comparisons will be conducted against LLM-based systems equipped with web browsing, tool use, or fully integrated RAG workflows, including multi-turn and interactive prompting strategies.

Fourth, expert scoring was performed manually, introducing subjectivity despite inter-rater consistency checks. Future studies will incorporate automated evaluation methods, such as rule-based matching and semantic alignment metrics, to improve reproducibility and reduce dependence on manual assessments.

Fifth, although CanRisk-RAG includes an LLM-based literature summarization module, the current evaluation is primarily centered on retrieval and recommendation performance. The adopted metrics assess the relevance and reliability of retrieved models and metadata but do not explicitly quantify the factual accuracy or hallucination rate of the generated summaries. The summarization component is intended to serve as an assistive tool for facilitating user comprehension and synthesizing key information from retrieved literature, rather than to provide precise quantitative data; future work will incorporate dedicated evaluation protocols to systematically assess the quality and faithfulness of the generative component.

Finally, a key limitation of the present study is that the CanRisk-RAG knowledge base was constructed using literature retrieved prior to December 2022, and therefore does not include cancer risk prediction models published from 2023 onward. This temporal constraint may temporarily place CanRisk-RAG at a relative disadvantage when compared with contemporary LLMs that potentially have access to more recent biomedical literature. Beyond general concerns regarding data currency, this cutoff has specific implications for users seeking state-of-the-art computational or deep learning–based cancer risk prediction models, as recent years have witnessed a rapid proliferation of AI-driven approaches incorporating deep neural networks and multimodal data integration. As a result, the current version of CanRisk-RAG may be less suitable for users whose primary objective is to identify the most recent model architectures. To address this limitation and enhance temporal coverage, we have developed a semiautomated update pipeline that integrates automated literature retrieval, rule-based metadata extraction, and LLM-assisted curation, while retaining expert review to ensure data quality. The knowledge base is planned to be updated on a semiannual basis. Future evaluations will incorporate newly published models to assess how expanded temporal coverage influences retrieval performance, and to enable more temporally aligned comparisons with evolving LLM-based systems, particularly in the context of rapidly advancing AI-driven cancer risk modeling.

Conclusion

This study introduced CanRisk-RAG, a knowledge-driven retrieval framework for cancer risk prediction models that integrates structured knowledge with semantic reasoning. Through systematic evaluation, CanRisk-RAG demonstrated superior performance in relevance, reliability, and data consistency compared with existing retrieval systems. These results indicate that combining structured knowledge representation with LLM-based reasoning can substantially improve the precision and transparency of model discovery in biomedical research, supporting reproducible and evidence-based applications in precision oncology.

Acknowledgments

The authors used ChatGPT (OpenAI, GPT-4) for language editing purposes during the preparation of this manuscript. The authors confirm that all final manuscript decisions, including content approval and submission, were made exclusively by the authors.

Funding

This work was supported by the National Natural Science Foundation of China (32570773, 32270690) and the Medical Research Projects in Chengdu City (2025434).

Data Availability

The source code for data processing is available at GitHub [52,53].

Authors' Contributions

Conceptualization: BS (lead), SR (equal)
Data curation: SR
Formal analysis: SR (lead), XZ (supporting)
Funding acquisition: BS (lead), SR (supporting)
Investigation: SR
Methodology: BS (lead), SR (equal), J Zhang (supporting)
Project administration: BS (lead), SR (equal)
Resources: BS
Supervision: BS
Validation: SR (lead), XZ (supporting), JD (supporting), YZ (supporting), CB (supporting)
Visualization: SR (lead), J Zhao (supporting)
Writing—original draft: SR (lead), XZ (supporting)

XSL•FO
RenderX

Writing—review & editing: SR (lead), JS (supporting), J Zhang (supporting), HL (supporting), ZF (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods and details.
[DOCX File, 1745 KB - medinform_v14i1e78519_app1.docx]

Multimedia Appendix 2
Raw scores of experiments.
[XLSX File, 73 KB - medinform_v14i1e78519_app2.xlsx]

References

1. Vickers AJ. Prediction models in cancer care. CA Cancer J Clin 2011;61(5):315-326. [doi: 10.3322/caac.20118] [Medline: 21732332]
2. Lophatananon A, Usher-Smith J, Campbell J, et al. Development of a cancer risk prediction tool for use in the UK primary care and community settings. Cancer Prev Res (Phila) 2017 Jul;10(7):421-430. [doi: 10.1158/1940-6207.CAPR-16-0288] [Medline: 28559460]
3. Engel C, Fischer C. Breast cancer risks and risk prediction models. Breast Care (Basel) 2015 Feb;10(1):7-12. [doi: 10.1159/000376600] [Medline: 25960719]
4. Tutsoy O, Sumbul HE. A novel deep machine learning algorithm with dimensionality and size reduction approaches for feature elimination: thyroid cancer diagnoses with randomly missing data. Brief Bioinform 2024 May 23;25(4):bbae344. [doi: 10.1093/bib/bbae344] [Medline: 39007597]
5. Chen X, Jansen L, Guo F, Hoffmeister M, Chang-Claude J, Brenner H. Smoking, genetic predisposition, and colorectal cancer risk. Clin Transl Gastroenterol 2021 Mar 1;12(3):e00317. [doi: 10.14309/ctg.0000000000000317] [Medline: 33646204]
6. Byrne S, Boyle T, Ahmed M, Lee SH, Benyamin B, Hyppönen E. Lifestyle, genetic risk and incidence of cancer: a prospective cohort study of 13 cancer types. Int J Epidemiol 2023 Jun 6;52(3):817-826. [doi: 10.1093/ije/dyac238] [Medline: 36651198]
7. Su YR, Sakoda LC, Jeon J, et al. Validation of a genetic-enhanced risk prediction model for colorectal cancer in a large community-based cohort. Cancer Epidemiol Biomarkers Prev 2023 Mar 6;32(3):353-362. [doi: 10.1158/1055-9965.EPI-22-0817] [Medline: 36622766]
8. Bruno V, Betti M, D'Ambrosio L, et al. Machine learning endometrial cancer risk prediction model: integrating guidelines of European Society for Medical Oncology with the tumor immune framework. Int J Gynecol Cancer 2023 Nov 6;33(11):1708-1714. [doi: 10.1136/ijgc-2023-004671] [Medline: 37875322]
9. Choi J, Ha TW, Choi HM, et al. Development of a breast cancer risk prediction model incorporating polygenic risk scores and nongenetic risk factors for Korean women. Cancer Epidemiol Biomarkers Prev 2023 Sep 1;32(9):1182-1189. [doi: 10.1158/1055-9965.EPI-23-0064] [Medline: 37310812]
10. Tutsoy O, Koç GG. Deep self-supervised machine learning algorithms with a novel feature elimination and selection approaches for blood test-based multi-dimensional health risks classification. BMC Bioinformatics 2024 Mar 8;25(1):103. [doi: 10.1186/s12859-024-05729-2] [Medline: 38459463]
11. Hassan FH, Wye AYK, Yusof SSS, Xiang TY. Predicting risk of getting smoking-related cancer: a comparison of three prediction models. Presented at: LOPAL '18: Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications; May 2-5, 2018. [doi: 10.1145/3230905.3230950]
12. Bae CY, Kim BS, Jee SH, Lee JH, Nguyen ND. A study on survival analysis methods using neural network to prevent cancers. Cancers (Basel) 2023 Sep 27;15(19):4757. [doi: 10.3390/cancers15194757] [Medline: 37835451]
13. Mertens E, Barrenechea-Pulache A, Sagastume D, Vasquez MS, Vandevijvere S, Peñalvo JL. Understanding the contribution of lifestyle in breast cancer risk prediction: a systematic review of models applicable to Europe. BMC Cancer 2023 Jul 21;23(1):687. [doi: 10.1186/s12885-023-11174-w] [Medline: 37480028]
14. Lee CPL, Irwanto A, Salim A, et al. Breast cancer risk assessment using genetic variants and risk factors in a Singapore Chinese population. Breast Cancer Res 2014 Jun 18;16(3):R64. [doi: 10.1186/bcr3678] [Medline: 24941967]
15. Liang H, Zhou X, Zhu Y, et al. Association of outdoor air pollution, lifestyle, genetic factors with the risk of lung cancer: a prospective cohort study. Environ Res 2023 Feb 1;218:114996. [doi: 10.1016/j.envres.2022.114996] [Medline: 36481370]
16. Li Y, Xiao X, Han Y, et al. Genome-wide interaction study of smoking behavior and non-small cell lung cancer risk in Caucasian population. Carcinogenesis 2018 Mar 8;39(3):336-346. [doi: 10.1093/carcin/bgx113] [Medline: 29059373]
17. Markowetz F. All models are wrong and yours are useless: making clinical prediction models impactful for patients. NPJ Precis Oncol 2024 Feb 28;8(1):54. [doi: 10.1038/s41698-024-00553-6] [Medline: 38418530]

XSL•FO
RenderX

18. Colditz GA, Wei EK. Risk prediction models: applications in cancer prevention. Curr Epidemiol Rep 2015 Dec;2(4):245-250. [doi: 10.1007/s40471-015-0057-1]
19. Waters EA, Taber JM, McQueen A, Housten AJ, Studts JL, Scherer LD. Translating cancer risk prediction models into personalized cancer risk assessment tools: stumbling blocks and strategies for success. Cancer Epidemiol Biomarkers Prev 2020 Dec;29(12):2389-2394. [doi: 10.1158/1055-9965.EPI-20-0861] [Medline: 33046450]
20. White J. PubMed 2.0. Med Ref Serv Q 2020;39(4):382-387. [doi: 10.1080/02763869.2020.1826228] [Medline: 33085945]
21. Risk assessment macros, software programs, and tools. National Cancer Institute. URL: https://dceg.cancer.gov/tools/risk-assessment [accessed 2025-09-13]
22. Prediction tools - a tool for doctors and patients. Memorial Sloan Kettering Cancer Center. URL: https://www.mskcc.org/nomograms [accessed 2025-09-13]
23. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
24. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023 Nov 16;6(1):210. [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
25. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
26. Li Y, Tao W, Li Z, et al. Artificial intelligence-powered pharmacovigilance: a review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. J Biomed Inform 2024 Apr;152:104621. [doi: 10.1016/j.jbi.2024.104621] [Medline: 38447600]
27. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023 Aug 1;34(8):1302-1304. [doi: 10.1681/ASN.0000000000000166] [Medline: 37254254]
28. Day T. A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. Prof Geogr 2023 Nov 2;75(6):1024-1027. [doi: 10.1080/00330124.2023.2190373]
29. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]
30. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 2025 Apr 1;32(4):605-615. [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
31. Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. J Am Med Inform Assoc 2024 Sep 1;31(9):2030-2039. [doi: 10.1093/jamia/ocae129] [Medline: 38857454]
32. Aftab W, Apostolou Z, Bouazoune K, Straub T. Optimizing biomedical information retrieval with a keyword frequency-driven prompt enhancement strategy. BMC Bioinformatics 2024 Aug 27;25(1):281. [doi: 10.1186/s12859-024-05902-7] [Medline: 39192204]
33. Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report. arXiv. Preprint posted online on Dec 27, 2024. [doi: 10.48550/arXiv.2412.19437]
34. Galli C, Donos N, Calciolari E. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. Information 2024;15(2):68. [doi: 10.3390/info15020068]
35. Lis S, Kubkowski M, Borkowska O, Serwa D, Kuparnik J. Analyzing credit risk model problems through natural language processing-based clustering and machine learning: insights from validation reports. J Risk Model Valid 2024;18. [doi: 10.21314/JRMV.2024.006]
36. Wilianto D, Suganda Girsang A. Automatic short answer grading onhigh school's e-learning using semantic similarity methods. TEM Journal 2023 Feb 27:297-302. [doi: 10.18421/TEM121-37]
37. Kiesslich T, Beyreis M, Zimmermann G, Traweger A. Citation inequality and the Journal Impact Factor: median, mean, (does it) matter? Scientometrics 2021 Feb;126(2):1249-1269. [doi: 10.1007/s11192-020-03812-y]
38. Pockett RD, Castellano D, McEwan P, Oglesby A, Barber BL, Chung K. The hospital burden of disease associated with bone metastases and skeletal-related events in patients with breast cancer, lung cancer, or prostate cancer in Spain. Eur J Cancer Care (Engl) 2010 Nov;19(6):755-760. [doi: 10.1111/j.1365-2354.2009.01135.x] [Medline: 19708928]
39. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492] [Medline: 30207593]
40. Kim G, Bahl M. Assessing risk of breast cancer: a review of risk prediction models. J Breast Imaging 2021;3(2):144-155. [doi: 10.1093/jbi/wbab001] [Medline: 33778488]
41. Lee A, Mavaddat N, Cunningham A, et al. Enhancing the BOADICEA cancer risk prediction model to incorporate new data on RAD51C, RAD51D, BARD1 updates to tumour pathology and cancer incidence. J Med Genet 2022;59(12):1206-1218. [doi: 10.1136/jmedgenet-2022-108471] [Medline: 36162851]
42. Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer 2019 Jul;121(1):76-85. [doi: 10.1038/s41416-019-0476-8] [Medline: 31114019]

XSL•FO
RenderX

43. Lee YCA, Al-Temimi M, Ying J, et al. Risk prediction models for head and neck cancer in the US population from the INHANCE consortium. Am J Epidemiol 2020 Apr 2;189(4):330-342. [doi: 10.1093/aje/kwz259] [Medline: 31781743]
44. Katki HA, Kovalchik SA, Petito LC, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. Ann Intern Med 2018 Jul 3;169(1):10-19. [doi: 10.7326/M17-2701] [Medline: 29800127]
45. Zhang X, Rice M, Tworoger SS, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: a nested case-control study. PLoS Med 2018 Sep;15(9):e1002644. [doi: 10.1371/journal.pmed.1002644] [Medline: 30180161]
46. Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A. Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. JAMA Netw Open 2023 Jun 1;6(6):e2318495. [doi: 10.1001/jamanetworkopen.2023.18495] [Medline: 37318804]
47. Guan Z, Huang T, McCarthy AM, et al. Combining breast cancer risk prediction models. Cancers (Basel) 2023 Feb 8;15(4):1090. [doi: 10.3390/cancers15041090] [Medline: 36831433]
48. Törnberg P. Chatgpt-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv. Preprint posted online on Apr 13, 2023. [doi: 10.48550/arXiv.2304.06588]
49. ScholarAI. 2024. URL: https://scholarai.io [accessed 2024-08-04]
50. Waisberg E, Ong J, Masalkhi M, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye 2024 Mar;38(4):642-645. [doi: 10.1038/s41433-023-02760-0]
51. Ethical Review Measures for Life Science and Medical Research Involving Human Subjects (National Health and Family Planning Commission Document No. 4 [2023]). Sichuan University. URL: https://kyy.scu.edu.cn/info/1416/5946.htm [accessed 2026-03-03]
52. Shuminren/llms_clinicalriskmodel_tool. GitHub. URL: https://github.com/Shuminren/LLMs_ClinicalRiskModel_Tool [accessed 2026-02-16]
53. Shuminren/CanRisk-RAG. GitHub. URL: https://github.com/Shuminren/CanRisk-RAG [accessed 2026-03-02]

Abbreviations

AUC: area under the curve
JIF: journal impact factor
LLM: large language model
MID: model ID
RAG: retrieval-augmented generation

Edited by A Coristine; submitted 04.Jun.2025; peer-reviewed by M Al-Agil, O Tutsoy, Z Li; revised version received 27.Jan.2026; accepted 29.Jan.2026; published 09.Mar.2026.

Please cite as:
Ren S, Zheng X, Zhao J, Du J, Zhang Y, Bi C, Song J, Zhang J, Lang H, Zhang F, Shen B
Knowledge-Guided Explainable Recommendation Tool for Cancer Risk Prediction Models Using Retrieval-Augmented Large Language Models: Development and Validation Study
JMIR Med Inform 2026;14:e78519
URL: https://medinform.jmir.org/2026/1/e78519
doi:10.2196/78519

© Shumin Ren, Xin Zheng, Jing Zhao, Jiale Du, Yuxin Zhang, Cheng Bi, Jie Song, Jinyi Zhang, Hongmei Lang, Fan Zhang, Bairong Shen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Age on Hospital Outcomes Following Minimally Invasive Posterior Lumbar Interbody Fusion: Retrospective Analysis of the Nationwide Inpatient Sample Database from 2016 to 2020

Yu-Jun Lin[1,2], MD; Fu-Yuan Shih[1,2], MD; Jin-Fu Huang[1], MD; Chun-Wei Ting[1], MD; Yu-Chin Tsai[1], MD, PhD; Lin Chang[1], MD; Yu-Hua Huang[1], MD, PhD; Ming-Jung Chuang[1], MD

[1]Department of Neurosurgery, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung City, Taiwan, Taiwan

[2]Department of Neurosurgery, Kaohsiung Municipal Ta-Tung Hospital, Kaohsiung City, Taiwan

Corresponding Author:
Yu-Jun Lin, MD
Department of Neurosurgery
Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine
Number 123, Da-Pei Road
Niao-Song District
Kaohsiung City, Taiwan, 83301
Taiwan
Phone: 886 7 7317123 ext 8011
Fax: 886 7 7318762
Email: linyujun1022@gmail.com

Abstract

Background: Minimally invasive posterior lumbar interbody fusion (MIS-PLIF) is commonly performed to treat degenerative lumbar spinal conditions. Patients of advanced age often present with multiple comorbidities and reduced physiological reserves, influencing surgical risks and recovery. The growing aging population has led to a rising demand for care for older adults, posing significant challenges for health care systems worldwide.

Objective: This study aimed to identify the associations between different age groups and MIS-PLIF outcomes.

Methods: This study retrospectively analyzed data from the United States Nationwide Inpatient Sample collected between 2016 and 2020. Patients aged ≥60 years who underwent MIS-PLIF were eligible for inclusion in this study. Patients were categorized into age groups (60-69, 70-79, and ≥80 y). Logistic and linear regressions were used to determine the associations between the study variables and outcomes, including in-hospital mortality, complications, nonroutine discharge, and length of stay.

Results: A total of 785 patients aged ≥60 (mean age 69.4, SD 0.2) years who underwent MIS-PLIF were included in the analysis, and 18.7% (147/785) experienced at least one complication. After adjustment, compared with patients aged 60 to 69 years, the risk of nonroutine discharge was significantly increased in patients aged 70 to 79 years (adjusted odds ratio 2.33, 95% CI 1.57-3.46; *P*<.001) and ≥80 years (adjusted odds ratio 4.79, 95% CI 2.64-8.67; *P*<.001). No significant differences in the risk of complications or length of hospital stay were observed across the age groups.

Conclusions: In older patients undergoing MIS-PLIF, advanced age is an independent predictor of nonroutine discharge. Furthermore, our findings suggest that age alone is not an independent risk factor for complications or extended hospital stays among older patients. These findings underscore that MIS-PLIF is a viable option for older patients, for whom extra attention may still be needed for postoperative care. Implementing age-stratified management for older patients undergoing MIS-PLIF may have important clinical policy implications.

(JMIR Med Inform 2026;14:e76424) doi:10.2196/76424

KEYWORDS

age; minimally invasive posterior lumbar interbody fusion; MIS-PLIF; Nationwide Inpatient Sample; NIS; in-hospital outcome

XSL•FO
RenderX

Introduction

Minimally invasive posterior lumbar interbody fusion (MIS-PLIF) represents a surgical cornerstone in addressing lumbar spine pathology [1,2]. This innovative procedure accesses the lumbar spine through a posterior approach to treat damaged intervertebral discs and fuse adjacent vertebrae [3]. The procedure can achieve results similar to those of open procedures while minimizing disruption to the surrounding tissues.

With the increase in population age, the incidence of degenerative spine disorders has also increased, leading to an increased demand for surgical treatments. Conditions such as osteoarthritis, spinal stenosis, and degenerative disc disease are more common in older individuals, resulting in a growing need for procedures such as spinal fusion, laminectomy, and discectomy to manage pain and improve mobility [4]. The relationship between advancing age and spinal health is paramount, given the potential implications of age-related changes in bone density [5,6]. These changes may not only influence the procedural success of MIS-PLIF but also impact the overall safety and effectiveness of the procedure.

Surgery for degenerative spinal conditions presents specific challenges in the older population seeking relief from symptoms [7,8]. Beyond the structural considerations of bone density, older patients often have additional health concerns and comorbidities, adding increased risk to the surgical procedure and the subsequent recovery process [9]. Understanding the relationship between age and surgical outcomes is important for tailoring surgical approaches to ensure optimal outcomes and postoperative recovery in the older population.

Despite MIS-PLIF being a common and generally well-tolerated procedure, limited research has specifically focused on the impact of age on postoperative outcomes [10,11]. The increasing prevalence of degenerative spinal diseases in older individuals makes it important to understand how age may influence the success and safety of MIS-PLIF. Furthermore, age-stratified management is becoming increasingly important due to the rapidly growing older population worldwide. This study aimed to deepen the understanding of the effect of age on outcomes by analyzing the relationship between advancing age and postoperative outcomes of MIS-PLIF using a large, nationally representative dataset from the United States [12]. Using a large, nationally representative dataset, this study provides the statistical power necessary to detect clinically meaningful differences across age groups that smaller studies may have missed. This study may offer valuable implications for both clinical practice and policy development.

Methods

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki (1975) and its amendments. The protocol for this study was submitted for review to the Institutional Review Board of Chang Gung Medical Foundation (202301168B0), which granted an exemption from formal Institutional Review Board oversight. This study complies with the terms of the Nationwide Inpatient Sample (NIS) data-use agreement. The data used in this study were obtained through the Online Health Care Cost and Utilization Project (HCUP) Central Distributor. Given the anonymized nature of the data, the requirement for informed consent was consequently waived. No compensation was provided to participants.

Data Source

This investigation drew upon data extracted from the NIS for the period of 2016 to 2020. The NIS, an initiative of the HCUP under the aegis of the Agency for Healthcare Research and Quality [13], encompasses a 20% stratified sample of inpatient admissions across 45 states and 1051 participating hospitals. The database meticulously records a wealth of patient information at the time of discharge, including primary and secondary diagnoses and procedures, dates of admission and discharge, discharge status, demographic details, and duration of hospital stay. Moreover, the NIS is equipped with statistical weights designed to facilitate the extrapolation of these findings to national patient volumes.

Study Design

This study was designed as a population-based, retrospective analysis. All data were sourced through a request to the Online HCUP Central Distributor, which manages the database [14] (certificate HCUP-4T28I74JW). This research strictly adhered to the data-use protocols stipulated by the NIS under the HCUP. The investigation used deidentified secondary data from the NIS database; hence, there was no direct involvement of patients or the public.

Study Population

The NIS database was searched for patients aged ≥60 years who underwent MIS-PLIF during the study period. The exclusion criteria were (1) traumatic injury; (2) injury in traffic accidents; and (3) missing values for weight, outcomes of interest, and covariates. All diagnoses and procedures were identified by the *International Classification of Diseases, Ninth Revision and Tenth Revision, Clinical Modification* (*ICD-9-CM* and *ICD-10-CM*) codes, as detailed in Table S1 in Multimedia Appendix 1. Following previous literature, we stratified age into 3 groups (ie, 60-69 y, 70-79 y, and ≥80 y) to perform comparative analyses across these categories [15,16].

Main Outcomes and Variables

The primary outcomes included in-hospital mortality, nonroutine discharge (ie, discharged to long-term care facilities), prolonged length of stay (LOS; defined as ≥75th percentile of LOS in the study population), and complications. Complications assessed included surgical (eg, infection, dural tear, hemorrhage, hematoma or seroma, and postoperative anemia) and medical complications (eg, pneumonia, pulmonary embolism, acute kidney injury, acute myocardial infarction, and urinary retention) based on insights from our literature review and clinical experience.

Demographic variables analyzed in this study encompassed patient age, sex, race or ethnicity (categorized as White, Black, Hispanic, and other), household income, insurance status

XSL•FO
RenderX

(identified by the primary payer), and whether admission occurred over a weekend. Household income quartiles were derived from the NIS, which estimates income based on the ZIP code of each patient's residence [17]. Comorbidities were identified using *ICD-9* and *ICD-10* codes and included diabetes mellitus, osteoporosis, obesity (defined as a BMI ≥30 kg/m²), chronic obstructive pulmonary disease, renal disease, hypertension, coronary artery disease, and heart failure. Relevant codes for each comorbidity are provided in Table S1 in Multimedia Appendix 1. Charlson Comorbidity Index was calculated from individual comorbidities to represent patients' overall severity of comorbid conditions.

Additionally, characteristics related to the hospital, such as bed size, location and teaching status, and regional location, were also collected as part of the extensive dataset available for all patients.

Statistical Analysis

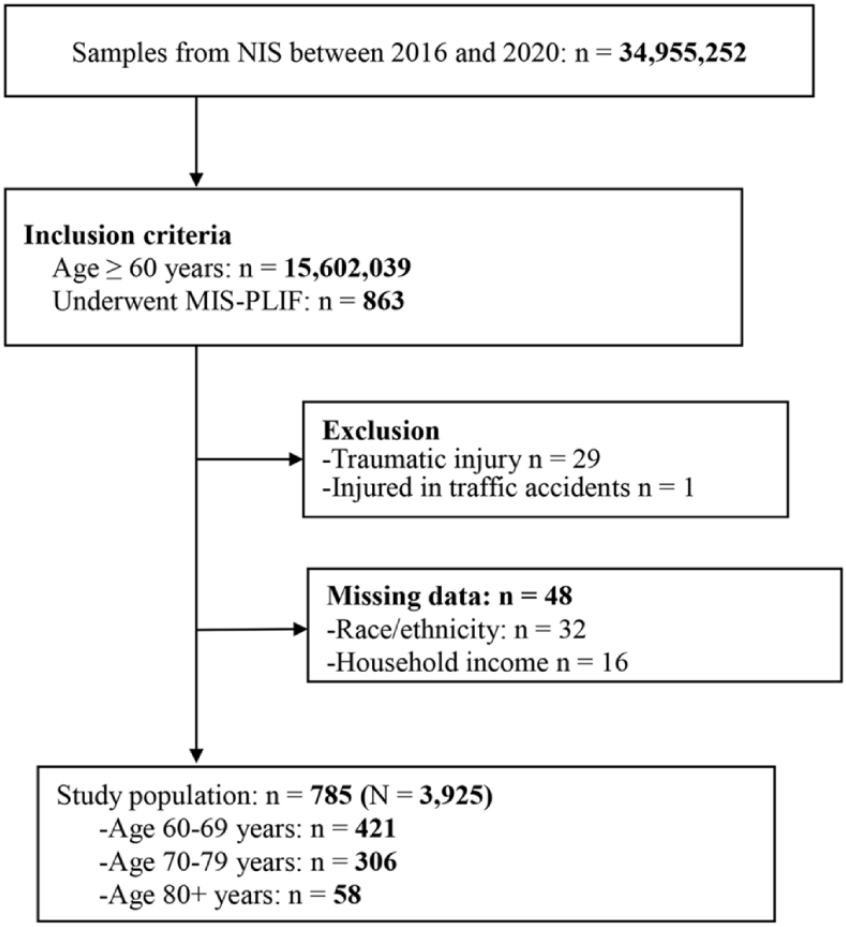
The NIS database comprises a 20% stratified sample of US annual inpatient admissions, using weighted samples, strata, and clusters to facilitate national estimates in all analyses. Data analysis was conducted using the SURVEY procedure in SAS, which is optimized for sample survey data. Descriptive statistics for patients are presented as counts (n) and weighted percentages (%) or as means with SEs. Group comparisons for categorical variables were conducted using the Rao-Scott chi-square test, whereas weighted mean differences for continuous variables were analyzed using survey methods that account for stratification, clustering, and sampling weights, ensuring valid and robust statistical inferences within the context of complex survey designs. Logistic regression analyses were performed to assess the associations between study variables and binary outcomes, with results presented as odds ratios (ORs) and 95% CI. Linear regression was performed to assess the associations between the study variables and LOS. Multivariable regression was adjusted for variables that were significant (P<.05) in the univariate analysis. All statistical tests were 2-sided, with P<.05, deemed to indicate statistical significance. The analyses were conducted using SAS (version 9.4; SAS Institute Inc).

Results

Patient Selection

The patient selection process is illustrated in Figure 1. Between 2016 and 2020, 863 patients aged ≥60 years who underwent MIS-PLIF were identified from the NIS database. After excluding 29 patients with traumatic injuries, 1 patient with a traffic accident injury, and 48 patients with missing information on race, ethnicity, and household income, 785 patients were included in the study. Using the sample weight from the NIS database, this cohort corresponds to a total of 3925 hospitalized patients across the United States, as illustrated in Figure 1.

Figure 1. Flow diagram of patient selection. MIS-PLIF: minimally invasive posterior lumbar interbody fusion; NIS: Nationwide Inpatient Sample.



XSL•FO
RenderX

Characteristics of the Study Population

In the study population, the mean age was 69.4 (SD 0.2) years, 60.9% (478/785) were female individuals, 79.2% (622/785) were White, and 72.9% (572/785) had insurance covered by Medicare or Medicaid. A total of 66% (518/785) of the patients were diagnosed with spondylolisthesis. Patients aged 60 to 69 years had a significantly higher percentage of obesity, patients aged 70 to 79 years had a significantly higher frequency of diabetes mellitus and heart failure than the other age groups, and patients aged ≥80 years had the highest frequency of renal disease and hypertension (Table 1).

XSL•FO
RenderX

Table 1. Characteristics of the study population by age group.

Characteristics	All patients (N=785)	Age (y)			*P* value
		60-69 (n=421)	70-79 (n=306)	≥80 (n=58)	
Patient characteristics					
Age (y), mean (SD)	69.4 (0.2)	64.5 (0.1)	73.6 (0.1)	82.7 (0.1)	*<.001*[a]
Sex, n (%)					.95
Male	307 (39.1)	163 (38.7)	121 (39.5)	23 (39.7)	
Female	478 (60.9)	258 (61.3)	185 (60.5)	35 (60.3)	
Race and ethnicity, n (%)					*<.001*[a]
Black	93 (11.8)	63 (15)	29 (9.5)	1 (1.7)	
Hispanic	41 (5.2)	26 (6.2)	13 (4.2)	2 (3.4)	
White	622 (79.2)	314 (74.6)	256 (83.7)	52 (89.7)	
Other	29 (3.7)	18 (4.3)	8 (2.6)	3 (5.2)	
Household income, n (%)					*<.001*[a]
Q1[b]	196 (25)	117 (27.8)	63 (20.6)	16 (27.6)	
Q2	203 (25.9)	98 (23.3)	96 (31.4)	9 (15.5)	
Q3	198 (25.2)	113 (26.8)	67 (21.9)	18 (31)	
Q4	188 (23.9)	93 (22.1)	80 (26.1)	15 (25.9)	
Medicare or Medicaid insurance, n (%)					*<.001*[a]
No	213 (27.1)	182 (43.2)	29 (9.5)	2 (3.4)	
Yes	572 (72.9)	239 (56.8)	277 (90.5)	56 (96.6)	
Admission type, n (%)					.62
Elective	713 (90.8)	383 (91)	279 (91.2)	51 (87.9)	
Emergent	72 (9.2)	38 (9)	27 (8.8)	7 (12.1)	
Spondylolisthesis, n (%)					.71
No	267 (34)	147 (34.9)	100 (32.7)	20 (34.5)	
Yes	518 (66)	274 (65.1)	206 (67.3)	38 (65.5)	
Hospital characteristics					
Hospital bed size, n (%)					.31
Small	158 (20.1)	87 (20.7)	61 (19.9)	10 (17.2)	
Medium	267 (34)	149 (35.4)	102 (33.3)	16 (27.6)	
Large	360 (45.9)	185 (43.9)	143 (46.7)	32 (55.2)	
Hospital location or teaching status, n (%)					.11
Rural	45 (5.7)	22 (5.2)	18 (5.9)	5 (8.6)	
Urban nonteaching	174 (22.2)	91 (21.6)	67 (21.9)	16 (27.6)	
Urban teaching	566 (72.1)	308 (73.2)	221 (72.2)	37 (63.8)	
Hospital region, n (%)					.27
Northeast	155 (19.7)	79 (18.8)	63 (20.6)	13 (22.4)	
Midwest	160 (20.4)	86 (20.4)	59 (19.3)	15 (25.9)	
South	387 (49.3)	216 (51.3)	148 (48.4)	23 (39.7)	
West	83 (10.6)	40 (9.5)	36 (11.8)	7 (12.1)	
Comorbidities, n (%)					
Diabetes mellitus	199 (25.4)	100 (23.8)	91 (29.7)	8 (13.8)	*.003*[a]

XSL•FO
RenderX

Characteristics	All patients (N=785)	Age (y)			*P* value
		60-69 (n=421)	70-79 (n=306)	≥80 (n=58)	
Osteoporosis	50 (6.4)	23 (5.5)	22 (7.2)	5 (8.6)	.33
Obesity	198 (25.2)	116 (27.6)	72 (23.5)	10 (17.2)	*.03*[a]
Chronic obstructive pulmonary disease	74 (9.4)	40 (9.5)	29 (9.5)	5 (8.6)	.96
Renal disease	66 (8.4)	20 (4.8)	29 (9.5)	17 (29.3)	*<.001*[a]
Hypertension	574 (73.1)	284 (67.5)	243 (79.4)	47 (81)	*<.001*[a]
Coronary heart disease	116 (14.8)	54 (12.8)	50 (16.3)	12 (20.7)	.06
Heart failure	26 (3.3)	9 (2.1)	15 (4.9)	2 (3.4)	*<.001*[a]
Charlson Comorbidity Index, n (%)					*<.001*[a]
0	372 (47.4)	205 (48.7)	142 (46.4)	25 (43.1)	
1	212 (27)	128 (30.4)	74 (24.2)	10 (17.2)	
2	111 (14.1)	52 (12.4)	45 (14.7)	14 (24.1)	
≥3	90 (11.5)	36 (8.6)	45 (14.7)	9 (15.5)	

[a]*P*<.05.

[b]Quartiles (Q1-Q4) were derived by dividing the distribution of the variable into four equal parts, with Q1 representing the lowest and Q4 the highest range.

In-Hospital Outcomes

Only 0.1% (1/785) of the study population died in the hospital, whereas 18.7% (147/785) had at least one complication, of which 13.8% (108/785) had a least one surgical complication, and 7.3% (57/785) had at least one medical complication. Among the 3 age groups, patients aged ≥80 years had a significantly higher frequency of nonroutine discharge (20/58, 35% in the ≥80 y age group vs 52/421, 12.4% in the 60-69 y age group and 76/306, 24.9% in the 70-79 y age group; *P*<.001). There were no significant differences in complications (*P*=.68) and LOS (*P*=.11) between the 3 age groups (Table 2).

XSL•FO
RenderX

Table 2. In-hospital outcomes of the study population by age group.

Outcomes	All patients (N=785)	Age groups (y)			*P* value
		60-69 (n=421)	70-79 (n=306)	≥80 (n=58)	
All-cause in-hospital mortality, n (%)	1 (0.1)	0 (0)	1 (0.3)	0 (0)	—[a]
Any complications, n (%)	147 (18.7)	72 (17.1)	63 (20.6)	12 (20.7)	.31
Any surgical complication, n (%)	108 (13.8)	57 (13.5)	43 (14.1)	8 (13.8)	.97
Infection	15 (1.9)	11 (2.6)	4 (1.3)	0 (0)	—
Dural tear	12 (1.5)	4 (1)	6 (2)	2 (3.4)	.22
Hemorrhage, hematoma, and seroma	2 (0.3)	2 (0.5)	0 (0)	0 (0)	—
Postoperative anemia	81 (10.3)	41 (9.7)	34 (11.1)	6 (10.3)	.77
Any medical complication, n (%)	57 (7.3)	25 (5.9)	27 (8.8)	5 (8.6)	.10
Pneumonia	3 (0.4)	1 (0.2)	1 (0.3)	1 (1.7)	.01[b]
Pulmonary embolism	2 (0.3)	1 (0.2)	1 (0.3)	0 (0)	—
Acute kidney injury	13 (1.7)	8 (1.9)	5 (1.6)	0 (0)	—
Acute myocardial infarction	2 (0.3)	0 (0)	1 (0.3)	1 (1.7)	—
Retention of urine	43 (5.5)	17 (4)	22 (7.2)	4 (6.9)	.04[b]
Nonroutine discharge[c], n (%)	148 (18.9)	52 (12.4)	76 (24.9)	20 (34.5)	<.001[b]
Length of stay (d), mean (SD)	2.9 (0.1)	2.7 (0.1)	3.2 (0.2)	3.1 (0.1)	.11

[a]Not applicable.

[b]*P*<.05.

[c]Excluding patients who died in the hospital.

Associations Between Age and Inpatient Outcomes

Table 3 shows the univariate and multivariable analyses of the associations between patient and in-hospital outcomes. After adjustment, the multivariable analysis showed that compared with patients aged 60 to 69 years, those aged 70 to 79 years (adjusted OR 2.33, 95% CI 1.57-3.46; *P*<.001) and those aged ≥80 years (adjusted OR 4.79, 95% CI 2.64-8.67; *P*<.001) had a significantly higher risk of nonroutine discharge (Table 3).

Table 3. Associations between any surgical or medical complication, nonroutine discharge, and age.

Outcomes	Age (y)							
	70-79 vs 60-69				≥80 vs 60-69			
	OR[a] (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value	OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
Any surgical complication[b]	1.04 (0.71-1.54)	.83	1.05 (0.72-1.55)	.79	1.02 (0.55-1.90)	.95	0.98 (0.53-1.83)	.95
Any medical complication[c]	1.53 (0.99-2.38)	.06	1.16 (0.69-1.94)	.57	1.49 (0.75-2.96)	.25	0.81 (0.28-2.33)	.69
Nonroutine discharge[d,e]	2.36 (1.66-3.34)	*<.001* [f]	2.33 (1.57-3.46)	*<.001* [f]	3.73 (2.17-6.43)	*<.001*	4.79 (2.64-8.67)	*<.001*

[a]OR: odds ratio.

[b]Adjusted for variables that were significant (*P*<.05) in the univariate analysis, including sex, admission type, spondylolisthesis, hospital location, teaching status, and hospital region.

[c]Adjusted for variables that were significant (*P*<.05) in the univariate analysis, including sex, Medicare or Medicaid, admission type, spondylolisthesis, hospital location, teaching status, renal disease, heart failure, and Charlson Comorbidity Index (CCI).

[d]Adjusted for variables that were significant (*P*<.05) in the univariate analysis, including race, Medicare or Medicaid, admission type, spondylolisthesis, hospital location, teaching status, hospital region, diabetes mellitus, obesity, chronic obstructive pulmonary disease, renal disease, hypertension, coronary heart disease, heart failure, and CCI.

[e]Excluding patients who died in the hospital.

[f]*P*<.05.

XSL•FO
RenderX

In addition, compared to patients aged 60 to 69 years, the LOS was not significantly different in the other 2 age groups after adjustment for confounders in the multivariable analysis (70-79 vs 60-69, adjusted β=0.33, *P*=.14; ≥80 vs 60-69, adjusted β=0.14, *P*=.61). Similar findings were observed when the LOS analysis was restricted to home-discharge patients (Table S2 and S3 in Multimedia Appendix 1). Similar outcome patterns were observed when age was modeled as a continuous variable (Tables S4 and S5 in Multimedia Appendix 1).

Discussion

Principal Findings

This study used the NIS data to determine the impact of age on outcomes following MIS-PLIF in patients aged ≥60 years. The results revealed a low in-hospital mortality rate (only 0.1%) and a complication rate of 18.7%, with no significant differences across age groups and a similar LOS between the 3 age groups. Nonroutine discharge is a binary variable that indicates a deviation from the standard discharge pathway, such as transfer to another hospital with better medical resources, and is often associated with greater clinical severity. Notably, older age was associated with an increased risk of nonroutine discharge, with a 2.3-fold increase in patients aged 70 to 79 years and a 4.8-fold increase in those aged ≥80 years. However, older age did not significantly influence LOS or complication risk. These findings offer valuable insights into the relationship between age and postoperative outcomes after MIS-PLIF, suggesting that while MIS-PLIF is a viable and generally safe option for older patients, extra attention should be given to postoperative care, especially for those aged >70 years.

Minimally invasive procedures have become the method of choice due to decreased surgical trauma and faster recovery. Studies have compared the outcomes of MIS-PLIF with those of open procedures. For example, Mimura et al [18] compared adjacent segment pathology in patients who received a minimally invasive procedure and those who received an open approach for degenerative spondylolisthesis. The results showed that compared to conventional open PLIF, MIS-PLIF was associated with a higher disease-free survival rate and a lower incidence of adjacent segment pathology. Goldstein et al [19] performed a systematic review and meta-analysis examining adverse events of minimally invasive versus open posterior lumbar fusion. All outcomes of interest were similar between the 2 groups, although patients who underwent the minimally invasive procedures had a slightly lower rate of postoperative medical complications. However, the authors noted that the overall quality of the evidence was low.

Among the various outcomes assessed in our study, only nonroutine discharge was significantly associated with older age. Other studies have also reported that increasing age is associated with higher rates of nonroutine discharge in patients undergoing lumbar fusion; however, these studies did not further stratify older patients by age [20,21]. While there can be many reasons for this association, a possible reason is that older patients are in general more likely to be transferred to specialized care facilities after discharge because they have greater needs than their younger counterparts [22]. In addition, patients with a more advanced age may need more care than family members can provide at home or specialized rehabilitation services that are available in such facilities.

While advanced age is associated with adverse events for many surgical procedures [12,23], our results clearly showed that increased age did not significantly affect LOS or complication rate following MIS-PLIF. Luo et al [24] compared the surgical outcome of multilevel anterior cervical discectomy and fusion in older and younger patients with myelopathy. Overall, outcomes were similar between different age groups, although the older patients tended to have a lower recovery rate and higher short-term complication rate. Liu et al [25] reported that in patients undergoing PLIF, factors associated with the need for a perioperative blood transfusion were 3 or more fusion segments, low hemoglobin preoperatively, hypertension, lower BMI, and more advanced age. Our results challenge common perceptions about the risks associated with surgical procedures in older populations and highlight the potential viability of MIS-PLIF for patients of advanced age, even for those aged >80 years.

Clinical Implications

The findings of this study have important clinical implications for the management of older patients undergoing MIS-PLIF. Notably, while MIS-PLIF is generally safe for older adults, the increased likelihood of nonroutine discharges in patients aged ≥70 years underscores the need for tailored postoperative care plans. Clinicians should consider enhanced recovery protocols and more comprehensive discharge planning to address the specific needs of older patients. This could involve early involvement of multidisciplinary teams, including geriatric specialists, physical therapists, and social workers, to facilitate smoother transitions to either home or rehabilitative settings.

Strengths and Limitations

The primary strength of this study lies in its analysis of a large and nationally representative patient database, which enhances the generalizability of the conclusions to the broader population in the United States. This extensive scope ensures that the findings reflect a wide demographic profile, providing valuable insights applicable to the general population. The study's relatively large sample size enabled the assessment of various complications, including infrequent complications.

Nevertheless, this study has several limitations. This was a retrospective study, which may have introduced selection bias. As with other studies relying on *ICD* coding systems, this study may have been affected by potential coding inaccuracies. Additionally, it lacks long-term postdischarge follow-up data and preadmission status information, which are essential for a comprehensive assessment of outcomes. Furthermore, comorbidities may vary in severity; however, the NIS database lacks detailed clinical parameters, such as glomerular filtration rate, glycated hemoglobin levels, and precise BMI values, limiting our ability to assess severity or perform precise staging for conditions such as diabetes, chronic kidney disease, and obesity. Complications were identified throughout the hospitalization period, without the ability to distinguish between intraoperative and postoperative events. Preoperative physical

XSL•FO
RenderX

performance status, such as the American Society of Anesthesiologists Physical Status Classification, was not available. We could not confirm the timing of postoperative infections due to the absence of time-stamped data. Certain outcomes, such as changes in neurological function, altered mental status, or pain, are not captured in this administrative database, as it lacks such granular clinical and laboratory details. Minor complications, such as intolerance to pain medication, are also difficult to accurately identify, as they may not be consistently coded in electronic medical records. Although this study provides a broad overview of the outcomes using a large cohort, these limitations warrant cautious interpretation of the findings and highlight the need for future clinical studies to address more detailed outcome measures.

Conclusions

In conclusion, this study on the impact of age on MIS-PLIF outcomes demonstrated that the procedure is relatively safe for patients aged ≥60 years, as evidenced by the low in-hospital mortality rate and low complication rate, regardless of age. However, patients aged ≥70 years had a higher risk of nonroutine discharge. Although age alone did not significantly affect LOS and complications, targeted postoperative care for older patients, particularly those aged ≥70 years, is still recommended.

These findings support the notion that MIS-PLIF is a viable surgical option for older adults, including those aged ≥80 years. Importantly, age alone should be viewed as a factor that necessitates tailored postoperative strategies. Implementing age-stratified management pathways—including early engagement of geriatric care teams, customized rehabilitation plans, and structured discharge protocols—may optimize recovery, reduce health care burden, and enhance patient outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: YJL
Data curation: YJL, FYS
Formal analysis: JFH
Funding acquisition: YJL
Investigation: CWT
Methodology: YJL, FYS, JFH
Project administration: YCT, LC
Resources: YJL
Software: JFH, CWT
Supervision: YHH, MJC
Validation: YJL, LC
Visualization: YCT, LC
Writing – original draft: YJL, FYS, JFH, CWT
Writing – review and editing: YHH, MJC
All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Tables S1 to S5.
[DOCX File , 24 KB - medinform_v14i1e76424_app1.docx]

References

1. Logroscino CA, Proietti L, Pola E, Scaramuzzo L, Tamburrelli FC. A minimally invasive posterior lumbar interbody fusion for degenerative lumbar spine instabilities. Eur Spine J 2011 May 29;20 Suppl 1(Suppl 1):S41-S45 [FREE Full text] [doi: 10.1007/s00586-011-1762-1] [Medline: 21445617]

XSL•FO
RenderX

2. Sidhu GS, Henkelman E, Vaccaro AR, Albert TJ, Hilibrand A, Anderson DG, et al. Minimally invasive versus open posterior lumbar interbody fusion: a systematic review. Clin Orthop Relat Res 2014 Jun;472(6):1792-1799. [doi: 10.1007/s11999-014-3619-5] [Medline: 24748069]
3. Xu DS, Walker CT, Godzik J, Turner JD, Smith W, Uribe JS. Minimally invasive anterior, lateral, and oblique lumbar interbody fusion: a literature review. Ann Transl Med 2018 Mar;6(6):104. [doi: 10.21037/atm.2018.03.24] [Medline: 29707553]
4. Wu PH, Kim HS, Jang IT. Intervertebral disc diseases PART 2: a review of the current diagnostic and treatment strategies for intervertebral disc disease. Int J Mol Sci 2020 Mar 20;21(6):2135 [FREE Full text] [doi: 10.3390/ijms21062135] [Medline: 32244936]
5. Grote C, Reinhardt D, Zhang M, Wang J. Regulatory mechanisms and clinical manifestations of musculoskeletal aging. J Orthop Res 2019 Jul 03;37(7):1475-1488 [FREE Full text] [doi: 10.1002/jor.24292] [Medline: 30919498]
6. Sharma N, Reche A. Unraveling the relationship between osteoporosis, treatment modalities, and oral health: a comprehensive review. Cureus 2023 Nov;15(11):e49399 [FREE Full text] [doi: 10.7759/cureus.49399] [Medline: 38146583]
7. Agarwal N, Goldschmidt E, Taylor T, Roy S, Dunn SC, Bilderback A, et al. Impact of frailty on outcomes following spine surgery: a prospective cohort analysis of 668 patients. Neurosurgery 2021 Feb 16;88(3):552-557. [doi: 10.1093/neuros/nyaa468] [Medline: 33372214]
8. Al Jammal OM, Shahrestani S, Delavar A, Brown NJ, Gendreau JL, Lien BV, et al. Demographic predictors of treatments and surgical complications of lumbar degenerative diseases: an analysis of over 250,000 patients from the National Inpatient Sample. Medicine (Baltimore) 2022 Mar 18;101(11):e29065 [FREE Full text] [doi: 10.1097/MD.0000000000029065] [Medline: 35356929]
9. Damluji AA, Forman DE, van Diepen S, Alexander KP, Page RL2, Hummel SL, et al. Older adults in the cardiac intensive care unit: factoring geriatric syndromes in the management, prognosis, and process of care: a scientific statement from the American Heart Association. Circulation 2020 Jan 14;141(2):e6-32 [FREE Full text] [doi: 10.1161/CIR.0000000000000741] [Medline: 31813278]
10. Mohan S, Cha ED, Lynch CP, Geoghegan CE, Jadczak CN, Singh K. Impact of advanced age on postoperative outcomes following transforaminal lumbar interbody fusion. J Am Acad Orthop Surg 2021 May 17;29(17):e869-e879. [doi: 10.5435/jaaos-d-20-01382]
11. Tong Y, Fernandez L, Bendo JA, Spivak JM. Enhanced recovery after surgery trends in adult spine surgery: a systematic review. Int J Spine Surg 2020 Aug 31;14(4):623-640 [FREE Full text] [doi: 10.14444/7083] [Medline: 32986587]
12. Watt J, Tricco AC, Talbot-Hamon C, Pham B, Rios P, Grudniewicz A, et al. Identifying older adults at risk of harm following elective surgery: a systematic review and meta-analysis. BMC Med 2018 Jan 12;16(1):2 [FREE Full text] [doi: 10.1186/s12916-017-0986-2] [Medline: 29325567]
13. Healthcare cost and utilization project user support. Overview of the Nationwide Inpatient Sample (NIS). URL: https://hcup-us.ahrq.gov/nisoverview.jsp [accessed 2023-12-12]
14. HCUP-US central distributor availability of databases. Agency for Healthcare Research and Quality. URL: https://hcup-us.ahrq.gov/db/availability_public.jsp [accessed 2025-05-29]
15. Garfein AJ, Herzog AR. Robust aging among the young-old, old-old, and oldest-old. J Gerontol B Psychol Sci Soc Sci 1995 Mar 01;50(2):S77-S87. [doi: 10.1093/geronb/50b.2.s77] [Medline: 7757843]
16. Reinert SS, Kinney AL, Jackson K, Diestelkamp W, Bigelow K. Age stratification and sample entropy analysis enhance the limits of stability tests for older adults. J Appl Biomech 2017 Dec 01;33(6):419-423. [doi: 10.1123/jab.2016-0019] [Medline: 28605220]
17. NIS Description of data elements. Agency for Healthcare Research and Quality. URL: https://hcup-us.ahrq.gov/db/vars/zipinc_qrtl/nisnote.jsp [accessed 2025-05-29]
18. Mimura T, Tsutsumimoto T, Yui M, Takahashi J, Kuraishi S, Misawa H. Adjacent segment pathology following posterior lumbar interbody fusion for lumbar degenerative spondylolisthesis: a comparison between minimally invasive and conventional open approach. Spine J 2021 Aug;21(8):1297-1302. [doi: 10.1016/j.spinee.2021.03.027] [Medline: 33785474]
19. Goldstein CL, Macwan K, Sundararajan K, Rampersaud YR. Perioperative outcomes and adverse events of minimally invasive versus open posterior lumbar fusion: meta-analysis and systematic review. J Neurosurg Spine 2016 Mar;24(3):416-427. [doi: 10.3171/2015.2.SPINE14973] [Medline: 26565767]
20. Best MJ, Buller LT, Falakassa J, Vecchione D. Risk factors for nonroutine discharge in patients undergoing spinal fusion for intervertebral disc disorders. Iowa Orthop J 2015;35:147-155 [FREE Full text] [Medline: 26361458]
21. Cabrera A, Bouterse A, Nelson M, Razzouk J, Ramos O, Bono CM, et al. Accounting for age in prediction of discharge destination following elective lumbar fusion: a supervised machine learning approach. Spine J 2023 Jul;23(7):997-1006. [doi: 10.1016/j.spinee.2023.03.015] [Medline: 37028603]
22. Sokas CM, Cowan J, Dalton MK, Coogan K, Bader A, Bernacki R, et al. Association between patient-reported frailty and non-home discharge among older adults undergoing surgery. J Am Geriatr Soc 2020 Dec 08;68(12):2909-2913. [doi: 10.1111/jgs.16846] [Medline: 33031587]
23. Massarweh NN, Legner VJ, Symons RG, McCormick WC, Flum DR. Impact of advancing age on abdominal surgical outcomes. Arch Surg 2009 Dec 21;144(12):1108-1114. [doi: 10.1001/archsurg.2009.204] [Medline: 20026827]

XSL•FO
RenderX

24. Luo CA, Lim AS, Lu ML, Chiu PY, Lai PL, Niu CC. The surgical outcome of multilevel anterior cervical discectomy and fusion in myelopathic elderly and younger patients. Sci Rep 2022 Mar 16;12(1):4495 [FREE Full text] [doi: 10.1038/s41598-022-08243-8] [Medline: 35296700]
25. Liu B, Pan J, Zong H, Wang Z. Establishment and verification of a perioperative blood transfusion model after posterior lumbar interbody fusion: a retrospective study based on data from a local hospital. Front Surg 2021 Aug 30;8:695274. [doi: 10.3389/fsurg.2021.695274] [Medline: 34527694]

Abbreviations

HCUP: Health Care Cost and Utilization Project
ICD-9-CM: International Classification of Diseases, Ninth Revision
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
LOS: length of stay
MIS-PLIF: minimally invasive posterior lumbar interbody fusion
NIS: Nationwide Inpatient Sample
OR: odds ratio

Edited by A Benis; submitted 23.Apr.2025; peer-reviewed by T Asada, P Menon Naliyatthaliyazchayil; comments to author 16.Jun.2025; revised version received 31.Jul.2025; accepted 30.Nov.2025; published 06.Jan.2026.

Please cite as:
Lin YJ, Shih FY, Huang JF, Ting CW, Tsai YC, Chang L, Huang YH, Chuang MJ
Impact of Age on Hospital Outcomes Following Minimally Invasive Posterior Lumbar Interbody Fusion: Retrospective Analysis of the Nationwide Inpatient Sample Database from 2016 to 2020
JMIR Med Inform 2026;14:e76424
URL: https://medinform.jmir.org/2026/1/e76424
doi:10.2196/76424
PMID:

©Yu-Jun Lin, Fu-Yuan Shih, Jin-Fu Huang, Chun-Wei Ting, Yu-Chin Tsai, Lin Chang, Yu-Hua Huang, Ming-Jung Chuang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mortality Prediction Among People Living With HIV on Antiretroviral Therapy in Public Health Facilities in Gondar City Administration, Northwest Ethiopia: Machine Learning–Based Study

Andualem Enyew Gedefaw[1], MSc; Getaye Tizazu Biwota[2], MPH; Abraraw Gebre Mesele[1], MPH; Abraham Keffale Mengistu[2], MSc; Gizaw Hailiye Teferi[2], MPH

[1]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, P.O. Box 196, Hospita, Gondar, Ethiopia

[2]Department of Health Informatics, College of Medicine and Health Sciences, Debre Markos University, Debre Markos, Ethiopia

Corresponding Author:
Andualem Enyew Gedefaw, MSc
Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, P.O. Box 196, Hospita, Gondar, Ethiopia

Abstract

Background: Predicting mortality among people living with HIV enables clinicians to implement timely, targeted, and preventive interventions at the start of antiretroviral therapy (ART). However, prognostic models must rely strictly on baseline predictors to avoid look-ahead bias and ensure scientific validity. This study evaluates machine-learning (ML) algorithms for baseline mortality prediction using routine electronic medical record data.

Objective: This study aims to predict mortality among people living with HIV receiving ART using baseline clinical and sociodemographic characteristics through ML models in public health facilities of Gondar City Administration, Northwest Ethiopia.

Methods: The retrospective cohort study was conducted using electronic medical record data from 12,871 people living with HIV on ART (2005 - 2024). Seven base classifiers were evaluated using stratified 10-fold cross-validation. Synthetic minority oversampling technique (SMOTE)–balanced variants were used only for sensitivity analysis. SMOTE oversampling was applied only to training folds; the final evaluation used the original imbalanced test data. Shapley Additive Explanations (SHAP) analysis identified key baseline predictors.

Results: Gradient boosting on the original data achieved superior performance (accuracy 87.0%, F_1-score 0.619, area under the receiver operating characteristic curve 0.859), outperforming extreme gradient boosting (F_1-score 0.609, area under the receiver operating characteristic curve 0.835) and SMOTE variants. The SHAP analysis identified education level, lack of formal education (+0.84), and a low baseline cluster of differentiation 4 (CD4; a type of immune cell count) count of 140 cells/mm³ (+0.54) as substantially increasing predicted mortality risk. Urban residence (−0.35) and working functional status (−0.12) showed protective effects, whereas age (45 y; −0.02) had minimal influence in the illustrated case. Globally, lower CD4 counts and the absence of formal education were consistently associated with increased mortality risk.

Conclusions: Ensemble ML models demonstrated moderate-to-strong discrimination for predicting mortality among people living with HIV using strictly baseline routine electronic medical record data. SHAP-based interpretability confirmed that educational attainment and baseline CD4 count were the dominant determinants of predicted mortality risk, underscoring the combined influence of socioeconomic vulnerability and immunological status at ART initiation. These findings support the potential utility of interpretable ML models for early risk stratification and targeted clinical decision-making in resource-limited settings; however, external validation is required before routine clinical implementation.

(JMIR Med Inform 2026;14:e78770) doi:10.2196/78770

KEYWORDS

HIV mortality prediction; machine learning; ensemble algorithms; Shapley Additive Explanations; SHAP analysis; antiretroviral therapy; baseline predictors; electronic medical records

XSL•FO
RenderX

Introduction

Recent advances in machine learning (ML) have enabled the development of predictive models using routinely collected electronic medical record (EMR) data to support clinical decision-making and risk stratification. Several studies demonstrated the feasibility of applying ML algorithms, such as logistic regression, random forest, gradient boosting, and extreme gradient boosting (XGBoost), to predict mortality and other adverse outcomes across diverse clinical settings using EMR or electronic health record data. Systematic reviews highlighted that ML-based mortality prediction models often outperform traditional statistical approaches, while also emphasizing the importance of appropriate validation strategies and the prevention of data leakage through baseline-only predictors [1-3]. EMR-based ML models have been successfully applied to predict short-term and in-hospital mortality among patients with conditions, such as cancer, heart failure, and COVID-19, using structured clinical and laboratory variables [4-6]. These studies underscore the growing role of explainable artificial intelligence (XAI) methods, such as SHAP (Shapley Additive Explanations), in improving the interpretability and clinical relevance of complex ML models, which is a key consideration for adoption in real-world health care settings [5,7].

HIV remains a major public health concern globally and in sub-Saharan Africa, where mortality among people living with HIV persists despite widespread ART coverage. Since the early 2000s, when the global HIV epidemic peaked, new infections and HIV-related deaths have declined. Expanded antiretroviral therapy (ART) coverage has reduced HIV-related mortality by almost 47%, from 2.1 million in 2004 to 630,000 in 2023 [8]. However, the epidemic remains a significant global health issue, especially in regions with limited access to preventive, diagnostic, and treatment services. Currently, an estimated 39.9 million individuals are living with HIV, according to the UNAIDS (Joint United Nations Program on HIV/AIDS) 2024 Global HIV and AIDS Statistics Report.

Sub-Saharan Africa bears the highest global HIV burden, accounting for approximately 67% of HIV cases and over 60% of HIV-related deaths [8]. Southern and Eastern African nations have made progress in reducing mortality and increasing access to ART. However, challenges, such as poverty, stigma, violence, and health care inequality, hinder efforts to decrease HIV-related fatalities. In 2023, approximately 390,000 Africans died from HIV-related causes. While ART rollouts have significantly reduced these numbers, the disease remains a serious public health concern.

Ethiopia has one of the largest HIV epidemics in East Africa, with approximately 610,000 people living with HIV in 2023 [2]. Despite improvements in health care infrastructure and ART coverage, significant challenges persist, especially for individuals with advanced HIV or treatment failure. In 2023, Ethiopia recorded approximately 10,000 HIV-related deaths, with rural areas facing higher mortality due to limited health care access [9]. HIV-related mortality remains high in sub-Saharan Africa, including Ethiopia, even though ART has transformed HIV into a chronic, manageable condition. Key contributors include late ART initiation, poor adherence, medication resistance, and virological failure, which prevent effective viral suppression [10-12]. Virological failure accelerates disease progression and increases the risk of AIDS-related death. Reducing mortality among people living with HIV requires early diagnosis, improved ART regimens, and better patient monitoring.

A 2022 study at the University of Gondar Comprehensive and Specialized Hospital reported a virological failure rate of 14% among people living with HIV on ART [12]. This highlights the urgent need for predictive tools to identify high-risk patients early and improve clinical outcomes. Mortality prediction plays a crucial role in health care, epidemiology, insurance, and policymaking. It helps health care providers identify high-risk individuals, enabling early interventions and personalized treatments that improve patient outcomes [13]. It also supports epidemiological research by analyzing disease patterns and assessing intervention effectiveness [8. Advanced methodologies, such as statistical models, ML, and deep learning, have significantly improved the accuracy and applicability of mortality predictions [14,15].

ML offers a data-driven solution for predicting mortality risk in people living with HIV on ART. ML algorithms analyze large datasets to detect complex patterns, risk factors, and interactions that traditional statistical methods may overlook [16]. This study develops interpretable ML models using strictly baseline predictors (recorded at ART initiation) from 12,871 people living with HIV EMRs to enable prospective mortality risk stratification. There is a pressing need to implement advanced predictive tools, such as ML models, to identify individuals at greater risk of treatment failure or death [17]. These tools can enable health care providers to take proactive measures, improve treatment outcomes, and reduce HIV-related mortality in Ethiopia.

This study aimed to develop and validate ensemble ML models predicting mortality among people living with HIV on ART in Gondar, Ethiopia, using baseline-only EMR data, with the SHAP analysis to identify clinically actionable risk factors. Several studies in Africa and Asia have applied ML for HIV-related outcomes, including mortality and treatment interruption. A meta-analysis across 24 studies found that ML models achieved a C-index of ~0.83 for HIV mortality prediction [18]. In China, an XGBoost and random forest ensemble achieved an area under the receiver operating characteristic (ROC) curve (AUC) of ~0.98 for in-hospital mortality among patients with HIV/AIDS with cryptococcal infection [19]. In Nigeria, a large EMR-based ML study of 41,394 people living with HIV reported an AUC of ~0.8 for treatment interruption [20]. However, these studies often used hospital-based cohorts or follow-up predictors, focused on treatment interruption or virological failure rather than long-term mortality, and lacked comprehensive interpretability analyses. Our study addresses these gaps through large-scale public health facility EMR data (n=12,871), strict baseline-only prediction, rigorous temporal validation, and SHAP-based clinical interpretability.

XSL•FO
RenderX

Traditional methods of monitoring and predicting patient outcomes rely on periodic clinical assessments, laboratory tests, and physician judgment. These approaches often detect complications only after symptoms have worsened, leading to delayed interventions that may not prevent adverse outcomes. Additionally, reliance on manual record keeping and retrospective data analysis limits the ability to identify high-risk patients early, reducing the effectiveness of timely medical interventions. In resource-constrained settings, such as the Gondar City Administration, integrating ML models could improve health care delivery, optimize resource distribution, and enhance proactive patient management [21]. This approach aligns with Ethiopia's national HIV/AIDS control strategy and global health goals aimed at reducing HIV-related mortality.

Methods

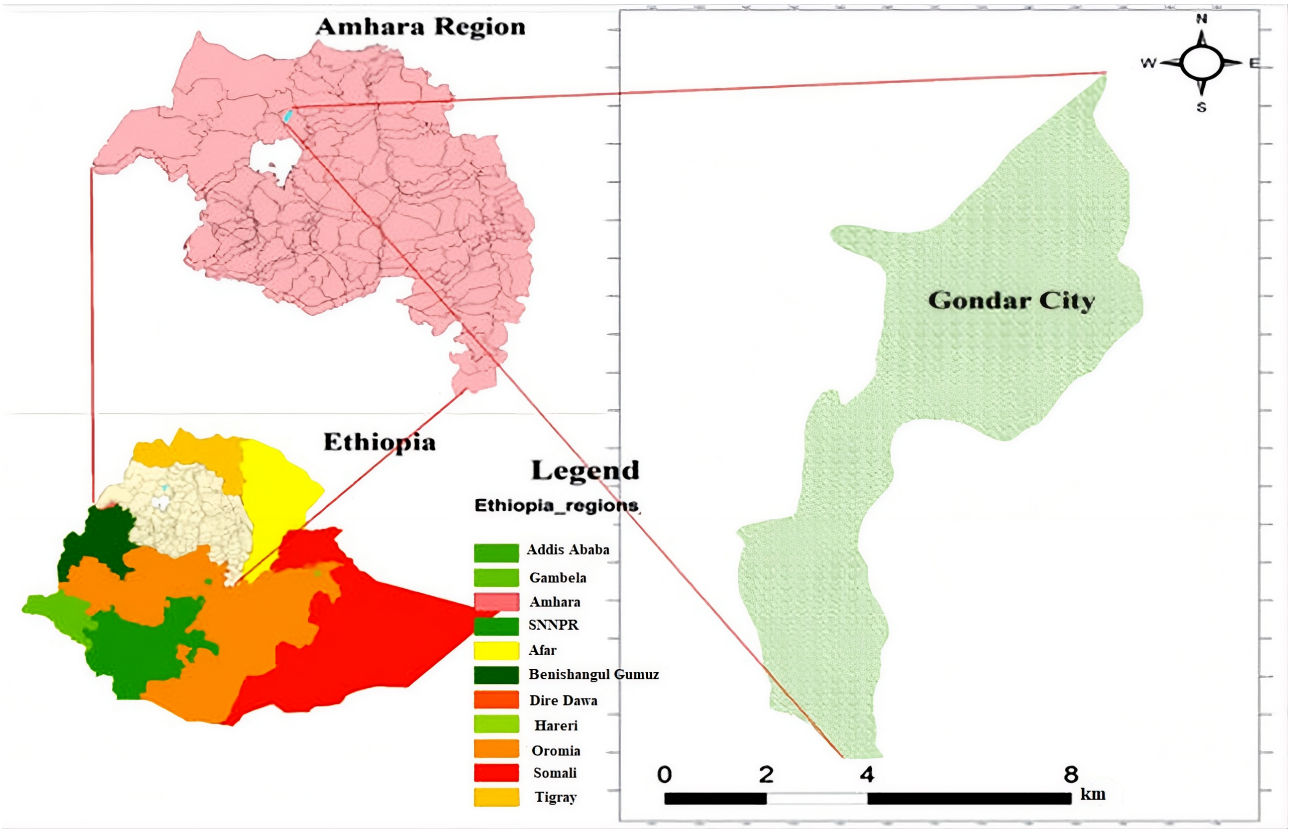
Study Area and Period

Data were extracted from the ART EMRs covering the period from 2005 to 2024. Only patients on ART for ≥6 months before analysis were included to ensure baseline predictor availability.

This study was conducted at public health facilities in Gondar City Administration (Amhara Region, Ethiopia) from October 10 to December 10, 2024. Gondar City (population 457,938) is located 748 km northwest of Addis Ababa. In the Gondar City Administration, several health care facilities provide ART services to people living with HIV. Among them, the University of Gondar Specialized and Comprehensive Hospital serves as a key provider of ART services, offering specialized care to patients. Additionally, multiple health centers, including Azezo Health Center, Gondar Health Center, Maraki Health Center, Mintwab Health Center, St. Gebriel Health Center, Teda Health Center, and Woleka Health Center, also provide ART services, ensuring accessibility to treatment at various levels of health care.

Health facility mapping was conducted to provide context to the data source; however, facility service gaps were not the focus of this predictive modeling study (see Figure 1).

Figure 1. Geographic location of the study area showing the Gondar City Administration, Amhara National Regional State, Northwest Ethiopia, where a retrospective cohort study of adults living with HIV receiving antiretroviral therapy (ART) was conducted using electronic medical records from public health facilities, 2005 - 2024. The map of the study area was adopted from Tamiru AT, Rade BK et al. (2020) [22].

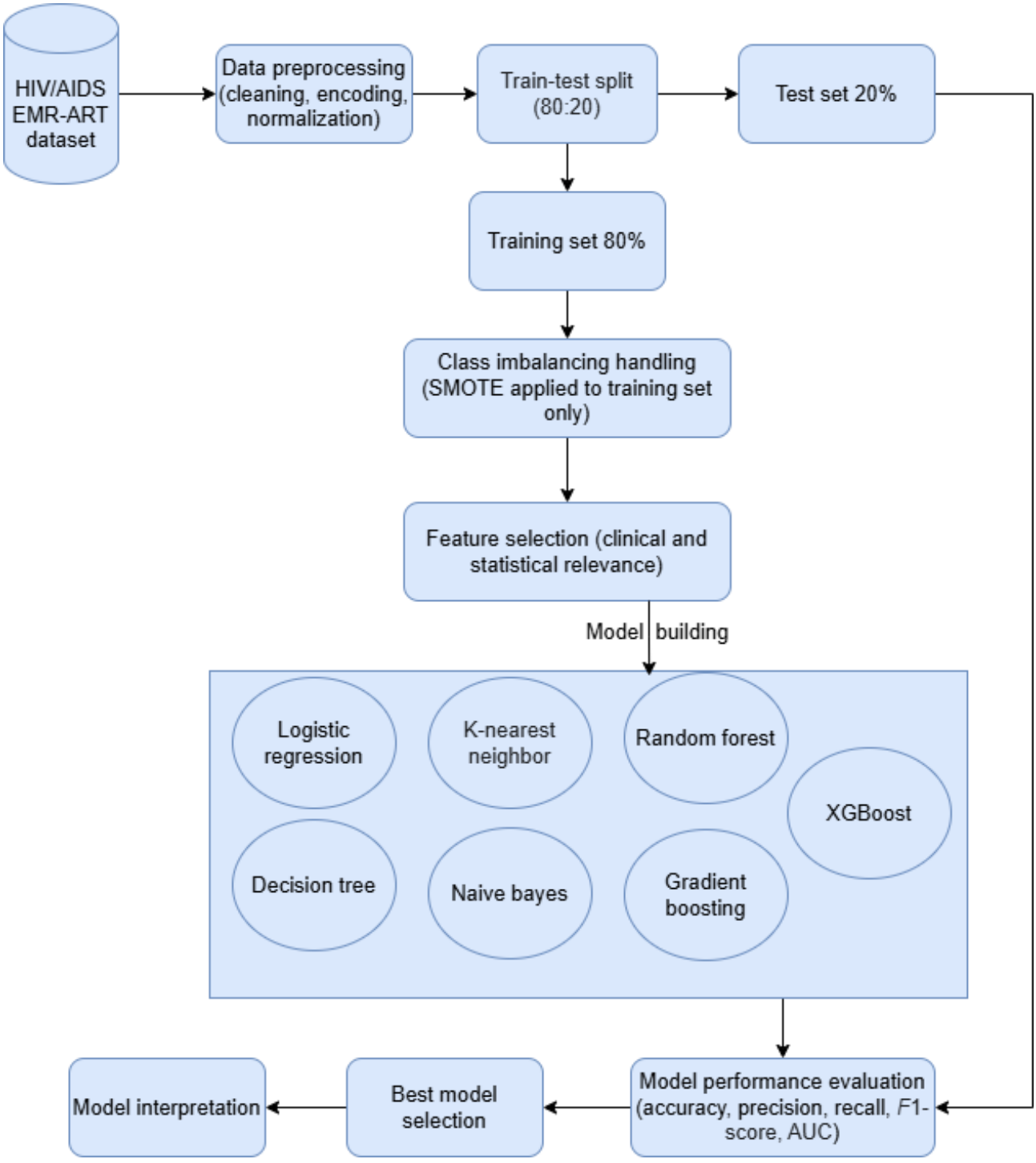


Study Design

The retrospective cohort study analyzed baseline-only EMR data (ART initiation records) to predict prospective mortality (see Figure 2).

XSL•FO
RenderX

Figure 2. Workflow diagram of the machine-learning pipeline used to predict mortality among adults living with HIV on antiretroviral therapy (ART) in public health facilities of Gondar City Administration, Northwest Ethiopia, 2024. AUC: area under the receiver operating characteristic curve; EMR: electronic medical record; SMOTE: synthetic minority oversampling technique; XGBoost: extreme gradient boosting.



Data Source

Data are sourced from EMR-ART databases across Gondar City Administration public health facilities, containing baseline sociodemographic, clinical, laboratory, and treatment initiation data.

Description of the Dataset

The dataset used in this study focused on people living with HIV undergoing ART in public health facilities within the Gondar City Administration, Ethiopia. Data collection covered the entire period since the start of ART services at the facilities.

Inclusion and Exclusion Criteria

Eligible participants were adults aged ≥18 years of age with ≥6 months of ART exposure. Participants were included if baseline variables were recorded at ART initiation; limited missing baseline values (<10%) were addressed using imputation. We excluded patients with incomplete ART start dates, those transferred to other facilities without outcome documentation,

XSL•FO
RenderX

and non-HIV–related deaths (eg, accidents and trauma) to ensure outcome validity.

Predictor Variables

The binary outcome was mortality (follow-up status: alive=0, dead=1). Baseline predictors (all recorded at ART initiation) included 10 variables across 3 domains: sociodemographic (age, sex, marital status, educational level, residence, and religion), clinical (tuberculosis [TB] screening result and functional status), and immunological (baseline cluster of differentiation 4 [CD4] count). The baseline CD4 count was modeled as a continuous numerical feature in all ML algorithms. Descriptive statistics are therefore presented as median and IQR. Variable selection used recursive feature elimination with cross-validation applied exclusively to training data. The subsequent SHAP analysis confirmed the clinical plausibility of retained predictors.

Operational Definitions

Mortality, the outcome variable of interest in the study, is defined as the death of people living with HIV receiving ART. This is directly or indirectly related to the progression of HIV, opportunistic infections, or complications arising from the virus or treatment.

A low CD4 count of less than 200 cells/mm³ indicates a severely weakened immune system, placing the individual at high risk for opportunistic infections and HIV-related complications. A normal CD4 count between 200 and 500 cells/mm³ suggests a moderately functioning immune system. A high CD4 count greater than 500 cells/mm³ is considered within the normal range and reflects a healthier immune system.

Comorbidities are the presence of other chronic health conditions in people living with HIV alongside HIV infection.

Data Collection Tool and Procedures

The data collection tool used in this study was an electronic data extraction form specifically designed to retrieve relevant patient information from the EMR-ART database. The form was developed via a standardized template based on key study variables to ensure data consistency, completeness, and accuracy. The data collection process began with obtaining ethical approval and permission from relevant health authorities. Data collectors were trained in using the extraction tool and procedures to ensure uniformity in data retrieval. Data were extracted by reviewing patient records from the EMR-ART database using selected features and exported into Excel. Each record was thoroughly reviewed for accuracy, and a second data collector cross-verified entries to maintain data consistency.

Ethical Considerations

Ethical approval for this study was obtained from the College of Medicine and Health Sciences Institutional Research Ethics Review Committee (CMHS IRERC) at Debre Markos University (reference number RCSTTD/403/01/17). Due to the retrospective design, the requirement for informed consent was waived by the committee in accordance with national research ethics guidelines.

All data used were deidentified before analysis to ensure participant privacy and confidentiality. No personally identifiable information (eg, names or medical record numbers) was extracted. Data were stored on password-protected computers accessible only to authorized researchers.

There was no financial compensation to participants, as this study relied on existing EMRs. No images or identifiable information about individual participants are presented in this paper. Ethical approval was obtained from the institutional review board of the Amhara Public Health Institute (APHI) (approval number: APHI/322/007). The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and complied with the Ethiopian National Research Ethics Review Guideline [23].

Data Quality Control

Data quality control procedures were applied throughout the study to ensure reliability and validity. These procedures involved consistency checks and the correction of missing entries. Data entry was supervised by experienced clerks, and predefined validation rules flagged inconsistencies.

Data Management and Analysis

Association Rule Mining

Association rule mining was conducted to identify frequent co-occurrence patterns among clinical and treatment-related factors associated with mortality. The Apriori algorithm implemented in the mlxtend Python library was used. The mlxtend Python library was developed by Sebastian Raschka and provides useful extensions for machine learning and data science tasks. Key metrics included support (frequency of rule occurrence), confidence (conditional probability of the consequent given the antecedent), and lift (degree of dependence between antecedent and consequent). Rules with support ≥0.065, confidence ≥0.64, and lift >4.6 were retained for interpretation. These analyses were used to explore unadjusted co-occurrence patterns and were not intended to estimate causal or adjusted effects.

Data Cleaning

This phase involved handling missing values, detecting and removing outliers via Excel filters, and addressing class imbalance. The dataset contained less than 10% missing values for most variables. Median imputation was applied to numerical variables, such as CD4 count, while mode imputation was used for categorical baseline variables, such as marital status, residence, and religion. These simple imputation methods were chosen due to the relatively small proportion of missingness and to maintain data interpretability whereas mode imputation was applied to categorical data. One-hot encoding transforms categorical variables, creating separate columns for each category. To handle class imbalance, various techniques have been tested, with SMOTE (synthetic minority oversampling technique) providing the best accuracy. Other balancing techniques, such as random undersampling and cost-sensitive learning, were initially tested. However, they resulted in reduced sensitivity for the minority class (deceased patients). SMOTE provided the best balance between precision and recall, ensuring robust generalization across folds.

XSL•FO
RenderX

Feature Engineering

By converting raw data into meaningful features, feature engineering enhances model performance. One-hot encoding was used for categorical variables, and normalization ensured comparable feature scales [24].

Feature Selection and Dimensionality Reduction

Feature selection was performed exclusively using recursive feature elimination with cross-validation on the training data to identify the most informative subset of predictors while minimizing redundancy and overfitting.

Data Splitting

A training-test split and K-fold cross-validation ensured robust evaluation. The dataset was divided into training and testing sets, with K-fold validation reducing overfitting and improving generalizability.

Model Training and Evaluation

To prevent data leakage, the dataset was first split into training and hold-out test sets. All preprocessing steps, including imputation, feature scaling, encoding, and feature selection, were conducted exclusively within the training data using a cross-validation framework. Model performance was assessed using stratified 10-fold cross-validation on the training set and subsequently evaluated on the original, imbalanced hold-out test set. Sensitivity analyses were conducted by applying SMOTE to the training data only to assess the robustness of model performance under class imbalance.

Data Leakage Assessment and Stress-Test Validation

To ensure the robustness and validity of the predictive modeling framework, a series of predefined stress tests were conducted to assess potential data leakage, proxy outcome variables, and look-ahead bias.

Stress Test: Broken Split Check (Demographic-Only Model)

To verify independence between training and testing datasets, an XGBoost model was trained using only demographic variables (age and sex). The resulting model achieved an AUC of 0.631 on the hold-out test dataset, consistent with expected performance for demographic-only mortality prediction models. This finding confirmed that the data split procedure did not introduce artificial performance inflation due to overlapping observations or improper sampling procedures.

Model Selection

After the model was trained, several classifiers were evaluated to identify the most suitable model for predicting mortality among people living with HIV users. Given that the outcome variable is categorical and falls into 2 mutually exclusive groups, the problem was framed as a binary classification task. To select the best model, we compared the performance of these classifiers via evaluation metrics, such as accuracy, precision, recall, F_1-score, and AUC [25]. During cross-validation, we examined the consistency of each model's performance across different subsets of the data. The model with the best trade-off between predictive power and generalization to unseen data was chosen for final model selection. This approach ensured that the selected model was the right model for accurately predicting mortality among people living with HIV.

Evaluation Criteria

In this study, the performance of the predictive models was evaluated by testing a dataset within a training-test split and cross-validation. The performance of the trained models was subsequently evaluated on the test set based on the criteria of accuracy score, ROC curve, precision, recall, and *F*-measure. The confusion matrix, which is a matrix of N×N, where N is the number of predicted classes, displaying the number of correct and incorrect predictions made by the classification model, was used in this study. Because mortality prevalence was 20%, F_1-score and AUC were prioritized over accuracy for model comparison.

Results

Description of Sociodemographic Characteristics

A total of 12,871 study participants were included in this study; the majority were between the ages of 38 and 47 years (n=4468, 34.7%), suggesting a largely middle-aged population; 7688 (59.7%) of the sample were female, indicating a gender imbalance; the majority were married (n= 4921, 51.2%) and had completed secondary school education (n=2953, 31.3%), indicating a moderate level of education; a sizable portion lived in urban areas (n=8878, 77.1%), suggesting greater representation of urban areas; and the vast majority were Orthodox Christians (n=8714, 91.5%), representing a highly homogeneous religious composition (see Table 1).

XSL•FO
RenderX

Table . Sociodemographic characteristics of adult people living with HIV receiving antiretroviral therapy in public health facilities of Gondar City Administration, Northwest Ethiopia, based on a retrospective cohort extracted from electronic medical records, 2005 - 2024 (N=12,871).

Features and categories	Values, n (%)
Age, y	
18 - 27	783 (6.1)
28 - 37	1741 (13.5)
38 - 47	4468 (34.7)
48 - 57	3998 (31.1)
58 and above	1881 (14.6)
Sex	
Female	7688 (59.7)
Male	5183 (40.3)
Marital Status	
Married	4921 (51.2)
Divorced	2243 (23.3)
Never married	1449 (15.1)
Widowed	995 (10.4)
Education level	
Higher education	1448 (15.3)
Secondary education	2953 (31.3)
Primary education	2602 (27.5)
No education	2443 (25.9)
Residence	
Urban	8878 (77.1)
Rural	2635 (22.9)
Religion	
Orthodox	8714 (91.5)
Muslim	684 (7.2)
Protestant	96 (1)
Catholic	23 (0.24)
Other	6 (0.06)

Baseline Clinical and Immunological Characteristics

Among 12,871 people living with HIV at ART initiation, most exhibited preserved functional status (working: n=11,970, 93.1%) with low TB prevalence (n=219, 1.7% positive), consistent with late ART presentation in Ethiopian cohorts (see Table 2). The median baseline CD4 count was 225 (IQR 131 - 367) cells/mm³.

XSL•FO
RenderX

Table . Baseline clinical and immunological characteristics of people living with HIV receiving antiretroviral therapy in public health facilities, Gondar City Administration, Northwest Ethiopia, 2024.

Feature and category	Values
Tuberculosis screening result, n (%)	
No	12,655 (98.3)
Yes	216 (1.7)
Functional status, n (%)	
Working	11,978 (93.1)
Ambulatory	527 (4.1)
Bedridden	366 (2.8)
Baseline CD4[a] count (cells/mm³), median (IQR)	225 (131-367)

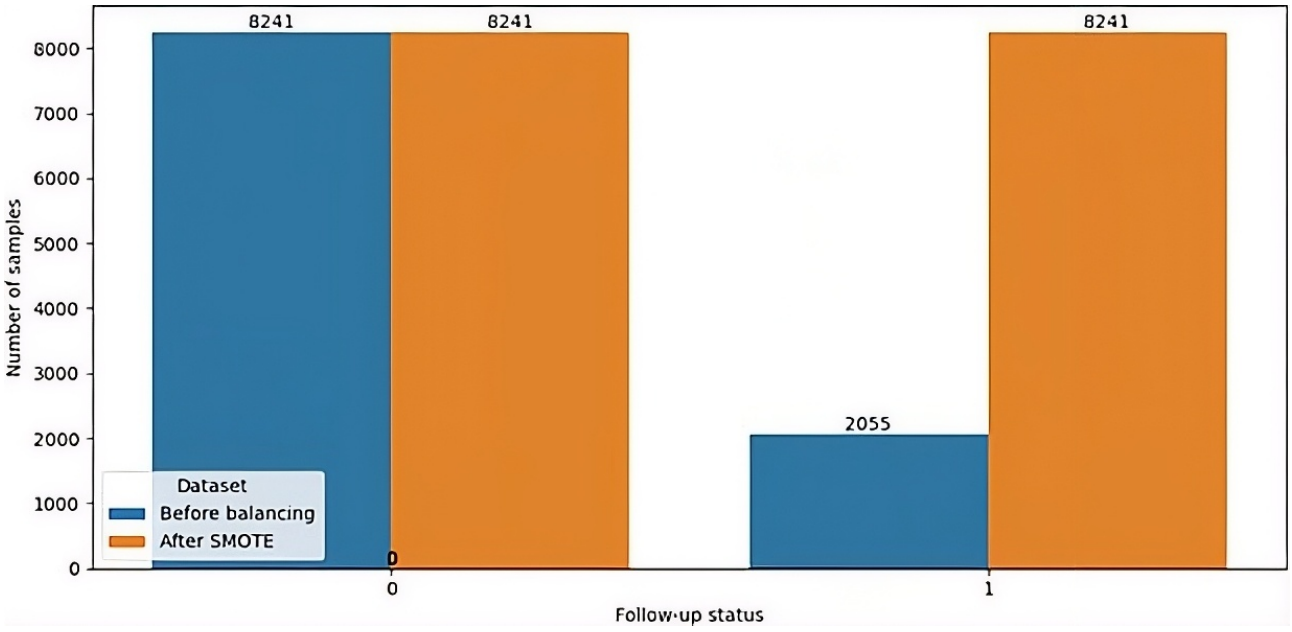
[a]CD4: cluster of differentiation 4 (a type of immune cell count).

Balancing Dataset

As seen in the descriptive statistics, the prevalence of mortality prediction among people living with HIV was 20%. This shows that the dataset was imbalanced; most observations (80%) were concentrated in the majority class (ie, alive). The SMOTE oversampling strategy added 6186 synthetic observations from the minority group (ie, dead) to balance the unbalanced distribution of the outcome variable. Therefore, the class distribution for mortality prediction among people living with HIV was balanced using SMOTE, resulting in a symmetric training dataset with 8241 observations in each class (alive and dead) to support the development of reliable and robust predictive models (see Figure 3). To avoid data leakage, the dataset was first split into training (80%) and testing (20%) subsets. SMOTE oversampling was then applied only to the training data, while the test set remained in its original, imbalanced form. This ensured unbiased performance evaluation on unseen data.

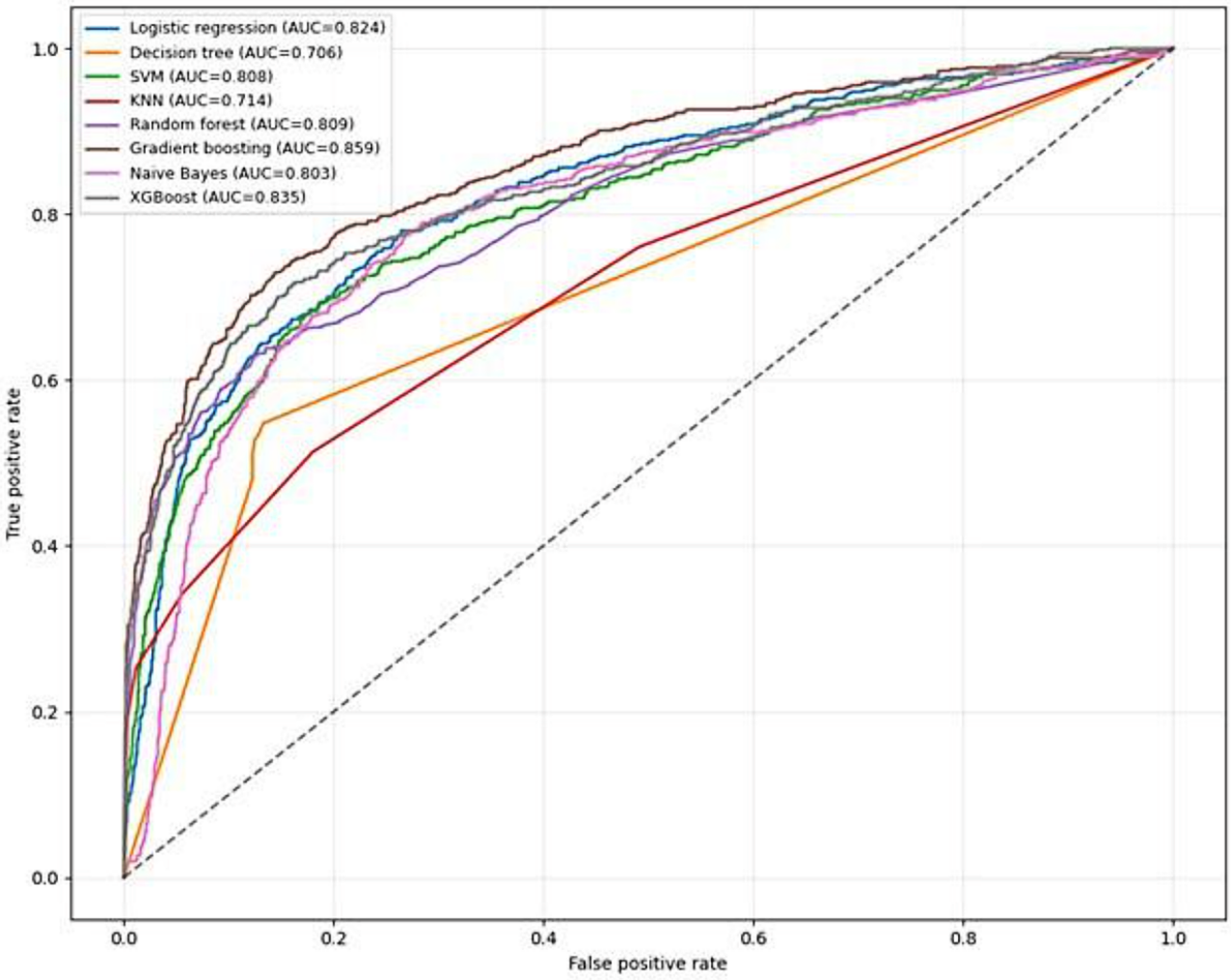
ROC curves for 7 ML models trained on the original imbalanced dataset (20% mortality prevalence) demonstrated clinically realistic discrimination (AUC range: 0.70 - 0.86). Gradient boosting achieved superior performance (AUC=0.859), followed by XGBoost (AUC=0.835) and logistic regression (AUC=0.824). Random forest (AUC=0.809) and Naive Bayes (AUC=0.803) showed moderate discrimination, while k-nearest neighbor (KNN) and decision tree showed lower discrimination compared to ensemble models (see Figure 4).

Figure 3. Class distribution of mortality outcomes (alive vs deceased) among adults living with HIV receiving antiretroviral therapy (ART) in Gondar City Administration, Northwest Ethiopia, showing the original imbalanced dataset and the training set after SMOTE (synthetic minority oversampling technique) was applied to address class imbalance for model development, 2005 - 2024.



XSL•FO
RenderX

Figure 4. Receiver operating characteristic (ROC) curves for 7 machine-learning classifiers evaluated on the original imbalanced test dataset for mortality prediction among adults living with HIV on antiretroviral therapy (ART) in public health facilities of Gondar City Administration, Northwest Ethiopia, reporting discrimination performance before class balancing, 2005 - 2024. AUC: area under the receiver operating characteristic curve; KNN: k-nearest neighbor; SVM: support vector machine; XGBoost: extreme gradient boosting.



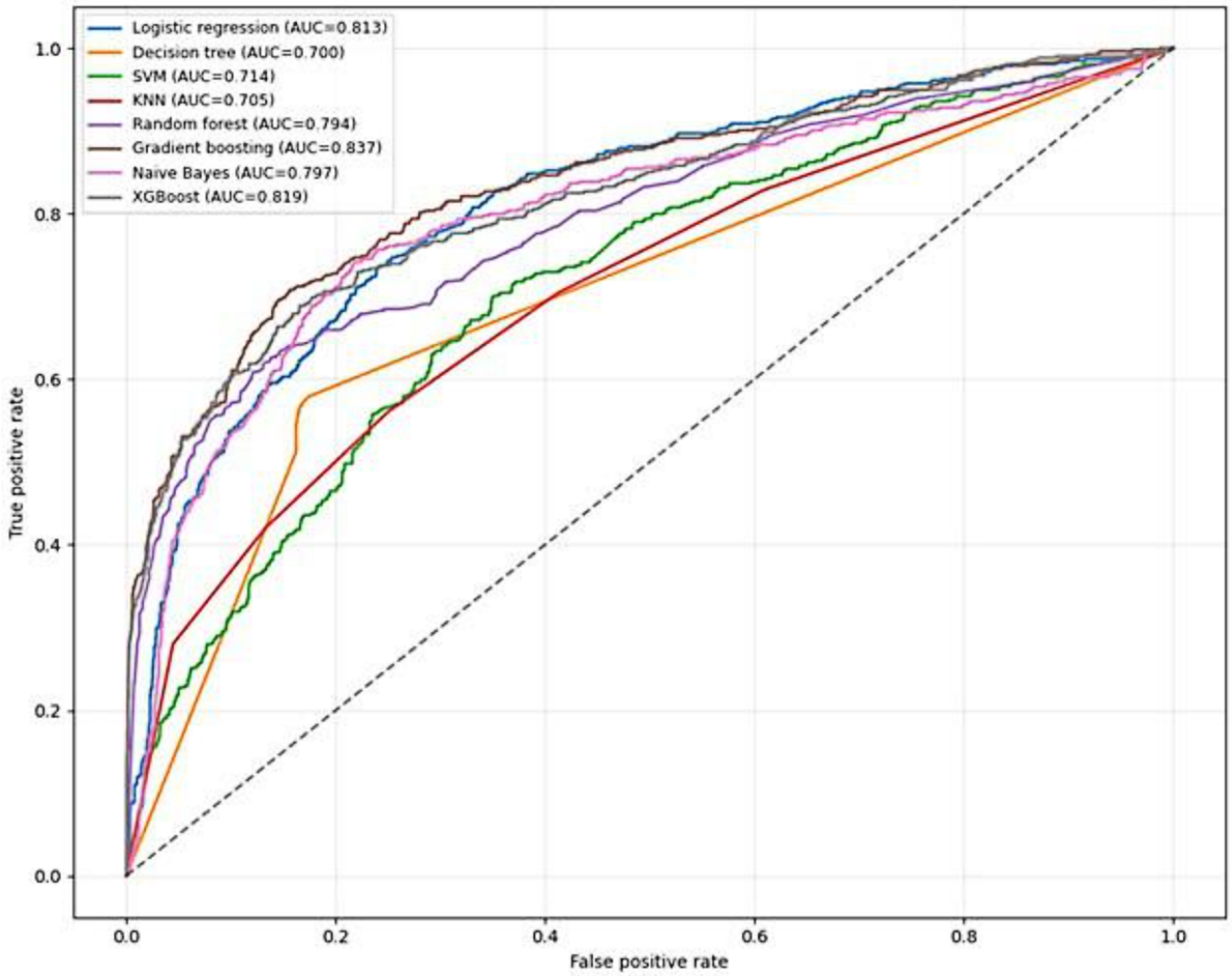
SMOTE-balanced models (training only, evaluated on the original imbalanced test set) demonstrated stable discrimination with modest recall improvements but precision trade-offs typical of oversampling. Logistic regression achieved AUC=0.813 (gain +0.0 from original), decision tree AUC=0.700 (−0.006), and KNN AUC=0.705 (−0.009). Tree-based ensembles maintained robustness: andom forest AUC=0.794 (−0.015), gradient boosting AUC=0.837 (−0.022), XGBoost AUC=0.819 (−0.016), and Naive Bayes AUC=0.797 (−0.006). Original configurations outperformed SMOTE variants (median ΔAUC=−0.013 [IQR -0.019 to 0.008]), favoring models trained on unaltered data for potential future clinical implementation, pending external validation (see Figure 5).

Model performance comparison across original and SMOTE-balanced configurations demonstrates gradient boosting original's superiority (F_1-score=0.619, AUC=0.859), outperforming XGBoost original (F_1-score=0.609, AUC=0.835). SMOTE-trained models showed recall gains but precision losses: XGBoost SMOTE (F_1-score=0.592, AUC=0.819; ΔF_1-score=−0.017), random forest SMOTE (F_1-score=0.576, AUC =0.794; ΔF_1-score=−0.021), and gradient boosting SMOTE (F_1-score=0.608, AUC=0.837; ΔF_1-score=−0.011). Ensemble methods maintained robust discrimination after balancing (AUC>0.79), whereas simpler models showed greater variability. KNN and decision tree exhibited the lowest performance across configurations, reaffirming ensemble learning's advantage for high-dimensional, imbalanced HIV mortality prediction.

Original data configurations consistently outperformed SMOTE variants (median ΔF_1-score=−0.011, median ΔAUC=−0.013), supporting potential future clinical implementation pending external validation without synthetic oversampling. This finding aligns with real-world prevalence modeling priorities where precision preservation exceeds recall optimization for resource allocation in ART programs (see Figure 6).

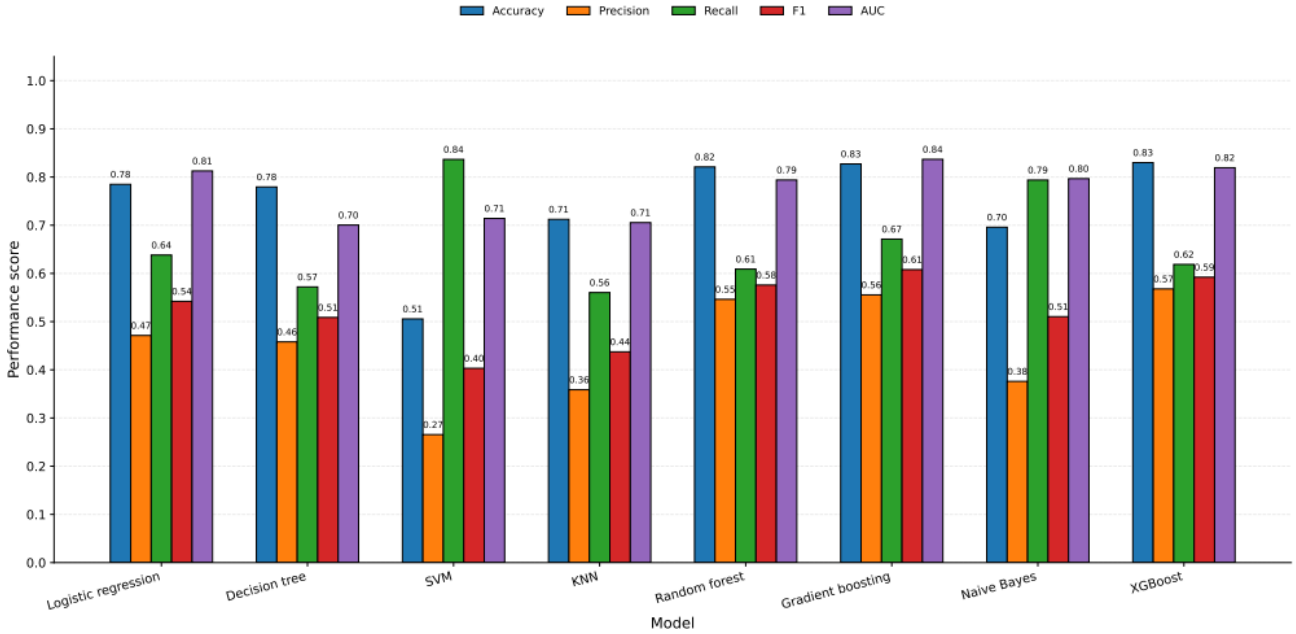
XSL•FO
RenderX

Figure 5. Receiver operating characteristic (ROC) curves for 7 machine-learning classifiers trained on SMOTE (synthetic minority oversampling technique)-balanced training data and evaluated on the original imbalanced test dataset for mortality prediction among adults living with HIV on antiretroviral therapy (ART) in Gondar City Administration, Northwest Ethiopia, illustrating postbalancing model discrimination, 2024. AUC: area under the receiver operating characteristic curve; KNN: k-nearest neighbor; SVM: support vector machine; XGBoost: extreme gradient boosting.



XSL•FO
RenderX

Figure 6. Comparative performance of machine-learning models (accuracy, precision, recall, F_1-score, and area under the receiver operating characteristic curve [AUC]) for mortality prediction among adults living with HIV receiving antiretroviral therapy (ART) in public health facilities of Gondar City Administration, Northwest Ethiopia, 2024. KNN: k-nearest neighbor; SMOTE: synthetic minority oversampling technique; SVM: support vector machine; XGBoost: extreme gradient boosting.



Model Building and Selection

The predictive performance of 7 ML models was comprehensively evaluated using accuracy, precision, recall, F_1-score (primary), and AUC on an 80:20 stratified train-test split. A critical distinction is that all preprocessing procedures, including SMOTE, were applied to training data only; the final evaluation used the original imbalanced test set (20% mortality prevalence) to reflect real-world deployment conditions.

On the original imbalanced test data, gradient boosting demonstrated superior balanced performance (accuracy=87.0%, precision=74.5%, recall=52.9%, F_1-score=0.619, AUC=0.859), outperforming XGBoost (F_1-score=0.609, AUC=0.835) and random forest (F_1-score=0.597, AUC=0.809). Ensemble methods consistently exceeded single classifiers, followed by Naive Bayes (F_1-score=0.562) and logistic regression (F_1-score=0.553).

SMOTE sensitivity analysis (training only) yielded recall gains but precision losses: XGBoost recall improved by +8.2% (F_1-score=0.592, ΔF_1-score=−0.017) and gradient boosting by +14.2% (F_1-score=0.608, ΔF_1-score=−0.011). Original configurations outperformed SMOTE variants across top models (median ΔF_1-score=−0.011 [IQR 0.006-0.019], ΔAUC=−0.013 [IQR 0.008-0.019]), confirming unaltered EMR data superiority for clinical precision.

Gradient boosting (original) was selected as optimal due to the highest F_1-score, cross-validation stability, and SHAP interpretability, establishing moderate-to-strong discrimination performance for 20% prevalence mortality prediction in resource-limited settings (see Table 3).

XSL•FO
RenderX

Table . Performance comparison of 7 machine-learning classifiers for binary mortality prediction among adults living with HIV on antiretroviral therapy in Gondar City Administration, Northwest Ethiopia, showing accuracy, precision, recall, F_1-score, and area under the receiver operating characteristic curve (AUC) evaluated on the original imbalanced hold-out test dataset, with models trained on unbalanced and SMOTE (synthetic minority oversampling technique–balanced training data), 2024.

Model and data	Accuracy	Precision	Recall	F_1-score	AUC	ΔF_1-score
Gradient boosting						
Original	0.870	0.745	0.529	0.619	0.859	—[a]
SMOTE	0.827	0.556	0.671	0.608	0.837	−0.011
XGBoost[b]						
Original	0.863	0.704	0.537	0.609	0.835	—
SMOTE	0.830	0.568	0.619	0.592	0.819	−0.017
Random forest						
Original	0.855	0.672	0.537	0.597	0.809	—
SMOTE	0.821	0.546	0.609	0.576	0.794	−0.021
Logistic regression						
Original	0.852	0.696	0.459	0.553	0.824	—
SMOTE	0.785	0.471	0.638	0.542	0.813	−0.011
Naive Bayes						
Original	0.821	0.548	0.576	0.562	0.803	—
SMOTE	0.696	0.376	0.794	0.510	0.797	−0.052
Decision tree						
Original	0.806	0.513	0.527	0.520	0.706	—
SMOTE	0.779	0.458	0.572	0.509	0.700	−0.011
KNN[c]						
Original	0.824	0.605	0.342	0.437	0.714	—
SMOTE	0.712	0.359	0.560	0.437	0.705	0.000

[a]Not available.

[b]XGBoost: extreme gradient boosting.

[c]KNN: k-nearest neighbor.

Association Rule Mining

Association rule mining using the Apriori algorithm (mlxtend; rules with support ≥0.065, confidence ≥0.64, and lift >4.6 were retained for interpretation) identified high-risk sociodemographic profiles among baseline characteristics. The dominant pattern—"rural residence + age 38‐47 years + no formal education + low baseline CD4 → mortality"—exhibited support=0.0686 (6.9% prevalence), confidence=68.3%, and lift=4.76, indicating that this subgroup showed a 4.76-fold higher co-occurrence with mortality compared with the baseline mortality prevalence (20% baseline prevalence).

In addition, demographic context (Christian religion, TB-negative status) maintained high lift values (4.62‐4.74), indicating that this subgroup exhibited the strongest co-occurrence pattern with mortality in the unadjusted association analysis (see Table 4). These association rules reflect co-occurrence patterns and should not be interpreted as causal risk factors.

The gradient boosting model demonstrates strong and clinically plausible classification performance (see Figure 7), which presents the confusion matrix of the optimized gradient boosting classifier.

XSL•FO
RenderX

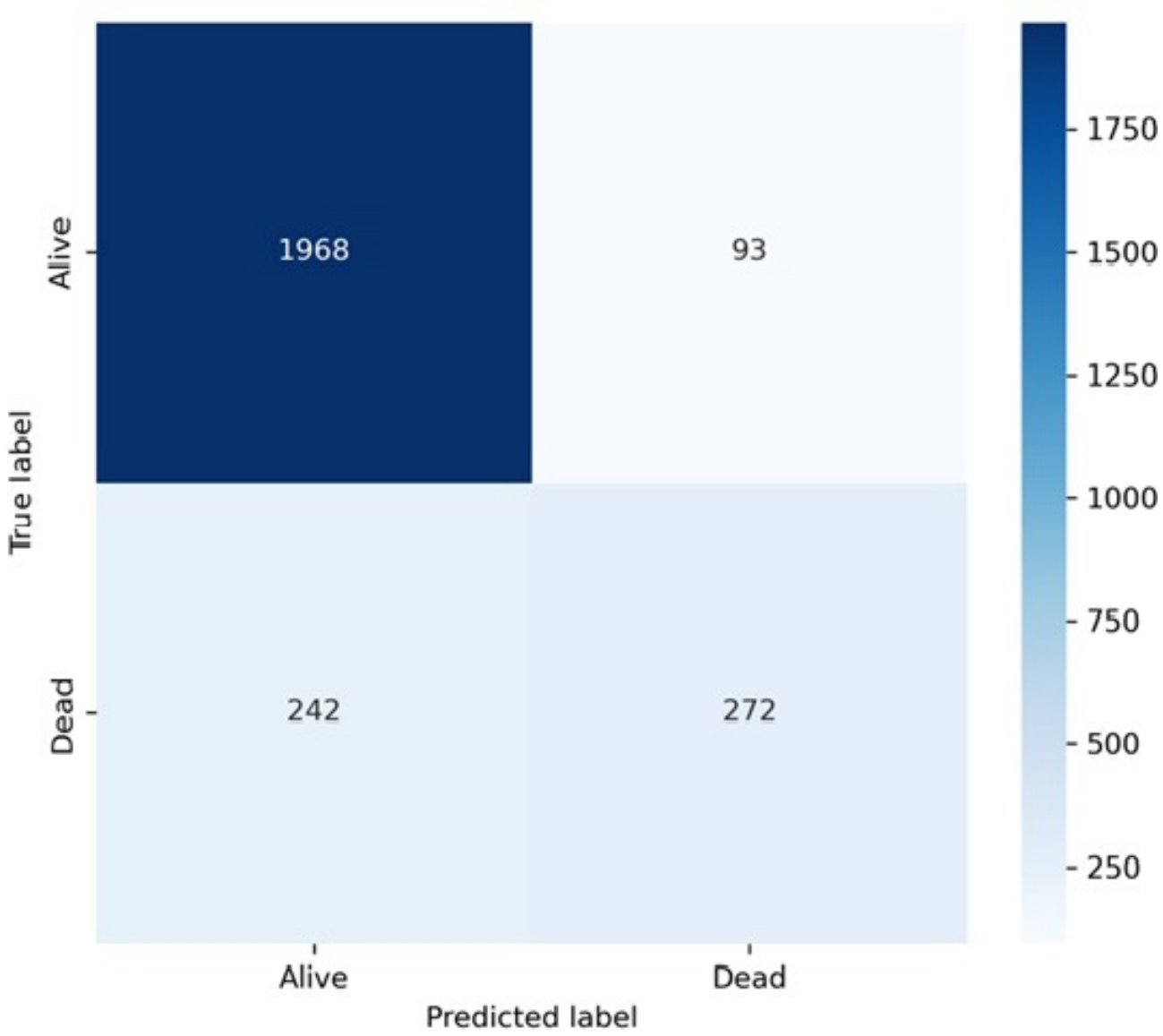
Table . Association rule mining results using the Apriori algorithm showing frequent baseline sociodemographic factor combinations associated with mortality among adults living with HIV on antiretroviral therapy (ART) in Gondar City Administration, Northwest Ethiopia, 2005 - 2024 (N=12,871).

Antecedent	Consequent	Support	Confidence	Lift
Rural+age 38 - 47+no education+low CD4[a]	Mortality	0.0686	0.683	4.76
Christian+rural+age 38 - 47+no education+low CD4	Mortality	0.0686	0.692	4.72
TB[b]-+rural+age 38 - 47+no education+low CD4	Mortality	0.0651	0.680	4.74
Rural+age 38 - 47+no education	Mortality	0.0707	0.668	4.65
Christian+rural+age 38 - 47+no education	Mortality	0.0707	0.677	4.62
TB-+rural+age 38 - 47+no education	Mortality	0.0672	0.664	4.63

[a]CD4: cluster of differentiation 4 (a type of immune cell count).

[b]TB: tuberculosis.

Figure 7. Confusion matrix for the optimized gradient boosting classifier predicting mortality among adults living with HIV receiving antiretroviral therapy (ART) in public health facilities of Gondar City Administration, Northwest Ethiopia, 2024.

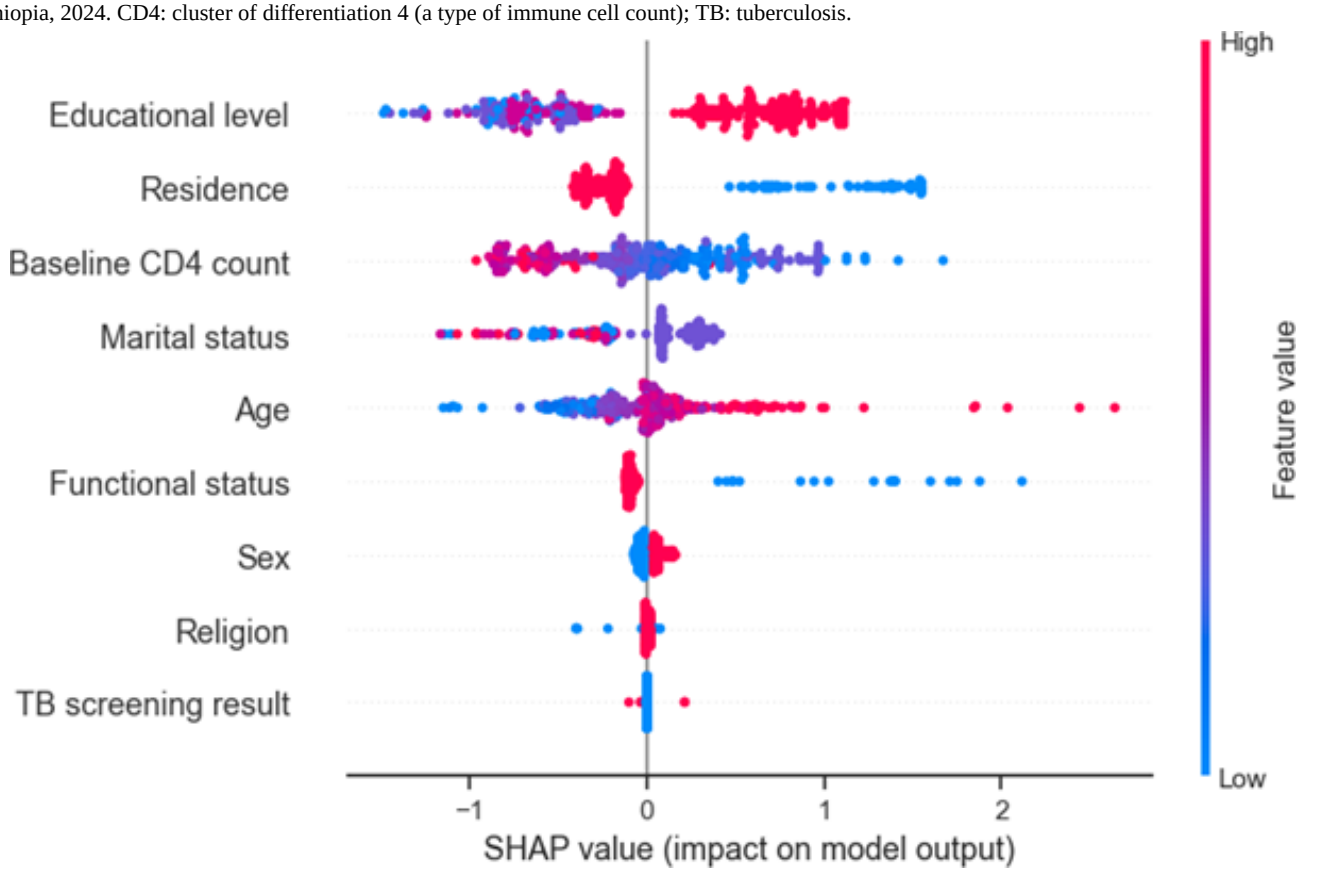


XSL•FO
RenderX

Feature Ranking

Gradient boosting trained on the original dataset was selected as the optimal model, and the SHAP summary analysis was identified as the most influential baseline predictors of mortality among 12,871 people living with HIV at ART initiation. The educational level was ranked as the strongest predictor, with lower education associated with positive SHAP values (increased mortality risk) and higher education demonstrating consistent protective effects. Residence was the second most important factor, with rural residence contributing to higher predicted mortality and urban residence showing protective effects. The baseline CD4 count exhibited a clear biological gradient: lower CD4 values clustered on the positive SHAP side, indicating increased mortality risk, whereas higher CD4 counts were associated with negative SHAP values, reflecting protection. Marital status and age showed moderate influence, with older age generally increasing predicted risk. Functional status also contributed meaningfully, as ambulatory or bedridden states increased mortality risk compared to working status. In contrast, sex, religion, and TB screening results demonstrated relatively small SHAP magnitudes, indicating limited impact on overall prediction. Overall, the SHAP ranking confirms that socioeconomic vulnerability and immunological status at ART initiation are the dominant determinants of predicted mortality risk (see Figure 8).

Figure 8. Global feature importance ranking based on mean absolute Shapley Additive Explanations (SHAP) values from the gradient boosting model, identifying the top predictors of mortality among adults living with HIV on antiretroviral therapy (ART) in Gondar City Administration, Northwest Ethiopia, 2024. CD4: cluster of differentiation 4 (a type of immune cell count); TB: tuberculosis.



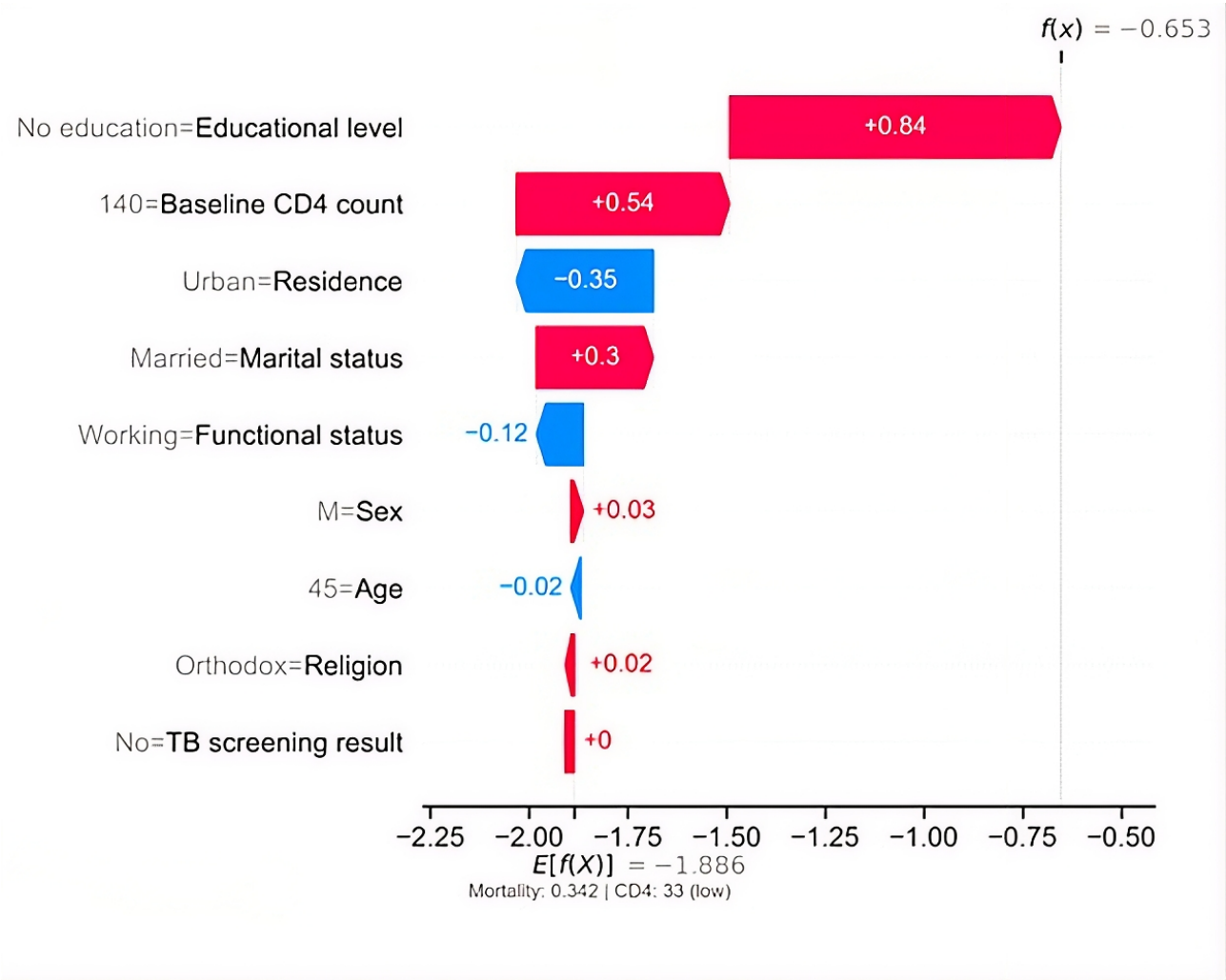
Waterfall Plot

The SHAP waterfall plot illustrates the individual prediction for a 45-year-old patient. The model output is expressed on the log-odds scale for the mortality class (dead=1), with the cohort baseline expected value of $E[f(X)]=-1.886$, representing the average predicted mortality risk. For this patient, the cumulative contribution of key baseline features shifts the model output to $f(x)=-0.653$. Because this value is substantially higher than the cohort baseline, it indicates an elevated predicted mortality risk relative to the average patient, although the absolute probability remains below 0.5.

Positive SHAP values (red bars) indicate features that increase predicted mortality risk, whereas negative SHAP values (blue bars) indicate protective effects. The strongest contributor to increased mortality risk was a lack of formal education (+0.84), followed by a low baseline CD4 count of 140 cells/mm³ (+0.54). Married as marital status (+0.30) and male as sex (+0.03) provided additional, smaller risk contributions. In contrast, urban residence (−0.35) and working functional status (−0.12) exerted protective effects by reducing the predicted mortality risk. Age (45 y; −0.02), religion (+0.02), and TB screening results (~0) had minimal influence on the individual prediction (see Figure 9).

XSL•FO
RenderX

Figure 9. Shapley Additive Explanations (SHAP) waterfall plot illustrating an individual-level prediction from the gradient boosting mortality model for an adult living with HIV on antiretroviral therapy (ART) in Gondar City Administration, Northwest Ethiopia, 2024. CD4: cluster of differentiation 4 (a type of immune cell count); TB: tuberculosis.



Discussion

Principal Findings

To our knowledge, this represents one of the first Ethiopia-based applications of interpretable ensemble ML (gradient boosting combined with SHAP) using baseline EMR data to predict mortality among 12,871 people living with HIV, achieving moderate but clinically realistic performance (F_1-score=0.619, AUC=0.859). Unlike earlier exploratory analyses, the final model was developed strictly using predictors recorded at ART initiation, eliminating postbaseline variables to avoid look-ahead bias and ensure valid prognostic modeling. This study contributes new evidence by demonstrating the feasibility of applying XAI models to routine EMR data to support proactive risk stratification and clinical decision-making in resource-limited settings. However, the model should be interpreted as a decision-support tool rather than a standalone clinical triage system, given its moderate sensitivity. Consistent with findings from studies published in JMIR journals, our results demonstrate the feasibility of integrating XAI models into routine clinical care to support proactive risk stratification and decision-making, particularly in low-resource settings where the efficient use of existing EMR data is critical [1-3].

Gradient boosting trained on verified baseline-only predictors achieved the best overall balance between discrimination and generalization performance (F_1-score=0.619, AUC=0.859), outperforming logistic regression, Naive Bayes, KNN, and decision tree models. While XGBoost and random forest also demonstrated competitive performance, the corrected baseline-only gradient boosting model provided the most stable and clinically interpretable results for the 20% mortality prevalence setting.

SHAP values were calculated with respect to the mortality class (dead=1). Accordingly, positive SHAP values indicate an increase in predicted mortality risk, whereas negative SHAP values represent a reduction in predicted mortality risk (protective effect).

The most influential baseline predictors included educational level, baseline CD4 count, residence, marital status, and functional status. In the illustrated case, lack of formal education and a low baseline CD4 count (140 cells/mm^3) exerted strong positive contributions, substantially increasing the predicted mortality risk. Urban residence and working functional status demonstrated protective contributions.

XSL•FO
RenderX

Importantly, the baseline CD4 count was modeled as a continuous predictor in the ML algorithms, and descriptive statistics have been revised to report a median of 225 (IQR 131-367) cells/mm³ to ensure methodological consistency. Unlike preliminary analyses that yielded inflated performance estimates, the corrected baseline-only model produced moderate but clinically realistic discrimination (AUC=0.859), consistent with previously published EMR-based mortality prediction studies [4-7].

Comparison to Prior Work

These results are consistent in accuracy with studies conducted in Nigeria and Thailand, where ensemble models achieved high predictive accuracy in the clinical context [20,26]. Several studies have explored the use of ML algorithms for mortality prediction among people living with HIV. Similarly, a study from China utilizing a support vector machine reported an accuracy of 86%, which aligns with this study's pre-SMOTE support vector machine performance [16].

The SHAP analysis provided valuable insights into the factors influencing mortality prediction.

The SHAP global importance analysis identified the educational level as the most influential baseline predictor of mortality, followed by residence, baseline CD4 count, marital status, age, and functional status. Specifically, a lack of formal education and lower baseline CD4 values were associated with an increased predicted mortality risk, whereas urban residence and working functional status were protective.

This hierarchy reflects the combined influence of socioeconomic vulnerability and biological immunosuppression at ART initiation, underscoring that both structural and clinical determinants contribute substantially to mortality risk prediction. For instance, a study conducted in Kenya found that patients on ART for extended periods had significantly lower mortality rates, reinforcing the critical role of early and consistent treatment adherence in improving survival [27].

Ensemble methods, such as XGBoost, random forest, and gradient boosting, consistently outperformed simpler models, such as logistic regression and Naive Bayes. Similar findings were reported in studies from Northern Thailand, East Africa, and Nigeria, where ensemble models achieved superior performance in health prediction tasks [28,29].

The model comparison plot demonstrated that gradient boosting achieved an accuracy of 87.0%, precision of 74.5%, recall of 52.9%, F_1-score of 0.619, and AUC of 0.859. Although discrimination was strong, the recall of 52.9% indicates that nearly half of the mortality cases were not identified at the default classification threshold. Therefore, while the model demonstrates predictive capability, its sensitivity limits immediate deployment as a high-stakes early-warning system without threshold optimization.

Limitations

This study used retrospective secondary EMR data, which may contain incomplete or inaccurately recorded variables, introducing potential information bias. Only 1 baseline record per patient was retained to avoid correlated observations, limiting longitudinal analysis. Patients who transferred to other facilities were excluded, potentially introducing selection bias.

Although the cohort required documented baseline ART initiation records, limited variable-level missingness (<10%) was addressed using median (continuous variables) and mode (categorical variables) imputation. Therefore, this was not a strict complete-case analysis. Excluding individuals on ART for less than 6 months may have introduced survivorship bias, potentially inflating performance metrics and limiting generalizability to newly initiated ART populations.

SMOTE-based oversampling may not fully replicate real-world population distributions. External validation using independent datasets was not performed, and model performance may vary in other geographic or clinical contexts. Data were drawn from a single city administration, limiting external validity.

Finally, given the moderate sensitivity (52.9%, 272/514), the model may miss a substantial proportion of high-risk patients at the default threshold. Future implementations should consider threshold calibration based on clinical priorities, particularly if minimizing false negatives is paramount.

Future Directions

Future studies should use multicenter or national-level data to enhance generalizability and external validity. Prospective validation studies are needed to evaluate real-world performance. Threshold optimization strategies should be explored to improve sensitivity where early mortality detection is prioritized. Integrating temporal modeling approaches may capture longitudinal treatment dynamics. Additionally, embedding explainable models within clinical workflows could support structured risk assessment, provided that implementation is accompanied by clinician oversight and contextual calibration.

Conclusion

This study demonstrates that interpretable ensemble ML models trained exclusively on baseline EMR data achieved moderate yet clinically meaningful discrimination in predicting mortality risk among people living with HIV initiating ART. SHAP-based interpretation confirmed that educational attainment and baseline CD4 count were the most influential predictors, followed by residence, age, and functional status, highlighting the combined impact of socioeconomic vulnerability and immunological status at treatment initiation. These findings suggest that transparent and explainable ML approaches may support early risk stratification and targeted intervention planning within HIV care programs. Nevertheless, external validation and prospective evaluation are required before integration into routine clinical decision-making.

XSL•FO
RenderX

Acknowledgments

All authors declared that they had insufficient funding to support open-access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Vasudevan L, Kibria MG, Kucirka LM, et al. Machine learning models to predict risk of maternal morbidity and mortality from electronic medical record data: scoping review. J Med Internet Res 2025 Aug 14;27:e68225. [doi: 10.2196/68225] [Medline: 40811480]
2. Tahir N, Jung CR, Lee SD, Azizah N, Ho WC, Li TC. Federated learning-based model for predicting mortality: systematic review and meta-analysis. J Med Internet Res 2025 Jul 21;27:e65708. [doi: 10.2196/65708] [Medline: 40690657]
3. Lu SC, Xu C, Nguyen CH, Geng Y, Pfob A, Sidey-Gibbons C. Machine learning-based short-term mortality prediction models for patients with cancer using electronic health record data: systematic review and critical appraisal. JMIR Med Inform 2022 Mar 14;10(3):e33182. [doi: 10.2196/33182] [Medline: 35285816]
4. Ikemura K, Bellin E, Yagi Y, et al. Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. J Med Internet Res 2021 Feb 26;23(2):e23458. [doi: 10.2196/23458] [Medline: 33539308]
5. Vaid A, Somani S, Russak AJ, et al. Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. J Med Internet Res 2020 Nov 6;22(11):e24018. [doi: 10.2196/24018] [Medline: 33027032]
6. Lv H, Yang X, Wang B, et al. Machine learning-driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: retrospective study. J Med Internet Res 2021 Apr 19;23(4):e24996. [doi: 10.2196/24996] [Medline: 33871375]
7. Diwan S, Gandhi V, Baidya Kayal E, Khanna P, Mehndiratta A. Explainable machine learning models for mortality prediction in patients with sepsis in tertiary care hospital ICU in low- to middle-income countries. Intensive Care Med Exp 2025 Jun 3;13(1):56. [doi: 10.1186/s40635-025-00765-5] [Medline: 40459817]
8. Global HIV & AIDS statistics — Fact sheet. UNAIDS. URL: https://www.unaids.org/en/resources/fact-sheet [accessed 2026-04-14]
9. Kloos H, Converse P, Mariam DH, Mulatu MS, Kaba M, Mekonnen W. Bibliography on HIV/AIDS in Ethiopia and Ethiopians in the diaspora: the 2017 update. Ethiop J Health Dev 2018;32(4) [FREE Full text]
10. Vitalis D. Adherence to Antiretroviral Therapy among Perinatal Women in Guyana: Challenges and Lessons for Developing Nations: Springer; 2021:183-199. [doi: 10.1007/978-981-15-3974-9]
11. Argaw GS, Gelaye KA, Lakew AM, et al. Survival and predictors of mortality among HIV-infected adults after initiation of antiretroviral therapy in Eastern Ethiopia Governmental hospitals, from January 2015 to December 2021 (multi-center retrospective follow-up study). BMC Infect Dis 2024;24(1). [doi: 10.1186/s12879-024-10225-2]
12. Mamo DN, Yilma TM, Tewelgne MF, et al. Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. BMC Med Inform Decis Mak 2023 Apr 21;23(1):75. [doi: 10.1186/s12911-023-02167-7] [Medline: 37085851]
13. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. Cell 2020 Apr 2;181(1):92-101. [doi: 10.1016/j.cell.2020.03.022] [Medline: 32243801]
14. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
15. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018 Feb 26;18(1):24. [doi: 10.1186/s12874-018-0482-1] [Medline: 29482517]
16. Shi M, Lin J, Wei W, et al. Machine learning-based in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China. PLoS Negl Trop Dis 2022 May;16(5):e0010388. [doi: 10.1371/journal.pntd.0010388] [Medline: 35507586]

XSL•FO
RenderX

17. Gebremedhin A, Gebremariam S, Haile F, Weldearegawi B, Decotelli C. Predictors of mortality among HIV infected children on anti-retroviral therapy in Mekelle Hospital, Northern Ethiopia: a retrospective cohort study. BMC Public Health 2013 Nov 6;13(1):1047. [doi: 10.1186/1471-2458-13-1047] [Medline: 24517533]
18. Li Y, Feng Y, He Q, et al. The predictive accuracy of machine learning for the risk of death in HIV patients: a systematic review and meta-analysis. BMC Infect Dis 2024 May 6;24(1):474. [doi: 10.1186/s12879-024-09368-z] [Medline: 38711068]
19. Zhan B, Wei W, Xie Z, et al. Machine learning-based prognostic prediction for hospitalized HIV/AIDS patients with cryptococcus infection in Guangxi, China. BMC Infect Dis 2024 Oct 8;24(1):1121. [doi: 10.1186/s12879-024-10013-y] [Medline: 39379851]
20. Ijaiya M, Troncoso E, Mutloatse M, et al. Use of machine learning in predicting continuity of HIV treatment in selected Nigerian states. PLOS Glob Public Health 2025;5(4):e0004497. [doi: 10.1371/journal.pgph.0004497] [Medline: 40273279]
21. Odukoya O, Nwaneri S, Odeniyi I, et al. Development and comparison of three data models for predicting diabetes mellitus using risk factors in a Nigerian population. Healthc Inform Res 2022 Jan;28(1):58-67. [doi: 10.4258/hir.2022.28.1.58] [Medline: 35172091]
22. Tamiru AT, Rade BK, Taye EB, et al. Community Level of COVID-19 Information Exposure and Influencing Factors in Northwest Ethiopia. . .
23. Ethiopia's national research ethics guidelines. URL: https://www.scribd.com/document/450446123/national-research-ethics-review-guidline [accessed 2026-04-15]
24. Sarkar D, Bali R, Sharma T. Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems: Apress; 2018:25-30 URL: https://link.springer.com/book/10.1007/978-1-4842-3207-1 [accessed 2026-03-27]
25. Kuhn M, Johnson K. Applied Predictive Modeling: Springer; 2013.
26. Soha K, Phuthomdee S, Srichai T, Kittiratanawasin L, Han WM, Teeraananchai S. Evaluating machine learning algorithms for predicting HIV status among young Thai men who have sex with men. BMJ Health Care Inform 2025 May 15;32(1):e101189. [doi: 10.1136/bmjhci-2024-101189] [Medline: 40379266]
27. Mugo CW, Shkedy Z, Mwalili S, et al. Modelling trends of CD4 counts for patients on antiretroviral therapy (ART): a comprehensive health care clinic in Nairobi, Kenya. BMC Infect Dis 2022 Jan 4;22(1):29. [doi: 10.1186/s12879-021-06977-w] [Medline: 34983418]
28. Quan VM, Vongchak T, Jittiwutikarn J, et al. Predictors of mortality among injecting and non-injecting HIV-negative drug users in northern Thailand. Addiction 2007 Mar;102(3):441-446. [doi: 10.1111/j.1360-0443.2006.01709.x] [Medline: 17298652]
29. Kibuuka H, Musingye E, Mwesigwa B, et al. Predictors of all-cause mortality among people with human immunodeficiency virus (HIV) in a prospective cohort study in East Africa and Nigeria. Clin Infect Dis 2022 Sep 10;75(4):657-664. [doi: 10.1093/cid/ciab995] [Medline: 34864933]

Abbreviations

ART: antiretroviral therapy
AUC : area under the receiver operating characteristic curve
CD4: cluster of differentiation 4 (a type of immune cell count)
EMR: electronic medical record
KNN: k-nearest neighbor
ML: machine learning
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority oversampling technique
TB : tuberculosis
UNAIDS: Joint United Nations Program on HIV/AIDS
XAI: explainable artificial intelligence
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by A Coristine; submitted 09.Jun.2025; peer-reviewed by R Mpofu, WD Wei; revised version received 02.Mar.2026; accepted 12.Mar.2026; published 20.Apr.2026.

Please cite as:
Gedefaw AE, Biwota GT, Mesele AG, Mengistu AK, Teferi GH
Mortality Prediction Among People Living With HIV on Antiretroviral Therapy in Public Health Facilities in Gondar City Administration, Northwest Ethiopia: Machine Learning–Based Study
JMIR Med Inform 2026;14:e78770
URL: https://medinform.jmir.org/2026/1/e78770
doi:10.2196/78770

© Andualem Enyew Gedefaw, Getaye Tizazu Biwota, Abraraw Gebre Mesele, Abraham Keffale Mengistu, Gizaw Hailiye Teferi. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Detection of Interpretable and Fine-Grained Brain Tumor Magnetic Resonance Imaging Based on Progressive Pruning: Machine Learning Model Development and Validation Study

Yupeng Liu[1], PhD; Shuwei Song[1], ME; Shibo Lian[1], ME; Xiaochen Zhang[2], PhD

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China

[2]Heilongjiang Institute of Technology, No. 999 Hongqi Street, Daowai District, Harbin City, Heilong, Harbin, China

Corresponding Author:
Xiaochen Zhang, PhD
Heilongjiang Institute of Technology, No. 999 Hongqi Street, Daowai District, Harbin City, Heilong, Harbin, China

Abstract

Background: Brain tumor is one of the most malignant diseases of the central nervous system, and early accurate detection is of great significance for improving patient survival rate. However, the heterogeneity of brain tumors in terms of morphology, size, and location on magnetic resonance imaging (MRI) image, as well as their similarity to surrounding normal brain tissue, poses significant challenges for tumor detection.

Objective: This study aims to develop a high-performance brain tumor detection framework that integrates feature enhancement, channel attention, and progressive pruning, achieving an optimal balance between detection accuracy, model efficiency, and interpretability for slice-level MRI tumor localization tasks.

Methods: This paper proposes a convolution Prewitt-and-pooling–based preprocessing (CSPP) approach, based on the "you only look once" version 11 (YOLOv11) framework, which highlights important structural detail more effectively than traditional statistics. A dynamic convolution–based C3k2 (DCC) module was integrated to more efficiently capture both local and global features. A channel prior convolutional attention (CPCA) module was introduced before the detection head, enabling the network to specifically focus on information-rich channels and key spatial regions. Through a progressive hybrid pruning strategy (PHPS), the model was optimized for efficient inference. Furthermore, Eigen-class activation mapping (Eigen-CAM) was used to interpret the prediction result, making them more transparent.

Results: Extensive experiments on 3 brain tumor MRI datasets demonstrated the superior performance of CDCP-YOLO (CSPP-DCC-CPCA-PHPS–YOLO). On Br35H, the mean average precision (mAP) at an intersection-over-union (IoU) threshold of 0.5 ($mAP_{0.5}$) increased by 2.6%, average mAP over several IoU thresholds (0.50-0.95; $mAP_{0.5:0.95}$) increased by 5.9%, and number of floating-point operations (×10 ; GFLOPs) decreased by 47.7%. On Roboflow, $mAP_{0.5}$ increased by 19.5%, $mAP_{0.5:0.95}$ increased by 7.7%, and GFLOPs decreased by 47.7%. On Capstone, $mAP_{0.5}$ increased by 6.9%, $mAP_{0.5:0.95}$ increased by 5.8%, and GFLOPs decreased by 47.7%.

Conclusions: The proposed CDCP-YOLO framework achieves an optimal balance between accuracy, efficiency, and interpretability, providing a lightweight and reliable solution for slice-level brain tumor detection in MRI images.

(JMIR Med Inform 2026;14:e84095) doi:10.2196/84095

KEYWORDS

magnetic resonance imaging; MRI; brain tumor; convolution Prewitt-and-pooling–based preprocessing; progressive hybrid pruning strategy; dynamic convolution-based C3k2; feature fusion; medical imaging; brain tumor detection; deep learning; class activation mapping; Eigen-CAM; lightweight model

Introduction

Background

Brain tumors are a type of highly complex and rapidly progressing major disease within the central nervous system. Their malignancy is often characterized by extreme invasiveness and recurrence, posing a serious threat to patients' lives and health. Medical research indicates that accurate identification and timely diagnosis of tumor types in the early stages can significantly extend patient survival and improve quality of life. Magnetic resonance imaging (MRI) has become the gold standard for preoperative imaging diagnosis of most brain tumors due to its noninvasive, high-resolution, and multimodal imaging advantages, playing a key role in the auxiliary diagnosis and efficacy evaluation of brain tumors [1-3]. Although MRI

XSL•FO
RenderX

provides rich information at the image level, brain tumor detection still faces several challenges:

1. The tumor shape is complex and diverse, with blurred boundaries, and its grayscale distribution often overlaps with normal brain tissue, making it difficult to distinguish and locate accurately.
2. The anatomical structure of the brain is complex and varies significantly between individuals, making it difficult for traditional methods to establish a universally applicable and robust expression model.
3. Manual image segmentation is not only time-consuming and labor-intensive but also susceptible to subjective factors, necessitating efficient and intelligent auxiliary diagnostic tools.

In recent years, an increasing number of researchers have begun to focus on intelligent methods for brain tumor detection. The main research is based on digital image processing and machine learning to assist in identifying brain tumor regions [4]. This type of method usually first uses MRI image acquisition to obtain brain image data; then uses image processing to extract key visual features, such as gray scale, shape contours, and texture patterns; and finally inputs these features into support vector machines, random forests, or k-nearest neighbors for judging the tumor site and type. This type of method, which relies on handcrafted features, is effective for the detection of target lesions, but when faced with complex organizational structures and highly heterogeneous tumor morphologies, it often exhibits poor robustness, sensitivity to image noise, and difficulty in generalization, which limits its widespread application in clinical practice. With the rapid development of the convolutional neural network (CNN), target detection has entered a new stage with end-to-end learning as its core. Girshick et al [5] proposed region-based CNN (R-CNN), which first extracts candidate target regions and then uses support vector machines for classification, significantly improving detection performance. Subsequently, Fast R-CNN and Faster R-CNN [6,7] further accelerated detection speed and improved accuracy by sharing features and a region proposal network. To achieve real-time detection, Redmon et al [8] proposed the "you only look once" (YOLO) series of algorithms, which transformed target detection into a regression problem, simultaneously predicting multiple bounding boxes and categories in a single forward pass, greatly increasing the detection speed. Subsequent versions of YOLO have continuously improved the detection accuracy and network structure and are widely used in scenarios such as autonomous driving and medical image analysis. In addition, single shot multibox detector (SSD) [9] also achieves a balance between speed and accuracy by detecting at multiple feature scales. These deep learning–based algorithms not only overcome the limitations of traditional manual feature engineering but also greatly improve the modeling accuracy and practical application value, becoming the mainstream direction of current image target detection, including brain tumor detection.

Related Works

YOLO-Based Detection on MRI

In recent years, target detection models based on YOLO have been widely applied and continuously optimized in medical imaging. In the area of noninvasive disease detection, researchers have primarily focused on introducing attention mechanism, multitask learning, and lightweight model optimization. Chen et al [10] proposed the MSA-YOLOv5, which combines multiple attention mechanisms to focus on the automatic detection of lesions in acute ischemic stroke. In a multimodal MRI model, the detection of small lesions and embolic signals is enhanced, improving detection accuracy while reducing the number of parameters. Tang et al [11] developed a YOLOv5 model based on the squeeze-and-excitation attention mechanism, which improved the detection of Parkinson disease. The model enhances detection accuracy by adaptively focusing on key features, especially in distinguishing substantia nigra and red nucleus lesions in T2-weighted MRI. Wang et al [12] introduced an improved YOLOv5 architecture for the diagnosis and grading of lumbar disc herniation. This method adds an attention module to the Cross Stage Partial part and enhances the Spatial Pyramid Pooling-Fast part, achieving multitask learning for both classification and grading of Pfirrmann grades, highlighting prominent features in the intervertebral space and high-intensity zones.

In the context of brain tumor MRI detection, the research on improving YOLO has also advanced rapidly. Kang et al [13] proposed a new YOLO architecture named RCS-YOLO, which is optimized for brain tumor detection in medical imaging. RCS-YOLO uses reparameterized convolution combined with channel shuffle (RCS) to enhance the model's computational efficiency and detection accuracy. Kang et al [14] developed BGF-YOLO, an improved YOLOv8 that enhances brain tumor MRI detection performance by introducing multilevel feature fusion, a dynamic attention mechanism, and an additional detection head. Kang et al [15] proposed a new YOLO architecture, pretrained knowledge-guided YOLO (PK-YOLO), optimized for brain tumor detection in multiplanar MRI slices. PK-YOLO uses a pretrained RepViT backbone network combined with sparse mask modeling technology and Focaler-intersection-over-union (IoU) regression loss to improve the detection performance of small targets, making it the first YOLO object detector to introduce pretrained knowledge guidance. Dixit et al [16] proposed a brain tumor detection method based on YOLOv4-tiny, which uses transfer learning and a fine-tuning technique optimized for MRI images. This method leverages features from the pretrained Common Objects in Context (COCO) dataset with the 29-layer YOLOv4-tiny architecture, improving the model's computational efficiency and detection accuracy through precise hyperparameter tuning. Abdusalomov et al [17] proposed an improved YOLOv7 architecture optimized for brain tumor detection in MRI images. This method integrates a Convolutional Block Attention Module, a Spatial Pyramid Pooling Fast+ layer, and a Bidirectional Feature Pyramid Network to improve the detection accuracy for glioma, meningioma, and pituitary tumors.

XSL•FO
RenderX

Pruning Method

As the complexity of the deep neural network continues to increase, how to reduce computational and storage requirements while maintaining high performance has become a key challenge in medical imaging processing. Pruning techniques offer an effective solution to this problem and have demonstrated considerable potential in various medical image analysis applications. Fernandes et al [18] proposed a generative adversarial pruning method based on an evolutionary strategy, specifically optimizing medical image diagnosis. This method selects options through minimal Wasserstein distance. Wu et al [19] developed the FairPrune method, a new technique that achieves fairness by pruning, specifically applied to the diagnosis of skin diseases. This method prunes based on differences in parameter importance, significantly improving the fairness of the model across different demographic groups. Adnan et al [20] proposed a structured pruning method specifically for optimizing the U-Net architecture. This method addresses the pruning complexity between the encoder and decoder in U-Net, compressing the model by assessing the importance of individual channels and tasks. Fernandes et al [21] further developed a structured pruning framework that integrates multitask learning and pruning. This method uses iterative pruning and block-based network deepening, optimizing the model with a policy-based and multiobjective decision-making process. Cocosco et al [22] proposed a fully automated and nonparametric brain tissue classification method. This method uses a nonparametric implementation, training sample selection through the minimum spanning tree method and stereotaxic space priors, showing significant performance improvement in subjects with large morphological variations. Xuan et al [23] introduced a pruning method for k-space subsampling and reconstruction based on a generative model. This method is inspired by network pruning, starting with a fully sampled k-space model and iteratively removing less important k-space phase encoding, demonstrating good performance in single-coil and multicoil MRI reconstruction. Graziani et al [24] developed an interpretable pruning strategy on CNN specifically for scale-variant features in medical images. This method uses deep learning interpretability techniques to analyze the hierarchical scale coding of the InceptionV3 and ResNet50 architectures, finding that scale information peaks in the middle layers and decreases near the softmax layer. This discovery leads to a pruning strategy that significantly improves the performance of nucleus regression and mitosis classification in histopathological images. Holste et al [25] systematically analyzed the impact of the pruning on medical image classification across various long-tailed multilabel disease datasets for the first time. This study in the chest X-ray image diagnosis experimentally demonstrated that the pruning has differential effects on different diseases, with rare diseases being more susceptible to being “forgotten” than common diseases. The study also introduces the concept of pruning-identified exemplars, revealing through human reader studies that pruning-identified exemplars often have more label noise, lower image quality, and higher diagnostic uncertainty. Saleh et al [26] investigated the effectiveness of different network architectures (GoogLeNet, ResNet, and EfficientNet) combined with transfer learning and the network pruning algorithm for medical image classification. They validated the effectiveness of these techniques in both brain tumor classification and chest X-ray inflammation detection. Jaiswal et al [27] proposed a pruning-assisted, self-supervised image localization method. This method uses a “learning by forgetting” training scheme, which significantly improves skin disease localization performance under unsupervised, weakly supervised, and sparsely supervised settings.

Visualization Based on Class Activation Mapping

In recent years, visualization methods based on class activation mapping (CAM) have been widely applied in interpretability research on deep neural networks. The classic gradient-weighted CAM (Grad-CAM) [28] calculated specific category information relative to the feature maps of each layer by determining the weights of each channel, thereby generating a heat map that is closely related to the target category. However, when dealing with multitarget and complex background scenes, Grad-CAM sometimes struggles to capture fine-grained details and local features. To address this limitation, Grad-CAM++ [29] further improved the high-order information, significantly enhancing sensitivity to small targets and overlapping regions. On the other hand, Eigen-CAM [30], which got rid of the reliance on gradient calculation, was based on principal component analysis to extract convolution features, generating class-independent visual interpretations in an unsupervised manner. Furthermore, to further enhance the detail and perceptual awareness of heat maps, LayerCAM [31] was proposed to fuse the activation maps from the deep and shallow convolution layers, resulting in a more pixel-level local correspondence and generating finer-grained heat maps that better express boundary and structural information. To validate the CAM method in medical diagnosis, many recent studies have explored its use. Windisch et al [32] proposed a ResNet50-based brain tumor detection model and used Grad-CAM for model interpretability analysis, used for the identification of meningiomas and gliomas in MRI slices. Shawon et al [33] proposed a cost-sensitive deep neural network that integrated multiple interpretable techniques (including Grad-CAM, LIME, and Score-CAM) for model interpretation, used for brain tumor detection under imbalanced data conditions. Dasanayaka et al [34] proposed a deep learning model based on U-Net and DenseNet, which used Grad-CAM to generate heat maps for brain tumor segmentation and classification. This combined strategy not only achieves effective segmentation and classification but also provides effective visualization. Zeineldin et al [35] proposed an enhanced EfficientNetv2 that integrates a global attention mechanism and efficient channel attention and uses Grad-CAM visualization for model interpretation, applied to MRI-based brain tumor classification. Guluwadi et al [36] proposed a brain MRI detection method that combines ResNet50 with Grad-CAM.

Objectives

In recent years, the rapid advancement of deep learning has significantly propelled the development of brain tumor detection, achieving remarkable results. However, current brain tumor detection models still face several challenges:

XSL•FO
RenderX

1. The boundaries of brain tumors are often blurry and the structures are irregular. The shallow features extracted by traditional convolutions are insufficient to effectively capture the key regions, leading to high rates of missed detection and false positives.
2. The parameter size of the target detection model is massive, and the computational overhead is high, which is not conducive to deployment on clinical edge devices, thus limiting their practicality. Traditional pruning methods, typically based on a single-weight threshold for coarse-grained pruning, can easily damage key model structures. At the same time, the sparse fine-tuning process, due to its inefficiency, struggles to meet the demands of high efficiency and deployability in medical scenarios.
3. Most target detection methods lack interpretability, making it difficult to clearly present the decision-making basis of the model to doctors, which limits its credibility in auxiliary clinical decision-making.

This paper proposes an interpretable, fine-grained brain tumor MRI detection method based on progressive pruning to address these challenges in an integrated manner. Importantly, CDCP-YOLO (CDC-YOLO with PHPS) is not a simple stacking of existing techniques but a task-driven co-design framework specifically tailored to the structural complexity, computational constraints, and interpretability requirements of brain tumor MRI detection. Its methodological distinctions are summarized as follows. First, unlike RCS-YOLO, which primarily accelerates inference through convolution reparameterization and channel shuffling, CDCP-YOLO introduces structure-aware input-level modeling (convolution Prewitt-and-pooling–based preprocessing [CSPP]) that explicitly embeds classical edge priors into the network. This design targets blurred and ambiguous tumor boundaries at the earliest stage of feature extraction—an issue that reparameterization alone cannot effectively address. Second, whereas BGF-YOLO improves detection performance mainly through multilevel feature fusion and additional detection heads, CDCP-YOLO adopts dynamic convolution–based C3k2 (DCC) feature adaptation, enabling input-dependent kernel generation. This strategy allows the network to flexibly adapt to heterogeneous tumor morphologies and scales without introducing detection head redundancy. Third, while PK-YOLO relies on external pretrained knowledge and sparse mask modeling to enhance small-target detection, CDCP-YOLO does not depend on any external pretraining paradigm. Instead, it improves representation capacity through internal, data-adaptive mechanisms, ensuring robustness and generalizability under limited or domain-specific medical datasets. Fourth, existing pruning-based detectors typically use single-criterion or one-shot pruning strategies that may disrupt critical feature pathways. In contrast, the proposed progressive hybrid pruning strategy (PHPS) jointly considers global channel sparsity and local structural dependency in a staged manner, enabling aggressive model compression while preserving detection-critical structures. Finally, unlike prior works that treat Grad-CAM–style visualization as a post hoc analysis tool, CDCP-YOLO tightly integrates Eigen-CAM into the detection head, treating interpretability as a core design objective rather than an auxiliary component. This integration ensures semantic consistency between detection results and visual explanations, which is essential for clinical decision support.

The main contributions of this work are summarized as follows:

1. Multiscale feature enhancement mechanism: a CSPP module was introduced at the model's input stage, integrating Prewitt edge detection with pooling operations to strengthen the construction of tumor edges and structures. A DCC module was introduced into the backbone and neck networks to achieve adaptive expression for different tumor forms and sizes. A channel prior convolutional attention (CPCA) module was introduced before the detection head to guide the network to focus on the key region of the most discriminative brain tumor.
2. Lightweight pruning strategy: a progressive hybrid pruning strategy was proposed, combining L1-norm and GroupNorm feature statistics. The pruning process is carried out in stages, prioritizing the pruning of redundant channels while maximally preserving key information flow. This method significantly reduces parameter size and computational overhead while effectively mitigating the performance degradation caused by large-scale pruning.
3. Interpretable confidence enhancement: a gradient-free principal component analysis method was adopted, which can generate clear saliency heat maps, visually demonstrating the most important basis for the brain tumor MRI model's decision-making. This significantly improves the transparency, credibility, and practicality of the model in clinical applications.

Methods

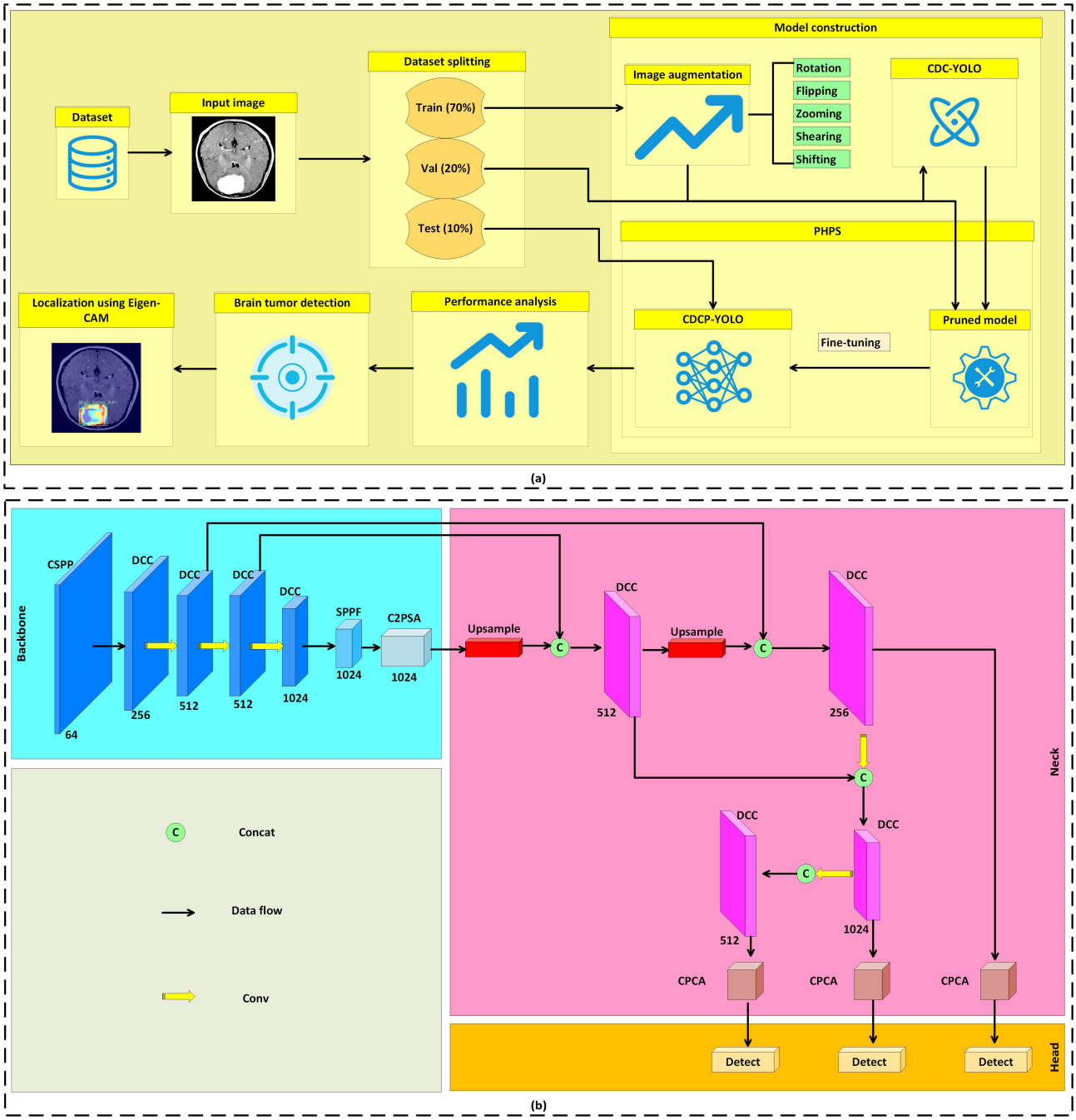
Overall Framework

This paper proposes an interpretable and fine-grained brain tumor detection model based on progressive pruning. As shown in Figure 1A, the overall framework was as follows: MRI images were trained on the improved CDC-YOLO network, and an efficient CDCP-YOLO model was obtained through a progressive hybrid pruning strategy and fine-tuning. As shown in Figure 1B, CDC-YOLO is composed of three main structures:

1. Backbone stage: the CSPP module is used to replace the first 2 convolutional layers. CSPP combines Prewitt edge detection with pooling operations to enhance the model's initial feature extraction capability for edges and textures. Subsequently, the main network uses multiple DCC modules, using dynamic convolution to enhance the perceptual ability for multiscale tumor regions.
2. Neck stage: multiscale feature maps are fused through upsampling and concatenation. The C3k2 module introduces dynamic convolution and key location fusion, and the CPCA module is introduced, combining channel and spatial attention mechanisms to guide the model to focus on key diagnostic areas and improve detection accuracy.
3. Head stage: multiple detection heads of different sizes are used to detect targets of different sizes.

XSL•FO
RenderX

Figure 1. Overall framework and network architecture of the proposed CDCP-YOLO model for brain tumor magnetic resonance imaging detection. (A) The complete training and pruning pipeline, including feature enhancement, progressive hybrid pruning, and fine-tuning. (B) Detailed architecture of the CDC-YOLO backbone, neck, and detection head, where CSPP enhances edge features at the input stage, DCC adapts to multiscale tumor morphology, and channel prior convolutional attention (CPCA) guides the detection head to focus on discriminative channels and spatial regions. C2PSA, C2 block integrated with partial self-attention; CAM, class activation mapping; CDC-YOLO, cross-scale dynamic convolution–based YOLO; CDCP-YOLO, CDC-YOLO with PHPS; CSPP, convolution Prewitt-and-pooling–based preprocessing; Concat, concatenation; Conv, convolution; DCC, dynamic convolution–based C3k2 module; SPPF, spatial pyramid pooling-fast; Val, validation subset; YOLO, "you only look once" framework.

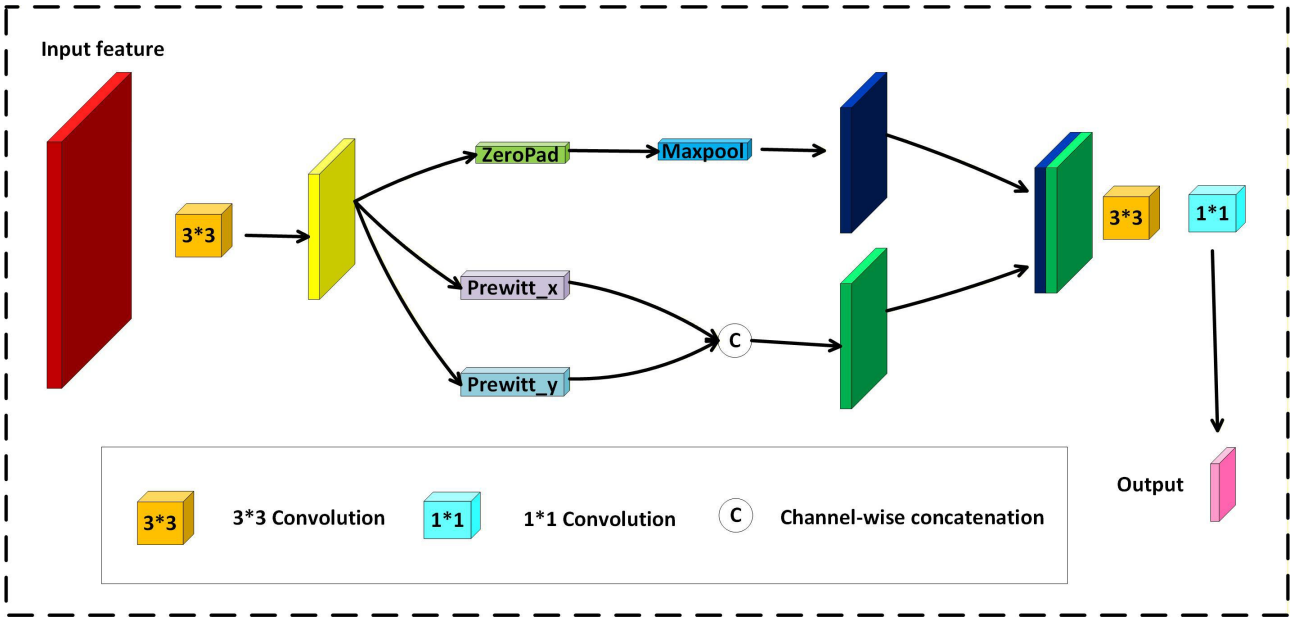


Convolution Prewitt-and-Pooling–Based Preprocessing

The head of YOLOv11 includes 2 initial convolution layers, primarily used to extract initial features from the input image. By extracting deeper features layer by layer, the network can construct sufficient information for effective target detection. To enhance the model's perceptual ability for edge structures, the first 2 standard convolutional layers in YOLOv11 are replaced with a self-designed CSPP (Figure 2), which more effectively extracts contour information and local texture features from the image. As the input feature processing unit of the entire CDCP-YOLO, CSPP integrates standard convolution, Prewitt edge detection, and multiscale pooling operations, aiming to enhance the model's perceptual ability for brain tumor edge features and its adaptability to different shapes.

XSL•FO
RenderX

Figure 2. Structural illustration of the convolution Prewitt-and-pooling–based preprocessing (CSPP) module. The module integrates standard convolution, Prewitt edge detection (horizontal and vertical), and max pooling to enhance tumor boundary and texture representation at the input stage. Maxpool, maximum pool.



First, the input image X undergoes a 3×3 convolution operation to obtain the initial feature map X1; this operation completes spatial compression and channel expansion.

(1)X1=Conv(X)

Next, the Prewitt edge detection operator is applied to X1, filtering in the horizontal direction Gx and the vertical direction Gy to extract the edge response in the image.

(2)XPrewitt=PrewittConv(X1)

The output of the Prewitt operator is obtained by calculating the Euclidean distance to combine the responses from both directions, resulting in the complete edge information XPrewitt.

(3)XPrewitt=(X1×Gx)2+(X1×Gy)2

Simultaneously, another branch inputs X1 into a max-pooling layer, and the spatial dimensions are kept unchanged through zero-padding, generating the pooled feature Xpool. The 2 output features, XPrewitt and Xpool, are concatenated along the channel dimension to generate the feature X2. This fusion retains both structural edges and spatial context information.

(4)X2=Concat(XPrewitt,Xpool)

The fused feature map X2 is then passed through a 3×3 convolution to further extract high-level features.

(5)X3=Conv3×3(X2)

Finally, the number of channels is compressed, and the dimensions are adjusted using a 1×1 convolution to generate the final feature Y, which serves as the input to the backbone structure.

(6)Y=Conv1×1(X3)

CSPP uses the Prewitt operator in both horizontal and vertical directions to automatically extract prominent edge contours in the image, significantly enhancing the detection capability for blurry contours and unclear tumor regions. At the same time, the output of Prewitt is fused with the original feature map, which not only preserves the local details but also strengthens the expression of structural information. In addition, the pooling operation compresses the spatial dimensions while keeping the feature map size unchanged, providing the network more robustness to variations in tumor size and location.

The CSPP module is designed to introduce deterministic structural priors into the early stages of feature extraction. While standard convolutional layers learn kernels stochastically, the inclusion of a fixed Prewitt operator provides explicit edge-sensitive cues that are critical for delineating tumor boundaries. We acknowledge that in medical imaging physics, difference operators are typically sensitive to high-frequency noise. Therefore, rather than using the Prewitt output directly, the CSPP module integrates it through a multistage fusion strategy. Specifically, the gradient maps are processed via max pooling to perform local maximum selection, which emphasizes strong structural edges while suppressing isolated noise spikes. These features are then fused with learnable convolutional features using a 1×1 convolution layer and batch normalization. This design allows the network to adaptively weight the explicit structural priors against learnable representations, ensuring that the model captures fine-grained morphological details without amplifying imaging artifacts.

Dynamic Convolution-Based C3k2

The C3k2 module is a core feature extraction unit in YOLOv11, and its design goal is to achieve the best balance between feature representation capability and computational efficiency. This module uses a multibranch structure and residual connections to enhance the multiscale feature extraction capability while maintaining training stability. C3k2 supports two configuration modes:

1. In the C3k=True mode, a lightweight C3k branch is used. The input features are divided into multiple groups for separate processing and then fused through concatenation. This group-wise convolution strategy significantly reduces

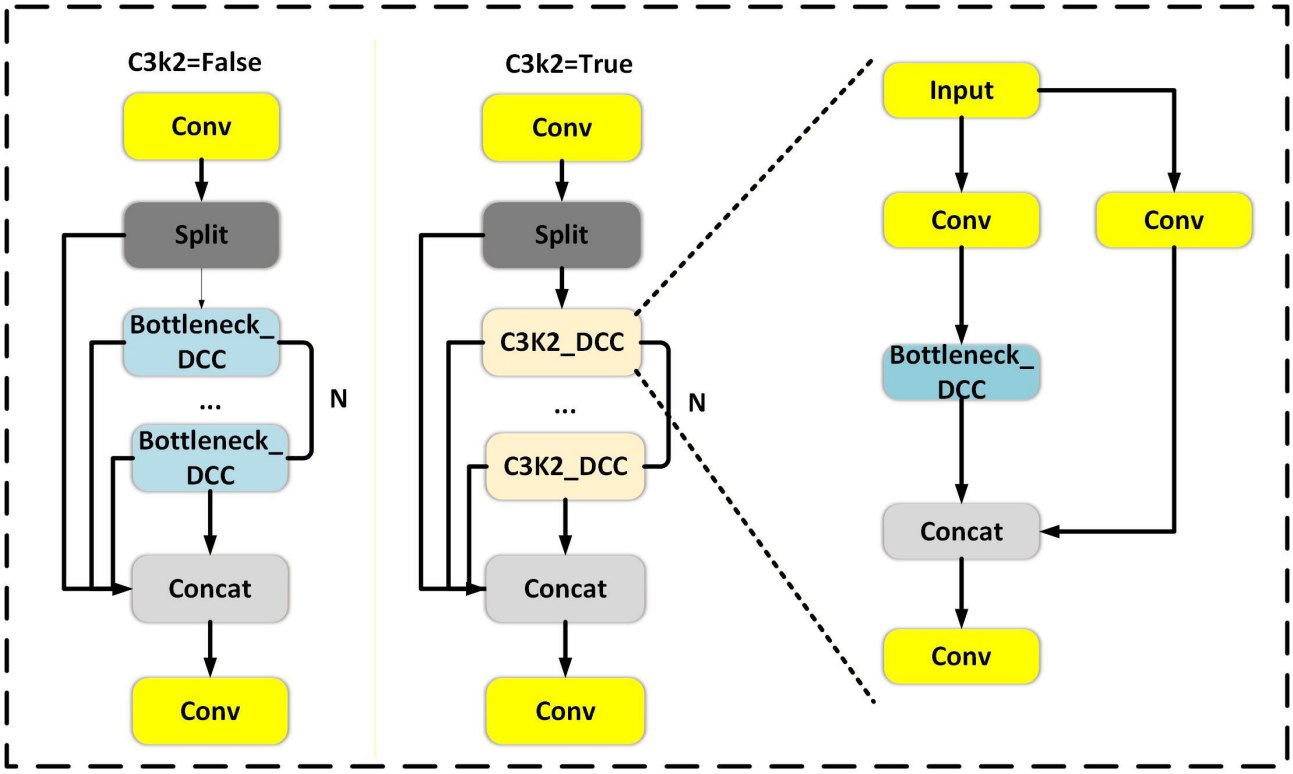
XSL•FO
RenderX

computational complexity, making it particularly suitable for applications with high real-time requirements.

2. In the C3k=False mode, a more complex bottleneck structure is introduced, combined with additional convolution and activation layers to extract deeper and more discriminative features, thus showing superiority in more accuracy-sensitive tasks.

To enhance the network's modeling capability for complex tumor regions, in this study, we built a DCC module (Figure 3) based on the original C3k2 module by introducing a dynamic convolution [37]. The DCC module enhances the model's capabilities in multiscale structure modeling, detail preservation, and contextual awareness by replacing the fixed convolution with the input-adaptive dynamic convolution.

Figure 3. Architecture of the dynamic convolution–based C3k2 (DCC) module. The convolution kernel is dynamically generated according to the global context of the input feature map, enabling adaptive modeling of tumors with diverse sizes and shapes. Concat, concatenation; Conv, convolution.



Given an input feature X∈RCin×H×W, the DCC module uses a dynamic kernel *W(X)* to perform the convolution operation:

(7)Y=W(X)×X

The generation process of the convolution kernel WX depends on the global context information of the input image. To achieve this weight regeneration, a global average pooling layer is first used to compute the channel-wise global vector:

(8)g=1H×W∑i=1H∑j=1WX[:,i,j]

Then, 2 fully connected layers and a nonlinear activation function are used to generate the dynamic attention weights:

(9)θ=FC2(σ(FC1(g)))

Here σ() is the ReLU activation function, and FC1 and FC2 are the first and second fully connected layers, respectively. The final parameter θ determines the weighting of the dynamic convolution kernel WX, thereby adapting to the feature distribution of the current input image for optimal perception.

The C3k2 module mainly extracts higher-level features through multiple bottleneck and convolution layers. This module stacks multiple bottleneck units, replacing standard static convolution operations, enabling the network to adaptively select more suitable feature representations at different spatial scales, and thereby improving the modeling capability for diverse tumor features. Unlike fixed convolution kernels, the dynamic convolution mechanism automatically adjusts the kernel weights based on the input features, making it adaptable to different image contexts. Especially when dealing with tumor regions with blurry boundaries and irregular shapes, dynamic convolution can more finely capture local structures and texture changes. In addition, dynamic convolution also enhances the model's ability to resist interference from complex backgrounds and low signal-to-noise ratios, reducing the occurrence of missed and false detections. At the same time, the DCC module introduces an input-dependent kernel selection strategy, which endows the C3k2 module with stronger language modeling and morphological adaptation capabilities, providing richer and more discriminative feature representations for subsequent detection heads.

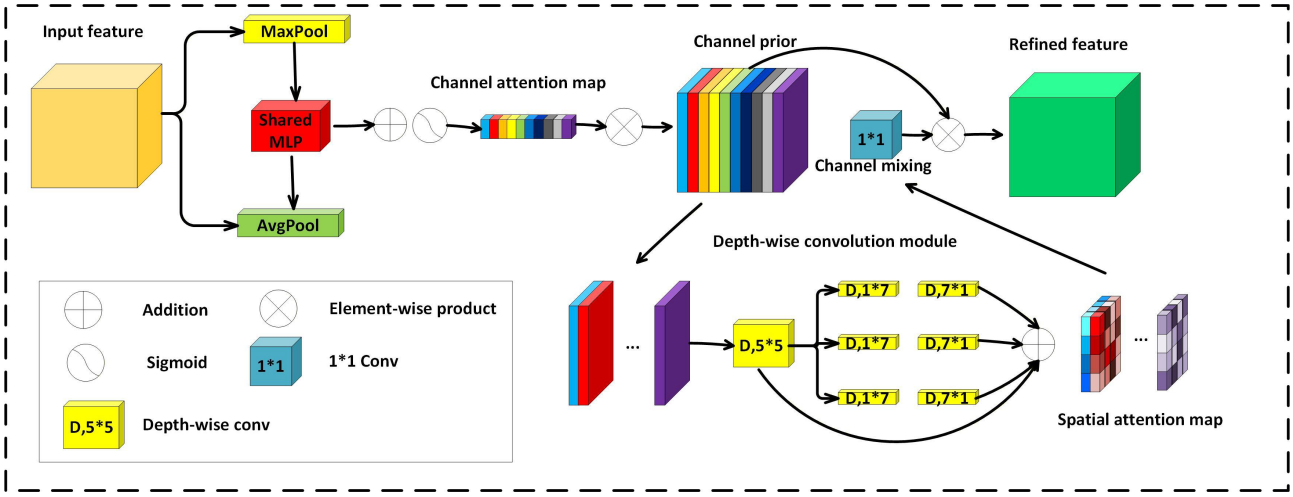
Channel Prior Convolutional Attention

The detection head is the key part for the final bounding box regression. However, the detection head relies on the feature maps from the previous layer, and these feature maps may contain a large amount of information, where many channels may be redundant or irrelevant to the target detection. The attention mechanism is proposed to focus on important information. The CPCA module (Figure 4) [38] uses multiscale depth-wise separable convolutions to maintain the channel prior

XSL•FO
RenderX

while extracting spatial relationships, enabling the network to focus on information-rich channels and key spatial regions. The CPCA module includes the sequential placement of channel attention and spatial attention. The spatial information of the feature maps is aggregated by channel attention through operations such as average pooling and max pooling. The spatial attention is then processed through a shared multilayer perceptron and added to generate a channel attention map. The channel prior is obtained by element-wise multiplication of the input feature and the channel attention map. Subsequently, the channel prior is input into a depth-wise convolution block to generate a spatial attention map. The convolution block receives the spatial attention map to perform channel mixing. Finally, the channel mixing result is element-wise multiplied with the channel prior to obtain the optimized feature as output. The channel mixing process helps to enhance the feature representation.

Figure 4. Structure of the channel prior convolutional attention (CPCA) module, which sequentially applies channel attention and spatial attention to emphasize tumor-relevant channels and spatial regions before the detection head. Conv, convolution; MaxPool, maximum pool; AvgPool, average pool.



Given an intermediate feature map F∈ RC×H×W, a 1D channel attention map Mc∈ RC×1×1 is first inferred through the channel attention module. Then, Mc is element-wise multiplied with the input feature F to obtain the channel attention–optimized feature map Fc∈ RC×H×W. Subsequently, the spatial attention (SA) processes Fc to generate a 3D spatial attention map Ms∈ RC×H×W. The final output feature map F^ is obtained by element-wise multiplication of Msand Fc:

(10)Fc=CA(F)⊗F

(11)F^=SA(Fc)⊗Fc

, where ⊗ represents element-wise multiplication. Channel attention (CA) is obtained as follows:

(12)CA(F)=σ(MLP(AvgPool(F))+MLP(MaxPool(F)))

, where σ represents the sigmoid function. Spatial attention (SA) is obtained as follows:

(13)SA(F)=Conv1×1(∑i=03Branchi(DwConv(F)))

, where DwConv represents depth-wise convolution, Branchi represents the *i*th branch, and Branch0 is an identity connection.

Applying CPCA to the predetection layer optimizes the semantic aggregation capability of the predetection feature layer. The detection head, as the key component for classification and bounding box regression, heavily relies on the quality of the preceding feature layers. However, the original feature maps often contain a large amount of redundant channels or low-response regions, which interfere with the model's discrimination process, especially in medical images with blurry tumor boundaries and complex backgrounds. CPCA performs adaptive channel modeling, enabling it to automatically identify and enhance significant channels related to tumors, thereby effectively improving the discriminability of features. Furthermore, this mechanism, combined with a spatial attention strategy, guides the network to focus on key areas of the image, enhancing the perceptual ability for fine-grained features such as lesion edges and structural changes. By introducing CPCA before the detection head, the model can perform inference based on more refined, robust, and contextually sensitive features.

Progressive Hybrid Pruning Strategy

To address the dual requirements of high-precision modeling for complex brain tumor structures and efficient inference, we propose the PHPS, as summarized in Textbox 1. The PHPS first uses an L1-norm–based channel importance evaluation to perform coarse-grained pruning, enabling the rapid removal of globally redundant feature channels. Subsequently, GroupNorm-based grouping information is incorporated to conduct fine-grained structural pruning, which corrects channel importance by considering local structural dependencies and avoids distortions caused by sole reliance on weight magnitude.

XSL•FO
RenderX

Textbox 1. Pruning processing

- Input: pretrained model parameters W, overall pruning ratio s, pruning threshold SC, step pruning ratio M.
- Output: pruned and fine-tuned model W1

1. Initialize CPC-YOLO model.
2. Set target sparsity s; define thresholds SC.
3. Initialize current global sparsity S←0.
4. While S<s, do:
5. For each prunable layer l in the model, do:
6. If the current layer's sparsity SL < SC, then:
7. Compute importance score: ILI(wi).
8. Sort channels in ascending order.
9. Prune the lowest-ranked channels with ratio M.
10. Else:
11. Compute importance score:IGN(Gj)=∑wi∈Gjwi2.
12. Sort groups by IGN(Gj) in ascending order.
13. Prune the lowest-ranked groups with ratio M.
14. End if
15. End for
16. Update the global sparsity S.
17. End while
18. Fine-tune the pruned model on the training dataset.
19. Return the final pruned model W1.

Through this dual evaluation mechanism of global coarse-grained pruning combined with local fine-grained structural correction, the PHPS can more accurately preserve discriminative feature channels under challenging conditions such as blurred tumor boundaries and high intertumor heterogeneity. Importantly, the PHPS adopts a progressive pruning process, in which sparsity is gradually increased toward a predefined global target. Model fine-tuning is performed only once after the entire pruning process is completed, allowing effective recovery of detection performance while avoiding excessive training overhead. This design achieves a balanced optimization between model compression and precision retention.

Interpretability

In the detection head output stage, the feature map is directly sent to the Eigen-CAM branch while being used for bounding box regression. The most representative convolution activation regions are extracted through principal component analysis. This design not only ensures the semantic consistency between interpretation and detection but also allows for reverse influence on model optimization during training. The extraction of saliency regions enables the network to focus more on real lesion boundaries and key structures, enhancing the model's adaptability to blurry edges and complex backgrounds. Unlike traditional Grad-CAM, which relies on backpropagation, Eigen-CAM is based on forward features, is class agnostic, lightweight, and efficient, making it more suitable for brain tumor MRI detection.

In this work, Eigen-CAM was applied to the high-level convolutional feature maps before the detection head, rather than to class-specific logits. Although Eigen-CAM is a class-agnostic visualization method that does not rely on gradients with respect to a specific category, the visualized activations are still strongly task driven. Since the detector is trained to localize tumor regions using bounding box regression and objective supervision, the learned high-level features inherently emphasize spatial regions that are most relevant to tumor detection. As a result, the generated heat maps naturally concentrate on tumor areas instead of irrelevant anatomical structures such as the skull or eyes.

For a certain convolution feature layer $F \in RC \times H \times W$, it is first reshaped into a 2D matrix:

$$F' \in RC \times (H \times W) \quad (14)$$

Then the covariance matrix is calculated:

$$\Sigma = F'F'T \quad (15)$$

Next, eigenvalue decomposition is performed on it, and taking the eigenvector u1 corresponding to the largest eigenvalue, a heat map is finally generated by weighting the principal components:

$$M = \sum c=1 Cu1[c] \cdot Fc \quad (16)$$

XSL•FO
RenderX

Ethical Considerations

This study is based entirely on publicly available brain tumor MRI datasets, including Br35H [39], Roboflow dataset [40], and Capstone Brain Tumor dataset [41]. All datasets were obtained from open-access platforms (Kaggle and Roboflow) under their respective terms of use. Each dataset contains anonymized medical images that do not include any personally identifiable information, patient metadata, or clinical identifiers. Therefore, no additional ethical approval or informed consent was required for the use of these datasets.

The datasets were used solely for academic and noncommercial research purposes, strictly following ethical research guidelines and data-sharing policies. All experimental procedures comply with the principles of the Declaration of Helsinki and institutional data protection standards.

Results

Experimental Setup

Br35H [39] was used as the main dataset. This dataset is one of the most representative and widely used public benchmarks in brain tumor MRI detection, covering multiple tumor types and MRI modalities. However, it should be noted that the specific clinical metadata regarding the exact MRI sequences (eg, T1-weighted, T2-weighted, or fluid-attenuated inversion recovery) and the specific tumor subtypes are not provided in the original dataset repository. Despite the absence of these specific labels, its high-quality annotations, imaging diversity, and clinical coverage make it a standard for performance evaluation and comparison in automatic brain tumor detection research. To verify the robustness of CDCP-YOLO, the experiments were also conducted on 2 other datasets, Roboflow [40] and Capstone [41]. We strictly adhered to the original dataset split, as detailed in Table 1.

Table . Dataset splits used for experiments on the Br35H, Roboflow, and Capstone brain tumor magnetic resonance imaging (MRI) datasets.[a]

Dataset	Total number of images, n	Training set, n	Validation set, n	Test set, n
Br35H	801	500	201	100
Roboflow	300	210	60	30
Capstone	911	638	182	91

[a]The table reports the total number of images and the corresponding training, validation, and test splits for each dataset. These official splits were strictly followed in all experiments to ensure fair evaluation and reproducibility of the reported results.

To enhance the robustness and generalizability of the model, data augmentation was performed during the training phase, including geometric transformations such as rotation, flipping, scaling, and shearing of MRI images, as well as random adjustments to the brightness and contrast. This strategy effectively expands the diversity of the training samples, thereby reducing the risk of model overfitting. This enhancement operation was generated in real time during each training epoch, ensuring that the model obtained “new” samples for training in different epochs, improving its adaptability to complex scenes and diverse tumor morphology. The specific data augmentation hyperparameters used during training included mosaic image composition (mosaic=1.0), geometric transformations (random rotation [degress=10], scaling [scale=0.5], and horizontal flipping [fliplr=0.5]), and color-space perturbations in the hue-saturation-value (HSV) domain (hsv_h=0.015, hsv_s=0.7, and hsv_v=0.4). These augmentation strategies are widely adopted in YOLO-based detection frameworks to enhance data diversity and improve model robustness against variations in object scale, orientation, and imaging conditions.

The experiment was conducted using an Intel Xeon Gold 5320 CPU @ 2.20GHz and an NVIDIA A40 48GB GPU. The training was conducted for 300 epochs, with a batch size of 16, using a stochastic gradient descent (SGD) optimizer, and a patience parameter of 50. The learning rate was set to 0.01. The software environment consisted of Ubuntu 20.04 (Canonical), CUDA 11.8 (NVIDIA), and PyTorch 2.1.0 (Meta AI Research).

After the completion of the initial full training stage, the PHPS was applied to the converged model. Specifically, the base model was first trained for 300 epochs using the standard training configuration described above. Once convergence was achieved, structured pruning was performed according to the PHPS criteria to remove redundant channels and blocks.

Following pruning, the resulting compact model was subjected to a dedicated fine-tuning stage, which was treated as an independent optimization phase and clearly distinguished from the initial training process. During this postpruning fine-tuning, the model was trained for 300 epochs under the same hardware and software environment as the initial training stage, using the SGD optimizer. The batch size was set to 8, the learning rate was fixed at 0.01, and a patience parameter of 50 was applied for early stopping.

Evaluation Metrics

Precision is the ratio of the number of correctly predicted positive samples to the total number of the detected samples, as shown below:

(17)Precision=TPTP+FP×100%

Recall is the ratio of the number of correctly predicted positive samples to the number of the actual positive samples, as shown below:

(18)Recall=TPTP+FN×100%

Mean average precision (mAP) is the result obtained by averaging the average precision of all categories, used to measure the detection performance of the model across all categories.

(19)mAP=∑APNC×100%

XSL•FO
RenderX

, where $mAP_{0.5}$ represents the mAP at an IoU threshold of 0.5 and $mAP_{0.5:0.95}$ represents the average mAP over several IoU thresholds (0.50-0.95). Params represents parameter size in the model, the number of floating-point operations (×10 ; GFLOPs) signify computational complexity, and frames per second (FPS) were used to measure inference complexity.

SD was used to measure the variability of a given evaluation metric for the same method across multiple repeated experiments. It was computed as follows:

$$SD=\frac{1}{N-1}\sum_{i=1}^{N}(M_i-\bar{M})^2 \quad (20)$$

, where Mi denotes the value of the evaluation metric obtained in the ith experiment, and $\bar{M}$ represents the mean value of the corresponding metric over N experiments.

Comparative Experiments

This paper compares between a variety of mainstream detection models on Br35H, including the classic YOLO series models (such as YOLOv3, YOLOv5n, YOLOv10n, and YOLOv11n); 2-stage detectors (such as Faster R-CNN and Cascade R-CNN); and a 1-stage detection method, TOOD (Table 2). To ensure a fair and unbiased comparison, all baseline models reported in Table 2 were trained from scratch under a unified experimental protocol.

Table . Performance comparison of different object detection models on the Br35H brain tumor magnetic resonance imaging dataset.[a]

Model	Precision[b]	Recall	$mAP_{0.5}$[c]	$mAP_{0.5:0.95}$[d]	Params (M)[e]	GFLOPs[f]	FPS[g]
YOLOv3-tiny	0.932	0.836	0.898	0.569	12.12	18.9	240.8
YOLOv3	0.926	0.836	0.861	0.602	103.67	282.2	70.2
YOLOv5n	0.891	0.869	0.938	0.601	2.50	7.1	134
YOLOv6n	0.91	0.844	0.931	0.578	4.23	11.8	145
YOLOv8n	0.928	0.843	0.914	0.604	3	8.1	144
TOOD	0.888	0.895	0.925	0.630	32.02	144	13.1
YOLOv10n	0.919	0.84	0.914	0.608	2.27	6.5	131
YOLOv11n	0.904	0.853	0.918	0.585	2.58	6.3	142.2
Faster R-CNN	0.854	0.836	0.896	0.578	41.35	155	15.2
Cascade R-CNN	0.86	0.895	0.874	0.584	69.15	183	14
DINO	0.939	0.75	0.84	0.563	47.54	205	10.5
RCS-YOLO	0.953	0.828	0.824	0.548	45.70	94.5	215.2
BGF-YOLO	0.964	0.885	0.959	0.648	84.30	568.9	32.5
CDC-YOLO	0.881	0.906	0.946	0.660	3.64	7	130
CDCP-YOLO	0.918	0.904	0.944	0.644	2.07	3.3	152

[a]All models were trained and evaluated under the same experimental protocol to ensure a fair comparison. Higher values indicate better performance for accuracy and speed metrics, whereas lower values indicate better efficiency for model complexity metrics.

[b]Precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$ were used to evaluate detection accuracy.

[c]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[d]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[e]Params (M) was used to measure model size.

[f]GFLOPs: number of floating-point operations (×10); used to measure computational complexity.

[g]FPS: frames per second; used to measure inference efficiency.

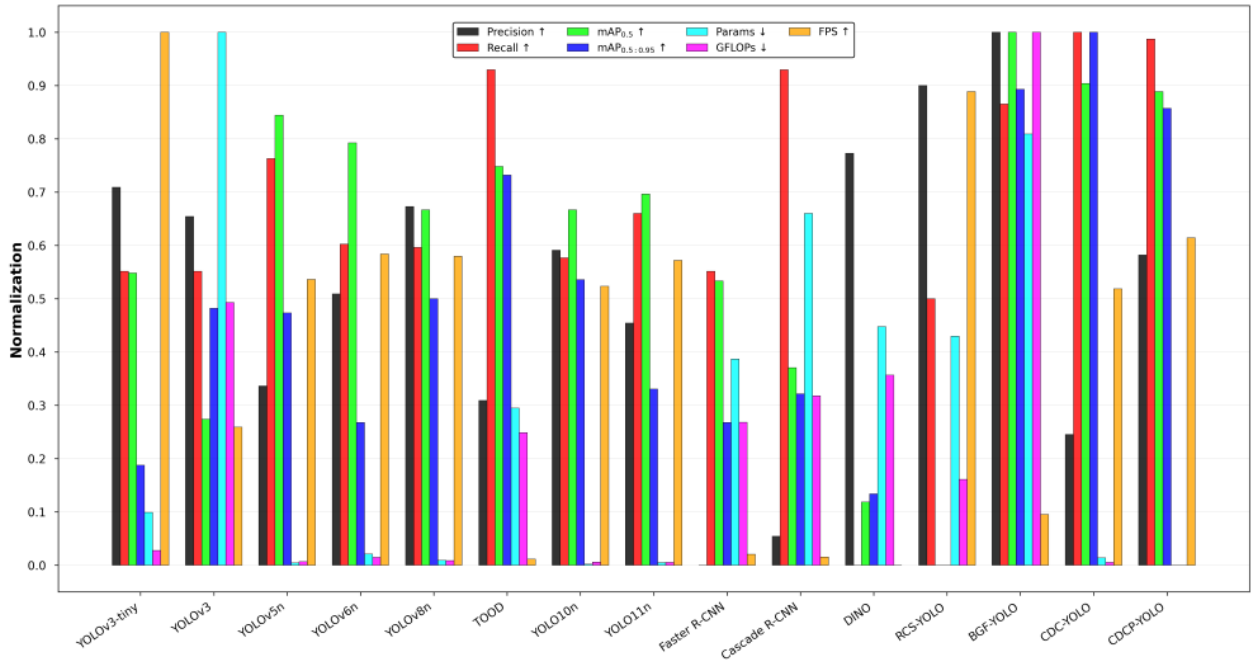
CDCP-YOLO reached 0.946 in $mAP_{0.5}$, approaching CDC-YOLO's 0.944, significantly outperforming traditional models such as YOLOv3-tiny (0.898) and YOLOv3 (0.861). At the same time, it achieved a high score of 0.644 on the more challenging $mAP_{0.5:0.95}$ metric, second only to CDC-YOLO (0.660). In addition, CDCP-YOLO's recall (0.904) and precision (0.918) were both at a leading level, indicating its excellent detection capabilities. In terms of model complexity, the parameter count of CDCP-YOLO was only 2.07M, which is about 1/33 of that of Cascade R-CNN (69.15M) and 1/20 of that of Faster R-CNN (41.35M), and was also smaller than that of YOLOv3 (103.67M), achieving the goal of lightweight design. Its computational complexity was 3.3 GFLOPs, which was also significantly lower than that of YOLOv5n (7.1 GFLOPs) and YOLOv6n (11.8 GFLOPs), demonstrating good computational efficiency. In terms of inference speed, CDCP-YOLO reached 152 FPS, which is significantly faster than Cascade R-CNN (14 FPS), Faster R-CNN (15.2 FPS), and even some lightweight YOLO models such as YOLOv5n (134 FPS) and YOLOv10n (131 FPS), showing significant advantages in practical deployment.

As seen in Figure 5, CDCP-YOLO was in the optimal or near-optimal state in terms of accuracy, model size, computational complexity, and inference speed, demonstrating

XSL•FO
RenderX

the advantage of balancing accuracy and efficiency. It is particularly noteworthy that CDCP-YOLO significantly surpassed YOLOv5n (0.644 vs 0.601) and was superior to TOOD (0.644 vs 0.630) on $mAP_{0.5:0.95}$, while it had a much lower parameter count and computational load than these models. This indicates that the model still has stable performance at high IoU thresholds, possessing good robustness and generalizability. CDCP-YOLO achieved high detection accuracy and inference speed under the premise of maintaining a small model size and low computational overhead.

Figure 5. Multimetric normalization comparison of different object detection models on the Br35H brain tumor magnetic resonance imaging dataset. All evaluation metrics are normalized to the range [0, 1] to enable fair comparison across models with different scales. The compared models included YOLOv3-tiny, YOLOv3, YOLOv5n, YOLOv6n, YOLOv8n, YOLOv10n, YOLOv11n, Faster R-CNN, Cascade R-CNN, DINO, RCS-YOLO, BGF-YOLO, CDC-YOLO, and the proposed CDCP-YOLO model. The metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). Arrows indicate optimization direction: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). FPS, frames per second; GFLOPs, number of floating-point operations (×10); mAP, mean average precision; $mAP_{0.5}$, mAP at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mAP over several intersection-over-union thresholds (0.50-0.95); R-CNN, region-based convolutional neural network; YOLO, “you only look once” framework.



Ablation Experiment

Using YOLOv11n as a baseline, the effectiveness of each module is compared (Table 3).

XSL•FO
RenderX

Table . Ablation study results of individual modules on the Br35H brain tumor magnetic resonance imaging dataset, showing the incremental impact of the CSPP[a], DCC[b], CPCA[c], and PHPS[d] modules on detection performance and model efficiency.

Model	CSPP	DCC	CPCA	PHPS	Precision[e]	Recall	$mAP_{0.5}$[f]	$mAP_{0.5:0.95}$[g]	Params (M)[h]	GFLOPs[i]	FPS[j]
YOLOv11n	—[k]	—	—	—	0.904	0.853	0.918	0.585	2.58	6.3	142.2
C-YOLO	✓[l]	—	—	—	0.904	0.852	0.921	0.639	2.58	6.5	140
CD-YOLO	✓	✓	—	—	0.900	0.893	0.941	0.660	3.46	6.3	133
CDC-YOLO	✓	✓	✓	—	0.881	0.906	0.946	0.660	3.64	7.0	130
CDCP-YOLO	✓	✓	✓	✓	0.918	0.904	0.944	0.644	2.07	3.3	152

[a]CSPP: convolution Prewitt-and-pooling–based preprocessing.

[b]DCC: dynamic convolution–based C3k2.

[c]CPCA: channel prior convolutional attention.

[d]PHPS: progressive hybrid pruning strategy.

[e]Precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$ were used to evaluate detection accuracy.

[f]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[g]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[h]Params (M) was used to measure model size.

[i]GFLOPs: number of floating-point operations (×10); used to measure computational complexity.

[j]FPS: frames per second; used to measure inference efficiency.

[k]Indicates that the corresponding module is absent in the model.

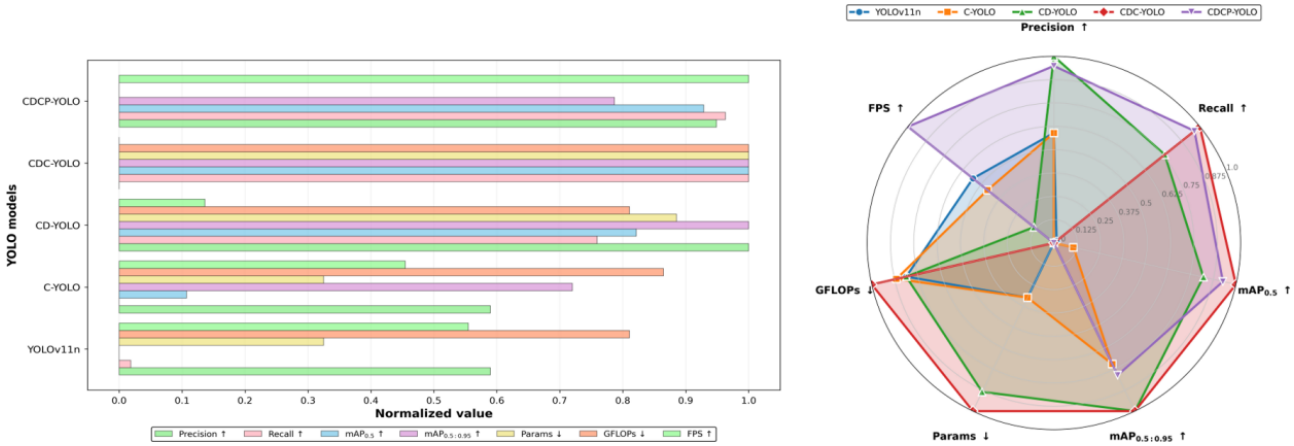
[l]A check mark (✓) indicates that the corresponding module is included in the model.

The $mAP_{0.5}$ of the original YOLOv11n was 0.918, $mAP_{0.5:0.95}$ was 0.585, parameter size was 2.58M, computational complexity was 6.3 GFLOPs, and inference speed was 142.2 FPS. After adding the CSPP module, the $mAP_{0.5}$ of C-YOLO slightly increased to 0.921 and its $mAP_{0.5:0.95}$ significantly increased to 0.639, while the number of parameters and computational complexity remained unchanged, indicating that the CSPP structure has a good effect on improving the model's ability to capture fine-grained features. After introducing the DCC module, the detection accuracy was further improved: the $mAP_{0.5}$ of CD-YOLO increased to 0.941, and the $mAP_{0.5:0.95}$ increased to 0.660. Although the parameter size slightly increased to 3.46M, computational complexity remained at 6.3 GFLOPs, and inference speed dropped to 133 FPS, the model's improvement in accuracy effectively enhanced its contextual modeling capability. Subsequently, with the addition of the CPCA module to obtain CDC-YOLO, the $mAP_{0.5}$ reached 0.946 (the highest value), and the $mAP_{0.5:0.95}$ remained unchanged at 0.660. Although parameter size slightly increased to 3.64M and inference speed dropped to 130 FPS, CDC-YOLO still performed outstandingly in terms of detection accuracy, further verifying the performance improvement effect of CPCA. On the basis of CDC-YOLO, PHPS was introduced to obtain CDCP-YOLO, which achieved a rebalance of model lightweighting and performance. This model reached 0.944 in $mAP_{0.5}$ and 0.644 in $mAP_{0.5:0.95}$, with only 2.07M parameters and computation complexity of 3.3 GFLOPs. The inference speed significantly increased to 152 FPS. This result shows that the PHPS basically retains high performance while compressing the model size.

As Figure 6 shows, CDCP-YOLO achieved a good balance in terms of accuracy, model size, computational load, and speed, showing a high degree of practicality and deployment advantages. Compared with other improved models, CDCP-YOLO had an absolute advantage in the 2 dimensions "fewest parameters" and "fastest inference," while also maintaining excellent performance in detection accuracy.

XSL•FO
RenderX

Figure 6. Ablation study results of the proposed CDCP-YOLO on the Br35H brain tumor magnetic resonance imaging dataset. The figure on the left shows the multimetric normalization bar chart, where all evaluation metrics are normalized to the range [0, 1] for fair comparison across models. The metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). The figure on the right presents a radar-based performance analysis plot summarizing the overall accuracy-efficiency trade-off of different ablation variants (YOLOv11n, C-YOLO, CD-YOLO, CDC-YOLO, and CDCP-YOLO). Arrows indicate metric preference: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). This figure demonstrates how each module (convolution Prewitt-and-pooling–based preprocessing [CSPP], dynamic convolution–based C3k2 [DCC], channel prior convolutional attention [CPCA], and progressive hybrid pruning strategy [PHPS]) progressively improves detection accuracy while reducing model complexity. FPS, frames per second; GFLOPs, number of floating-point operations (×10); mAP, mean average precision; $mAP_{0.5}$, mAP at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mAP over several intersection-over-union thresholds (0.50-0.95); YOLO, "you only look once" framework.



Performance on Different Datasets

To comprehensively evaluate the performance of CDCP-YOLO, the performance of YOLOv11, CDC-YOLO, and CDCP-YOLO was compared on different datasets (Table 4).

Table . Performance comparison of YOLOv11, CDC-YOLO, and CDCP-YOLO on 3 brain tumor magnetic resonance imaging datasets: Br35H, Roboflow, and Capstone.[a]

Dataset and model	Precision	Recall	$mAP_{0.5}$[b]	$mAP_{0.5:0.95}$[c]	Params (M)[d]	GFLOPs[e]	FPS[f]
Br35H							
YOLOv11	0.904	0.853	0.918	0.585	2.58	6.3	142.2
CDC-YOLO	0.881	0.906	0.946	0.660	3.64	7.0	130
CDCP-YOLO	0.918	0.904	0.944	0.644	2.07	3.3	152
Roboflow							
YOLOv11	0.615	0.467	0.561	0.276	2.58	6.3	117.6
CDC-YOLO	0.706	0.716	0.77	0.361	3.64	7.0	115
CDCP-YOLO	0.699	0.729	0.756	0.353	2.07	3.3	130.5
Capstone							
YOLOv11	0.768	0.806	0.840	0.567	2.58	6.3	132.2
CDC-YOLO	0.877	0.828	0.876	0.625	3.64	7.0	130.3
CDCP-YOLO	0.897	0.845	0.909	0.625	2.07	3.3	150

[a]The table reports detection accuracy metrics (precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$) and efficiency metrics (Params, GFLOPs, and FPS) to evaluate both effectiveness and deployability across different datasets.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[d]Params (M) was used to measure model size.

[e]GFLOPs: number of floating-point operations (×10).

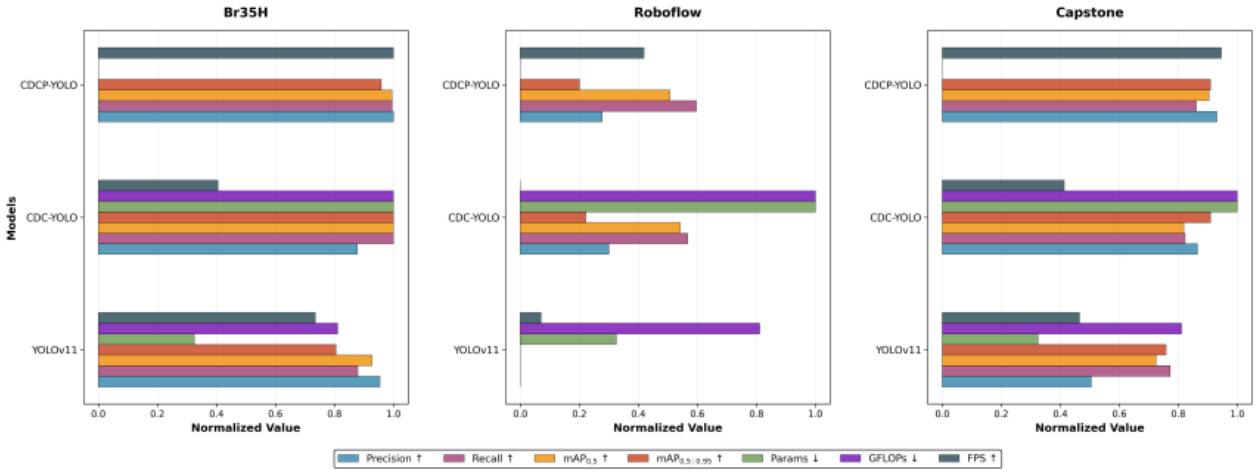
[f]FPS: frames per second.

XSL•FO
RenderX

On Br35H, CDCP-YOLO was superior to other models in terms of precision (0.918) and inference speed (152 FPS). Although its $mAP_{0.5}$ was slightly lower than that of CDC-YOLO (0.944 vs 0.946), the outstanding performance of CDCP-YOLO in terms of parameter size (2.07M) and computational complexity (3.3 GFLOPs) makes it more suitable for practical application scenarios with limited resources. In addition, the $mAP_{0.5:0.95}$ metric showed that it achieved a good trade-off between performance and computational efficiency. The Roboflow dataset represented a more challenging general-purpose detection scenario, resulting in lower overall detection accuracy than that achieved on the Br35H dataset. CDC-YOLO achieved the highest $mAP_{0.5}$ (0.770) and $mAP_{0.5:0.95}$ (0.361), while CDCP-YOLO performed best in recall (0.729), and it had a higher inference speed (130.5 FPS) than CDC-YOLO (115 FPS). The overall performance of YOLOv11 was the weakest, especially with obvious shortcomings in recall (0.467) and $mAP_{0.5:0.95}$ (0.276), indicating that the original model lacked sufficient generalizability on this dataset. On Capstone, CDCP-YOLO once again demonstrated the best overall performance. It also showed the highest precision (0.897), $mAP_{0.5}$ (0.909), and FPS (150), indicating that the model has stronger stability and efficiency in actual inference tasks. Although CDC-YOLO and CDCP-YOLO were comparable in $mAP_{0.5:0.95}$ (both 0.625), CDCP-YOLO demonstrated higher comprehensive advantages by virtue of its smaller model size and faster speed.

As seen in Figure 7, CDCP-YOLO performed outstandingly in terms of "fewest parameters," "lowest computational load," and "fastest FPS," while also maintaining a leading position in detection accuracy metrics, demonstrating an extremely high balance and deployment advantage. YOLOv11, although fast in some scenarios, has lower precision and recall. CDC-YOLO had a slight advantage in precision, but its computational complexity and parameter size were higher.

Figure 7. Multimetric normalization analysis of CDCP-YOLO and baseline models on 3 brain tumor magnetic resonance imaging datasets: Br35H, Roboflow and Capstone. Each graph corresponds to 1 dataset, showing the normalized performance of YOLOv11, CDC-YOLO, and CDCP-YOLO across multiple evaluation metrics. All metrics are normalized to the range [0, 1] to enable fair comparison across datasets with different scales. The evaluated metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). Arrows indicate the optimization direction: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). FPS, frames per second; GFLOPs, number of floating-point operations (×10); mAP, mean average precision; $mAP_{0.5}$, mAP at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mAP over several intersection-over-union thresholds (0.50-0.95); YOLO, "you only look once" framework.



Impact of Different Attention Mechanisms

This paper introduces these modules including MixStructureBlock [42], MSCAttention [43], MSPABlock [44], and CPCA for comparing different attention mechanisms (Table 5).

XSL•FO
RenderX

Table . Comparative analysis of different attention mechanisms on the Br35H brain tumor magnetic resonance imaging dataset.[a]

Attention mechanism	Precision	Recall	$mAP_{0.5}$[b]	$mAP_{0.5:0.95}$[c]	Params (M)[e]	GFLOPs[d]	FPS[f]
+MixStructureBlock	0.827	0.665	0.754	0.354	67.90	11	120
+MSCAttention	0.891	0.877	0.933	0.624	2.00	3.3	117.6
+MSPABlock	0.883	0.863	0.924	0.637	2.50	3.7	110
+CPCA[g]	0.918	0.904	0.944	0.644	2.07	3.3	152

[a]The table reports detection accuracy metrics (precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$) and efficiency metrics (Params, GFLOPs, and FPS) to evaluate the effectiveness and computational cost of different attention modules.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[d]GFLOPs: number of floating-point operations (×10).

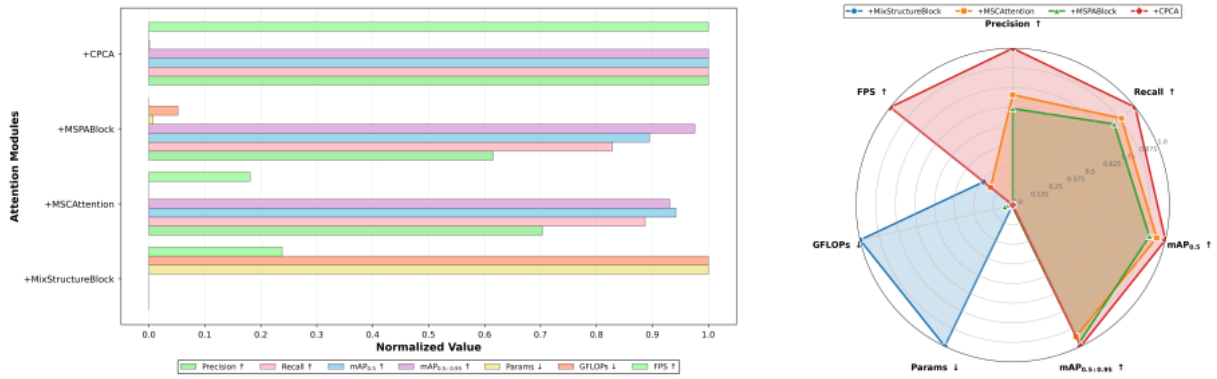
[e]Params (M) was used to measure model size.

[f]FPS: frames per second.

[g]CPCA: channel prior convolutional attention.

As Figure 8 shows, CPCA performed excellently in terms of accuracy, efficiency, and resource consumption, making it the most cost-effective attention choice. In contrast, although MixStructureBlock has a complex structure, it was in a disadvantageous position in all performance indicators, indicating that the redundancy in its design was not converted into performance gains.

Figure 8. Comparative analysis of different attention mechanisms on the Br35H brain tumor magnetic resonance imaging dataset. The subfigure on the right shows the multimetric normalization bar chart, where all evaluation metrics are normalized to the range [0,1] for fair comparison among different attention modules. The compared attention mechanisms include MixStructureBlock, MSCAttention, MSPABlock, and the proposed channel prior convolutional attention (CPCA) module. The evaluated metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). The subfigure on the right presents a radar-based performance analysis plot, summarizing the overall accuracy-efficiency trade-offs of different attention mechanisms. Arrows indicate metric preference: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). FPS, frames per second; GFLOPs, number of floating-point operations (×10); $mAP_{0.5}$, mean average precision at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mean average precision over several intersection-over-union thresholds (0.50-0.95).



Impact of Different Loss Functions

Comparative experiments were conducted by introducing loss functions such as GIoU [45], DIoU [46], EIoU [47], SIoU, ShapeIoU, PIoU, WIoU, and CIoU (Table 6).

XSL•FO
RenderX

Table . Performance comparison of different bounding box regression loss functions on the Br35H brain tumor magnetic resonance imaging dataset.[a]

Loss function	Precision	Recall	$mAP_{0.5}$[b]	$mAP_{0.5:0.95}$[c]	Params (M)[d]	GFLOPs[e]	FPS[f]
GIoU	0.893	0.820	0.926	0.635	2.07	3.3	150
DIoU	0.880	0.877	0.933	0.629	2.07	3.3	146.2
EIoU	0.867	0.787	0.875	0.546	2.07	3.3	126.1
SIoU	0.948	0.753	0.872	0.564	2.07	3.3	145
ShapeIoU	0.945	0.842	0.930	0.642	2.07	3.3	136.1
PIoU	0.890	0.885	0.923	0.646	2.07	3.3	124.8
WIoU	0.900	0.889	0.939	0.643	2.07	3.3	143
CIoU	0.899	0.873	0.944	0.644	2.07	3.3	152

[a]The table reports detection accuracy metrics (precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$) and efficiency metrics (Params, GFLOPs, and FPS) to evaluate the impact of different loss functions on detection performance and computational efficiency.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

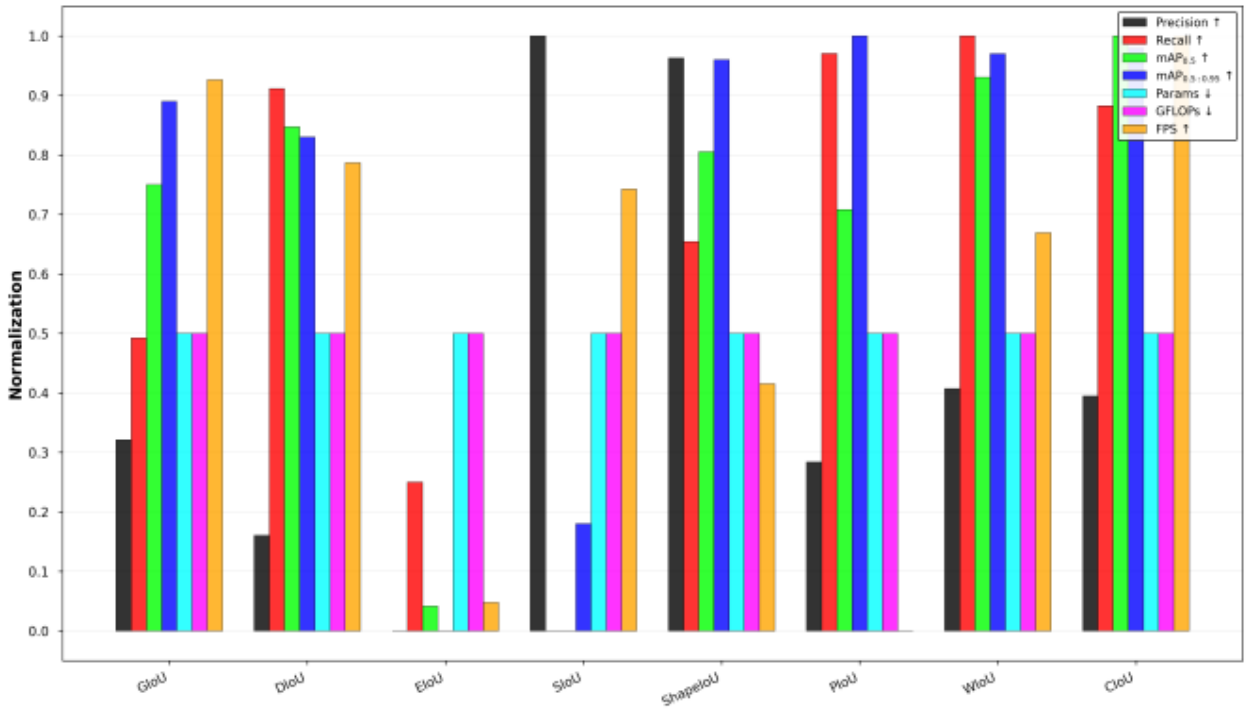
[d]Params (M) was used to measure model size.

[e]GFLOPs: number of floating-point operations (×10).

[f]FPS: frames per second.

As Figure 9 shows, CIoU, as a new generation of bounding box regression loss function, significantly improved the accuracy and stability of target detection under the premise of ensuring model lightweighting, making it the most practical and valuable choice at present. At the same time, ShapeIoU and PIoU also had advantages in scenarios with high precision requirements, while WIoU provided good recall and generalization performance.

Figure 9. Multimetric normalization comparison of different bounding box regression loss functions on the Br35H brain tumor magnetic resonance imaging dataset. All evaluation metrics are normalized to the range [0,1] to enable fair comparison among different loss functions. The compared loss functions include GIoU, DIoU, EIoU, SIoU, ShapeIoU, PIoU, WIoU, and CIoU. The evaluated metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). Arrows indicate optimization direction: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). FPS, frames per second; GFLOPs, number of floating-point operations (×10); mAP, mean average precision; $mAP_{0.5}$, mAP at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mAP over several intersection-over-union thresholds (0.50-0.95).



XSL•FO
RenderX

Pruning Experiment

Five common pruning methods were systematically evaluated, including LAMP, L1, Random, GroupNorm, and PHPS (Table 7). All methods were kept consistent in terms of the number of parameters (2.07M) and computational complexity (3.3 GFLOPs) to facilitate a fair comparison of their performance differences.

Table . Performance comparison of different pruning strategies on the Br35H brain tumor magnetic resonance imaging dataset under the same parameter budget.[a]

Pruning method	Precision	Recall	$mAP_{0.5}$[b]	$mAP_{0.5:0.95}$[c]	Params[d]	GFLOPs[e]	FPS[f]
Lamp	0.858	0.891	0.941	0.640	2.07	3.3	137.1
L1	0.893	0.902	0.934	0.640	2.07	3.3	147
Random	0.866	0.902	0.940	0.637	2.07	3.3	150
GroupNorm	0.909	0.861	0.944	0.632	2.07	3.3	146.5
PHPS[g]	0.899	0.873	0.944	0.644	2.07	3.3	152

[a]The table reports detection accuracy metrics (precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$) and efficiency metrics (Params, GFLOPs, and FPS) to evaluate the impact of different pruning methods on detection performance and inference efficiency.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[d]Params (M) was used to measure model size.

[e]GFLOPs: number of floating-point operations (×10).

[f]FPS: frames per second.

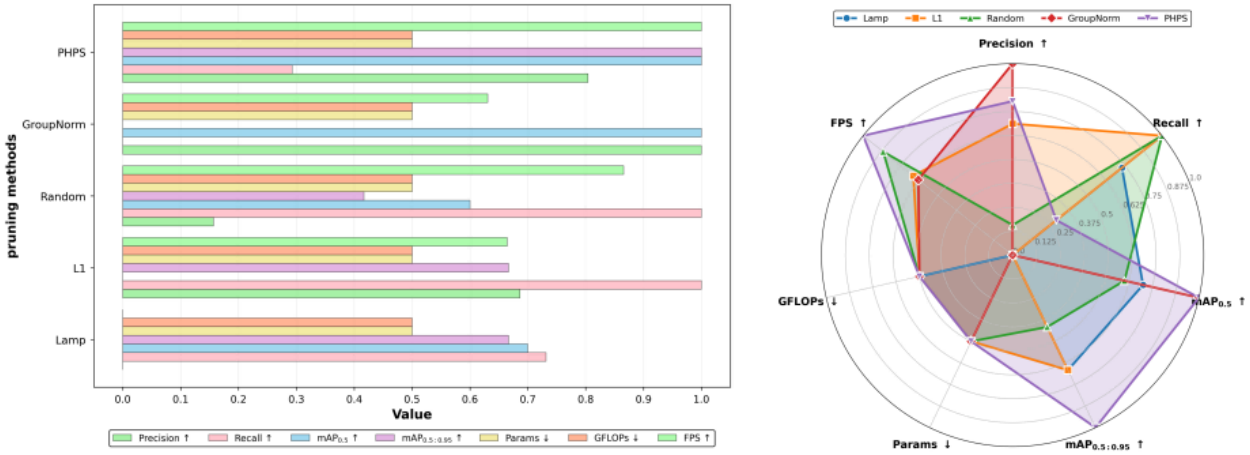
[g]PHPS: progressive hybrid pruning strategy.

The performance of PHPS was balanced and optimal across all performance indicators. The $mAP_{0.5}$ was the same as the GroupNorm method at 0.944, but PHPS reached the highest value of 0.644 on $mAP_{0.5:0.95}$ and also had the fastest inference speed (152 FPS), showing the best balance of precision and speed. GroupNorm had an advantage in precision (0.909), but recall rate was low (0.861), and $mAP_{0.5:0.95}$ was not as good as that of PHPS, indicating that its strength in boundary fitting may affect the completeness of target detection. Among the other methods, L1 had a slight advantage in recall (0.902), and the overall precision was stable. Random pruning is simple and has not undergone structural optimization, but it performed well in terms of inference speed (150 FPS). LAMP was inferior in all indicators, especially precision (0.858) and $mAP_{0.5:0.95}$ (0.640), indicating that its pruning method failed to effectively retain key features, weakening the model's detection ability.

Figure 10 shows that PHPS was in the outermost layer in dimensions such as "precision," "recall," "mAP" and "inference speed," representing that it scored the highest on each metric and had extremely strong practicality and deployment advantages. In contrast, the normalized evaluation of LAMP and GroupNorm showed obvious shortcomings, with deficiencies in detection accuracy and speed, respectively, making it difficult to meet the requirements of high-performance real-time tasks.

XSL•FO
RenderX

Figure 10. Comparative analysis of different pruning strategies on the Br35H brain tumor magnetic resonance imaging dataset under the same parameter budget. The left subfigure shows the multimetric normalization bar chart, where all evaluation metrics are normalized to the range [0,1] for fair comparison. The compared pruning methods include LAMP, L1-norm pruning, random pruning, GroupNorm-based pruning, and the proposed progressive hybrid pruning strategy (PHPS). The evaluated metrics include precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, number of parameters (Params), computational complexity (GFLOPs), and inference speed (FPS). The right subfigure presents a radar-based performance analysis plot summarizing the overall accuracy-efficiency trade-offs of different pruning strategies. Arrows indicate optimization direction: ↑ denotes that higher values are better (precision, recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$, and FPS), while ↓ denotes that lower values are better (Params and GFLOPs). FPS, frames per second; GFLOPs, number of floating-point operations (×10); mAP, mean average precision; $mAP_{0.5}$, mAP at an intersection-over-union threshold of 0.5; $mAP_{0.5:0.95}$, average mAP over several intersection-over-union thresholds (0.50-0.95).



To assess robustness, we fixed the same fully trained baseline checkpoint and repeated the pruning procedure 5 times with different random seeds. For each run, we pruned the model to the same parameter budget as that reported in Table 7 and performed a single final fine-tuning stage. These results demonstrated that, although the mean mAP gap between PHPS and random pruning was limited, PHPS provided significantly improved robustness and stability, which is particularly important for reliable deployment in medical imaging applications (Table 8).

Table . Stability comparison of different pruning strategies under the same parameter budget on the Br35H brain tumor magnetic resonance imaging dataset.[a]

Pruning method	SD ($mAP_{0.5}$[b]) $\times10^{-3}$	SD ($mAP_{0.5:0.95}$[c]) $\times10^{-3}$
Lamp	3.35	5.66
L1	1.41	3.54
Random	7.07	2.12
GroupNorm	2.83	1.42
PHPS[d]	0.71	0.71

[a]The table reports the SD of $mAP_{0.5}$ and $mAP_{0.5:0.95}$ (scaled by 10^{3}) across multiple training runs, which reflects the robustness and training stability of each pruning method.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[d]PHPS: progressive hybrid pruning strategy.

Although different pruning strategies vary in their effectiveness, they also differ in practical implementation costs. Unstructured pruning often introduces limited pruning time overhead but requires sparse inference support, whereas structured pruning incurs additional pruning and fine-tuning costs during training while enabling direct acceleration at inference. The proposed PHPS performs pruning offline during training and does not introduce extra computational overhead at inference, making it suitable for deployment without modifying existing inference pipelines.

Five-Fold Cross-Validation

To further evaluate the statistical reliability of the performance gain between CDCP-YOLO and YOLOv11, we conducted a 5-fold cross-validation. The dataset was partitioned into 5 mutually exclusive folds. In each fold, 4 folds were used for training and the remaining one for testing, ensuring that each sample was evaluated exactly once.

All models were trained from scratch under identical training settings for each fold. The final performance was reported as the mean (SD) across the 5 folds.

XSL•FO
RenderX

The mean (SD) values of precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$ across the 5 folds are reported in Table 9. The results showed that CDCP-YOLO consistently outperformed YOLOv11 across different folds, with lower variance and more stable performance, confirming that the observed improvement was statistically reliable and reproducible, rather than an artifact of random initialization.

Table . Five-fold cross-validation results of YOLOv11 and CDCP-YOLO on the Br35H brain tumor magnetic resonance imaging dataset.[a]

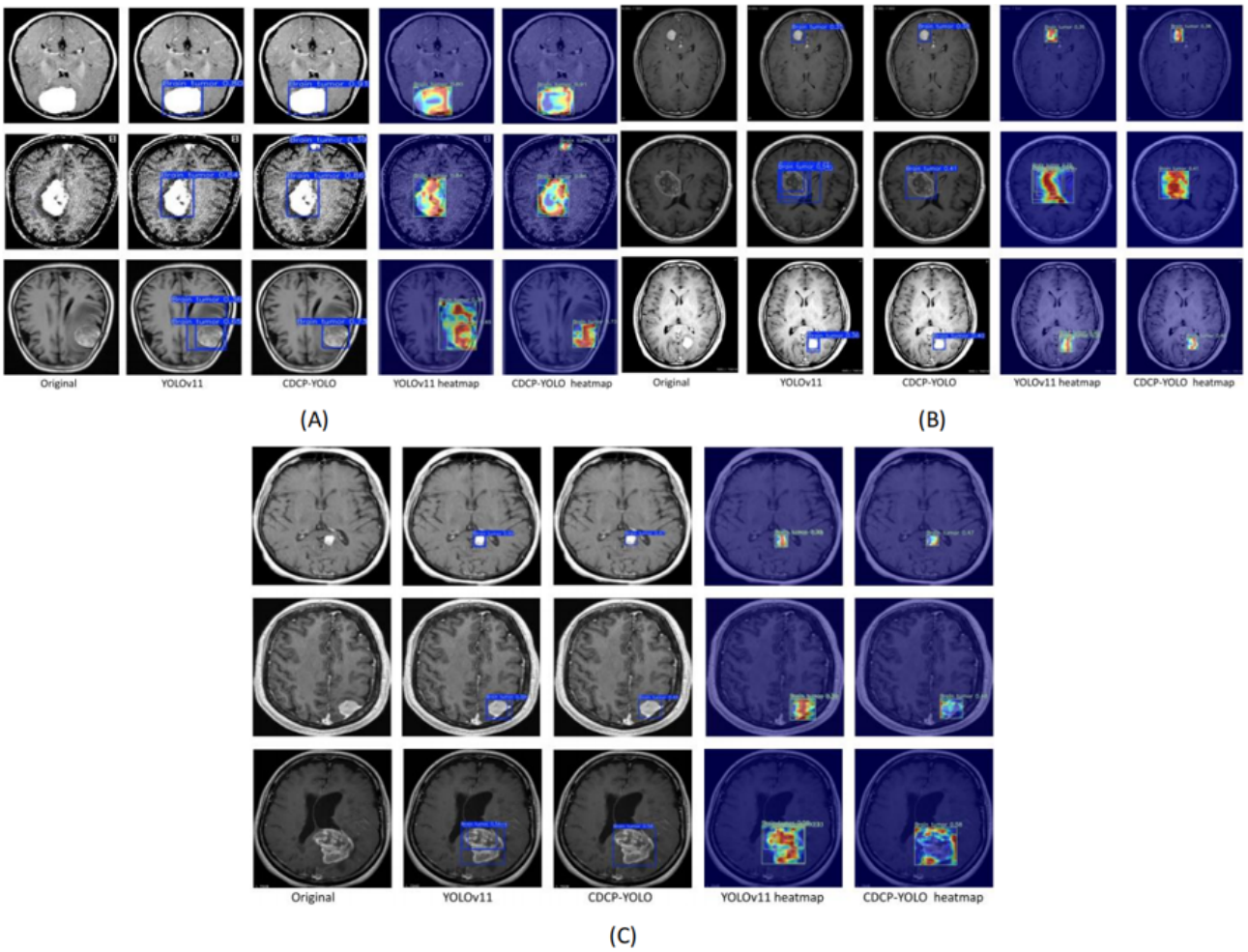
Model	Precision, mean (SD $\times10^{-3}$)	Recall, mean (SD $\times10^{-3}$)	$mAP_{0.5}$, mean (SD $\times10^{-3}$)	$mAP_{0.5:0.95}$, mean (SD $\times10^{-3}$)
YOLOv11	0.907 (17.880)	0.854 (2.492)	0.920 (3.030)	0.588 (11.500)
CDCP-YOLO	0.918 (1.410)	0.906 (2.610)	0.943 (2.590)	0.644 (1.410)

[a]The table reports the mean values of precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$ across 5 folds, together with the corresponding SD ($\times10^{3}$) values, which reflect the training stability and robustness of each model.

Interpretability

In Figure 11, for each dataset, the first column shows the original MRI image, and the remaining columns show the detection boxes generated by the YOLOv11 and CDCP-YOLO frameworks and the corresponding Eigen-CAM heat maps, respectively. The overall comparison results showed that CDCP-YOLO performed with higher detection confidence, more compact detection boxes, and more focused responses on different datasets, significantly outperforming the YOLOv11 baseline.

Figure 11. Qualitative visual comparison of detection and interpretability results between YOLOv11 and the proposed CDCP-YOLO model on 3 brain tumor magnetic resonance imaging (MRI) datasets. (A) Br35H dataset, (B) Roboflow dataset, and (C) Capstone dataset. For each dataset, the columns (from left to right) show the original MRI slice, detection results of YOLOv11, detection results of CDCP-YOLO, Eigen-CAM heat map of YOLOv11, and Eigen-CAM heat map of CDCP-YOLO. The heat map color gradient ranges from blue to red, indicating low to high activation intensity, respectively, where warmer colors correspond to regions that contribute most to the detection decision. Compared with YOLOv11, CDCP-YOLO produces more compact and accurate bounding boxes and generates more focused and lesion-aligned activation maps, especially for small, blurry, or irregular tumor regions. These results demonstrate the improved detection reliability and interpretability of the proposed framework. Eigen-CAM, Eigen-class activation mapping; YOLO, “you only look once” framework.



XSL•FO
RenderX

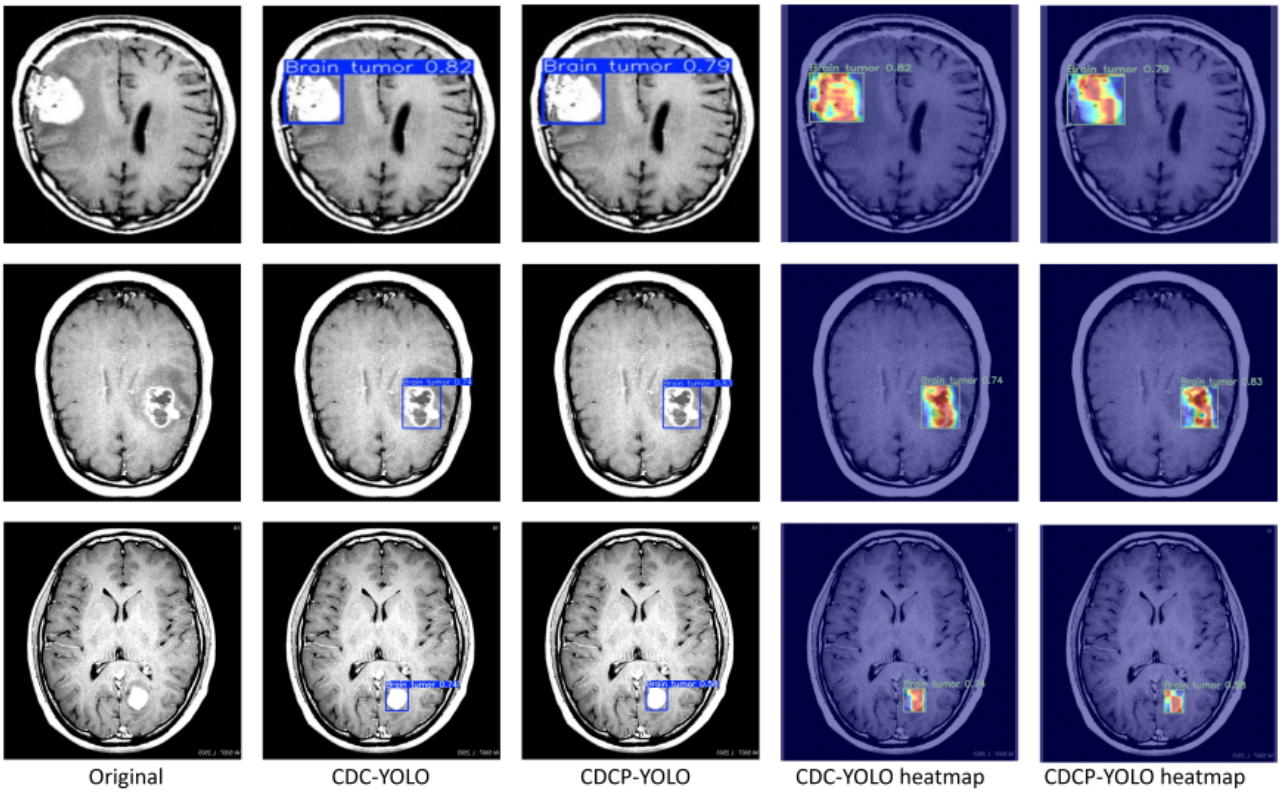
From both quantitative and qualitative perspectives: on Br35H, the detection confidence of CDCP-YOLO in the first image was much higher than YOLOv11; in the second image, YOLOv11 failed to detect the tumor on the left, while CDCP-YOLO successfully identified it; in the third image, YOLOv11 produced 2 bounding boxes with inaccurate positioning, while the detection box of CDCP-YOLO was closer to the lesion outline; the Eigen-CAM heat map also mainly covered the high-signal core and extended to the indistinct boundary. On Roboflow, the confidence level of CDCP-YOLO was also significantly higher than that of YOLOv11, and it was more accurate in terms of boundary fit and shape depiction. On Capstone, the 3 sets of images consistently showed that CDCP-YOLO not only has higher detection confidence but its bounding boxes were also more consistent with the actual lesion, and the thermal response was tighter and more focused. In MRI with complex backgrounds, YOLOv11 is often interfered with by shadows and anatomical structures (such as blood vessels and choroid plexus), leading to boundary offsets or missed detections. The heat map of CDCP-YOLO can stably focus on the tumor lesion area, effectively suppressing the interference of shadows and structural noise. This advantage stems from the deep integration of multiscale feature enhancement modules (CSPP and DCC) and CPCA, which significantly enhances the sensitivity and discrimination of tumor regions during feature extraction and judgment.

Clinical Sensitivity and False Negative Analysis

In medical diagnosis, false negatives—particularly missed detections of small or visually ambiguous lesions—pose a critical risk. To evaluate the potential impact of pruning on clinical sensitivity, we further analyzed the behavior of the proposed PHPS from both quantitative and qualitative perspectives.

Quantitatively, the recall remained stable across different pruning configurations, indicating that the progressive pruning process did not significantly impair the model's ability to detect tumor regions. This suggests that the proposed strategy effectively preserved detection-critical channels during compression. Qualitatively, visualization results on representative cases involving small-scale or blurred tumor boundaries showed that the pruned model maintained consistent activation patterns compared with the unpruned baseline. These findings indicate that PHPS mitigates the risk of increased false negatives while achieving substantial model compression, thereby enhancing its suitability for clinical deployment. Despite significant model compression, the pruned CDCP-YOLO preserves consistent tumor localization and activation patterns compared with the unpruned baseline, especially for small-scale or low-contrast lesions. These results indicate that the proposed PHPS did not increase false negative rates and maintained clinical sensitivity under challenging conditions (Figure 12).

Figure 12. Qualitative comparison of detection sensitivity on small or ambiguous brain tumor regions before and after pruning on the Br35H brain tumor magnetic resonance imaging (MRI) dataset. From left to right, each column shows the original MRI slice, detection result of the unpruned CDC-YOLO, detection result of the pruned CDCP-YOLO, Eigen-CAM heat map of CDC-YOLO, and Eigen-CAM heat map of CDCP-YOLO. The heat map color gradient ranges from blue to red, indicating low to high activation intensity, respectively, where warmer colors correspond to regions that contribute most to the detection decision. Eigen-CAM, Eigen-class activation mapping; YOLO, "you only look once" framework.



To statistically substantiate the "fine-grained" detection capability and robustness to small-scale tumors, we further evaluated the models using the AP_{small} metric, following the standard COCO evaluation protocol. As summarized in Table

XSL•FO
RenderX

10, CDCP-YOLO exhibited a superior sensitivity to small lesions compared to all baseline models. Specifically, CDCP-YOLO achieved an AP_{small} of 0.487, outperforming the YOLOv11n baseline (0.412) by a substantial margin of 18.2%. This quantitative improvement indicates that the integration of the CSPP and DCC modules for edge and feature enhancement, alongside the CPCA mechanism for precise spatial attention, effectively preserves the subtle morphological features of early-stage or small-scale tumors—details that are frequently overlooked by standard architectures. These results provide rigorous evidence that the proposed framework is not only efficient but also highly reliable for detecting clinically significant small lesions.

Table . Quantitative evaluation of small-lesion detection performance, AP_{small} (the baseline YOLOv11n and the proposed CDCP-YOLO model), on the Br35H brain tumor magnetic resonance imaging dataset using the standard Common Objects in Context (COCO)–style detection metrics.

Model	$mAP_{0.5}$[a]	AP_{small}[b]	Relative gain
YOLOv11n	0.918	0.412	—[c]
CDCP-YOLO	0.944	0.487	+18.2%

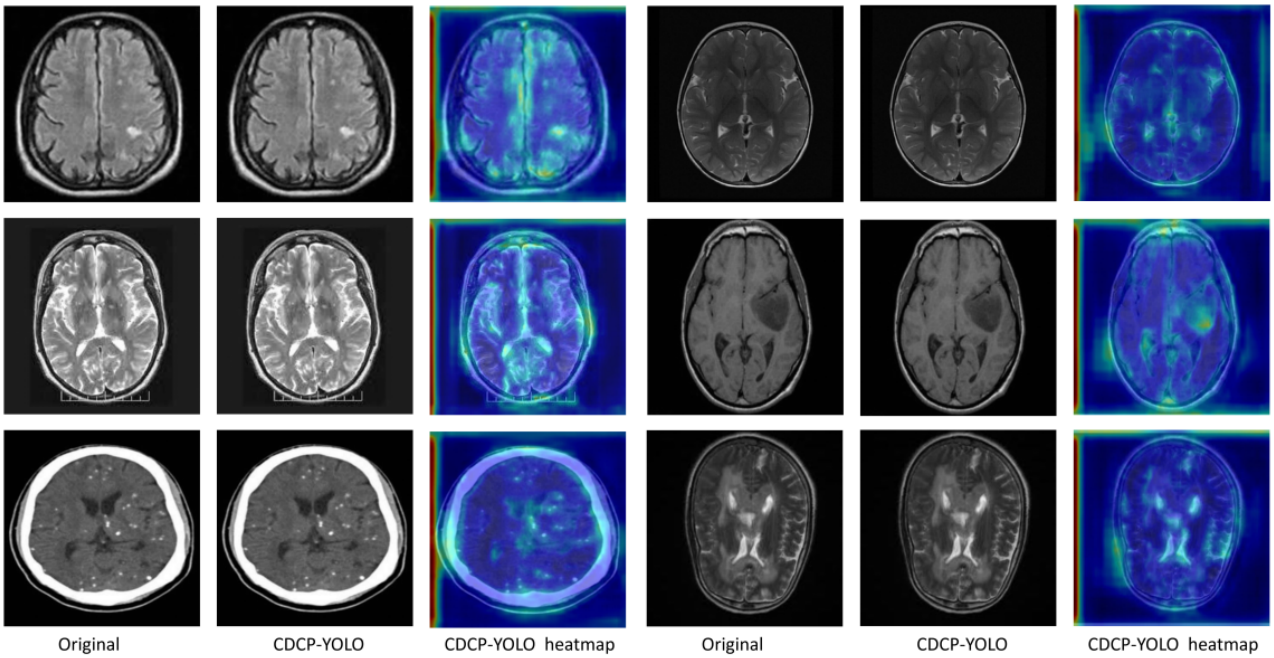
[a]$mAP_{0.5}$ denotes the mean average precision at an intersection-over-union threshold of 0.5.

[b]AP_{small} denotes the average precision computed only for small objects, following the COCO definition (ie, objects whose bounding-box area falls within the "small" size range specified by the COCO evaluation protocol).

[c]Not applicable.

To further assess the reliability and specificity of CDCP-YOLO, we conducted an inference-only validation on a negative control cohort of 50 healthy brain MRI slices obtained from the brain tumor dataset [48]. This test aimed to verify that the model does not produce "hallucinated" detections on normal anatomical structures. As shown in Figure 13, the model successfully classified all healthy slices as negative, producing no false positive detection boxes. Moreover, the Eigen-CAM visualizations showed only very weak background-level activations (cool colors) within healthy brain regions, without the high-intensity highlighted responses typically observed in tumor areas. This negative control experiment provided empirical evidence for the model's robust safety profile, ensuring that its high tumor sensitivity is not achieved at the expense of misdiagnosing healthy anatomy.

Figure 13. Representative examples of negative-control validation on healthy brain magnetic resonance imaging (MRI) slices using inference-only CDCP-YOLO and Eigen-CAM visualization. This figure presents an inference-only negative control experiment conducted on healthy brain MRI slices without tumor lesions, used to evaluate the specificity and safety of the proposed CDCP-YOLO model. The healthy slices were obtained from a publicly available brain MRI dataset and were not included in the training process. For each example, the first column shows the original MRI slice, the second column shows the detection result produced by CDCP-YOLO, and the third column presents the corresponding Eigen-CAM heat map overlaid on the original image. The color scale of the heat map represents the relative contribution of image regions to the model's prediction, where warm colors (red/yellow) indicate higher activation intensity, and cool colors (blue) indicate low or background-level responses. Eigen-CAM, Eigen-class activation mapping; YOLO, "you only look once" framework.



XSL•FO
RenderX

Hyperparameter Sensitivity Analysis of PHPS

To further assess the robustness of the proposed method, we conducted a sensitivity analysis on key training and pruning hyperparameters involved in PHPS. Specifically, we analyzed the sparsity coefficient (SC) and the learning rate (LR) schedule, as these parameters directly influence pruning behavior and convergence stability.

In this analysis, only one hyperparameter was varied at a time while all other settings were kept fixed, allowing us to isolate the effect of each parameter. The results are summarized in Table 11. As shown, the proposed method exhibited stable detection performance across a reasonable range of SC values, indicating that PHPS is not overly sensitive to the pruning threshold. In addition, different learning rate schedules led to comparable performance, suggesting that the training process remained robust under different optimization dynamics.

Table . Sensitivity analysis of key hyperparameters in the progressive hybrid pruning strategy (PHPS) on the Br35H brain tumor magnetic resonance imaging dataset.[a]

Hyperparameter	Setting	Precision	Recall	$mAP_{0.5}$[b]	$mAP_{0.5:0.95}$[c]
SC[d]	0.2	0.912	0.898	0.939	0.637
	0.3	0.918	0.904	0.944	0.644
	0.4	0.910	0.895	0.936	0.631
LR[e] schedule	StepLR	0.914	0.899	0.941	0.639
	CosineAnnealingLR	0.918	0.904	0.944	0.644
	CosineAnnealingWarm-Restarts	0.916	0.902	0.940	0.635

[a]The table reports the detection performance under different settings of the sparsity coefficient (SC) and learning rate (LR) scheduling strategies, including StepLR, CosineAnnealingLR, and CosineAnnealingWarmRestarts. Detection accuracy was evaluated using precision, recall, $mAP_{0.5}$ and $mAP_{0.5:0.95}$.

[b]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[c]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[d]SC: sparsity coefficient.

[e]LR: learning rate.

Ablation Experiments for Architectural and Pruning Analysis

To rigorously isolate the architectural contributions of the proposed CSPP, DCC, and CPCA modules and analyze the impact of pruning, a pruned baseline model (YOLOv11n-1) was included in the ablation study, as summarized in Table 12. Specifically, the standard YOLOv11n backbone was pruned to the same parameter budget (2.07M) using the same PHPS and fine-tuning protocol but without introducing any of the proposed architectural modules. This design allows us to decouple the effect of pruning from that of architectural enhancement. As shown in Table 12, applying PHPS to the standard YOLOv11n resulted in performance degradation, with $mAP_{0.5}$ dropping from 0.918 to 0.902. This outcome demonstrates that aggressive pruning on a baseline backbone without specialized modules inevitably disrupts critical feature pathways, challenging the notion that pruning alone can enhance model capability. In contrast, the proposed CDCP-YOLO model achieved a significantly higher $mAP_{0.5}$ of 0.944 under the same 2.07M parameter budget. These results clarify that the observed performance gains originated from the architectural enhancements—CSPP, DCC, and CPCA—which provided robust feature representation that successfully compensated for the pruning-induced losses. This confirms that jointly designing architectural enhancements and progressive pruning is essential for achieving a superior accuracy-efficiency trade-off in resource-constrained environments.

XSL•FO
RenderX

Table . Ablation study isolating the contributions of architectural modules and pruning, through comparison of (1) the unpruned baseline YOLOv11n; (2) a pruned baseline (YOLOv11n-1) obtained by applying the progressive hybrid pruning strategy (PHPS) to the standard YOLOv11n (without CSPP[a], DCC[b], or CPCA[c]) to match the parameter budget of 2.07M; and (3) the proposed CDCP-YOLO model, which integrated the proposed architectural modules (CSPP, DCC, and CPCA) and was pruned and fine-tuned under the same PHPS setting.

Model	Precision	Recall	$mAP_{0.5}$[d]	$mAP_{0.5:0.95}$[e]	Params (M)[f]	GFLOPs[g]	FPS[h]
YOLOv11n	0.904	0.853	0.918	0.585	2.58	6.3	142.2
YOLOv11n-1	0.905	0.855	0.902	0.580	2.07	2.8	145.3
CDCP-YOLO	0.918	0.904	0.944	0.644	2.07	3.3	152

[a]CSPP: convolution Prewitt-and-pooling–based preprocessing.

[b]DCC: dynamic convolution–based C3k2.

[c]CPCA: channel prior convolutional attention.

[d]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[e]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[f]Params (M) was used to measure model size.

[g]GFLOPs: number of floating-point operations (×10).

[h]FPS: frames per second.

Impact of Training Budget on Performance Gain

To verify that the performance gains were not merely a result of longer total training duration, we retrained the YOLOv11n baseline for 600 epochs (YOLOv11n-2), matching the combined 2-stage budget of CDCP-YOLO. Results demonstrated that despite doubling the convergence time, the $mAP_{0.5}$ of YOLOv11n only achieved a negligible increase of 0.001, remaining significantly below the value (0.944) achieved by CDCP-YOLO (Table 13). This finding underscored that the core contributions to detection accuracy stemmed from the proposed feature enhancement modules and the PHPS strategy rather than extended training time.

Table . Impact of training budget alignment on model performance on the Br35H brain tumor magnetic resonance imaging dataset, assessed through comparison of the detection performance under three experimental settings: (1) the standard YOLOv11n baseline trained for 300 epochs, (2) the YOLOv11n baseline retrained for 600 epochs to align with the total 2-stage training budget of CDCP-YOLO (denoted as YOLOv11n-2), and (3) the proposed CDCP-YOLO trained using a 2-stage training strategy (300 epochs of initial training followed by 300 epochs of postpruning fine-tuning).

Model	Total epochs, n	Precision	Recall	$mAP_{0.5}$[a]	$mAP_{0.5:0.95}$[b]	Params (M)[c]	GFLOPs[d]
YOLOv11n	300	0.904	0.853	0.918	0.585	2.58	6.3
YOLOv11n-2	600	0.905	0.855	0.919	0.580	2.58	6.3
CDCP-YOLO	600	0.918	0.904	0.944	0.644	2.07	3.3

[a]$mAP_{0.5}$: mean average precision at an intersection-over-union threshold of 0.5.

[b]$mAP_{0.5:0.95}$: average mean average precision over several intersection-over-union thresholds (0.50-0.95).

[c]Params (M) was used to measure model size.

[d]GFLOPs: number of floating-point operations (×10).

Computational Overhead and Deployment Considerations

Although different pruning strategies vary in their compression effectiveness, they also differ in practical implementation cost. The computational overhead of pruning methods can be analyzed from 3 aspects: pruning time overhead, fine-tuning cost, and inference-time efficiency.

Unstructured pruning typically introduces limited overhead during pruning but requires specialized sparse inference support at deployment. In contrast, structured pruning methods incur additional pruning and fine-tuning costs during training while enabling direct reductions in parameter size and number of floating-point operations (FLOPs) without modifying standard inference pipelines.

The proposed strategy PHPS performs pruning offline during training in a stage-wise manner. Consistent with the standard characteristics of structured pruning, PHPS physically removes redundant parameters from the network, resulting in a reduced model size that naturally avoids additional computational overhead during the inference phase. This lack of inference overhead is a direct consequence of the physical parameter removal inherent to structured pruning rather than a unique architectural capability of the PHPS itself.

This design makes PHPS particularly suitable for real-world deployment scenarios, such as edge devices or clinical systems, where inference efficiency and implementation simplicity are critical and modifying inference engines is impractical.

XSL•FO
RenderX

Discussion

Principal Findings

This study proposes CDCP-YOLO, an interpretable and lightweight brain tumor MRI detection framework that integrates feature enhancement, channel-prior attention, and progressive hybrid pruning. Experimental results across 3 public datasets demonstrated that CDCP-YOLO achieved a favorable balance between detection accuracy, computational efficiency, and model interpretability for slice-level brain tumor localization tasks. Notably, the proposed method consistently outperformed the YOLOv11 baseline while reducing parameters and GFLOPs by nearly half.

Furthermore, 5-fold cross-validation confirmed that the observed performance gains were statistically stable rather than artifacts of random initialization. The integration of Eigen-CAM enables visually consistent and task-aligned explanations that highlight tumor regions across MRI slices.

Strengths and Limitations

Compared with prior YOLO-based brain tumor detectors that primarily rely on deeper architectures, additional detection heads, or external pretraining, CDCP-YOLO adopts a task-driven co-design strategy. Unlike RCS-YOLO or BGF-YOLO, which emphasize reparameterization or feature fusion, the proposed method enhances edge-sensitive feature extraction at the input stage and introduces dynamic convolution to adapt to heterogeneous tumor morphologies. In contrast to PK-YOLO, which depends on external pretrained knowledge, CDCP-YOLO improves representation capacity through internal data-adaptive mechanisms, making it more suitable for limited or domain-specific medical datasets. Moreover, most existing pruning-based detectors use one-shot or single-criterion pruning, whereas the proposed PHPS method preserves detection-critical structures through progressive and structure-aware pruning.

Several limitations should be acknowledged. First, the Br35H dataset used in this study does not provide patient identifiers, and the train, validation, and test split is therefore performed at the image (slice) level following the official dataset partitioning. As a result, strict patient-level splitting cannot be enforced, which may limit the assessment of patient-level generalization. This limitation is inherent to the dataset itself and will be addressed in future work using datasets with explicit patient-level annotations. Second, all datasets used in this study (Br35H, Roboflow, and Capstone) consist exclusively of tumor-positive MRI slices and do not include healthy control images. Consequently, strict negative control visualization for CAM-based interpretability cannot be conducted without introducing external data. Future work will incorporate MRI cohorts with healthy subjects to enable negative control interpretability analysis and more rigorous validation of model behavior on normal anatomy. Third, while the proposed pruning strategy preserves recall performance, extremely small or highly ambiguous lesions may still pose challenges under aggressive compression. This limitation was partially addressed through qualitative sensitivity analysis, but further prospective validation is required. Finally, this study focused on 2D slice–based detection and did not explicitly model interslice spatial continuity, which could further improve robustness in volumetric MRI analysis.

Future Directions

Future work will focus on expanding the dataset through multi-institutional collaboration and evaluating the framework on volumetric MRI sequences. In addition, integrating efficient transformer–based modules within the 2-stage framework will be explored to further enhance contextual modeling while maintaining deployability. Clinical-oriented studies, including cost-benefit and workflow integration analyses, will also be conducted to support real-world adoption.

Conclusions

In conclusion, this study demonstrated that CDCP-YOLO achieves a well-balanced integration of detection accuracy, computational efficiency, and interpretability for slice-level brain tumor localization in MRI images. The proposed framework is built upon a task-driven co-design, incorporating structure-aware feature enhancement, dynamic convolution, and a channel-prior attention mechanism to strengthen discriminative feature representation. In particular, the attention module guides the network to focus on lesion-relevant tumor regions, improving robustness to small lesions and tumors with ambiguous boundaries.

Furthermore, the proposed strategy PHPS effectively compresses the model by jointly considering global sparsity and local structural dependency, enabling substantial reductions in parameters and computational cost while preserving detection-critical pathways. In combination with the integrated Eigen-CAM–based interpretability design, CDCP-YOLO provides visually consistent and task-aligned explanations that reflect the model's detection behavior, enhancing the transparency and reliability of the framework.

Overall, the experimental results confirmed the feasibility of deploying lightweight, attention-enhanced, and interpretable object detection models for efficient slice-level brain tumor analysis and screening in MRI images. This work provides a practical and reliable solution for resource-constrained imaging environments and represents a meaningful step toward methodological advancement and potential clinical translation, rather than a direct clinical diagnostic system.

Funding

This work was supported by the National Natural Science Youth Fund of China (grant No. 61300115), the China Postdoctoral Science Foundation (grant number 2014M561331), and the Science and Technology Research Project of Heilongjiang Provincial Education Department of China (grant number 12521073).

XSL•FO
RenderX

Data Availability

All datasets used in this study are publicly available and can be accessed through their original published sources. No private, restricted, or proprietary data were used in this work. The use of publicly accessible datasets ensures the transparency and reproducibility of the experimental results. To support the reproducibility of the reported results, the source code for the proposed CDCP-YOLO architecture and the progressive hybrid pruning strategy (PHPS) has been made publicly available on GitHub [49].

Authors' Contributions

LYP contributed to conceptualization, formal analysis, investigation, methodology, and writing – original draft. SSW contributed to data curation, formal analysis, investigation, methodology, and writing – original draft. LSB contributed to data curation and investigation. ZXC contributed to conceptualization, funding acquisition, project administration, and writing – review and editing and served as the corresponding author. LYP is the first author of this study. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Fitzgerald RC. Big data is crucial to the early detection of cancer. Nat Med 2020 Jan;26(1):19-20. [doi: 10.1038/s41591-019-0725-7] [Medline: 31932790]
3. Ramtekkar PK, Pandey A, Pawar MK. A comprehensive review of brain tumour detection mechanisms. Comput J 2024 Apr 14;67(3):1126-1152. [doi: 10.1093/comjnl/bxad047]
4. El-Dahshan ESA, Hosny T, Salem ABM. Hybrid intelligent techniques for MRI brain images classification. Digit Signal Process 2010 Mar;20(2):433-441. [doi: 10.1016/j.dsp.2009.07.002]
5. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR):580-587. [doi: 10.1109/CVPR.2014.81]
6. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision: IEEE; 2015:1440-1448. [doi: 10.1109/ICCV.2015.169]
7. Ren S, He K, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 2015. [doi: 10.1109/TPAMI.2016.2577031]
8. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition: IEEE; 2016:779-788. [doi: 10.1109/CVPR.2016.91]
9. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: European Conference on Computer Vision: Springer International Publishing; 2016:21-37. [doi: 10.1007/978-3-319-46448-0_2]
10. Chen S, Duan J, Zhang N, et al. MSA-YOLOv5: multi-scale attention-based YOLOv5 for automatic detection of acute ischemic stroke from multi-modality MRI images. Comput Biol Med 2023 Oct;165:107471. [doi: 10.1016/j.compbiomed.2023.107471] [Medline: 37716245]
11. Xiongfeng T, Yingzhi L, Xianyue S, et al. Automated detection of knee cystic lesions on magnetic resonance imaging using deep learning. Front Med (Lausanne) 2022;9:928642. [doi: 10.3389/fmed.2022.928642] [Medline: 36016997]
12. Wang A, Wang T, Liu X, et al. Automated diagnosis and grading of lumbar intervertebral disc degeneration based on a modified YOLO framework. Front Bioeng Biotechnol 2025;13:1526478. [doi: 10.3389/fbioe.2025.1526478] [Medline: 39912111]
13. Kang M, Ting CM, Ting FF, Phan RCW. RCS-YOLO: a fast and high-accuracy object detector for brain tumor detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer Nature Switzerland; 2023:600-610. [doi: 10.1007/978-3-031-43901-8_57]
14. Kang M, Ting CM, Ting FF, Phan RCW. BGF-YOLO: enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer Nature Switzerland; 2024:35-45. [doi: 10.1007/978-3-031-72111-3_4]
15. Kang M, Ting FF, Phan RCW, Ting CM. PK-YOLO: pretrained knowledge guided YOLO for brain tumor detection in multiplanar MRI slices. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV): IEEE; 2025:3732-3741. [doi: 10.1109/WACV61041.2025.00367]
16. Dixit A, Singh P. Brain tumor detection using fine-tuned YOLO model with transfer learning. In: Artificial Intelligence on Medical Data: Proceedings of International Symposium, ISCMM 2021: Springer Nature; 2022:363-371. [doi: 10.1007/978-981-19-0151-5_30]
17. Abdusalomov AB, Mukhiddinov M, Whangbo TK. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers (Basel) 2023 Aug 18;15(16):4172. [doi: 10.3390/cancers15164172] [Medline: 37627200]

XSL•FO
RenderX

18. Fernandes FE, Yen GG. Pruning of generative adversarial neural networks for medical imaging diagnostics with evolution strategy. Inf Sci 2021 May;558:91-102. [doi: 10.1016/j.ins.2020.12.086]
19. Wu Y, Zeng D, Xu X, et al. FairPrune: achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer Nature; 2022:743-753. [doi: 10.1007/978-3-031-16431-6_70]
20. Adnan M, Ba Q, Shaikh N, et al. Structured model pruning for efficient inference in computational pathology. In: International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis: Springer Nature; 2024:140-149. [doi: 10.1007/978-3-031-77786-8_14]
21. Fernandes FE, Yen GG. Automatic searching and pruning of deep neural networks for medical imaging diagnostic. IEEE Trans Neural Netw Learning Syst 2020;32(12):5664-5674. [doi: 10.1109/TNNLS.2020.3027308]
22. Cocosco CA, Zijdenbos AP, Evans AC. A fully automatic and robust brain MRI tissue classification method. Med Image Anal 2003 Dec;7(4):513-527. [doi: 10.1016/s1361-8415(03)00037-9] [Medline: 14561555]
23. Xuan K, Sun S, Xue Z, Wang Q, Liao S. Learning MRI k-space subsampling pattern using progressive weight pruning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer; 2020:178-187. [doi: 10.1007/978-3-030-59713-9_18]
24. Graziani M, Lompech T, Müller H, Depeursinge A, Andrearczyk V. Interpretable CNN pruning for preserving scale-covariant features in medical imaging. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing: Springer International Publishing; 2020:23-32. [doi: 10.1007/978-3-030-61166-8_3]
25. Holste G, Jiang Z, Jaiswal A, et al. How does pruning impact long-tailed multi-label medical image classifiers? In: International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer Nature; 2023:663-673. [doi: 10.1007/978-3-031-43904-9_64]
26. Saleh L, Zhang L. Medical image classification using transfer learning and network pruning algorithms. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC): IEEE; 2023:1932-1938. [doi: 10.1109/SMC53992.2023.10393868]
27. Jaiswal A, Chen T, Rousseau JF, Peng Y, Ding Y, Wang Z. Attend who is weak: pruning-assisted medical image localization under sophisticated and implicit imbalances. : IEEE/CVF; 2023 Presented at: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision p. 4987-4996. [doi: 10.1109/WACV56688.2023.00496]
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of IEEE International Conference on Computer Vision: IEEE; 2017:618-626. [doi: 10.1109/ICCV.2017.74]
29. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV): IEEE; 2018:839-847. [doi: 10.1109/WACV.2018.00097]
30. Muhammad MB, Yeasin M. Eigen-CAM: class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN): IEEE; 2020:1-7. [doi: 10.1109/IJCNN48605.2020.9206626]
31. Jiang PT, Zhang CB, Hou Q, Cheng MM, Wei Y. LayerCAM: exploring hierarchical class activation maps for localization. IEEE Trans Image Process 2021;30:5875-5888. [doi: 10.1109/TIP.2021.3089943] [Medline: 34156941]
32. Windisch P, Weber P, Fürweger C, et al. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. Neuroradiology 2020 Nov;62(11):1515-1518. [doi: 10.1007/s00234-020-02465-1] [Medline: 32500277]
33. Shawon MTR, Shibli GMS, Ahmed F, Joy SKS. Explainable cost-sensitive deep neural networks for brain tumor detection from brain MRI images considering data imbalance. Multimed Tools Appl 2025;84(35):43615-43642. [doi: 10.1007/s11042-025-20842-x]
34. Dasanayaka S, Silva S, Shantha V, Meedeniya D, Ambegoda T. Interpretable machine learning for brain tumor analysis using MRI. In: 2022 2nd International Conference on Advanced Research in Computing (ICARC): IEEE; 2022:212-217. [doi: 10.1109/ICARC54489.2022.9754131]
35. Zeineldin RA, Karar ME, Elshaer Z, et al. Explainability of deep neural networks for MRI analysis of brain tumors. Int J Comput Assist Radiol Surg 2022 Sep;17(9):1673-1683. [doi: 10.1007/s11548-022-02619-x] [Medline: 35460019]
36. Mohamed MM, Mahesh TR, Vinoth KV, Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imaging 2024 May 11;24(1):107. [doi: 10.1186/s12880-024-01292-7] [Medline: 38734629]
37. Han K, Wang Y, Guo J, Wu E. ParameterNet: parameters are all you need for large-scale visual pretraining of mobile networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): IEEE/CVF; 2024:15751-15761. [doi: 10.1109/CVPR52733.2024.01491]
38. Huang H, Chen Z, Zou Y, et al. Channel prior convolutional attention for medical image segmentation. Comput Biol Med 2024 Aug;178:108784. [doi: 10.1016/j.compbiomed.2024.108784] [Medline: 38941900]
39. Hamada A. Br35H:: brain tumor detection 2020. Kaggle. 2020. URL: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection [accessed 2026-04-20]
40. YOLO computer vision model. Universe. URL: https://universe.roboflow.com/yolo-hz3ua/yolo-fj4s3 [accessed 2026-04-20]

XSL•FO
RenderX

41. Capstone computer vision dataset. Universe. URL: https://universe.roboflow.com/naufal-irfani/capstone-dk1nt [accessed 2026-04-20]
42. Lu L, Xiong Q, Xu B, Chu D. MixDehazeNet: mix structure block for image dehazing network. In: 2024 International Joint Conference on Neural Networks (IJCNN): IEEE; 2024:1-10. [doi: 10.1109/IJCNN60899.2024.10651326]
43. Shao H, Zeng Q, Hou Q, Yang J. MCANet: medical image segmentation with multi-scale cross-axis attention. Mach Intell Res 2025 Jun;22(3):437-451. [doi: 10.1007/s11633-025-1552-6]
44. Yu Y, Zhang Y, Cheng Z, Song Z, Tang C. Multi-scale spatial pyramid attention mechanism for image recognition: an effective approach. Eng Appl Artif Intell 2024 Jul;133:108261. [doi: 10.1016/j.engappai.2024.108261]
45. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): IEEE/CVF; 2019:658-666. [doi: 10.1109/CVPR.2019.00075]
46. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-iou loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence: 2020; 2020, Vol. 34:12993-13000. [doi: 10.1609/aaai.v34i07.6999]
47. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 2022 Sep;506:146-157. [doi: 10.1016/j.neucom.2022.07.042]
48. Backert T. Brain tumor dataset. Kaggle. 2024. URL: https://www.kaggle.com/datasets/tombackert/brain-tumor-mri-data [accessed 2026-04-20]
49. GitHub. Brain-tumor-detection: interpretable and fine-grained brain tumor MRI detection based on progressive pruning. URL: https://github.com/song68/Brain-Tumor-Detection [accessed 2026-04-25]

Abbreviations

CA: channel attention
CAM: class activation mapping
CNN: convolutional neural network
CPCA: channel prior convolutional attention
CSPP: convolution Prewitt-and-pooling–based preprocessing
DCC: dynamic convolution–based C3k2
DCM: disease classification module
FLOPs: number of floating-point operations
FPS: frames per second
GFLOPs: number of floating-point operations (×10)
Grad-CAM: gradient-weighted class activation mapping
HSV: hue-saturation-value
IoU: intersection over union
IQM: image quality module
mAP: mean average precision
MRI: magnetic resonance imaging
PHPS: progressive hybrid pruning strategy
R-CNN: region-based convolutional neural network
RCS: reparameterized convolution combined with channel shuffle
SA: spatial attention
SGD: stochastic gradient descent
SSD: single shot multibox detector
YOLO: you only look once

Edited by A Coristine; submitted 14.Sep.2025; peer-reviewed by HV Thanh, M Wang; revised version received 28.Feb.2026; accepted 28.Feb.2026; published 29.Apr.2026.

Please cite as:
Liu Y, Song S, Lian S, Zhang X
Detection of Interpretable and Fine-Grained Brain Tumor Magnetic Resonance Imaging Based on Progressive Pruning: Machine Learning Model Development and Validation Study
JMIR Med Inform 2026;14:e84095
URL: https://medinform.jmir.org/2026/1/e84095
doi:10.2196/84095

XSL•FO
RenderX

© Yupeng Liu, Shuwei Song, Shibo Lian, Xiaochen Zhang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparative Performance of 3 Analytical Models in Identifying Associated Factors of Pulmonary Dysfunction–Depression Comorbidity: China Health and Retirement Longitudinal Study–Based Nationwide Cross-Sectional Study

Qinglin Cheng[1,2*], MD, PhD; Qiancheng Cao[2*], MD; Weilin Teng[1], MD; Ruoqi Dai[1], MD; Qingjun Jia[1], MD; Xuexin Bai[1], MD; Qingchun Li[1], MD; Yifei Wu[1], MD; Yinyan Huang[1], MD

[1]Department of Tuberculosis Control and Prevention, Hangzhou Center for Disease Control and Prevention (Hangzhou Health Supervision Institution), Mingshi 568#, Shangcheng District, Hangzhou, Zhejiang, China

[2]School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Weilin Teng, MD
Department of Tuberculosis Control and Prevention, Hangzhou Center for Disease Control and Prevention (Hangzhou Health Supervision Institution), Mingshi 568#, Shangcheng District, Hangzhou, Zhejiang, China

Abstract

Background: Pulmonary dysfunctions are common and frequently co-occur with depressive symptoms, worsening outcomes, and increasing health care burden. Clinically usable models for identifying pulmonary dysfunction–depression comorbidity remain limited by suboptimal interpretability, inconsistent validation, and uncertain generalizability.

Objective: This study developed and compared logistic regression (LR), Bayesian network (BN), and Extreme Gradient Boosting (XGBoost) models for identifying factors associated with pulmonary dysfunction–depression comorbidity and evaluated their clinical usefulness across different decision thresholds.

Methods: Data were drawn from the 2011 and 2015 waves of the China Health and Retirement Longitudinal Study. The analytical sample comprised 1146 adults with confirmed pulmonary dysfunction, of whom 514 (44.9%) exhibited clinically significant depressive symptoms (10-item Center for Epidemiologic Studies Depression Scale [CESD-10] score of ≥10). Models incorporated demographic, biomarker, comorbidity, and behavioral variables. Performance was assessed via discrimination (area under the receiver operating characteristic curve [AUROC]), calibration (Hosmer-Lemeshow test), and decision curve analysis. Sensitivity analyses excluding psychiatric history addressed potential conceptual overlap with the outcome.

Results: LR and BN showed similar discrimination across cohorts (AUROC≈0.73), exceeding XGBoost (0.690 training; 0.650 validation). LR had the most balanced validation performance (specificity 0.721; sensitivity 0.647), whereas BN favored sensitivity (0.884) over specificity (0.401). Training calibration was good for LR or BN, but only LR remained acceptable in validation; XGBoost was miscalibrated. XGBoost's training net benefit did not generalize. Psychiatric history was the strongest factor (odds ratio 3.46 - 7.63), followed by nephropathy, arthritis, and gastropathy; BMI and household registration were inversely associated. Excluding psychiatric history modestly reduced AUROC. With 20 shared predictors, AUROCs converged (0.658 - 0.665), BN calibrated best, LR or BN remained sensitivity-forward, and XGBoost remained specificity-forward.

Conclusions: Using routinely available clinical and sociodemographic variables, LR and BN matched or exceeded XGBoost in externally validated performance and produced more reliable probability estimates. Model choice should align with intended use: BN (or LR) is preferable for sensitivity-forward screening, whereas XGBoost may be reserved for high-threshold confirmatory decisions only after recalibration. Across methods, psychiatric history, nephropathy, arthritis, gastropathy, household registration status, and BMI emerged as stable markers of vulnerability to depressive symptoms in pulmonary dysfunction.

(JMIR Med Inform 2026;14:e77940) doi:10.2196/77940

KEYWORDS

depression screening; pulmonary dysfunction; comorbidity; analytical models; Bayesian networks

XSL•FO
RenderX

Introduction

Pulmonary dysfunctions (PDs) are defined as a pathological state characterized by impaired ventilation (obstructive or restrictive patterns) and/or gas exchange abnormalities (diffusion limitation and ventilation-perfusion mismatch), leading to hypoxemia and/or hypercapnia [1,2]. Common etiologies include chronic obstructive pulmonary disease (COPD), asthma, interstitial lung disease, pulmonary fibrosis, and pulmonary embolism [1,2]. Pulmonary dysfunction–depression comorbidity (PDDC) constitutes a clinically significant syndemic, jointly contributing to accelerated morbidity, mortality, and increased health care burdens globally [3]. PDs affect more than 500 million individuals worldwide and are frequently complicated by depressive disorders, a comorbidity associated with 23% - 40% reductions in treatment adherence and a 1.8-fold increase in hospitalization rates [4-6]. The bidirectional relationship between PDs and depression has been extensively documented, with accumulating evidence demonstrating their synergistic impact on clinical outcomes. For instance, individuals with PDs, particularly pulmonary tuberculosis, COPD, and cystic fibrosis, exhibit a 1.5 - to 2.5-fold higher risk of developing depressive symptoms than the general population [7]. While the prevalence of depression in the general Chinese population is approximately 10.6% [8], it reaches up to 39.5% among patients with particularly pulmonary tuberculosis [9]. A matched cohort study further reported that the incidence of depression was 11.4 per 1000 person-years among patients with COPD, compared with 5.7 per 1000 person-years in non-COPD controls [10].

Longitudinal analyses confirm the progressive influence of depressive symptomatology on pulmonary deterioration [11,12]. Multivariable models further demonstrated that depression severity independently predicted reduced forced expiratory volume in 1 second as a percentage of the predicted value (FEV1%) at 2-year follow-up [11]. Conversely, impaired lung function has also been found to independently predict the onset of depression, potentially mediated by hypoxemia, systemic inflammation, and glucocorticoid dysregulation [12]. The underlying pathophysiological mechanisms involve neuroendocrine-immune cross talk, including proinflammatory cytokine release (eg,interleukin-6 and tumor necrosis factor–α) induced by chronic hypoxia and hypercapnia, which disrupt monoamine neurotransmission and hippocampal neurogenesis—hallmarks of depression [13,14]. Socioeconomic disparities further exacerbate these effects, with stronger associations observed between lung function decline and subsequent depression in populations of lower socioeconomic status [15]. Despite growing understanding of this bidirectional relationship, several critical gaps remain. First, most studies use cross-sectional designs, limiting causal inference regarding temporal dynamics. Second, although pulmonary rehabilitation and selective serotonin reuptake inhibitors have demonstrated efficacy in improving both psychological and pulmonary outcomes (eg, FEV1), the biological pathways mediating these dual benefits remain poorly characterized [16]. Third, standardized screening protocols for PDDC are inconsistently applied, resulting in underdiagnosis and undertreatment.

Clinically, the prognostic implications of this comorbidity are substantial. A German cohort study of patients with cystic fibrosis revealed that depressive symptoms interacted with baseline FEV1% to predict disease trajectories: individuals with preserved lung function (FEV1>70%) but comorbid depression experienced significantly steeper declines (ΔFEV1=–5.2% vs –1.3% in nondepressed counterparts over 2 years) [11]. These findings emphasize the importance of early psychological screening within pulmonary care frameworks. However, current predictive models remain insufficiently integrated and fail to capture the full complexity of this syndemic.

Traditional statistical methods, such as logistic regression (LR), have been widely used in early research due to their transparency and interpretability. Nevertheless, these approaches have notable limitations: univariate selection strategies may overlook important interaction effects (eg, the combined influence of biomarkers and psychosocial stress), and linear assumptions may be inadequate for modeling threshold phenomena in disease progression (eg, a sharp decline in lung function triggering psychological decompensation) [17,18]. In parallel, machine learning techniques—such as random forests, Bayesian networks (BNs), and Extreme Gradient Boosting (XGBoost)—have gained increasing traction in medical prediction [19]. These algorithms offer theoretical advantages in identifying complex patterns through automated feature engineering and nonparametric modeling. However, existing applications often neglect 3 key challenges: limited clinical interpretability (black box models hinder mechanistic understanding), sample size dependency (complex algorithms may perform suboptimally in small datasets), and insufficient validation strategies (lack of robust assessment across time and geographic regions).

Despite expanding literature, 3 unresolved issues persist. First, methodological constraints remain: current models prioritize predictive accuracy over clinical interpretability, thereby limiting their use in guiding mechanistic insights or informing therapeutic decisions. Second, generalizability remains questionable, as many studies rely on single-center or region-specific data, undermining external validity across diverse populations. Third, nonlinear relationships and synergistic interactions—such as FEV1 thresholds precipitating depression onset or the combined influence of biological and psychosocial factors—are inadequately addressed.

To address these limitations, the study pursues 3 objectives. First, model development: we propose a hybrid analytical framework that integrates traditional statistical methods (eg, LR), explainable machine learning (eg, BNs), and advanced machine learning algorithms (eg, XGBoost). This framework is designed to balance predictive performance with clinical interpretability, thereby supporting actionable decision-making. Second, framework exploration: BN analysis will be used to identify key mediators—such as biomarkers, sociodemographic characteristics, and comorbidity indicators—and to clarify their conditional dependencies within PDDC. Third, validation strategy: model performance will be rigorously assessed using bootstrap resampling and DeLong tests to evaluate discrimination across spatial and temporal dimensions while addressing sample size constraints.

XSL•FO
RenderX

Through these aims, the study seeks to develop a clinically actionable model for PDDC that can inform personalized screening and targeted interventions, with the potential to reduce the burden of this syndemic. Specifically, we aim to (1) compare the discriminative performance and clinical operating characteristics of LR, BN, and XGBoost models; (2) identify core factors consistently associated with PDDC across analytical approaches; and (3) assess the clinical usefulness of each model for screening versus confirmatory diagnostic applications. We hypothesize that although machine learning models may better capture complex nonlinear relationships, traditional statistical methods will achieve comparable discrimination with greater interpretability, and that chronic multimorbidity patterns and sociodemographic factors will emerge as dominant correlates across all modeling frameworks.

Methods

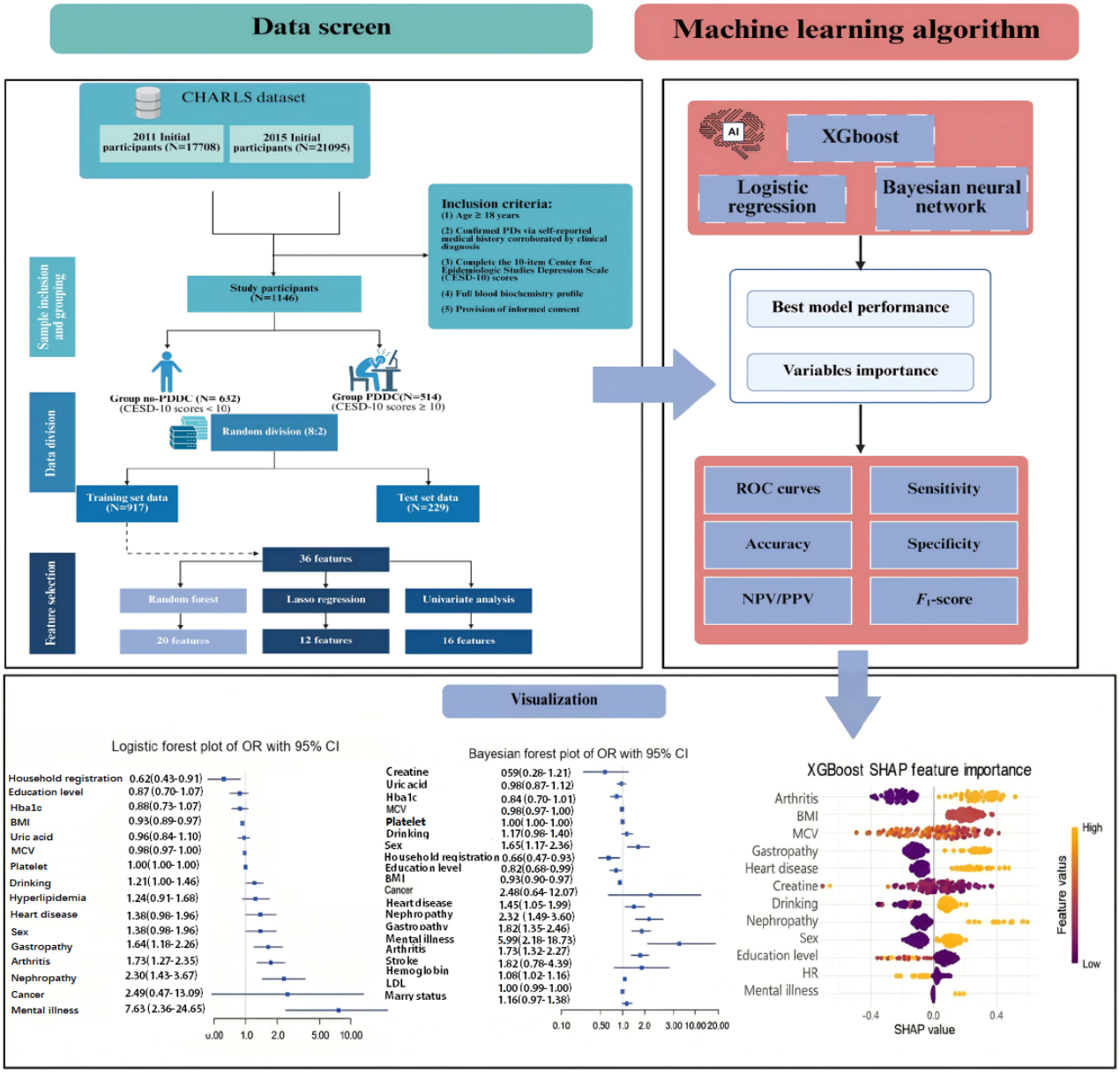
Data Source and Inclusion Criteria

The study used data from the 2011 and 2015 waves of the China Health and Retirement Longitudinal Study (CHARLS) [20], a nationally representative, longitudinal cohort study organized by the National School of Development at Peking University. The CHARLS database uses a multistage, stratified probability sampling design, spanning 28 provinces, municipalities, and autonomous regions, and encompasses 150 counties. It contains comprehensive health-related data for more than 20,000 individuals aged 45 years and older, including standardized assessments of demographic characteristics, health status and functional capacity, health care utilization and insurance coverage, and socioeconomic status, collected via structured questionnaires, clinical examinations, and laboratory assays.

Participants were selected through a rigorous multistage process. This study workflow was summarized in Figure 1. In the initial screening phase, all individuals with self-reported lung disease diagnoses or respiratory symptoms were identified from the database. From 38,803 CHARLS participants in the 2011 and 2015 waves, we applied sequential inclusion criteria: (1) age ≥18 years, (2) confirmed PD, (3) complete depression assessment, and (4) <20% missing predictor data. PD was defined using a composite criterion incorporating both self-reported physician diagnosis of chronic lung diseases (chronic bronchitis, emphysema, asthma, or pulmonary heart disease; corresponding to ICD-10 [*International Statistical Classification of Diseases, Tenth Revision*] codes J40-J47) and, where available, spirometry-confirmed airflow limitation (postbronchodilator FEV_1 /FVC <0.70 per Global Initiative for Chronic Obstructive Lung Disease criteria) [21]. This dual approach minimizes misclassification bias inherent in single-source definitions [22]. Exclusions included severe cognitive impairment (Mini-Mental State Examination <18) and terminal malignancies. The final analytical sample comprised 1146 participants.

XSL•FO
RenderX

Figure 1. This flowchart outlines the machine learning analysis pipeline based on the CHARLS dataset. AI: artificial intelligence; CESD-10: Center for Epidemiologic Studies Depression Scale-10; CHARLS: China Health and Retirement Longitudinal Study; HbA_{1c}: hemoglobin A_{1c}; HR: household registration; LDL: low-density lipoprotein; MCV: mean corpuscular volume; NPV/PPV: negative/positive predictive value; OR: odds ratios; PDs: pulmonary dysfunctions; PDDC: pulmonary dysfunction–depression comorbidity; ROC: receiver operating characteristic; SHAP: Shapley Additive Explanations; XGBoost: Extreme Gradient Boosting.



Variables and Data Collection

This analytical framework used multidimensional variables systematically categorized into four distinct domains: (1) Demographic parameters including age, sex, BMI, education level, household registration, and marital status. (2) Blood-based biomarkers comprising glucose, cholesterol, high-density lipoprotein, low-density lipoprotein, triglycerides, blood urea nitrogen, creatine, uric acid, high-sensitivity C-reactive protein, hemoglobin A_{1c} (HbA_{1c}), hematocrit, hemoglobin, white blood cell, mean corpuscular volume (MCV), and platelet. (3) Chronic disease comorbidity quantified through binary classification (yes/no) of 13 clinically significant noncommunicable conditions: hypertension, diabetes, cancer, hyperlipidemia, hepatopathy, heart disease, stroke, nephropathy, gastropathy, mental illness (defined as physician-diagnosed depression, anxiety, or other psychiatric conditions documented in medical history, distinct from current depressive symptoms assessed by 10-item Center for Epidemiologic Studies Depression Scale [CESD-10]), memory-associated disorders, arthritis, and asthma. Notably, while "mental illness" as a predictor and CESD-10–defined depression as an outcome are related constructs, they capture different dimensions: the former reflects historical clinical diagnosis, whereas the latter measures current symptom severity. (4) Behavioral determinants encompassing smoking and drinking. More details are reported in Table S1 in Multimedia Appendix 1.

Outcome Specification

Depressive symptomatology was assessed using the CESD-10 [23], a psychometrically validated screening instrument in population-based studies. Participants rated symptom frequency

XSL•FO
RenderX

over the preceding week (eg, depressed mood, loneliness, and fear) on a 4-point Likert scale (total score range: 0‐30). Consistent with established epidemiological thresholds, scores ≥10 were classified as clinically significant depressive symptoms, demonstrating robust sensitivity (91.93%) and specificity (92.76%) in prior validation studies [23].

Statistical Analysis

This study used a cross-sectional design using pooled data from 2 CHARLS waves (2011 and 2015) to develop and validate an analytical framework for identifying factors associated with prevalent depressive symptoms in patients with PDs. These waves were selected because they contain the most complete blood biomarker panels, which were essential for our multidimensional analytical approach. More recent CHARLS waves (2018 and 2020) have incomplete biomarker data due to changes in data collection protocols and COVID-19 pandemic disruptions. The pooled analytical sample comprised 1146 participants. The dataset was partitioned through random allocation, with 80% (917) of participants constituting the derivation cohort for model development, while the remaining 20% (229) served as an independent validation cohort to assess model generalizability.

Model Construction

Data Preprocessing

Systematic preprocessing was performed prior to model development. Missing values for continuous variables were imputed using multiple imputation techniques, while missing categorical variables were coded as a distinct category to retain maximal information.

LR Model Construction

We constructed a conventional LR model as a reference comparator, adopting a hierarchical analytical approach to variable selection. The methodology proceeded through two sequential phases: (1) Bonferroni-adjusted univariate screening ($\alpha \leq .05$) to identify candidate associated factors exhibiting preliminary associations with incident PDDC, followed by (2) an iterative refinement process using forward stepwise regression (entry threshold: $\alpha < .05$) to achieve optimal model parsimony while preserving discriminative capacity. The final LR model took the form:

$$\log (p1-p)=\beta 0+\sum i=1k\beta iXi$$

where p denotes the probability of depressive symptoms, β the intercept, and β the coefficients for predictors X.

BNs Model Construction

We implemented a BN model using the R package brms (BN Models using Stan, version 2.18.0). The modeling procedure comprised three principal components: (1) Associated-factor selection: twenty clinically relevant variables were identified through random forest–based feature importance analysis, spanning demographic parameters, blood-based biomarkers, chronic disease comorbidity, and behavioral determinants. (2) Prior distribution specification: weakly informative normal priors (N(0, 1)) were assigned to all regression coefficients, providing sufficient regularization while maintaining theoretical flexibility [24]. (3) Markov Chain Monte Carlo sampling configuration: the model used 4 parallel Markov Chain Monte Carlo chains, each executing 2000 iterations (1000 warm-up and 1000 sampling iterations) [25]. Convergence was optimized by setting the adaptation parameter (adapt_delta) to 0.95, effectively reducing divergent transitions while maintaining sampling efficiency.

XGBoost Model Construction

The XGBoost model construction used a 2-stage pipeline integrating Least Absolute Shrinkage and Selection Operator (LASSO) regularization for feature selection followed by gradient boosting for classification [26].

Feature Selection via LASSO

LASSO regression with L1 penalty was applied to all candidate variables. The optimal regularization coefficient (λ) was determined via 10-fold cross-validation, selecting the λ value within 1 standard error of the minimum cross-validated deviance (λ=0.023). This procedure yielded 12 variables with nonzero coefficients: mental illness, nephropathy, arthritis, gastropathy, household registration, BMI, sex, heart disease, drinking, education level, creatinine, and MCV. The relatively stringent feature selection reflects LASSO's tendency to select 1 representative variable among correlated predictors, which may have excluded potentially informative features.

Hyperparameter Optimization

XGBoost hyperparameters were optimized via grid search with 5-fold cross-validation on the training set. The parameter grid included max_depth (3, 4, 5, and 6), learning rate η (0.01, 0.05, 0.1, and 0.2), n_estimators (100, 200, and 300), subsample (0.7, 0.8, and 0.9), and colsample_bytree (0.7, 0.8, and 0.9). The optimal configuration maximizing cross-validated area under the receiver operating characteristic curve (AUROC) was max_depth=6, η=0.1, n_estimators=200, subsample=0.8, and colsample_bytree=0.8. Early stopping was implemented with a patience of 20 rounds to prevent overfitting. The objective function minimized:

$$L(\theta)=\sum i=1nl\ (yi,y\hat{}i) + \lambda\sum j=1m|wj|$$

where $l(y, \hat{y})$ denotes the logistic loss, w the feature weights, and λ the regularization strength.

Model Validation

All models were validated using a comprehensive set of metrics: (1) Primary metric: AUROC, with 95% CIs computed via the DeLong method. (2) Secondary metrics: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, each reported with point estimates and 95% CI. (3) Calibration: model calibration was assessed using calibration curves plotting observed versus predicted probabilities across deciles of predicted risk, with Hosmer-Lemeshow goodness-of-fit test statistics. (4) Clinical usefulness: decision curve analysis (DCA) was performed to evaluate net benefit across a range of clinically relevant threshold probabilities (0.1‐0.6), comparing each model with "treat all" and "treat none" strategies [27].

XSL•FO
RenderX

To facilitate fair model comparison, we additionally conducted a unified feature set analysis in which all 3 models were fit using the same prespecified predictor set (defined a priori based on clinical relevance and data availability). We reconstructed all 3 models using an identical 20-feature (ie, household registration, sex, education level, BMI, marital status, creatine, HbA_{1c}, MCV, platelet, LDL, uric acid, hemoglobin, cancer, nephropathy, heart disease, gastropathy, arthritis, mental illness, stroke, and drinking) input set. Performance metrics (discrimination, calibration, and decision-curve net benefit) were then compared under identical inputs. All analyses were conducted in R (version 4.4.3) using validated computational libraries (brms, rstan, dplyr, pROC, caret, xgboost, rmda, ggplot2, and forestplot), with complete reproducibility scripts archived in Table S2 in Multimedia Appendix 1.

Ethical Considerations

This study adhered to the principles of the Declaration of Helsinki and received ethical approval from the Biomedical Ethics Committee of Peking University (approval no. IRB00001052-11015). The requirement for informed consent was waived by the ethics committee because the study was based on a secondary analysis of routinely collected administrative data and did not involve any direct contact with individual participants. Both the original data collection and the current secondary analysis were approved without the need for additional informed consent. All data from the CHARLS were anonymized before analysis. Identifiable information had been removed by the data custodian, and the authors had no access to any personal information that could enable the identification of individual participants. Data access and analytical procedures were carried out in accordance with applicable data protection requirements. No financial or other incentives were provided to participants, as the study did not involve direct enrollment of or interaction with human subjects. Neither the manuscript nor the supplementary materials contains any images or other information that could reveal the identity of individual participants.

Results

Baseline Characteristics

The study cohort comprised 1146 participants (non-PDDC: n=632; PDDC: n=514), with baseline characteristics detailed in Table S3 in Multimedia Appendix 1. Demographic comparisons revealed no significant age difference between groups (mean 62.48, SD 9.93 vs mean 62.16, SD 9.67 years; *P*=.58), although marked disparities emerged in BMI (mean 23.20, SD 4.22 vs mean 22.54, SD 3.71 kg/m²; *P*=.01), sex distribution (male: 372/632, 58.9% vs 232/514, 45.1%; *P*<.001), educational attainment (high school or above: 169/632, 26.7% vs 65/514, 12.6%; *P*<.001), and urban household registration (215/632, 34.0% vs 100/514, 19.5%; *P*<.001). Marital status showed no statistical significance (*P*=.08). Hematologic profiling identified elevated uric acid (mean 4.81, SD 1.37 vs mean 4.43, SD 1.20 mg/dL; *P*<.001) and HbA_{1c} (mean 5.3, SD 1% vs mean 5.2, SD 1%; *P*=.002) in the non-PDDC group, alongside increased MCV (mean 91.59, SD 8.47 vs mean 90.25, SD 8.91 fL; *P*=.01) and reduced platelet counts (mean 194, SD 95 vs mean 202, SD 93×10 /L; *P*=.05). No significant differences were observed in glucose, lipid profiles, renal markers, or inflammatory indices. Chronic comorbidity analysis demonstrated substantially lower prevalence rates in the non-PDDC group for cancer (3/632, 0.5% vs 10/514, 1.9%; *P*=.02), hyperlipidemia (281/632, 44.5% vs 281/514, 54.7%; *P*=.001), heart disease (107/632, 16.9% vs 150/514, 29.2%; *P*<.001), and organ-specific pathologies (nephropathy: 46/632, 7.3% vs 88/514, 17.1%, *P*<.001; gastropathy: 130/632, 20.6% vs 211/514, 41.1%, *P*<.001). Behavioral assessments revealed differential alcohol consumption patterns (*P*<.001), with increased drinking frequency (≥ monthly: 171/632, 27.1% vs 98/514, 19.1%) and reduced nondrinkers (404/32, 64.0% vs 386/514, 75.1%) among patients with non-PDDC.

Nonunified Feature Set Analysis

Modeling Construction

LR Analysis

Sixteen significant associated factors (*P*<.05) were identified through univariate analysis, encompassing sociodemographic factors (BMI, sex, education level, and household registration), blood-based biomarkers (uric acid, HbA_{1c}, MCV, and platelet), chronic disease comorbidity (cancer, hyperlipidemia, heart disease, nephropathy, arthritis, gastropathy, and mental illness), and behavioral determinant (drinking) in the training dataset (Table 1). Multivariable LR analyses identified mental illness as the most robust associated factors of PDDC (odds ratios [ORs] 7.63, 95% CI 2.36 - 24.65), with subsequent clinical correlates demonstrating descending effect magnitudes: nephropathy (OR 2.30, 95% CI 1.43 - 3.67), arthritis (OR 1.73, 95% CI 1.27 - 2.35), gastropathy (OR 1.64, 95% CI 1.18 - 2.26), and drinking (OR 1.21, 95% CI 1.00 - 1.46). Notably, inverse associations emerged for household registration status (OR 0.62, 95% CI 0.43 - 0.91) and BMI (OR 0.93, 95% CI 0.89 - 0.97), suggesting potential protective effects (Figure 2A).

XSL•FO
RenderX

Table . Univariate analysis of the training dataset and validation dataset.

Variables[a,b]	Training dataset (N=917)			Validation dataset (N=229)		
	Group no-PDDC[c] (n=496)	Group PDDC (n=421)	*P* value	Group no-PDDC (n=136)	Group PDDC (n=93)	*P* value
Demographic parameters						
Age (years), mean (SD)	62.60 (10.07)	62.01 (9.72)	.37	62.07 (9.45)	62.83 (9.49)	.55
BMI, (kg/m²), mean (SD)	23.29 (4.37)	22.53 (3.81)	.01	22.85 (3.60)	22.61 (3.26)	.40
Sex, n (%)			<.001			.05
Male	294 (59.27)	192 (45.61)		78 (57.35)	40 (43.01)	
Female	202 (40.73)	229 (54.39)		58 (42.65)	53 (56.99)	
Education level, n (%)			<.001			.08
Primary school or below	290 (58.47)	328 (77.91)		78 (57.35	66 (70.97)	
Middle school	70 (14.11)	46 (10.93)		25 (18.38)	9 (9.68)	
High school or above	136 (27.42)	47 (11.16)		33 (24.26)	18 (19.35)	
Household registration, n (%)			<.001			.66
Rural area	327 (65.93)	349 (82.90)		90 (66.18)	65 (69.89)	
Urban area	169 (34.07)	72 (17.10)		46 (33.82)	28 (30.11)	
Marital status, n (%)			.05			.25
Married	432 (87.10)	348 (82.66)		116 (85.29)	74 (79.57)	
Separated and divorced	1 (0.20)	7 (1.66)		2 (1.47)	0 (0.00)	
Widowed	59 (11.90)	60 (14.25)		18 (13.24)	18 (19.35)	
Never married	4 (0.81)	6 (1.43)		0 (0.00)	1 (1.08)	
Blood-based biomarkers, mean (SD)						
Glucose (mg/dL)	106.40 (31.95)	104.71 (23.58)	.36	104.87 (27.93)	103.09 (21.11)	.58
Cholesterol (mg/dL)	190.55 (38.10)	189.67 (35.84)	.72	189.92 (36.36)	191.01 (39.90)	.83
HDL[d] (mg/dL)	53.67 (16.70)	53.22 (15.89)	.68	51.45 (14.54)	54.50 (13.72)	.11
LDL[e] (g/dL)	111.66 (33.08)	113.14 (29.84)	.48	112.33 (32.24)	112.58 (32.64)	.95
TG[f] (mg/dL)	122.83 (70.62)	122.26 (68.34)	.90	127.22 (78.07)	117.37 (67.64)	.31
BUN[g] (mg/dL)	16.09 (4.93)	15.69 (4.84)	.22	15.57 (4.35)	16.52 (5.07)	.14
Creatine (mg/dL)	0.84 (0.26)	0.78 (0.21)	<.001	0.79 (0.18)	0.76 (0.17)	.22
Uric acid (mg/dL)	4.83 (1.42)	4.46 (1.21)	<.001	4.71 (1.19)	4.29 (1.17)	.01
hsCRP[h] (mg/dL)	3.43 (9.10)	3.59 (8.03)	.77	4.39 (11.19)	2.64 (5.34)	.12
HbA_{1c}[i] (%)	5.44 (87.00)	5.26 (66.00)	<.001	5.44 (80.00)	5.39 (70.00)	.61
Hematocrit (mg/dL)	41.83 (6.04)	41.54 (6.06)	.47	42.67 (5.28)	41.41 (5.35)	.08
Hemoglobin (mg/dL)	14.27 (2.03)	14.42 (2.27)	.32	14.64 (1.99)	14.37 (1.73)	.27
WBC[j] (10^9/L)	6.30 (2.26)	6.49 (2.33)	.20	6.33 (2.04)	6.29 (1.94)	.88

XSL•FO
RenderX

Variables[a,b]	Training dataset (N=917)			Validation dataset (N=229)		
	Group no-PDDC[c] (n=496)	Group PDDC (n=421)	*P* value	Group no-PDDC (n=136)	Group PDDC (n=93)	*P* value
MCV[k] (fl)	91.41 (8.47)	90.48 (8.62)	.10	92.21 (8.46)	89.23 (10.12)	.02
Platelet (10^9/L)	201.32 (70.56)	211.70 (76.98)	.04	199.64 (70.83)	207.94 (70.53)	.38
Chronic disease comorbidity, n (%)						
Hypertension			.96			.70
No	327 (65.93)	276 (65.56)		86 (63.24)	62 (66.67)	
Yes	169 (34.07)	145 (34.44)		50 (36.76)	31 (33.33)	
Diabetes			.46			.90
No	428 (86.29)	355 (84.32)		119 (87.50)	80 (86.02)	
Yes	68 (13.71)	66 (15.68)		17 (12.50)	13 (13.98)	
Cancer			.03			.99
No	494 (99.60)	412 (97.86)		135 (99.26)	92 (98.92)	
Yes	2 (0.40)	9 (2.14)		1 (0.74)	1 (1.08)	
Hyperlipidemia			.004			.12
No	274 (55.24)	191 (45.37)		77 (56.62)	42 (45.16)	
Yes	222 (44.76)	230 (54.63)		59 (43.38)	51 (54.84)	
Hepatopathy			.47			.98
No	461 (92.94)	385(91.45)		130 (95.59)	88 (94.62)	
Yes	35 (7.06)	36 (8.55)		6 (4.41)	5 (5.38)	
Heart disease			<.001			.05
No	411 (82.86)	297 (70.55)		114 (83.82)	67 (72.04)	
Yes	85 (17.14)	124 (29.45)		22 (16.18)	26 (27.96)	
Stroke			.23			.41
No	484 (97.58)	404 (95.96)		136 (100.00)	92 (98.92)	
Yes	12 (2.42)	17 (4.04)		0 (0.00)	1 (1.08)	
Nephropathy			<.001			.04
No	457 (92.14)	346 (82.19)		129 (94.85)	80 (86.02)	
Yes	39 (7.86)	75 (17.81)		7 (5.15)	13 (13.98)	
Gastropathy			<.001			.02
No	396 (79.84)	245 (58.19)		106 (77.94)	58 (62.37)	
Yes	100 (20.16)	176 (41.81)		30 (22.06)	35 (37.63)	
Mental illness			.01			.16
No	491 (98.99)	404 (96.00)		135 (99.26)	89 (95.70)	
Yes	5 (1.01)	17 (4.00)		1 (0.74)	4 (4.30)	
Memory-associated disorders			.41			.99
No	488 (98.39)	410 (97.39)		132 (97.06)	90 (96.77)	
Yes	8 (1.61)	11 (2.61)		4 (2.94)	3 (3.23)	
Arthritis			<.001			.16
No	341 (68.75)	185 (43.94)		84 (61.76)	48 (51.61)	
Yes	155 (31.25)	236 (56.06)		52 (38.24)	45 (48.39)	
Asthma			.23			.24

XSL•FO
RenderX

Variables[a,b]	Training dataset (N=917)			Validation dataset (N=229)		
	Group no-PDDC[c] (n=496)	Group PDDC (n=421)	*P* value	Group no-PDDC (n=136)	Group PDDC (n=93)	*P* value
No	409 (82.46)	333 (79.10)		111 (81.62)	69 (74.19)	
Yes	87 (17.54)	88 (20.90)		25 (18.38)	24 (25.81)	
Behavioral determinants, n (%)						
Smoking			.87			.34
Nonsmoker	216 (43.55)	181 (42.99)		64 (47.06)	37 (39.78)	
Smoker	280 (56.45)	240 (57.01)		72 (52.94)	56 (60.22)	
Drinking			<.001			.40
Drink more than once a month	138 (27.82)	82 (19.48)		33 (24.26)	16 (17.20)	
Drink but less than once a month	46 (9.27)	24 (5.70)		10 (7.35)	6 (6.45)	
None of these	312 (62.90)	315 (74.82)		93 (68.38)	71 (76.34)	

[a]Continuous variables are presented as mean (SD). Normality was assessed using the Shapiro-Wilk test. Variables meeting normality assumptions (BMI, glucose, HDL, LDL, MCV, and hemoglobin) were compared using Student *t* test. Nonnormally distributed variables (TG, BUN, creatinine, uric acid, hsCRP, HbA_{1c}, hematocrit, WBC, and platelet) were compared using the Mann-Whitney *U* test.

[b]Categorical variables are presented as n (%) and compared using Pearson chi-square test or Fisher exact test (for expected cell counts <5).

[c]PDDC: pulmonary dysfunction–depression comorbidity.

[d]HDL: high-density lipoprotein.

[e]LDL: low-density lipoprotein.

[f]TG: triglycerides.

[g]BUN: blood urea nitrogen.

[h]hsCRP: high-sensitivity C-reactive protein.

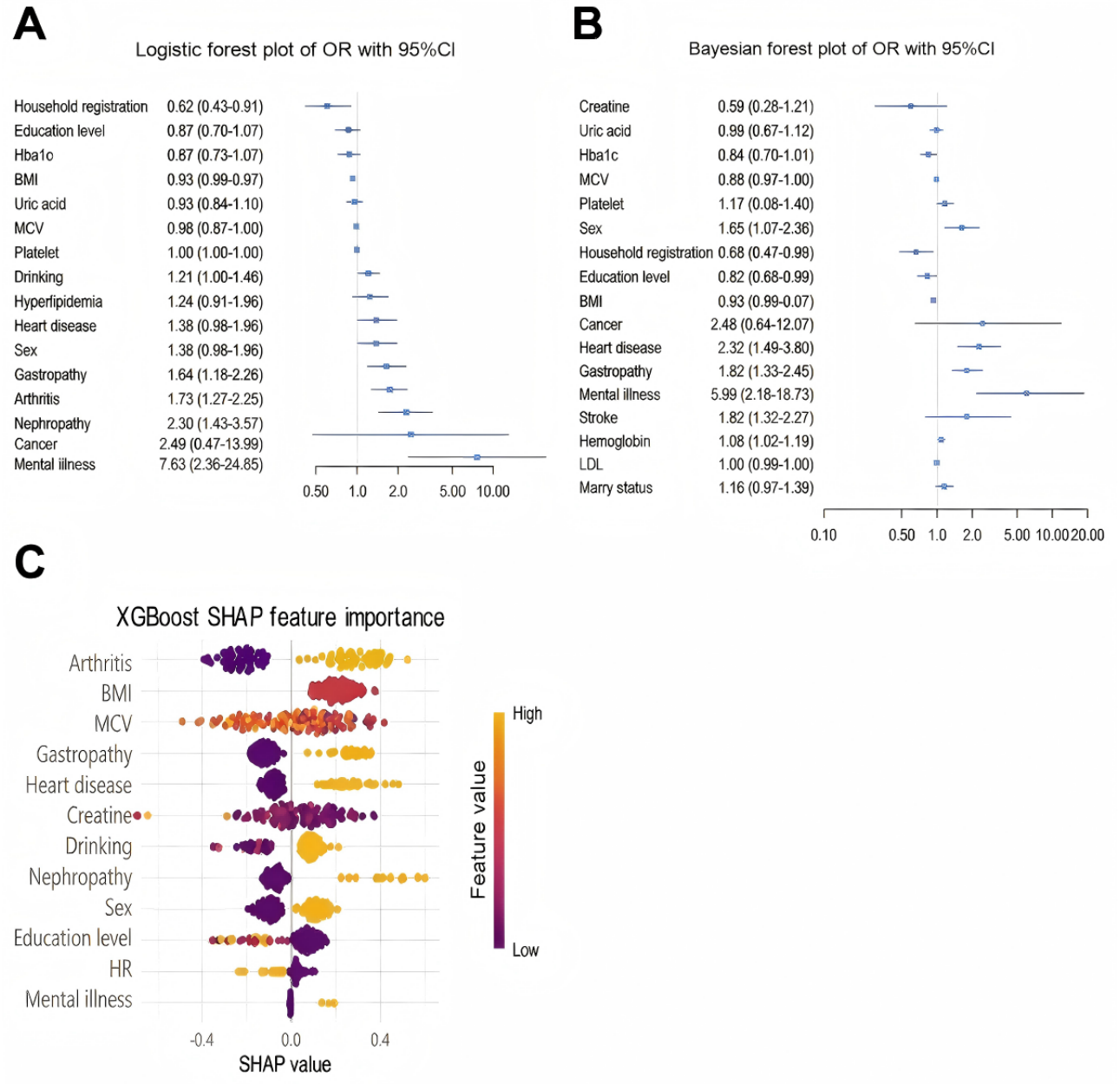
[i]HbA_{1c}: hemoglobin A_{1c}.

[j]WBC: white blood cell.

[k]MCV: mean corpuscular volume.

XSL•FO
RenderX

Figure 2. Visualization analyses of the 3 models in the nonunified feature set. (**A**) Logistic regression model, (**B**) Bayesian network model, and (**C**) Extreme Gradient Boosting model. HbA_{1c}: hemoglobin A_{1c}; HR: household registration; LDL: low-density lipoprotein; MCV: mean corpuscular volume; OR: odds ratios; SHAP: Shapley Additive Explanations.



BN Analysis

The BN model was constructed through 2 methodologically integrated phases. First, variable selection was conducted via random forest feature importance analysis, which identified 20 clinically significant associated factors across four domains: (1) demographic characteristics (household registration, sex, education level, BMI, and marital status), (2) hematologic biomarkers (creatine, HbA_{1c}, MCV, platelet, LDL, uric acid, and hemoglobin), (3) chronic comorbidities (cancer, nephropathy, heart disease, gastropathy, arthritis, mental illness, and stroke), and (4) behavioral determinants (drinking; Figure S1 in Multimedia Appendix 1). These variables demonstrated substantial analytical capacity. Second, the final model with Bayesian forest plot verification demonstrated mental illness as the strongest associated factor (OR 5.99, 95% CI 2.18 - 18.73), followed by nephropathy (OR 2.32, 95% CI 1.49 - 3.60), gastropathy (OR 1.82, 95% CI 1.35 - 2.46), arthritis (OR 1.73, 95% CI 1.32 - 2.27), sex (OR 1.65, 95% CI 1.17 - 2.36), heart disease (OR 1.45, 95% CI 1.05 - 1.99), and hemoglobin (OR 1.08, 95% CI 1.02 - 1.16). Protective associations were identified for household registration status (OR 0.66, 95% CI 0.47 - 0.93), education level (OR 0.82, 95% CI 0.68 - 0.99), and BMI (OR 0.93, 95% CI 0.90 - 0.97) (Figure 2B).

XGBoost Model Analysis

LASSO regression selected 12 robust associated factors, with 6 variables (mental illness, nephropathy, arthritis, gastropathy, household registration, and BMI; 50%) overlapping across all models. Arthritis, mental illness, gastropathy, sex, heart disease,

XSL•FO
RenderX

nephropathy, and drinking were consistently identified as positive associated factors, whereas BMI, household registration status, education level, creatine, and MCV demonstrated inverse associations (Figure S2 in Multimedia Appendix 1). Feature importance analysis via Shapley Additive Explanations (SHAP) values identified arthritis (mean |SHAP|=0.409), BMI (mean |SHAP|=0.376), and MCV (mean |SHAP|=0.350) as the 3 most salient associated factors of PDDC, with hierarchical dominance persisting across all model configurations (Figure 2C).

Sensitivity Analysis Excluding Mental Illness

To mitigate conceptual overlap between physician-diagnosed psychiatric history ("mental illness") and the CESD-10 outcome, we repeated the analyses after removing this covariate from all 3 models. In the LR model, nephropathy (OR 1.94, 95% CI 1.20 - 3.14), arthritis (OR 1.85, 95% CI 1.35 - 2.53), gastropathy (OR 1.93, 95% CI 1.38 - 2.70), household registration status (OR 0.67, 95% CI 0.46 - 0.98), and BMI (OR 0.92, 95% CI 0.88 - 0.96) remained significantly associated with the outcome. Similar associations were observed in the BN model: nephropathy (OR 2.10, 95% CI 1.30 - 3.54), arthritis (OR 1.95, 95% CI 1.40 - 2.74), gastropathy (OR 1.84, 95% CI 1.33 - 2.54), household registration status (OR 0.62, 95% CI 0.43 - 0.84), and BMI (OR 0.92, 95% CI 0.87 - 0.94). In the XGBoost model, the following features were identified as important contributors based on their SHAP values: arthritis (SHAP=0.260), gastropathy (SHAP=0.202), education level (SHAP=0.144), sex (SHAP=0.108), BMI (SHAP=0.106), household registration status (SHAP=0.079), MCV (SHAP=0.073), and creatine (SHAP=0.073). Model discrimination declined modestly (BN AUROC: 0.735-0.698, LR AUROC: 0.734-0.698, and XGBoost AUROC: 0.690-0.677) but remained acceptable, indicating that predictive performance was not primarily driven by the psychiatric history variable. The consistent, limited attenuation across models supports the independent predictive contribution of the remaining factors.

Discrimination, Calibration, and Classification Performance

To evaluate the discriminative capacity of analytical models in analyzing PDDC, receiver operating characteristic analysis was performed on both training and validation cohorts (Table 2 and Figure 3). In the training dataset, the LR model achieved an AUROC of 0.734 (95% CI 0.708 - 0.752) with balanced performance metrics (sensitivity 0.633 and specificity 0.735). The BN demonstrated comparable discrimination (AUROC 0.735, 95% CI 0.709 - 0.760) but showed distinct operational characteristics (sensitivity 0.894 vs specificity 0.389). The XGBoost algorithm yielded relatively lower predictive accuracy (AUROC 0.690, 95% CI 0.638 - 0.731) with moderate sensitivity (0.558) and specificity (0.444; Figure 3A). This performance pattern persisted in the test set, where LR maintained robust discrimination (AUROC 0.731, 95% CI 0.700 - 0.759) with preserved sensitivity (0.647) and specificity (0.721). BN showed similar AUROC performance (AUROC 0.733, 95% CI 0.709 - 0.770) but continued to exhibit high sensitivity (0.884) at the expense of specificity (0.401). XGBoost demonstrated reduced generalizability (AUROC 0.650, 95% CI 0.603 - 0.688) with suboptimal operational characteristics (Figure 3B; Table S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Comparison of discrimination metrics across 3 models.

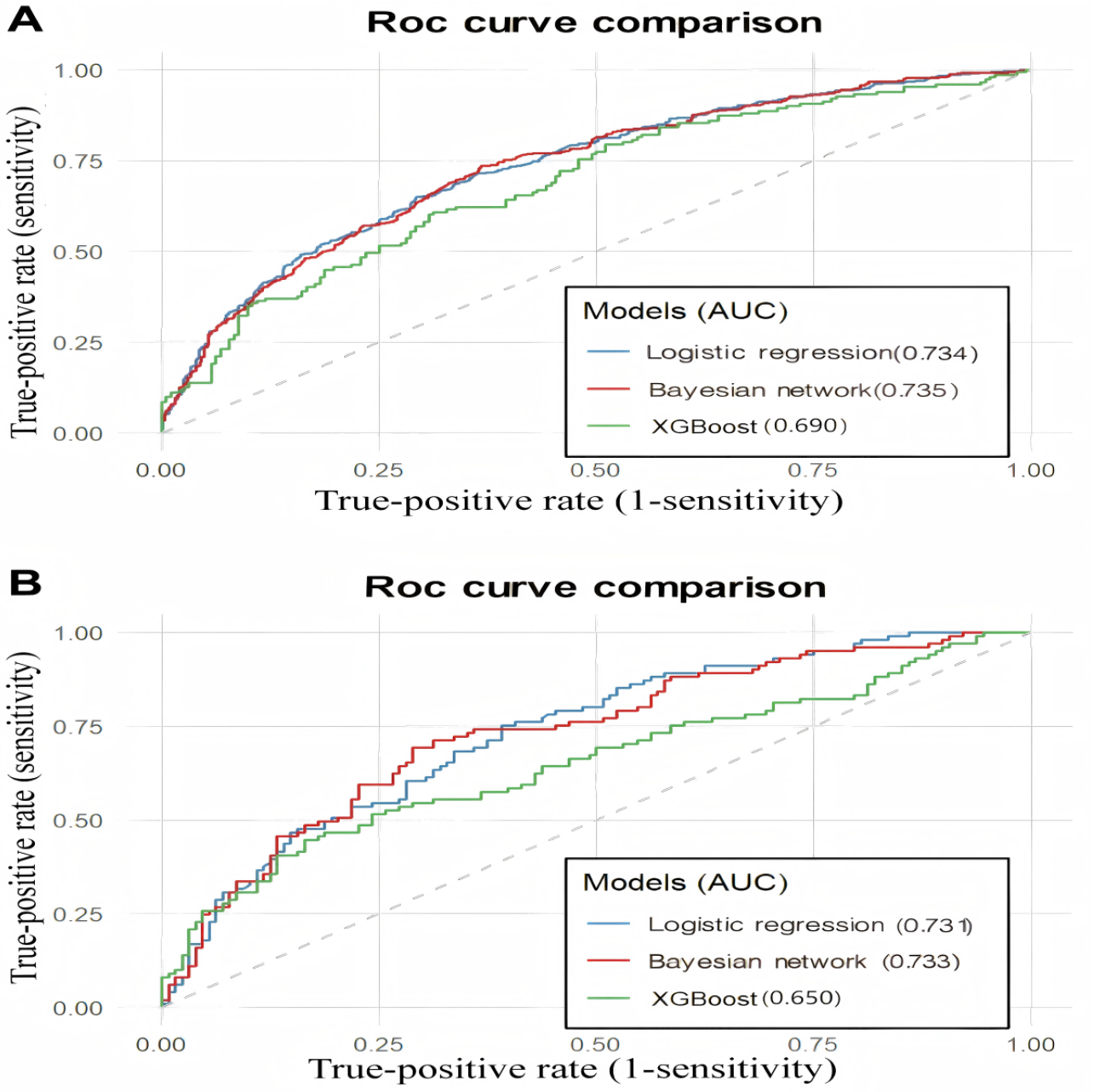
Metrics	Bayesian network model				Logistic regression model				Extreme Gradient Boosting model			
	Training set [a]		Test set [b]		Training set [a]		Test set [b]		Training set [a]		Test set [b]	
Threshold	0.612	0.527	0.591	0.375	0.407	0.493	0.364	0.373	0.504	0.468	0.515	0.509
Sensitivity	0.894	0.548	0.884	0.755	0.633	0.597	0.647	0.755	0.558	0.697	0.553	0.510
Specificity	0.389	0.814	0.401	0.635	0.735	0.763	0.721	0.635	0.444	0.792	0.561	0.833
Accuracy	0.668	0.695	0.664	0.689	0.679	0.688	0.682	0.689	0.499	0.750	0.491	0.689
Positive predictive value	0.643	0.706	0.621	0.626	0.746	0.672	0.731	0.626	0.482	0.732	0.503	0.712
Negative predictive value	0.749	0.689	0.773	0.762	0.620	0.699	0.687	0.762	0.520	0.762	0.517	0.677
F1.F1	0.748	0.618	0.651	0.684	0.685	0.632	0.665	0.684	0.517	0.714	0.437	0.594
Area under the receiver operating characteristic curve	0.735	0.749	0.733	0.661	0.734	0.753	0.731	0.665	0.690	0.844	0.650	0.658

[a]The first and second columns present the metrics for the nonunified and unified feature sets in the training set, respectively.

[b]The first and second columns present the metrics for the nonunified and unified feature sets in the test set, respectively.

XSL•FO
RenderX

Figure 3. Comparative performance of the 3 models in analyzing pulmonary dysfunction–depression comorbidity through receiver operating characteristic analysis in the nonunified feature set. (**A**) Discriminative capacity of logistic regression, Bayesian network, and Extreme Gradient Boosting models in the training dataset. (**B**) External validation of model generalizability in a test dataset. ROC: receiver operating characteristic; XGBoost: Extreme Gradient Boosting.



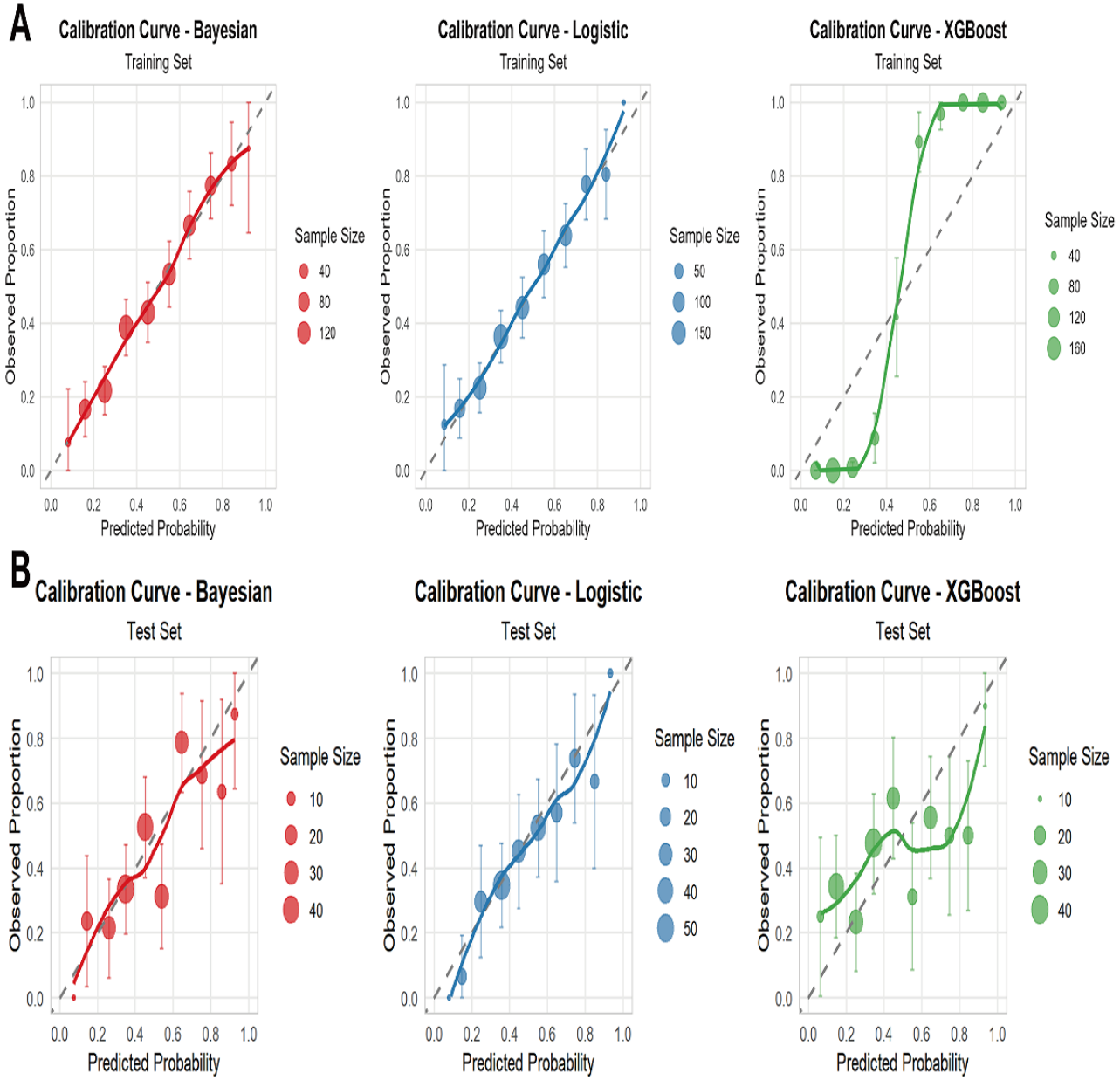
Calibration analyses for all 3 models are shown in Figure 4. In the training set, both LR and BN models demonstrated excellent calibration, with predicted probabilities closely aligned with observed outcomes (Hosmer-Lemeshow test: P=.98 and P=.99, respectively). In contrast, the XGBoost model exhibited significant miscalibration (P<.001), indicating overfitting (Figure 4A). In the test set, only the LR model maintained acceptable calibration (Hosmer-Lemeshow χ^2_8=10.71; P=.22), although moderate underestimation was observed in lower probability ranges (calibration slope=0.918). The BN model showed significant miscalibration (χ^2_8=21.18, P=.01; calibration slope=0.858), suggesting limited generalizability. The XGBoost model demonstrated the poorest calibration, with severe deviation across the probability spectrum (calibration slope=0.374; χ^2_8=48.93; P<.001; Figure 4B). Brier scores in the test set were 0.211 for LR, 0.213 for BN, and 0.256 for XGBoost, indicating comparable overall prediction accuracy between LR and BN. These findings suggest that while XGBoost achieves superior discrimination in the training set, its probability estimates require recalibration before clinical application. The LR model provides more reliable probability estimates with better preserved calibration upon external validation.

DCA revealed distinct threshold-dependent patterns between the training and test sets (Figure 5). In the training set, XGBoost demonstrated the highest net benefit across all threshold probabilities, outperforming both BN and LR models (Figure 5A). This pattern reversed in the test set, where XGBoost

XSL•FO
RenderX

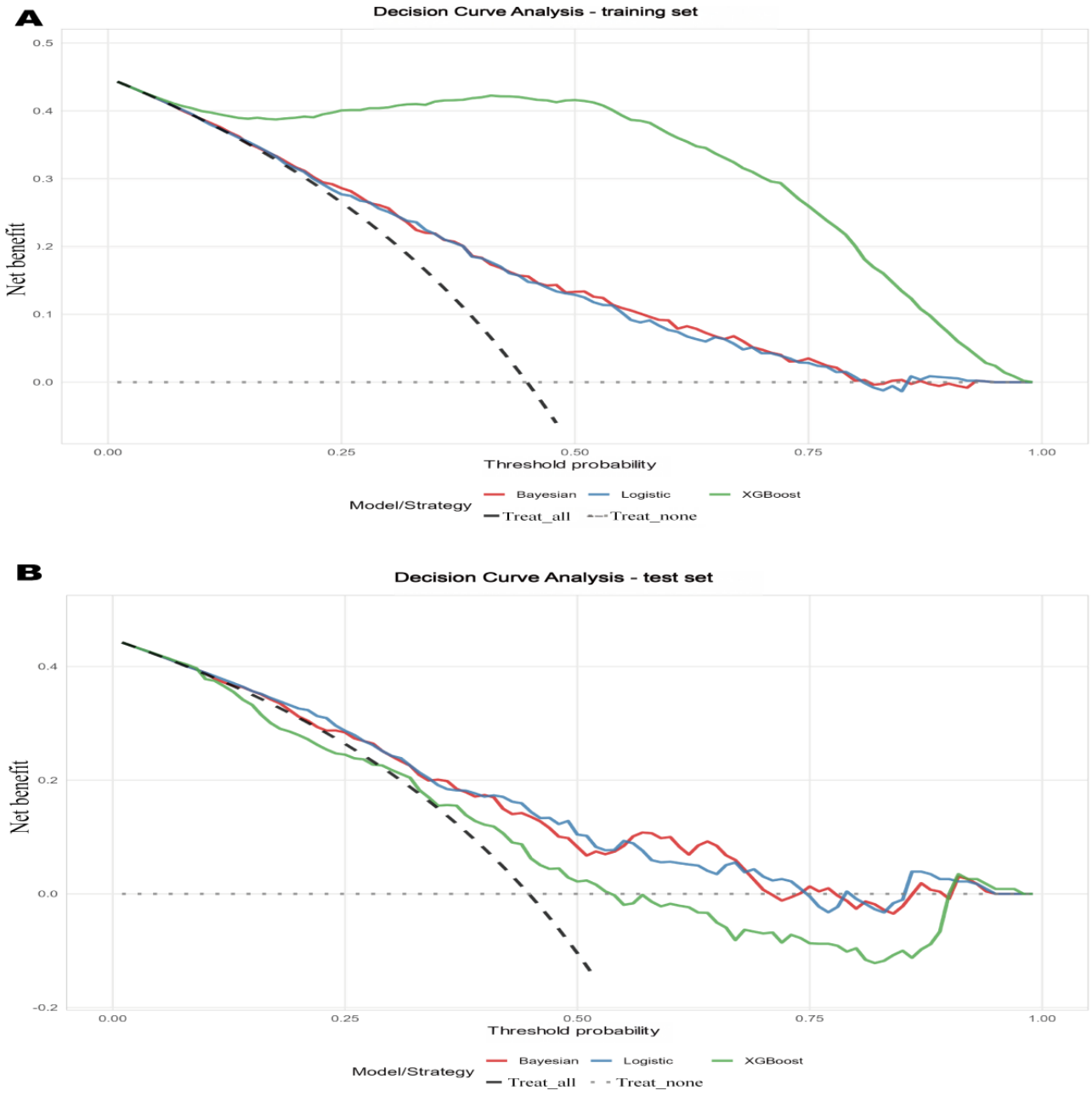
yielded lower net benefit than BN and LR across the entire threshold range, indicating limited generalizability (Figure 5B). Within the test set, BN and LR models provided more consistent clinical usefulness. At moderate-to-high thresholds (55% - 60%), the BN model showed marginally higher net benefit than LR. Overall, while XGBoost exhibited superior training set performance, and BN and LR achieved more stable net benefit upon external validation, consistent with overfitting by XGBoost.

Figure 4. Calibration curve analyses for logistic regression, Bayesian network, and Extreme Gradient Boosting models in analyzing pulmonary dysfunction–depression comorbidity in the nonunified feature set. (**A**) Training set. (**B**) Test set. XGBoost: Extreme Gradient Boosting.



XSL•FO
RenderX

Figure 5. Decision curve analyses for logistic regression, Bayesian network, and Extreme Gradient Boosting models in analyzing pulmonary dysfunction–depression comorbidity in the nonunified feature set. (**A**) Training set. (**B**) Test set. XGBoost: Extreme Gradient Boosting.



Unified Feature Set Analysis

Modeling Construction

LR Analysis

In multivariable LR, physician-diagnosed mental illness showed the strongest association with the outcome (OR 4.70, 95% CI 1.51 - 14.62). Other significant correlates, in descending magnitude, were nephropathy (OR 2.12, 95% CI 1.31 - 3.42), arthritis (OR 1.94, 95% CI 1.42 - 2.65), gastropathy (OR 1.84, 95% CI 1.32 - 2.58), male sex (OR 1.71, 95% CI 1.17 - 2.51), heart disease (OR 1.55, 95% CI 1.07 - 2.22), and hemoglobin (OR 1.10, 95% CI 1.02 - 1.20). Household registration status (OR 0.61, 95% CI 0.42 - 0.90) and BMI (OR 0.93, 95% CI 0.89 - 0.96) were inversely associated with the outcome.

BN Analysis

The BN model yielded a similar pattern. Mental illness remained the strongest predictor (OR 3.46, 95% CI 1.38 - 8.94), followed by nephropathy (OR 2.14, 95% CI 1.32 - 3.56), arthritis (OR 1.95, 95% CI 1.42 - 2.72), gastropathy (OR 1.86, 95% CI 1.35 - 2.56), male sex (OR 1.72, 95% CI 1.19 - 2.46), heart disease (OR 1.54, 95% CI 1.05 - 2.25), and hemoglobin (OR 1.11, 95% CI 1.02 - 1.20). Household registration status (OR 0.61, 95% CI 0.42 - 0.86) and BMI (OR 0.92, 95% CI 0.89 - 0.96) again showed inverse associations.

XGBoost Model Analysis

SHAP analyses from the XGBoost model ranked arthritis as the most influential feature (mean |SHAP| 0.260), followed by gastropathy (0.202), education level (0.144), sex (0.108), and BMI (0.106). These features contributed the largest marginal

XSL•FO
RenderX

effects on predictions and accounted for most of the model's explanatory signal (Figure S3 in Multimedia Appendix 1).

Sensitivity Analysis Excluding Mental Illness

To reduce potential conceptual overlap between physician-diagnosed psychiatric history ("mental illness") and the CESD-10 outcome, we repeated the unified feature set analyses after removing this covariate from all 3 models. Model discrimination remained largely unchanged, indicating that predictive performance was robust to this specification. The modest, consistent attenuation observed across models suggests that the remaining multimorbidity and sociodemographic predictors contribute independent prognostic information. Although mental illness was strongly associated with the outcome, the retained covariates continued to provide substantial predictive value (Table S5 in Multimedia Appendix 1).

Discrimination, Calibration, and Classification Performance

In the training set, LR achieved an AUROC of 0.753, with specificity of 0.763 and sensitivity of 0.597. The BN showed a similar AUROC (0.749) but a distinct operating profile, with higher specificity (0.814) and lower sensitivity (0.548). XGBoost demonstrated superior discrimination (AUROC 0.844), with sensitivity of 0.697 and specificity of 0.792. In the independent test cohort, discriminative performance was comparable across models: LR yielded an AUROC of 0.665 (95% CI 0.595 - 0.736), BN 0.661 (95% CI 0.590 - 0.731), and XGBoost of 0.658 (95% CI 0.585 - 0.731), with no significant pairwise differences (all P>.37). Despite similar AUROCs, threshold-dependent performance differed substantially. At their respective optimal thresholds, BN and LR favored sensitivity (0.755 for both), whereas XGBoost prioritized specificity (0.833; Table 2; Figure S4 and Table S4 in Multimedia Appendix 1).

Calibration results are shown in Figure S5 in Multimedia Appendix 1. In the training set, LR and BN exhibited near-perfect calibration (Hosmer-Lemeshow P=.84 and P=.93, respectively), whereas XGBoost was significantly miscalibrated (P<.001), consistent with overfitting. In the test set, only BN retained acceptable calibration (Hosmer-Lemeshow χ^2_8=14.23; P=.12), despite mild underestimation at lower probabilities (calibration slope=0.645). LR showed moderate miscalibration (P=.01; slope=0.613), and XGBoost performed worst, with systematic deviation across the probability range (slope=0.680) and unreliable Hosmer–Lemeshow statistics owing to sparse bins. Brier scores in the test set were 0.227 for XGBoost, 0.229 for LR, and 0.229 for BN, indicating marginally superior point prediction accuracy for XGBoost despite its calibration shortcomings. A composite calibration score integrating Brier score, calibration slope, and Hosmer-Lemeshow results ranked BN highest (0.493), followed by LR (0.182); XGBoost could not be fully scored owing to unreliable statistics. DCA revealed threshold-dependent net benefit (Figure S6 Multimedia Appendix 1). Between 0% and −40% risk threshold, LR and BN outperformed XGBoost, reflecting their higher sensitivity for ruling out disease in screening settings. Above −40%, XGBoost delivered the greatest net benefit, consistent with its high-specificity profile for confirmatory testing (except a narrow range near 60%, where LR was marginally superior). In summary, with identical features, the simpler LR and BN models matched XGBoost's discriminative performance while offering substantially better-calibrated probabilities. BN provided the most reliable risk estimates overall. For low-risk-threshold (screening) applications, BN or LR is preferred; for high-risk-threshold (confirmatory) decisions, XGBoost can be considered if its probabilities are recalibrated.

Discussion

Principal Findings

In a cohort of 1146 participants, PDDC was characterized by a higher burden of chronic comorbidities and less favorable sociodemographic profiles, while most routine laboratory markers showed limited separation between groups. Across modeling strategies, discrimination for PDDC risk was modest and broadly comparable, whereas calibration and threshold-specific operating behavior differed meaningfully. A small set of recurrent factors—psychiatric history, nephropathy, arthritis, gastropathy, household registration status, and BMI—emerged consistently across models and remained informative when psychiatric history was excluded to reduce conceptual overlap with the CESD-10 outcome.

Impact of Feature Selection

Different feature selection strategies (univariate screening for LR, random forest importance for BN, and LASSO- or SHAP-guided selection for XGBoost) yielded overlapping predictors, suggesting a stable signal rather than method-specific artifacts [28]. Notably, the unified feature set analysis narrowed performance differences and highlighted that model behavior was driven less by discrimination than by calibration and operating thresholds. This finding is practically important: when models are restricted to the same routinely available inputs, simpler approaches can match the discrimination of more complex learners, while often offering more interpretable structure and, in some evaluations, more reliable probabilities.

Comparative Performance of Analytical Models

Discriminative Capacity

In the development data, LR and BN achieved similar discrimination (AUROC=0.73) and consistently outperformed XGBoost (AUROC 0.69 in training, declining to 0.65 in validation), suggesting weaker generalizability for the more complex learner under the original specification. When evaluated in an independent test cohort with identical predictors, all 3 models converged to nearly identical AUROCs (0.66) and did not differ statistically. This convergence indicates that, with the available feature set, algorithmic complexity alone did not translate into improved discrimination; instead, the attainable signal for separating PDDC from non-PDDC appears constrained by the information content of routinely collected variables [29].

Clinical Operational Characteristics

Despite similar AUROCs, the models adopted distinct operating points at their optimized thresholds. LR tended to preserve a more balanced sensitivity-specificity trade-off in the development or validation evaluation, whereas BN consistently

XSL•FO
RenderX

favored sensitivity at the expense of specificity. This high-sensitivity profile may be advantageous in screening contexts where the primary objective is to minimize missed cases [30]. In contrast, XGBoost favored specificity in the independent test cohort, aligning better with confirmatory strategies where avoiding false positives is prioritized.

DCA underscored that clinical usefulness was threshold dependent and cohort sensitive [31]. In the development set, XGBoost showed higher net benefit across thresholds, but this advantage did not generalize; in the test set, LR and BN provided more stable net benefit. In the independent test cohort, LR or BN was preferred at lower risk thresholds, whereas XGBoost offered greater net benefit at higher thresholds, consistent with their sensitivity- versus specificity-forward profiles. These results argue against selecting a model solely on discrimination and support aligning model choice with the intended decision threshold range [32].

Limitations of Model Reliability

Calibration differentiated model reliability more clearly than discrimination [33]. In the development set, LR and BN were well calibrated, while XGBoost showed clear miscalibration consistent with overfitting. Calibration in held-out data was less stable and varied by evaluation cohort: in 1 test evaluation, LR retained acceptable calibration with mild underestimation at low predicted risks, whereas BN showed significant miscalibration; in the independent test cohort analysis with identical predictors, BN achieved the strongest overall calibration profile (despite underestimation at lower probabilities), LR was moderately miscalibrated, and XGBoost exhibited systematic deviation and unstable goodness-of-fit diagnostics due to sparse bins. Across these analyses, two points are consistent: (1) calibration is more fragile than discrimination and sensitive to dataset shift, and (2) XGBoost produced the least reliable probability estimates without recalibration, even when its point prediction error (Brier score) was competitive. Because risk stratification in practice relies on accurate probabilities rather than rankings alone, calibration performance should be a primary determinant of clinical readiness [34].

Clinical Significance of Key Associated Factors

Core Associated Factors Identified Across Models

Across analytic approaches, psychiatric history was the strongest associated factor, with nephropathy, arthritis, and gastropathy consistently ranking among the most influential clinical correlates. Heart disease and hemoglobin also emerged as contributors in multivariable models, reinforcing the relevance of broader multimorbidity and systemic health in PDDC risk profiles [35]. The repeated identification of these conditions across LR, BN, and XGBoost interpretability analyses (OR/credible intervals and SHAP rankings) supports their robustness as markers of vulnerability.

Crucially, removing psychiatric history to address conceptual overlap with CESD-10 produced only modest declines in discrimination and preserved the associations for nephropathy, arthritis, gastropathy, household registration status, and BMI. This pattern indicates that the predictive framework was not dominated by psychiatric history alone and that the remaining variables contributed independent prognostic information. By contrast, several laboratory biomarkers that differed between groups at baseline (eg, uric acid, HbA_{1c}, and platelet count) played a less consistent role in multivariable models, suggesting that their crude differences may be confounded or clinically modest in magnitude [36].

Impact of Sociodemographic Factors

Sociodemographic factors were consistently retained in the final models and were generally associated with lower PDDC risk. At baseline, rural household registration status and lower educational attainment were more common in the PDDC group; in adjusted analyses, both household registration status and education level were inversely associated with subsequent PDDC risk [37,38]. This pattern aligns with the influence of structural determinants—access to services, continuity of care, health literacy, and material constraints—on susceptibility to depressive symptoms among individuals with impaired pulmonary function [39,40].

The inverse association between BMI and PDDC risk contrasts with much of the Western literature, where obesity is often linked to higher depression risk. Several cohort- and context-specific considerations may account for this difference. In older Chinese adults, lower BMI more frequently reflects malnutrition, sarcopenia, or frailty, all of which are strongly associated with depressive symptoms. Moreover, an “obesity paradox” has been reported in Asian populations, whereby modestly higher BMI is associated with better outcomes in later life. In our cohort, BMI was relatively low overall (mean 22.9 kg/m²), and obesity was uncommon (BMI≥30 kg/m², approximately 4%), limiting power to detect nonlinear (U- or J-shaped) associations. Cultural norms may also shape the BMI-depression relationship by altering the social meaning of body weight and the extent of weight-related stigma [41,42]. Notably, BMI showed an inverse association across multiple models, consistent with the lower mean BMI observed in the PDDC group [43]; this likely captures vulnerability phenotypes—frailty, sarcopenia, or higher chronic disease burden—rather than a protective effect of lower weight per se [44]. Sex differences remained evident after adjustment in the unified analyses, suggesting heterogeneity in symptom reporting, comorbidity profiles, or access to care that merits further investigation.

Methodological Contributions and Innovations

This study contributes methodologically by evaluating 3 distinct modeling paradigms—linear (LR), probabilistic graphical (BN), and gradient-boosted trees (XGBoost)—under both model-specific and unified predictor settings. The evaluation emphasized not only discrimination but also calibration and DCA, aligning performance assessment with clinical decision-making rather than rank-based metrics alone. Interpretability was addressed through complementary tools (effect estimates for LR or BN and SHAP for XGBoost), enabling triangulation of core factors that remained consistent across methods. Finally, the sensitivity analysis excluding psychiatric history directly addressed a common concern in depression prediction and strengthened confidence that

XSL•FO
RenderX

multimorbidity and sociodemographic variables provide independent predictive signal.

Study Limitations

Several limitations should be considered. First, overall discrimination in external testing was modest, indicating that key determinants of PDDC risk remain unmeasured. Second, calibration was cohort dependent, and goodness-of-fit testing was unstable for XGBoost in the independent test cohort due to sparse bins, limiting inferential confidence from Hosmer-Lemeshow statistics. Third, CESD-10 is a screening tool rather than a diagnostic interview; outcome misclassification is possible. Fourth, comorbidity histories and behavioral measures may be subject to reporting or recording bias, and the cross-sectional design precludes causal interpretation. Fifth, our reliance on 2011 and 2015 CHARLS data, while necessary to obtain complete biomarker profiles, limits the temporal relevance of our findings. Health care accessibility, diagnostic practices, and population health characteristics may have evolved over the past decade. Finally, the feature set lacked clinically rich markers such as pulmonary function severity, symptom burden, medication exposure, and psychosocial stressors, which likely mediate or modify PDDC risk. Although the core biological and sociodemographic relationships we identified are likely to remain stable, validation with more recent cohort data is warranted when comprehensive biomarker panels become available.

Future Research Directions

Future work should focus on (1) prospective, multisite external validation to quantify calibration drift and transportability; (2) formal recalibration and model updating, particularly for tree-based learners, before clinical use; (3) incorporation of disease severity measures and longitudinal trajectories to distinguish incident from persistent depressive symptoms; and (4) evaluation of clinically actionable thresholds in workflow-based impact studies to determine whether model-guided screening or referral improves outcomes. Given the contribution of sociodemographic factors, fairness and subgroup performance analyses should also be prioritized.

Conclusions

Using routinely available clinical and sociodemographic data, LR and BN achieved discrimination comparable with XGBoost and, in several evaluations, provided more reliable probability estimates. Model selection should be driven by the intended clinical application: LR offers a balanced operating profile, BN is attractive when sensitivity is prioritized, and XGBoost may be considered for high-threshold decisions only with careful recalibration. Across modeling strategies, psychiatric history, nephropathy, arthritis, gastropathy, household registration status, and BMI emerged as stable associated factors, and the predictive signal persisted after excluding psychiatric history, underscoring the independent contribution of multimorbidity and social determinants to PDDC risk.

Acknowledgments

No generative artificial intelligence tools were used to generate the manuscript text, analyses, figures, tables, or multimedia appendices. All study design decisions, analyses, interpretations, and final wording were produced and verified by the authors. The authors would like to express their gratitude to the participants who kindly took part in this study.

Funding

This work is supported by the Basic Public Welfare Research Project of Zhejiang Province (grant number LGF21H260007), the Medical Science and Technology Project of Zhejiang Province (grant numbers 2020PY064 and 2021PY065), the Zhejiang Provincial Disease Prevention and Control Science and Technology Program Project (grant number 2026JKY172), and the Disease Prevention and Control Innovation Team of Zhejiang Province (project number 2026JKP-03).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

QC, WT, and CQC devised the analysis plan, conducted data and result analyses, and composed the final manuscript. QC and WT formulated the primary study framework. CQC contributed to the data analysis process and the interpretation of results. RD, QJ, XB, QL, YW, and YH reviewed and refined the final manuscript version, while RD, QJ, and XB collected relevant references. All authors have read and approved the published version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variable coding, statistical analysis, and model performance of the study on pulmonary dysfunction–depression comorbidity.
[DOCX File, 2666 KB - medinform_v14i1e77940_app1.docx]

XSL•FO
RenderX

References

1. Stanojevic S, Kaminsky DA, Miller MR, et al. ERS/ATS technical standard on interpretive strategies for routine lung function tests. Eur Respir J 2022 Jul;60(1):2101499. [doi: 10.1183/13993003.01499-2021] [Medline: 34949706]
2. The Asia Pacific COPD Roundtable Group. Global Initiative for Chronic Obstructive Lung Disease strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease: an Asia–Pacific perspective. Respirology 2005 Jan;10(1):9-17. [doi: 10.1111/j.1440-1843.2005.00692.x]
3. Hu W, Liu BP, Jia CX. Association and biological pathways between lung function and incident depression: a prospective cohort study of 280,032 participants. BMC Med 2024;22(1):160. [doi: 10.1186/s12916-024-03382-3]
4. Berk M, Köhler-Forsberg O, Turner M, et al. Comorbidity between major depressive disorder and physical diseases: a comprehensive review of epidemiology, mechanisms and management. World Psychiatry 2023 Oct;22(3):366-387. [doi: 10.1002/wps.21110] [Medline: 37713568]
5. Almeida SS, Zizzi FB, Cattaneo A, et al. Management and treatment of patients with major depressive disorder and chronic diseases: a multidisciplinary approach. Front Psychol 2020;11:542444. [doi: 10.3389/fpsyg.2020.542444] [Medline: 33101117]
6. Wang C, Chen H, Shang S. Association between depression and lung function in college students. Front Public Health 2023;11:1093935. [doi: 10.3389/fpubh.2023.1093935]
7. Guo L, Cao J, Cheng P, et al. Moderate-to-severe depression adversely affects lung function in Chinese college students. Front Psychol 2020;11:652. [doi: 10.3389/fpsyg.2020.00652] [Medline: 32351425]
8. Qin X, Wang S, Hsieh CR. The prevalence of depression and depressive symptoms among adults in China: Estimation based on a National Household Survey. China Economic Review 2018 Oct;51:271-282. [doi: 10.1016/j.chieco.2016.04.001]
9. Liu K, Zhang Y, Qu S, Yang W, Guo L, Zhang L. Prevalence and correlates of anxiety and depressive symptoms in patients with and without multi-drug resistant pulmonary tuberculosis in China. Front Psychiatry 2021;12:674891. [doi: 10.3389/fpsyt.2021.674891] [Medline: 34557116]
10. O'Toole J, Woo H, Putcha N, et al. Comparative impact of depressive symptoms and FEV1% on chronic obstructive pulmonary disease. Ann Am Thorac Soc 2022 Feb;19(2):171-178. [doi: 10.1513/AnnalsATS.202009-1187OC] [Medline: 34410883]
11. Fidika A, Herle M, Goldbeck L. Symptoms of depression impact the course of lung function in adolescents and adults with cystic fibrosis. BMC Pulm Med 2014 Dec 16;14:205. [doi: 10.1186/1471-2466-14-205] [Medline: 25515072]
12. Xie H, Jiang Y, Liu L, Peng H, Li J, Chen Z. Global prevalence and risk factors of depression in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis from 2000 to 2022. J Psychosom Res 2023 Dec;175:111537. [doi: 10.1016/j.jpsychores.2023.111537] [Medline: 37907038]
13. Bai Y, Cai Y, Chang D, Li D, Huo X, Zhu T. Immunotherapy for depression: recent insights and future targets. Pharmacol Ther 2024 May;257:108624. [doi: 10.1016/j.pharmthera.2024.108624] [Medline: 38442780]
14. Fang S, Wu Z, Guo Y, et al. Roles of microglia in adult hippocampal neurogenesis in depression and their therapeutics. Front Immunol 2023;14:1193053. [doi: 10.3389/fimmu.2023.1193053] [Medline: 37881439]
15. Rocha V, Fraga S, Moreira C, et al. Life-course socioeconomic disadvantage and lung function: a multicohort study of 70 496 individuals. Eur Respir J 2021 Mar;57(3):2001600. [doi: 10.1183/13993003.01600-2020] [Medline: 33214206]
16. Silva L, Maricoto T, Mota Â, et al. Effectiveness of a home-based pulmonary rehabilitation maintenance programme: the Rehab2Life study protocol. BMC Nurs 2024 May 21;23(1):338. [doi: 10.1186/s12912-024-01999-6] [Medline: 38773568]
17. Nzekwe CJ, Kim S, Mostafa SA. Interaction selection and prediction performance in high-dimensional data: a comparative study of statistical and tree-based methods. J Data Sci 2024;22(2):259-279. [doi: 10.6339/24-JDS1127]
18. St-Pierre J, Oualkacha K, Rai Bhatnagar S. Hierarchical selection of genetic and gene by environment interaction effects in high-dimensional mixed models. Stat Methods Med Res 2025 Jan;34(1):180-198. [doi: 10.1177/09622802241293768] [Medline: 39659138]
19. Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. IEEE Rev Biomed Eng 2021;14:116-126. [doi: 10.1109/RBME.2020.3007816] [Medline: 32746368]
20. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol 2014 Feb;43(1):61-68. [doi: 10.1093/ije/dys203] [Medline: 23243115]
21. Fang L, Gao P, Bao H, et al. Chronic obstructive pulmonary disease in China: a nationwide prevalence study. Lancet Respir Med 2018 Jun;6(6):421-430. [doi: 10.1016/S2213-2600(18)30103-6] [Medline: 29650407]
22. Osimo EF, Baxter LJ, Lewis G, Jones PB, Khandaker GM. Prevalent and incident depression in inflammatory bowel disease: a systematic review and meta-analysis of inflammation-depression links in somatic illness. Psychol Med 2019;49(12):1951-1963. [doi: 10.1017/S0033291719001454] [Medline: 31258105]
23. Fu H, Si L, Guo R. What is the optimal cut-off point of the 10-Item Center for Epidemiologic Studies Depression Scale for screening depression among Chinese individuals aged 45 and over? An exploration using latent profile analysis. Front Psychiatry 2022;13:820777. [doi: 10.3389/fpsyt.2022.820777] [Medline: 35360127]
24. Zhang Y, Pichon L, Roux S, Pellegrino A, Simonneau T, Tisseyre B. Why make inverse modeling and which methods to use in agriculture? A review. Comput Electron Agric 2024 Feb;217:108624. [doi: 10.1016/j.compag.2024.108624]

XSL•FO
RenderX

25. van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov Chain Monte-Carlo sampling. Psychon Bull Rev 2018 Feb;25(1):143-154. [doi: 10.3758/s13423-016-1015-8] [Medline: 26968853]
26. Xue X, Liu Z, Xue T, Chen W, Chen X. Machine learning for the prediction of acute kidney injury in patients after cardiac surgery. Front Surg 2022;9:946610. [doi: 10.3389/fsurg.2022.946610] [Medline: 36157418]
27. Zhang H, Miao Q, Fu Y, et al. Intratumoral and peritumoral radiomics based on automated breast volume scanner for predicting human epidermal growth factor receptor 2 status. Front Oncol 2025;15:1556317. [doi: 10.3389/fonc.2025.1556317]
28. Zhong W, Wang C, Wang J, Chen T. Machine learning models to further identify advantaged populations that can achieve functional cure of chronic hepatitis B virus infection after receiving Peg-IFN alpha treatment. Int J Med Inform 2025 Jan;193:105660. [doi: 10.1016/j.ijmedinf.2024.105660] [Medline: 39454328]
29. Cao Z, Xie X. Structure learning with consensus label information for multi-view unsupervised feature selection. Expert Syst Appl 2024 Mar;238:121893. [doi: 10.1016/j.eswa.2023.121893]
30. Balaban B, Demirci N, Yilgor C, et al. Predicting mechanical complications in adult spinal deformity patients with postoperative proportioned and moderately disproportioned alignment. Acta Orthop Traumatol Turc 2025 Jul 18;59(4):210-221. [doi: 10.5152/j.aott.2025.24146] [Medline: 40728025]
31. Zhang G, Yang H, Zhu X, et al. A CT-based radiomics nomogram to predict complete ablation of pulmonary malignancy: a multicenter study. Front Oncol 2022;12:841678. [doi: 10.3389/fonc.2022.841678]
32. Chen Y, Lv W, Liu X, et al. Development of a prediction model for hemorrhagic transformation after intravenous thrombolysis in patients with acute ischemic stroke: a retrospective analysis. BMC Med Inform Decis Mak 2025 Jul 1;25(1):227. [doi: 10.1186/s12911-025-03068-7] [Medline: 40598054]
33. Milan R, LeLorier J, Brouillette MJ, Holbrook A, Litvinov IV, Rahme E. Sex differences in the patterns of systemic agent use among patients with psoriasis: a retrospective cohort study in Quebec, Canada. Front Pharmacol 2022;13:810309. [doi: 10.3389/fphar.2022.810309] [Medline: 35242034]
34. Zhou A, Chen K, Wei Y, et al. Machine learning-based prediction of carotid intima–media thickness progression: a three-year prospective cohort study. Front Med 2025;12. [doi: 10.3389/fmed.2025.1593662] [Medline: 1593662]
35. Li M, Tang H, Liu X. Primary care team and its association with quality of care for people with multimorbidity: a systematic review. BMC Prim Care 2023 Jan 19;24(1):20. [doi: 10.1186/s12875-023-01968-z] [Medline: 36653754]
36. Chen S, Liu Y, Islam SMS, et al. A simple prediction model to estimate obstructive coronary artery disease. BMC Cardiovasc Disord 2018 Jan 16;18(1):7. [doi: 10.1186/s12872-018-0745-0] [Medline: 29338684]
37. Jiang H, Feng B, Wang Y, et al. Application of Anderson model to analyze the influencing factors of lung function test behavior in middle-aged and elderly people in China. Dialogues Health 2023 Dec;2:100138. [doi: 10.1016/j.dialog.2023.100138] [Medline: 38515471]
38. Guo J, Su M, Huang J, et al. Longitudinal bidirectional association between gastrointestinal disease and depression symptoms among middle-aged and older adults in China. Arch Public Health 2025 Jul 1;83(1):171. [doi: 10.1186/s13690-025-01671-8] [Medline: 40598682]
39. Lu Y, Feng L, Feng L, Nyunt MS, Yap KB, Ng TP. Systemic inflammation, depression and obstructive pulmonary function: a population-based study. Respir Res 2013 May 15;14(1):53. [doi: 10.1186/1465-9921-14-53] [Medline: 23676005]
40. Ghaemi Kerahrodi J, Brähler E, Wiltink J, et al. Association between medicated obstructive pulmonary disease, depression and subjective health: results from the population-based Gutenberg Health Study. Sci Rep 2019 Dec 27;9(1):20252. [doi: 10.1038/s41598-019-56440-9] [Medline: 31882715]
41. Zhang L, Liu K, Li H, et al. Relationship between body mass index and depressive symptoms: the “fat and jolly” hypothesis for the middle-aged and elderly in China. BMC Public Health 2016 Nov 29;16(1):1201. [doi: 10.1186/s12889-016-3864-5] [Medline: 27894296]
42. Zhou Q, Wang T, Basu K. Negative association between BMI and depressive symptoms in middle aged and elderly Chinese: results from a national household survey. Psychiatry Res 2018 Nov;269:571-578. [doi: 10.1016/j.psychres.2018.08.107] [Medline: 30199699]
43. Gosmanova EO, Molnar MZ, Alrifai A, et al. Impact of non-adherence on renal and cardiovascular outcomes in US veterans. Am J Nephrol 2015;42(2):151-157. [doi: 10.1159/000440685] [Medline: 26398170]
44. Çakmur H. Frailty among elderly adults in a rural area of Turkey. Med Sci Monit 2015 Apr 30;21:1232-1242. [doi: 10.12659/MSM.893400] [Medline: 25925800]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BN: Bayesian network
CESD-10: 10-item Center for Epidemiologic Studies Depression Scale
CHARLS: China Health and Retirement Longitudinal Study
COPD: chronic obstructive pulmonary disease
DCA: decision curve analysis
FEV1%: forced expiratory volume in 1 second as a percentage of the predicted value

XSL•FO
RenderX

HbA_{1c}: hemoglobin A_{1c}
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
LASSO: Least Absolute Shrinkage and Selection Operator
LR: logistic regression
MCV: mean corpuscular volume
OR: odds ratio
PD: pulmonary dysfunction
PDDC: pulmonary dysfunction–depression comorbidity
SHAP: Shapley Additive Explanations
XGBoost: Extreme Gradient Boosting

Edited by A Benis; submitted 22.May.2025; peer-reviewed by S Yang, W Yan; revised version received 08.Jan.2026; accepted 17.Feb.2026; published 09.Apr.2026.

Please cite as:
Cheng Q, Cao Q, Teng W, Dai R, Jia Q, Bai X, Li Q, Wu Y, Huang Y
Comparative Performance of 3 Analytical Models in Identifying Associated Factors of Pulmonary Dysfunction–Depression Comorbidity: China Health and Retirement Longitudinal Study–Based Nationwide Cross-Sectional Study
JMIR Med Inform 2026;14:e77940
URL: https://medinform.jmir.org/2026/1/e77940
doi:10.2196/77940

© Qinglin Cheng, Qiancheng Cao, Weilin Teng, Ruoqi Dai, Qingjun Jia, Xuexin Bai, Qingchun Li, Yifei Wu, Yinyan Huang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Code-Based Versus AutoML Methods for Pill Recognition in Clinical Settings: Comparative Performance Study

Amir Reza Ashraf[1], PharmD, PhD; Richárd Rádli[2], MSc; Zsolt Vörösházi[2], PhD; András Fittler[1], PharmD, PhD

[1]Department of Pharmaceutics, Faculty of Pharmacy, University of Pécs, Rókus utca 4, Pécs, Hungary

[2]Image Processing Research Laboratory, University of Pannonia, Veszprém, Hungary

Corresponding Author:
Amir Reza Ashraf, PharmD, PhD
Department of Pharmaceutics, Faculty of Pharmacy, University of Pécs, Rókus utca 4, Pécs, Hungary

Abstract

Background: Visual identification and verification of medications during dispensing and administration are prone to human error, particularly in high-pressure and high-volume clinical settings. Misidentification can lead to medication errors, posing risks to patient safety and placing a burden on health care systems. Recent advances in computer vision and object detection offer promising solutions for automated solid oral dosage form (pill) recognition. However, comprehensive studies comparing code-based and no-code (automated machine learning [AutoML]) approaches for pill recognition are lacking.

Objective: This study aimed to evaluate and compare the performance, cost, usability, and deployment feasibility of pill recognition models developed with Ultralytics YOLO11 and 3 cloud-based AutoML platforms (Amazon Rekognition Custom Labels, Google Vertex artificial intelligence [AI] AutoML Vision, and Microsoft Azure Custom Vision) using multiple datasets, including real-world clinical images.

Methods: Five training subsets of increasing size (1230, 3450, 7380, 14,400, and 26,880 images) from 30 commonly dispensed medications were used to train models on YOLO11 and 3 AutoML platforms. Models were evaluated on 6 datasets from different environments: clinical images from 3 hospitals, a verification dataset, a laboratory dataset, and an exhaustive testing set. Performance metrics, including accuracy, precision, recall, and mean average precision, were calculated. We evaluated the impact of training data size on performance and benchmarked training time, platform costs, and limitations.

Results: No single platform dominated across all test environments. On the verification dataset (optimal conditions), accuracy ranged from 80.83% (YOLO11) to 91.60% (Google Vertex AI) when trained with the full training dataset. YOLO11 showed consistent performance improvement with increasing training data (accuracy: 63.06%-80.83%) and achieved near-perfect precision and mean average precision scores (0.95 - 1.00). Google Vertex AI reached above 90% accuracy on 3 training subsets but showed unpredictable declines. Amazon Rekognition maintained near-perfect precision (0.92 - 1.00) but had the highest false negative rates (up to 0.74), missing many pills. Custom Vision demonstrated steady performance improvements (77.08%-85.62% accuracy) but lagged behind other AutoML platforms, probably due to its older YOLOv2-based architecture. On clinical datasets, accuracy fluctuated (20.62%-90%) depending on the dataset and platform. Training costs and time varied: YOLO11 (open-source), Microsoft Azure (US $9.50-US $28.60, allowed user-predefined training duration), Google Vertex AI (US $69.30 with consistent 2.5 - 3-hour training times), and Amazon Rekognition (US $5.43-US $43.89 with size-dependent training time scaling, reaching nearly 40 hours on the full 26,880-image dataset).

Conclusions: Each platform offers distinct advantages and trade-offs: YOLO11 provides the highest flexibility and lowest platform costs but requires technical expertise, while AutoML platforms can offer high performance at a higher cost but with limited user control, introducing unpredictability. The performance variations demonstrate that successful clinical deployment requires careful platform selection based on specific performance requirements, budget constraints, and available technical resources, followed by rigorous validation using real-world, representative data to ensure patient safety in clinical workflows.

(JMIR Med Inform 2026;14:e79160) doi:10.2196/79160

KEYWORDS

pill recognition; clinical pharmacy; medication safety; medication errors; clinical decision support; You Only Look Once; YOLO11; automated machine learning; AutoML; computer vision; object detection; deep learning

XSL•FO
RenderX

Introduction

Background

Medication errors are preventable failures in medical care that can potentially lead to patient harm and increased health care costs, with a global financial burden estimated at US $42 billion annually [1]. These errors are responsible for an estimated 7000 - 9000 deaths per year in the United States [1] and contribute to 1708 deaths in England [2], highlighting the urgent need for improved medication safety strategies. Medication errors can occur at various stages of therapy, from prescribing to dispensing and administration of medications [3], and are among the most common causes of death [4]. Medication therapy management in hospitals, clinics, and inpatient health care settings is often managed under the supervision of a clinical pharmacist, who is responsible for monitoring patients' treatment process and identifying any discrepancies [5]. During inpatient care, medication reconciliation, that is, the process in which pharmacists compare the medications a patient is currently taking or should be taking with newly ordered therapies, often requires visually distinguishing between numerous solid oral dosage forms (eg, pills, tablets, and capsules, often referred to collectively as "pills") based on the patient's therapy regimen. This process can be challenging because these products often show only subtle differences in their physical characteristics [6]. Therefore, even the most experienced professionals could occasionally make mistakes, as visual identification and verification are prone to human error, especially in high-pressure and high-volume environments such as hospitals and clinics [7]. These challenges and limitations have led to investments in automation technologies for unit-dose dispensing to improve hospital pharmacy services [8,9] and the exploration of leveraging novel techniques, such as object detection, to assist pharmacists with verification tasks and ultimately reduce dispensing errors [6]. However, it is important to note the specific clinical context for such tools. Unlike medication reconciliation at admission, where a patient may present a disorganized collection of loose pills for identification, during hospital pharmacy manual dispensing processes, staff typically use organized unit doses and dosette boxes. In this context, the primary challenge is usually not identifying a mystery pill but verifying that the pill in a specific compartment matches the prescribed therapy. Automated pill recognition systems could support this verification process, as well as additional workflows, including detection of dispensing errors before medications reach patients and identification of medications returned to the hospital pharmacy unit.

Recent advances in computer vision and object detection offer promising solutions for image analysis, including high-performance object detection models based on the You Only Look Once (YOLO) architecture, which allows fast and accurate localization and classification in a single step [10,11]. YOLO has proven to be an attractive choice for a variety of medical applications [11], including pill recognition [12]. YOLO has demonstrated strong performance in code-based pill recognition systems, particularly for preprocessing segmented images within multistream, 2-phase neural models using metric embedding [13]. One of the latest iterations in the Ultralytics YOLO series is YOLO11, released in September 2024. This version improved upon previous models by implementing a new backbone architecture that enhances feature extraction (identifying key patterns such as edges, textures, or shapes from raw image data) with improved accuracy and processing speed, making it suitable for complex visual recognition tasks [14]. These advancements create new opportunities for developing clinical decision support systems that assist health care professionals in the visual identification of pills.

In parallel with the evolution of code-based models, which provide flexibility and customization but require strong programming skills and longer development time, low-code or code-free automated machine learning (AutoML) platforms have emerged. These platforms enable relatively simple development and testing of specialized object detection models using easy-to-use cloud-based platforms [15,16].

Rationale

Although both traditional code-based models and cloud-based AutoML platforms have shown promise in medical image analysis [17], there is a lack of comprehensive, direct comparative studies evaluating the performance of these approaches in the context of pill recognition. Previous studies have commonly focused on controlled environments that do not reflect typical usage scenarios and conditions. The open-source, code-based deep learning object detection YOLO model and code-free, cloud-based AutoML platforms represent fundamentally different development concepts, each with its own associated advantages and limitations. This justifies a comprehensive comparison of these solutions to determine robustness, scalability, and implementation complexity in practical environments. Limited published, peer-reviewed studies compare the performance of traditional code-based and code-free pill recognition models using images captured in real-world clinical environments, which often differ substantially from standardized reference images. To our knowledge, this is the first study to directly compare these approaches using clinical data under such specific conditions.

Objectives

This study aimed to evaluate and compare the effectiveness of pill recognition models developed using YOLO11 with cloud-based AutoML platforms provided by Amazon Rekognition Custom Labels, Google Vertex artificial intelligence (AI) AutoML Vision, and Microsoft Azure Custom Vision. Factors, such as cost, ease of use, and deployment feasibility, as well as key performance metrics including accuracy, precision, recall, F_1-score, mean average precision (mAP), overall error rates (OERs), and false negative rates (FNRs), were evaluated on different test sets, including real-world images captured by clinical pharmacists.

Code-Based and Code-Free Object Detection

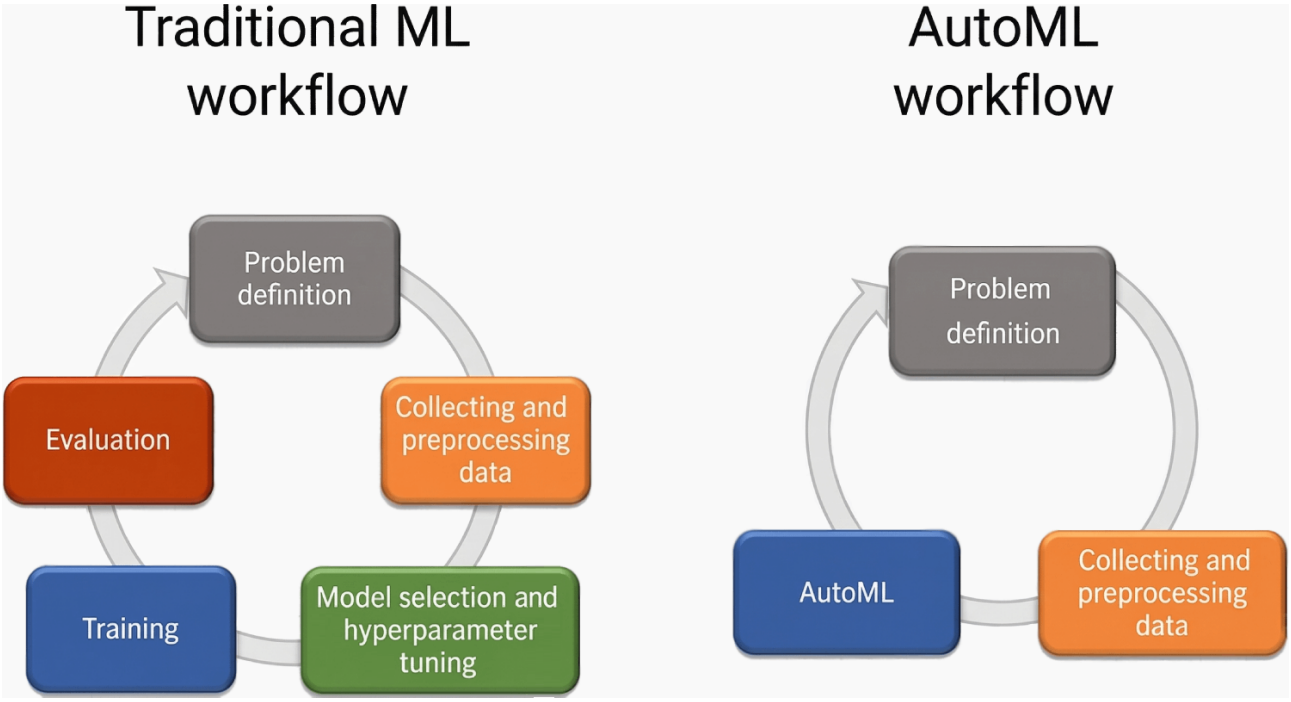
Traditional code-based deep learning frameworks, such as YOLO11, require programming and machine learning (ML) expertise but offer full control over model architecture and training parameters, whereas code-free development and training using cloud-based AutoML platforms provide simplified workflows with limited customization options. The YOLO

XSL•FO
RenderX

family of deep-learning–based object detectors uses a single-stage approach, which means that the entire input image is processed in a single forward pass, predicting bounding boxes and class probabilities simultaneously [18]. This allows YOLO models to balance inference speed and detection accuracy, making them one of the most widely used object detection algorithms. In our experiments, we used the YOLO11 real-time object detection model, which was designed to address a wide spectrum of application requirements and supports a wide range of computer vision tasks, including object detection [14,18]. A detailed description of the YOLO11 model structure is provided in Multimedia Appendix 1.

In a traditional deep learning ecosystem, model development involves several manual and iterative processes, such as data collection and preprocessing, feature engineering, model selection, hyperparameter tuning, training, and validation, as shown in Figure 1. This approach requires a high level of expertise and intensive experimentation to fine-tune the model's performance.

Figure 1. Traditional machine learning (ML) vs automated machine learning (AutoML) workflows.



Cloud-based, no-code AutoML platforms have emerged as suitable alternatives to traditional, often code-intensive model development approaches. These platforms automate most of the workflows involved in managing ML pipelines, thereby lowering the barrier to entry and making model development accessible even to those without programming skills or data science expertise [19]. In addition to enhancing accessibility, these platforms accelerate the model development lifecycle by automating repetitive and time-consuming tasks, allowing rapid prototyping, while reducing complexity by handling the underlying tasks associated with managing the computational infrastructure and ecosystem required for model development [19]. However, AutoML has inherent disadvantages. While AutoML democratizes model development by minimizing human effort, it requires significant computational resources, which can lead to expensive cloud service costs, particularly when developing models iteratively using large datasets. AutoML users typically have very limited control over the choice of base models, underlying training algorithms, and hyperparameter adjustments. In addition, AutoML platforms lack transparency in model architecture and hyperparameters and may raise privacy concerns [20]. Therefore, the choice of the optimal platform depends on a variety of factors.

This study aimed to provide empirical evidence to guide such platform selection decisions for pill recognition model development intended for deployment in clinical and hospital pharmacy settings, where accuracy and reliability are of utmost importance. We evaluated AutoML solutions offered by Amazon Web Services (AWS), Google Cloud, and Microsoft Azure, 3 of the most popular cloud-based providers for object detection, alongside code-based models trained with a traditional workflow using YOLO11.

Microsoft offers a wide range of AI and ML services on its Azure platform, including Custom Vision, a more specialized service focused on building, training, and deploying custom image classification and object detection models using a no-code or low-code approach [21]. It is accessible through a dedicated, code-free web interface and software development kits (SDKs) requiring minimal coding experience. Vertex AI AutoML Vision is part of Google Cloud's unified Vertex AI platform. Vertex AI allows users to train, deploy, and fine-tune a wide range of models with supervised learning, data- and application-specific training strategies, and tools to manage the entire lifecycle of ML [22]. Vertex AI offers choices for a manual approach that provides full control over the training pipeline using traditional coding development, as well as AutoML for code-free model development, which is accessible via the Google Cloud Console web interface and through SDKs compatible with popular open-source libraries such as TensorFlow (Google) and PyTorch (Meta AI) [23]. Amazon Rekognition Custom Labels is a feature

XSL•FO
RenderX

within AWS's broader Rekognition service that allows users to train custom object detection models using AutoML [24]. Although Amazon Rekognition follows a similar workflow to other platforms, including the option of manual labeling, it imposes stricter restrictions on the ML workflow.

Methods

Overview

This section details the methodologies used for data collection, preprocessing, and coded model development. We compared 2 fundamentally different approaches to model development and training: traditional code-based deep learning using YOLO11, and code-free development using 3 major cloud-based AutoML platforms (Amazon Rekognition Custom Labels, Vertex AI AutoML Vision, and Custom Vision). We performed 5 training runs per platform using 5 progressively larger subsets of our training images. The evaluation metrics used for benchmarking and comparative analysis are explained in detail in Multimedia Appendix 2.

Data Collection and Preprocessing

For model training and evaluation, we used a tailor-made proprietary dataset focusing on 30 commonly dispensed solid oral dosage forms in Hungary (Figure 2). Selection of medications and curation of the dataset followed a strict protocol described in our previous publication [15].

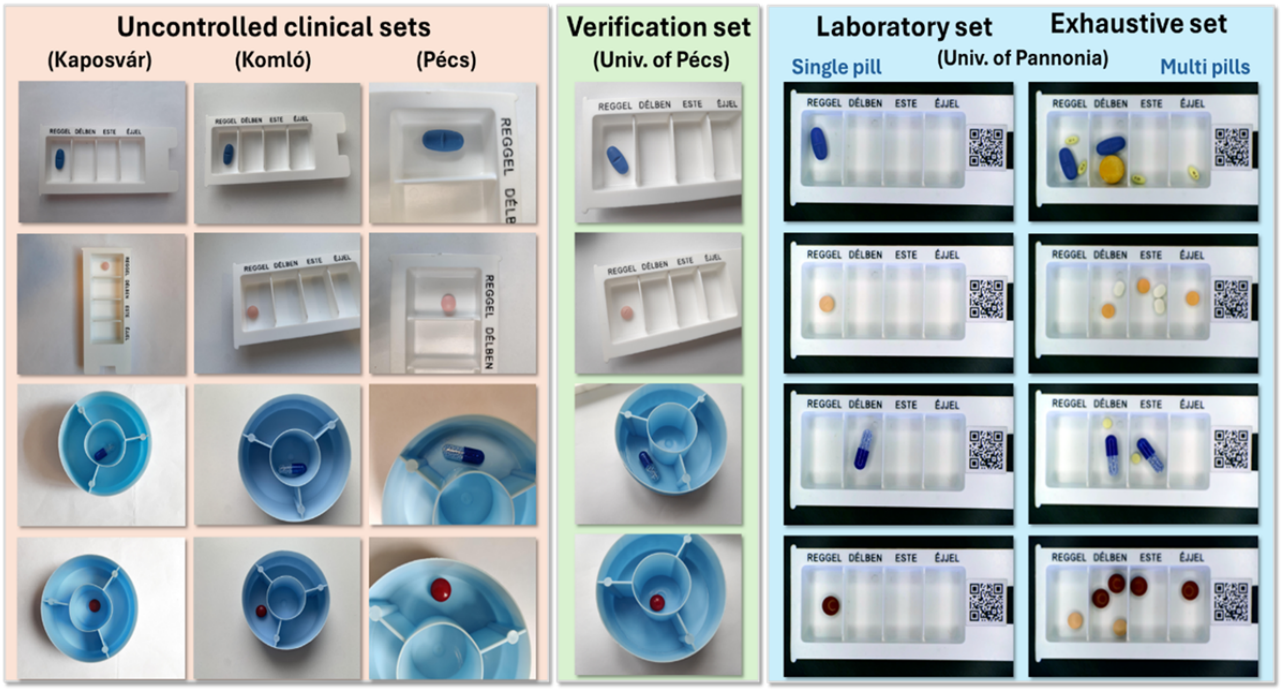
For testing datasets, images were captured with a single pill placed in blue-and-white dosette boxes routinely used for manual dispensing in Hungarian clinical settings, except for one dataset prepared for exhaustive performance evaluation, which contained multiple pills per image (refer to Figure 3).

XSL•FO
RenderX

Figure 2. The 30 medications selected for this study.

XSL•FO
RenderX

Figure 3. Illustration of testing datasets.



Training images had a resolution of 1200×900 pixels, while test images were at native device resolution (4032×3024 and 4000×3000 pixels) to better simulate real-world conditions where input images vary in size. All images were captured in JPEG format with consistent quality settings.

The study's datasets were organized as follows:

- The "full training dataset" comprised 26,880 images in total, provided by the University of Pécs, with 896 images per class for the 30 selected medications. To evaluate the impact of training data size on the model performance, 5 subsets were created from these images using a logarithmic binning strategy, each with a progressively increasing number of images:
  - Subset 1: 1230 images (41 images/class)
  - Subset 2: 3450 images (115 images/class)
  - Subset 3: 7380 images (246 images/class)
  - Subset 4: 14,400 images (480 images/class)
  - Subset 5: 26,880 images (full training dataset, 896 images/class)
- A "verification dataset" containing 1440 images (48/class) was created at the University of Pécs using the same location, conditions, and procedures as the training dataset images.
- Three "uncontrolled clinical test sets (UCTSs)" of 480 images each (16/class) were captured by clinical pharmacists at Kaposvár, Komló, and Pécs hospitals under authentic clinical conditions to assess model performance in uncontrolled health care environments. UCTS images were collected using purposive sampling to ensure balanced representation across all medication classes (16 images/class at each hospital site). This design enables fair per-class comparison of model performance, avoiding bias toward majority classes seen with imbalanced test sets. Similar balanced sampling approaches are used in established pill recognition benchmarks, such as the CURE dataset [25] and the National Library of Medicine Pill Image Recognition Challenge [26]. To create the dataset, clinical pharmacists photographed medications within their routine workspaces using the dosette boxes and imaging equipment available in their daily practice. Class identity recorded during image acquisition served as the ground truth label during subsequent evaluations. This protocol preserved authentic environmental and human variables affecting input image quality, including camera variability, lighting conditions, and operator technique, while maintaining balanced class representation for fair platform comparison. No image quality filtering or manipulation was applied; all captured images were included regardless of quality to ensure the dataset accurately reflected authentic real-world variability in clinical settings.
- Two additional test datasets were created in a controlled laboratory environment at the Image Processing Laboratory, University of Pannonia. The first dataset, referred to as "laboratory test set," contained 120 images with single pills per image, while the second dataset included 120 multipill images called "exhaustive test set" intended for exhaustive model testing. Medications were arranged in a white dosette box, and controllable top-mounted lighting was used to evenly illuminate the pills. Images were captured with a fixed camera position at 3840×2160 resolution, with a slight reduction in resolution following image undistortion.

YOLO11 Implementation

For our experiments, we selected the medium-sized pretrained model (YOLO11-m), which balances speed, accuracy, and training time. All hyperparameters remained at their default values, except for the early stopping patience parameter, which was set to 15 epochs to terminate training if the validation loss did not improve within 15 consecutive epochs. Standard data augmentation techniques (eg, rotation, brightness adjustment,

XSL•FO
RenderX

and so on) were not applied during training to ensure uniform conditions across platforms, as cloud-based AutoML platforms typically apply such augmentations automatically, with limited or no user control. All experiments ran for 100 epochs with a batch size of 20, using an NVIDIA Quadro RTX 5000 graphics processing unit (GPU) with 16 GB video random-access memory for both training and evaluation.

Cloud-Based AutoML Platform Implementation

Microsoft Azure Custom Vision

Custom Vision offers a streamlined workflow for training object detection models. Users must create Custom Vision training and prediction resources in their Azure subscription via the Azure portal. New object detection projects can be created through the Custom Vision portal or via SDK, which offers a choice of predefined domains optimized for specific scenarios and edge or mobile deployment. Training images can be uploaded directly via the web portal or SDK, with an integrated tagging interface for drawing bounding boxes and assigning labels. Custom Vision requires a minimum of 15 images per label for training, but recommends 50 or more images per label for better performance. Users can select either Quick training (completed in minutes) or Advanced training (with extended computation budget) [27]. The platform automatically selects the optimal base model, training, tuning, and augmentation settings based on uploaded data and the selected project domain. Each training run generates a new model iteration that can be published to a prediction endpoint accessible via Representational State Transfer application programming interfaces (APIs) or SDKs.

Google Vertex AI AutoML

Vertex AI requires an active Google Cloud project with images stored in Google Cloud Storage. The platform accepts various image formats, with data imported either from local machines or directly from Google Cloud. Annotations must be provided in comma-separated values or JSON Lines (JSONL) formats containing bounding box coordinates and corresponding labels. Vertex AI requires at least 10 annotated images per label for training and recommends providing 1000 or more for optimal model performance [28]. For unannotated datasets, users can apply annotations through the Google Cloud Console interface. Without user-specified splitting, Vertex AI automatically partitions datasets using an 80% - 10% - 10% split for training, validation, and testing [29]. Since our original XML annotations were incompatible with Vertex AI's input requirements, we converted them to a single JSONL file, with each line representing an annotated instance. We created the dataset by selecting object detection as the project objective and linking the JSONL file from a cloud storage bucket. Vertex AI's custom dataset splitting functionality was used to create appropriate training, validation, and testing sets. For optimization, we used the high-accuracy training option within Vertex AI. After performing initial experiments, we opted for a specific high-accuracy training configuration that significantly outperformed other available options.

Amazon Rekognition Custom Labels

Rekognition requires prelabeled data in Amazon SageMaker Ground Truth manifest format, a JSONL structure with pixel-based bounding box coordinates [30]. This format has a different internal structure compared to Vertex AI's requirements, necessitating the conversion of our annotations using Python scripts. Data management in Rekognition offers moderate flexibility; however, Amazon S3 is the mandatory cloud solution for image storage. If no test dataset is provided, Rekognition automatically splits the training data to create a test dataset using an 80% - 20% split [31].

Model training can be initiated through the AWS Console or the CreateProjectVersion API. However, Rekognition provides no hyperparameter control, limiting customization during training. Each successful training run generates a new project version with an associated Amazon Resource Name. Performance metrics, including F_1-score, precision, recall, and mAP, are available through the API and Rekognition.

Evaluation Metrics

Object detection evaluation differs from standard classification because it requires assessment of localization and classification, meaning models must simultaneously locate objects within an image and correctly identify each object. Therefore, metrics that ignore spatial accuracy of predictions, such as simple classification accuracy, precision, and recall without intersection over union (IoU) thresholds, or classification-only F_1-scores, are insufficient for rigorous model evaluation.

To comprehensively assess our pill recognition models for both localization precision and classification accuracy, we used metrics incorporating IoU to enable fair comparison across fundamentally different model architectures and development approaches.

IoU quantifies the spatial accuracy of object detection by measuring the overlap between predicted and ground-truth bounding boxes, calculated as the ratio of overlap area to union area between boxes. Higher values indicate better localization accuracy, with a threshold (typically 0.5) determining whether a detection is considered correct.

In object detection, the accuracy metric is adapted to account for both classification and localization performance. A prediction is considered correct only if it has the correct class label and the IoU with the ground-truth bounding box exceeds a specified threshold.

Precision and recall incorporate IoU thresholds for object detection. Precision measures the percentage of correct detections among all predictions, indicating how reliable the model's positive predictions are. Recall measures the percentage of ground-truth objects correctly detected, reflecting the model's ability to identify all relevant instances.

Average precision represents the area under the precision-recall curve, providing a comprehensive performance measure across different confidence thresholds for each class. mAP averages average precision values across all classes, providing an overall performance measure. Common variants include mAP@0.50 (using an IoU threshold of 0.5) and mAP@0.50 - 0.95

XSL•FO
RenderX

(averaging across multiple IoU thresholds from 0.5 to 0.95 in 0.05 increments), which provides a more rigorous evaluation.

Beyond standard performance metrics, a critical evaluation of model suitability for clinical deployment necessitates an analysis of error profiles. To understand the failures impacting patient safety, we analyzed OERs and FNRs for each model. OER and accuracy are inversely related; they measure the model's overall tendency to misidentify medications through missed detections or incorrect classifications. FNR specifically quantifies the proportion of pills that models fail to detect, a critical safety metric in clinical settings. A detailed description of evaluation metrics, mathematical definitions, and computational procedures is provided in Multimedia Appendix 2.

Analytical Approach

This study used a comparative benchmarking design to characterize platform performance across diverse deployment scenarios rather than hypothesis testing. We selected standard object detection metrics following established conventions in computer vision and pill recognition research to enable direct comparison with other published literature. We additionally reported OER and FNR due to their clinical relevance for patient safety. CIs and formal statistical comparisons across hospital sites were not performed, as the evaluated cloud-based AutoML platforms have major inherent limitations and do not consistently provide the data needed for these calculations.

Used Medicines

Our datasets included the 30 most commonly dispensed medications from 3 participating clinical centers [15]. The medications represent various therapeutic areas, including cardiovascular agents (eg, bisoprolol and perindopril), antibiotics (amoxicillin/clavulanic acid), analgesics (naproxen and tramadol), psychiatric medications (alprazolam and quetiapine), and essential vitamins and supplements.

These medications cover a diverse range of solid oral dosage forms, including film-coated and uncoated tablets (the most prevalent forms), enteric-coated tablets, hard and soft gelatin capsules, and one chewable tablet. A complete medication list with active pharmaceutical ingredients, dosage forms, distinctive features, colors, and shapes is provided in Multimedia Appendix 3.

Ethical Considerations

This study did not involve human participants, and ethics approval was therefore not required. No personally identifiable information was collected or processed. No informed consent was required, as no human participants were involved in this study.

Results

Overall Detection Performance

Our evaluation of pill recognition models across multiple test datasets revealed notable performance differences, with no single platform consistently dominating all test scenarios. We trained models using 5 progressively larger training subsets (from 1230 to 26,880 images) and evaluated them using standard metrics: accuracy, precision, recall, OER, FNR, F_1-score, mAP@0.50, and mAP@0.50 - 0.95 to comprehensively assess platform performance, strengths, and limitations. Table 1 presents the results of models trained on the full dataset of 26,880 images, with comprehensive results of all training runs and subsequent evaluations provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Summary of platform performance across all testing datasets (trained with the full training dataset of 26,880 images).

Platform and dataset	Accuracy (%)	Precision	Recall	OER[a]	FNR[b]	F_1-score	mAP@0[c].50	mAP@0.50 - 0.95
Ultralytics YOLO11								
Kaposvár	80.63	0.98	0.92	0.19	0.08	0.95	0.99	0.62
Komló	78.96	1.00	0.91	0.21	0.09	0.95	1.00	0.72
Pécs	64.58	0.99	0.91	0.35	0.09	0.95	0.99	0.76
Verification	80.83	1.00	0.96	0.19	0.04	0.98	1.00	0.89
Laboratory	40.00	1.00	0.87	0.60	0.13	0.93	0.96	0.48
Exhaustive	44.10	0.99	0.81	0.56	0.19	0.89	0.95	0.40
Google Vertex AI[d]								
Kaposvár	71.04	0.96	0.90	0.29	0.10	0.93	0.97	0.53
Komló	90.00	1.00	0.96	0.10	0.04	0.98	1.00	0.58
Pécs	65.42	1.00	0.81	0.35	0.19	0.90	0.97	0.67
Verification	91.60	1.00	0.98	0.08	0.02	0.99	1.00	0.79
Laboratory	67.50	1.00	0.86	0.33	0.14	0.92	1.00	0.54
Exhaustive	52.79	1.00	0.66	0.47	0.34	0.79	0.99	0.45
Microsoft Azure Custom Vision								
Kaposvár	62.71	0.98	0.80	0.37	0.20	0.88	0.99	0.58
Komló	77.71	0.99	0.92	0.22	0.08	0.95	1.00	0.67
Pécs	56.04	0.98	0.81	0.44	0.19	0.89	0.99	0.65
Verification	85.62	1.00	0.94	0.14	0.06	0.97	1.00	0.70
Laboratory	51.67	0.97	0.78	0.48	0.22	0.86	0.98	0.59
Exhaustive	33.61	0.92	0.54	0.66	0.46	0.68	0.77	0.33
Amazon Rekognition								
Kaposvár	70.63	1.00	0.76	0.29	0.24	0.86	0.76	0.45
Komló	77.71	1.00	0.80	0.22	0.20	0.89	0.79	0.53
Pécs	20.62	0.92	0.26	0.79	0.74	0.40	0.26	0.19
Verification	84.72	1.00	0.85	0.15	0.15	0.92	0.85	0.71
Laboratory	43.33	1.00	0.80	0.57	0.20	0.89	1.00	0.77
Exhaustive	41.97	0.99	0.76	0.58	0.24	0.86	0.98	0.67

[a]OER: overall error rate.

[b]FNR: false negative rate.

[c]mAP: mean average precision.

[d]AI: artificial intelligence.

Results are organized into 3 categories: UCTS (representing real-world hospital conditions), verification dataset (optimal imaging conditions replicating training images), and laboratory-controlled datasets (standardized research settings).

On the UCTS images from Kaposvár hospital, models trained with the full training dataset of 26,880 images exhibited varied performance (Table 1). YOLO11 demonstrated superiority, achieving the highest scores in accuracy (80.63%), recall (0.92), F_1-score (0.95), and mAP metrics. Amazon Rekognition achieved the highest precision (1) but also showed the highest FNR, indicating its conservative detection strategy. Vertex AI ranked second with 71.04% accuracy. Microsoft Azure did not achieve top results in any instance on this dataset, but it maintained consistent performance, with accuracy ranging from 56.87% to 62.92% across training set sizes (Multimedia Appendix 4).

On the UCTS images from Komló hospital, competition was balanced between YOLO11, Custom Vision, and Amazon Rekognition, while Vertex AI showed considerably higher performance compared to the others. Across all training configurations, YOLO11 dominated mAP metrics, achieving a perfect 1.0 mAP@0.50 with 14,400 and 26,880 training images. Vertex AI excelled in accuracy, reaching 90.62% with 7380 images and 90% with the full training subset of 26,880 images.

XSL•FO
RenderX

Custom Vision improved steadily from 69.58% to 77.71% accuracy with increasing training data, while Amazon Rekognition maintained perfect precision (1.0) across all 5 training scenarios while achieving 77.71% accuracy.

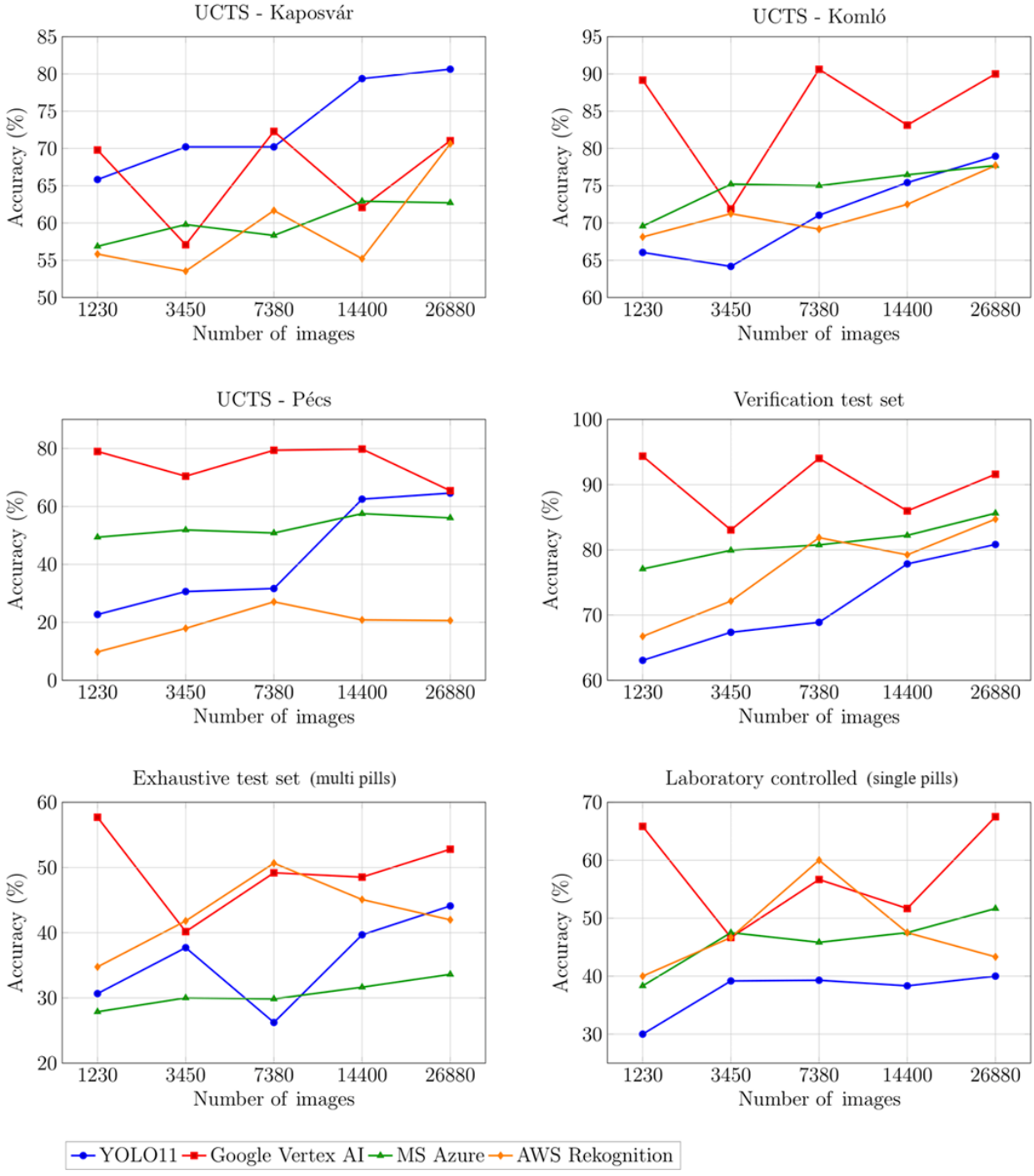
On the UCTS images from Pécs hospital, Vertex AI outperformed other platforms despite varying results, achieving 79.79% accuracy with the training subset of 14,400 images but 65.42% with the full dataset. The other AutoML solutions performed poorly: Microsoft Azure achieved 57.5% accuracy with the training subset of 14,400 images and 56.04% with the full training dataset, while Amazon Rekognition reached only as high as 27.08% accuracy with the training subset of 7380 images on this dataset, which fell to 20.62% when trained with the full training dataset (Multimedia Appendix 4). YOLO11 also showed strong results, particularly for the more challenging mAP@0.50‐0.95 threshold metric (0.76). This dataset demonstrated the challenging nature of certain clinical environments, where lighting conditions, camera positioning, and other external factors significantly impact model performance.

On the verification dataset from the University of Pécs, Vertex AI excelled across all metrics, achieving 91.60% accuracy, perfect 1.0 precision and mAP@0.50, and 0.98 recall with the full training dataset. It only slightly underperformed YOLO11 in the mAP@0.50‐0.95 metric (0.79 vs 0.89). Microsoft Azure achieved 85.62% accuracy, 1.0 precision, and 1.0 mAP@0.50, while Amazon Rekognition reached 84.72% accuracy with perfect precision (1.0) but lower recall (0.85). YOLO11 showed linear performance improvement with increasing training dataset size, reaching 80.83% accuracy with perfect precision (1.0) and a high recall of 0.96. This dataset demonstrated the potential upper bound on performance when testing conditions are well-matched to training data.

The laboratory-controlled (single pills) and the exhaustive (multipills) test sets from the University of Pannonia showed varying performance across different platforms under optimal imaging conditions in a fully controlled environment. When analyzing the exhaustive dataset, Vertex AI achieved the highest accuracy on the full training dataset (52.79%) with excellent mAP@0.50 (0.99), followed by YOLO11 (44.10%, 0.95), Amazon Rekognition (41.97%, 0.98), and Microsoft Azure (33.61%, 0.77). The lower accuracy scores in the exhaustive dataset reflect the increased complexity of detecting multiple objects within an image, with potential occlusions and varying orientations. On the laboratory dataset, Vertex AI achieved 67.50% accuracy with perfect 1.0 precision and mAP@0.50, followed by Amazon Rekognition, mainly due to its robust mAP metrics, which maintained good performance across other metrics, although still falling short of Vertex AI's performance. Figure 4 visualizes these trends across all datasets and training set sizes, illustrating nonlinear performance patterns and platform-specific variations.

XSL•FO
RenderX

Figure 4. Accuracy of object detection models across datasets. AI: artificial intelligence; AWS: Amazon Web Services; UCTS: uncontrolled clinical test set.



Influence of Training Dataset Size on Model Performance

Increasing training images generally improved model performance across all metrics, although improvements were nonlinear with platform-specific patterns. Table 2 illustrates these effects using the verification dataset, which, despite being captured under optimal conditions similar to training data, serves as a neutral baseline for comparing platform responses to increased training data because all models were trained and tested using the same datasets. Overall, YOLO11 and Custom Vision showed predictable linear improvement, with YOLO11's accuracy increasing from 63.06% to 80.83% and Microsoft Azure's from 77.08% to 85.62% as training data increased from 1230 to 26,880 images. Conversely, Amazon Rekognition and Vertex AI showed variable patterns with several fluctuations, suggesting dynamic model selection based on dataset characteristics, resulting in performance variations that do not follow the expected logarithmic improvement pattern with an eventual plateau typically seen in ML models. Notably, Vertex AI achieved near-optimal performance (94.37% accuracy, 1.0 precision, and 1.0 recall) with the smallest training subset at just 1230 images, though challenging clinical datasets required the full training set for acceptable performance.

XSL•FO
RenderX

Table . Effect of training dataset size on model performance on the verification dataset.

Platform and training subset	Accuracy	Precision	Recall	OER[a]	FNR[b]	F_1-score	mAP@0[c].50
Ultralytics YOLO11							
1230 images	63.06%	1.00	0.85	0.37	0.15	0.92	1.00
3450 images	67.36%	1.00	0.86	0.33	0.14	0.93	1.00
7380 images	68.89%	1.00	0.91	0.31	0.09	0.95	1.00
14,400 images	77.85%	1.00	0.96	0.22	0.04	0.98	0.95
26,880 images	80.83%	1.00	0.96	0.19	0.04	0.98	1.00
Google Vertex AI[d]							
1230 images	94.37%	1.00	1.00	0.06	0.00	1.00	1.00
3450 images	83.06%	1.00	0.98	0.17	0.02	0.99	1.00
7380 images	94.03%	1.00	0.99	0.06	0.01	0.99	1.00
14,400 images	85.97%	1.00	0.98	0.14	0.02	0.99	1.00
26,880 images	91.60%	1.00	0.98	0.08	0.02	0.99	1.00
Microsoft Azure Custom Vision							
1230 images	77.08%	1.00	0.91	0.23	0.09	0.95	1.00
3450 images	79.93%	1.00	0.92	0.20	0.08	0.96	1.00
7380 images	80.76%	1.00	0.90	0.19	0.10	0.95	1.00
14,400 images	82.22%	1.00	0.93	0.18	0.07	0.96	1.00
26,880 images	85.62%	1.00	0.94	0.14	0.06	0.97	1.00
Amazon Rekognition							
1230 images	66.74%	1.00	0.68	0.33	0.32	0.81	0.72
3450 images	72.15%	1.00	0.75	0.28	0.25	0.86	0.79
7380 images	81.87%	1.00	0.84	0.18	0.16	0.91	0.84
14,400 images	79.24%	1.00	0.81	0.21	0.19	0.89	0.82
26,880 images	84.72%	1.00	0.85	0.15	0.15	0.92	0.85

[a]OER: overall error rate.

[b]FNR: false negative rate.

[c]mAP: mean average precision.

[d]AI: artificial intelligence.

Cost, Usability, and Limitations of AutoML Platforms

Several important pricing and usability differences exist between platforms. Vertex AI and Amazon Rekognition Custom Labels charge continuously for endpoint availability, accumulating costs in minute increments even during idle periods. In contrast, Custom Vision charges only per request, eliminating costs during idle periods. Vertex AI maintains consistent flat-rate pricing regardless of the dataset size, charging US $3.465 per hour for training and US $2.002 per hour for deployment and online prediction [32]. Amazon Rekognition Custom Labels charges US $1 per hour for training and US $4 per hour for inference, but scales training time and resources with dataset size, resulting in proportionally higher costs for larger datasets [33]. Custom Vision charges US $10 per hour for training, but allows users to specify computation time budgets, providing better cost control. Inference costs US $2 per 1000 individual requests [34]. While Google and Amazon have regional pricing variations, Azure maintains uniform US dollar pricing that is directly converted to local currencies. Each platform offers limited free tiers: Azure provides 1 hour of free training per month plus 10,000 predictions; Google provides US $300 in free credit for new customers and includes 30 minutes of free online predictions; Amazon offers 1 free inference hour per month during its 12-month free tier. YOLO11, being open-source, has no platform costs but requires local computational resources.

Training times varied considerably. YOLO11 training times scaled linearly with dataset size, from 38 minutes for the 1230-image training subset to nearly 13 hours for the full dataset. Microsoft Azure completed all training within 3 hours, even when allocated 4 - 5 hours, indicating that the cloud-based AutoML platform determined that additional training was unnecessary. Amazon Rekognition required the longest time, reaching nearly 40 hours for the full dataset. Vertex AI

XSL•FO
RenderX

maintained consistent 2.5 - 3-hour training times regardless of dataset size. Table 3 provides an overview of training times and associated costs for each platform.

Table . Training time and cost comparison across platforms.

Platform and training subset	Training time (hour:minute)	Training cost (US $)
Ultralytics YOLO11		
1230 images	00:38	<1
3450 images	01:41	<1
7380 images	03:37	<1
14,400 images	06:53	<1
26,880 images	12:53	<1
Google Vertex AI[a]		
1230 images	02:38	69.30
3450 images	02:48	69.30
7380 images	02:53	69.30
14,400 images	02:58	69.30
26,880 images	02:59	69.30
Microsoft Azure Custom Vision		
1230 images	00:57	9.50
3450 images	01:03	10.50
7380 images	01:02	10.33
14,400 images	01:37	16.10
26,880 images	02:52	28.60
Amazon Rekognition		
1230 images	04:55	5.43
3450 images	09:28	10.48
7380 images	17:04	18.88
14,400 images	25:18	27.98
26,880 images	39:41	43.89

[a]AI: artificial intelligence.

Performance on Edge Device

We evaluated Microsoft Azure and YOLO11 on a Raspberry Pi 5 using the exhaustive test set. Only these platforms were tested since Microsoft Azure alone provided downloadable Open Neural Network Exchange (ONNX) models for edge deployment. Table 4 shows a significant performance trade-off between edge and cloud-based inference.

Microsoft Azure's performance decreased on-device (25.83% accuracy) compared to cloud inference (33.61% accuracy). YOLO11 maintained better consistency, achieving 43.33% accuracy with ONNX and 44.10% with the native .pt format. Inference times differed substantially: Microsoft Azure ONNX achieved 0.19 seconds per prediction, while YOLO11 required 1.38 seconds per prediction in ONNX format on the Raspberry Pi 5.

XSL•FO
RenderX

Table . Performance metrics comparison for edge deployment.

Platform and weight format[a]	Accuracy	Precision	Recall	F_1-score	mAP[b] @0.50	mAP@0.50 - 0.95	Average prediction time (seconds)
Ultralytics YOLO11							
ONNX[c]	43.33%	1.00	0.85	0.92	0.93	0.44	1.38
PyTorch (Meta AI; .pt)	44.10%	0.99	0.81	0.89	0.95	0.34	0.08
Microsoft Azure							
ONNX	25.83%	0.84	0.57	0.68	0.74	0.29	0.19
API[d] call	33.61%	0.92	0.54	0.68	0.77	0.33	0.83

[a]Models trained with the full dataset of 26,880 training images; metrics calculated at a confidence level of 0.5 and an intersection over union (IoU) threshold of 0.5.

[b]mAP: mean average precision.

[c]ONNX: Open Neural Network Exchange.

[d]API: application programming interface.

Discussion

Principal Findings

Our comprehensive evaluation of pill recognition models using code-based YOLO11 vs major code-free cloud-based AutoML platforms revealed that no single platform consistently outperforms others across all test scenarios, highlighting the importance of matching platform selection to specific deployment conditions and available training resources. The dramatic performance variations observed across clinical sites—for example, ranging from 20.62% accuracy (Amazon Rekognition) to 65.42% accuracy (Vertex AI) on the same clinical hospital dataset when both were trained on the full 26,880-image training dataset—demonstrate that imaging hardware, environmental, and personnel factors have a more tangible impact on model performance than platform choice alone.

Our results indicate that increasing the number of training images generally improves model performance across all metrics, although improvement patterns were not strictly linear and varied by platform, likely due to different underlying architectures and optimization strategies. Fair comparisons are challenging, especially for AutoML frameworks, which dynamically select the most suitable model architecture for each specific dataset and task. Consequently, different model types may be chosen (eg, a convolutional neural network in one instance and a vision transformer in another), even when trained on the same dataset.

Another critical finding is the substantial variation in how platforms balance the precision-recall trade-off across real-world clinical datasets. Amazon Rekognition's high precision across multiple clinical datasets often came at the cost of low recall, with values as low as 0.26 in one clinical dataset. This conservative approach resulted in the platform missing up to 74% of pills in the most challenging clinical dataset. Conversely, YOLO11's balanced precision-recall trade-off suggests suitability for scenarios where both false positives and false negatives carry clinical risk.

YOLO11 demonstrated the most consistent and predictable improvement trajectory with larger training datasets, achieving excellent mAP scores on most datasets, making it ideal for applications requiring precise object localization. The most dramatic improvement was observed on the most challenging clinical dataset, where accuracy increased from 22.71% to 64.58% and recall improved from 0.49 to 0.91. Consistently positive mAP trends across all datasets confirmed YOLO11's ability to effectively use additional training data.

Vertex AI presented more complex patterns. While generally improving with larger training datasets, performance fluctuations, including degradations, occurred in multiple instances with larger subsets. For instance, on the same clinical dataset, its accuracy dropped from 69.79% (1230-image training subset) to 57.08% (3450-image training subset), then recovered to 72.29% (7380-image training subset), dropped again to 62.08% (14,400-image training subset), and finally reached 71.04% (26,880-image full training dataset). This pattern likely reflects dynamic model architecture selection, where different model types are chosen for different dataset sizes. While resource-efficient, this behavior introduces unpredictability that could complicate training and deployment.

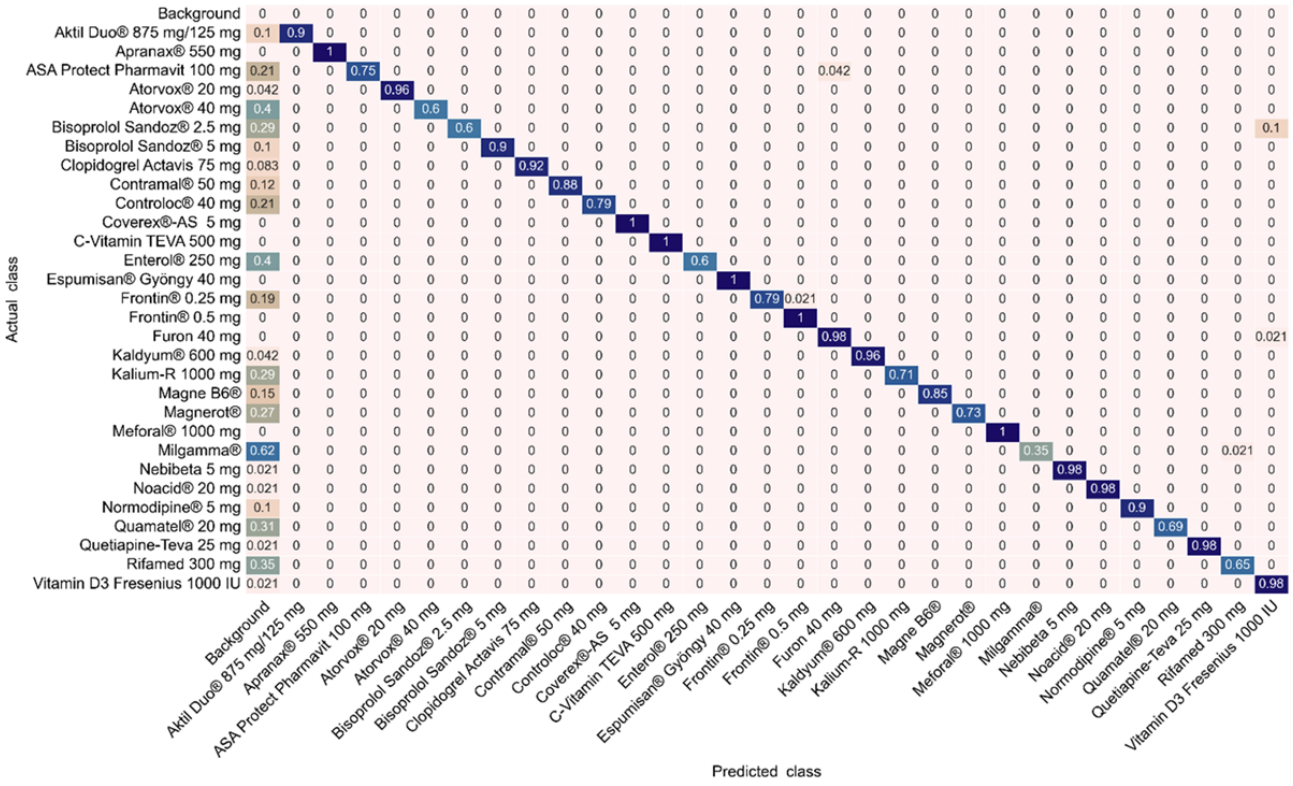
Custom Vision showed the most consistent behavior among cloud-based AutoML platforms, maintaining stable performance without degradation risks from model changes. Although Microsoft does not publicly disclose the model architecture used by Custom Vision, analyses of exported ONNX models suggest that the structure appears to be an older YOLOv2-based architecture, which would explain this consistency and the absence of architectural uncertainty we observed in other cloud-based AutoML platforms. Performance improvement trends paralleled YOLO11; however, the older architecture resulted in lower overall metrics.

Amazon Rekognition improved consistently with training data volume despite performing dynamic model selection during training. It demonstrated particularly strong performance on fully controlled datasets, especially in mAP values, with remarkable stability across IoU thresholds. Unlike other

XSL•FO
RenderX

platforms, it showed minimal degradation at the 0.50 IoU threshold and maintained strong performance on the challenging mAP@0.50 - 0.95 metric. It follows a conservative object detection strategy that prioritizes avoiding false positives, maintaining perfect precision (1.0) across multiple tests but with considerable false negative values indicative of missed detections, illustrated in the confusion matrix (Figure 5) with values listed in the background column. Analysis of confusion patterns revealed that the primary source of misclassification across all platforms was visual similarity among medications. Small, round, oblong, and oval white or off-white pills were most frequently confused with each other. Notably, these visually similar medications often represent therapeutically distinct drug classes (eg, beta-blockers, antiplatelet agents, and anxiolytics), where misidentification could have significant clinical consequences. Medications with distinctive visual characteristics, such as unique colors, shapes, or prominent surface markings, showed substantially lower misclassification rates across all platforms. The confusion matrices in Figure 5 and Multimedia Appendix 5 illustrate these visual and platform-specific confusion patterns in detail.

Figure 5. Confusion matrix showing Amazon Rekognition performance on the verification dataset (trained on the full training dataset of 26,880 images).



The cost-performance evaluation showed that although cloud-based AutoML platforms offer high accessibility, they also incur high platform costs, whereas traditional approaches, such as YOLO11, may provide comparable or better performance without platform costs. However, this does not account for the technical expertise required for YOLO11 implementation, which could offset the savings from platform costs. The primary cost consideration for YOLO11 implementation is human expertise rather than computational resources. Deployment requires proficiency in Linux system administration, Compute Unified Device Architecture, and CUDA Deep Neural Network library GPU driver configuration, Python programming, and deep learning frameworks such as PyTorch, typically representing senior AI and ML engineering expertise that substantially increases total implementation cost compared to the accessible web interfaces of cloud-based AutoML platforms. Our experiments used a workstation equipped with an NVIDIA Quadro RTX 5000 GPU, acquired for approximately US $3650 in 2021. The system consumes approximately 0.38 kW during training; consequently, the longest training run (approximately 13 hours) consumed an estimated 5 kWh of electricity, representing a negligible operational cost of less than US $1 at typical regional commercial electricity rates.

Among the cloud-based AutoML platforms, Vertex AI showed robustness with numerous comparative advantages, including high accuracy and F_1 performance values across various datasets, with relatively fast training time and good results on different training subsets, albeit with unpredictable fluctuations and significant cost barriers (higher flat-rate pricing and continued billing even when the endpoint is idle and the object detection model is not in use), which may limit its adoption in resource-constrained health care settings.

Limitations

The broader context of AI deployment in health care encompasses critical considerations beyond the scope of this study, including data security, privacy, and interoperability with existing hospital systems and protocols. Several limitations should be considered when interpreting our results. The dynamic model selection by cloud-based AutoML platforms prevented direct architectural comparisons, as different underlying models

XSL•FO
RenderX

were selected for different dataset sizes. Edge device testing was limited to Microsoft Azure and YOLO11, as other cloud-based AutoML platforms do not support ONNX model export. Performance metrics were evaluated at a conventional confidence threshold of 0.5, which may not be optimal for all platforms or deployment scenarios. The absence of CIs and formal statistical testing across hospital sites may be viewed as a limitation. However, the cloud-based AutoML platforms evaluated do not support the cross-validation procedures necessary for consistent variance estimation, and our study design prioritized assessment of the effects of real-world human and environmental factors on model performance. Several potential sources of bias should be considered when interpreting our results. The study was limited to 30 medications representing commonly dispensed products in Hungarian clinical settings, which does not encompass all possible visual characteristics encountered across different health care contexts or geographic regions. Performance may vary on larger medication databases, different medication types, and on medications with visual characteristics underrepresented in our dataset. While our clinical dataset was designed to assess real-world deployment challenges, it may confound interpretation of site-specific performance differences, as observed performance variations reflect the combined effects of platform capabilities and local imaging conditions rather than either factor in isolation. Collection of patient-based prescription images would face substantial ethical, regulatory, and practical barriers; therefore, in the creation of our clinical dataset, we used purposive sampling with balanced class representation to enable fair cross-platform comparison. While this approach is a standard methodology for ML benchmark studies and aligns with established datasets, it does not reflect natural prescription prevalence patterns. Future implementation studies should use representative sampling within appropriate ethical frameworks to validate model performance under authentic clinical workflow conditions.

Practical Implications

The substantial variation in performance observed across different clinical settings highlights the importance of environmental factors in model performance. The marked performance degradation observed in certain clinical settings indicates that local validation is essential before deployment.

When selecting platforms, users should consider each platform's unique characteristics, especially when working with smaller training datasets. YOLO11 offers competitive performance without platform costs and provides superior edge device compatibility for cost-sensitive applications requiring local processing. However, larger training datasets may be necessary for optimal performance. When evaluating the total cost of ownership, YOLO11's zero platform cost must be weighed against both hardware investment and the human capital required for implementation and ongoing maintenance after deployment. Consequently, cloud-based AutoML platforms, despite higher per-use costs, may prove more economical for some organizations and deployment scenarios. Vertex AI offers strong performance and ease of use despite higher costs, making it suitable for scenarios in which access to technical expertise is limited. Its fixed pricing model can be advantageous for situations requiring frequent processing of large volumes of images. Custom Vision and Amazon Rekognition are more economical for smaller-scale deployments, given their pricing structures. Custom Vision delivers predictable performance improvements, while Rekognition provides high-precision detection. Platform selection ultimately requires balancing technical requirements, budget constraints, and deployment scale to achieve optimal clinical pill recognition performance. For real-time point-of-care verification, response times below 2‐3 seconds generally integrate well into verification workflows without causing perceptible delays that could reduce productivity, introduce operator fatigue, or disrupt routine workflow rhythm. All prediction times observed in our study fall within this acceptable range, suggesting that computational speed is unlikely to be a limiting factor for any of the evaluated platforms.

The clinical implications of different error types warrant careful consideration for deployment planning. Missed detections and misidentifications pose a significant risk if undetected, potentially leading to missed therapeutic effects, unexpected adverse events, or dangerous drug interactions. Our findings reinforce the necessity of manual verification during clinical workflows for maintaining safety. Automated dispensing and verification systems should augment rather than replace pharmacist verification, with particular vigilance required for medications exhibiting high visual similarity to other products in the hospital formulary.

Conclusions

This study provides the first comprehensive comparison of a traditional code-based model and cloud-based AutoML approaches for pill recognition model training intended to be used in clinical settings. While no single platform dominated across all test scenarios, each demonstrated distinct advantages for specific use cases. YOLO11 proved reliable for model development and training due to its flexibility, cost-effectiveness, and edge deployment capabilities. Vertex AI delivered a strong overall performance with minimal technical expertise required. The substantial performance differences (20.62%-91.60% difference in accuracy) observed across UCTS underscore the critical importance of imaging conditions and the value of rigorous validation for real-world applications.

Platform selection for pill recognition should consider (1) available technical expertise and resources, (2) deployment environment characteristics, (3) budget constraints, and (4) edge deployment requirements. Future research should explore performance on larger medication databases and develop standardized imaging protocols to achieve consistent results for training pill recognition systems. As both traditional deep learning frameworks and cloud-based AutoML platforms continue to evolve, periodic systematic evaluation of these technologies using existing pill image datasets could provide valuable insights into how these technologies can be used to improve medication safety through automated verification systems.

XSL•FO
RenderX

Acknowledgments

The authors would like to sincerely thank the clinical pharmacists and staff at the Central Clinical Pharmacy of the University of Pécs, the Somogy County Kaposi Mór Teaching Hospital in Kaposvár, and the Healthcare Center in Komló for their assistance in collecting the clinical dataset images used in this study. We are grateful to NVIDIA Corporation for supporting our research through graphics processing units (GPUs) provided by the NVIDIA Hardware Grant Program.

Funding

This study was funded by the University Research Scholarship Program of the Hungarian National Research, Development, and Innovation Fund (EKÖP-24-4-I-PTE-57) and the Hungarian Scientific Research Fund grant (OTKA K-135729). The funder had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability

The data analyzed in this study are available from the corresponding author upon reasonable request. Requests should be sent to Amir Reza Ashraf at ashraf.amir.reza@pte.hu, specifying the type of data needed and its intended use. The corresponding author may provide the data subject to applicable conditions.

Authors' Contributions

ARA contributed to conceptualization, methodology, data curation, investigation, funding acquisition, project administration, writing the original draft, and reviewing and editing the manuscript. RR contributed to conceptualization, methodology, data curation, investigation, formal analysis, validation, writing the original draft, reviewing and editing the manuscript, and visualization. ZV contributed to methodology, visualization, funding acquisition, and reviewing and editing the manuscript. AF contributed to reviewing and editing the manuscript and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
General structure of the YOLO11 model.
[DOCX File, 73 KB - medinform_v14i1e79160_app1.docx]

Multimedia Appendix 2
Evaluation metrics.
[DOCX File, 343 KB - medinform_v14i1e79160_app2.docx]

Multimedia Appendix 3
List and properties of medications included in the study.
[DOCX File, 20 KB - medinform_v14i1e79160_app3.docx]

Multimedia Appendix 4
Performance metrics of pill recognition models across different datasets.
[DOCX File, 56 KB - medinform_v14i1e79160_app4.docx]

Multimedia Appendix 5
Additional confusion matrices.
[DOCX File, 4111 KB - medinform_v14i1e79160_app5.docx]

References

1. Global burden of preventable medication-related harm in health care: a systematic review. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240088887 [accessed 2025-06-16]
2. Elliott RA, Camacho E, Jankovic D, Sculpher MJ, Faria R. Economic analysis of the prevalence and clinical and economic burden of medication error in England. BMJ Qual Saf 2021 Feb;30(2):96-105. [doi: 10.1136/bmjqs-2019-010206] [Medline: 32527980]

XSL•FO
RenderX

3. Slight SP, Seger DL, Franz C, Wong A, Bates DW. The national cost of adverse drug events resulting from inappropriate medication-related alert overrides in the United States. J Am Med Inform Assoc 2018 Sep 1;25(9):1183-1188. [doi: 10.1093/jamia/ocy066] [Medline: 29939271]
4. Hodkinson A, Tyler N, Ashcroft DM, et al. Preventable medication harm across health care settings: a systematic review and meta-analysis. BMC Med 2020 Nov 6;18(1):313. [doi: 10.1186/s12916-020-01774-9] [Medline: 33153451]
5. Gillani SW, Gulam SM, Thomas D, et al. Role and services of a pharmacist in the prevention of medication errors: a systematic review. Curr Drug Saf 2021;16(3):322-328. [doi: 10.2174/1574886315666201002124713] [Medline: 33006539]
6. Lester CA, Li J, Ding Y, Rowell B, Kontar RA. Performance evaluation of a prescription medication image classification model: an observational cohort. NPJ Digit Med 2021 Jul 27;4(1):118. [doi: 10.1038/s41746-021-00483-8] [Medline: 34315995]
7. Lusk C, Catchpole K, Neyens DM, et al. Improving safety in the operating room: medication icon labels increase visibility and discrimination. Appl Ergon 2022 Oct;104:103831. [doi: 10.1016/j.apergo.2022.103831] [Medline: 35717790]
8. Batson S, Herranz A, Rohrbach N, Canobbio M, Mitchell SA, Bonnabry P. Automation of in-hospital pharmacy dispensing: a systematic review. Eur J Hosp Pharm 2021 Mar;28(2):58-64. [doi: 10.1136/ejhpharm-2019-002081] [Medline: 32434785]
9. Hänninen K, Ahtiainen HK, Suvikas-Peltonen EM, Tötterman AM. Automated unit dose dispensing systems producing individually packaged and labelled drugs for inpatients: a systematic review. Eur J Hosp Pharm 2023 May;30(3):127-135. [doi: 10.1136/ejhpharm-2021-003002] [Medline: 34795001]
10. Diwan T, Anirudh G, Tembhurne JV. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl 2023;82(6):9243-9275. [doi: 10.1007/s11042-022-13644-y] [Medline: 35968414]
11. Ragab MG, Abdulkadir SJ, Muneer A, et al. A comprehensive systematic review of YOLO for medical object detection (2018 to 2023). IEEE Access 2024;12:57815-57836. [doi: 10.1109/ACCESS.2024.3386826]
12. Tan L, Huangfu T, Wu L, Chen W. Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. BMC Med Inform Decis Mak 2021 Nov 22;21(1):324. [doi: 10.1186/s12911-021-01691-8] [Medline: 34809632]
13. Rádli R, Vörösházi Z, Czúni L. Metric ‐ based pill recognition with the help of textual and visual cues. IET Image Process 2024 Dec;18(14):4623-4638. [doi: 10.1049/ipr2.13273]
14. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv. Preprint posted online on Oct 23, 2024 URL: https://arxiv.org/abs/2410.17725 [accessed 2025-04-14] [doi: 10.48550/arXiv.2410.17725]
15. Ashraf AR, Somogyi-Végh A, Merczel S, Gyimesi N, Fittler A. Leveraging code-free deep learning for pill recognition in clinical settings: a multicenter, real-world study of performance across multiple platforms. Artif Intell Med 2024 Apr;150:102844. [doi: 10.1016/j.artmed.2024.102844] [Medline: 38553153]
16. Chen W, Tong J, He R, et al. An easy method for identifying 315 categories of commonly-used Chinese herbal medicines based on automated image recognition using AutoML platforms. Inform Med Unlocked 2021;25:100607. [doi: 10.1016/j.imu.2021.100607]
17. Wan KW, Wong CH, Ip HF, et al. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML vision in ultrasound breast lesions classification: a comparative study. Quant Imaging Med Surg 2021 Apr;11(4):1381-1393. [doi: 10.21037/qims-20-922] [Medline: 33816176]
18. Sapkota R, Qureshi R, Flores-Calero M, et al. YOLO11 to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series. arXiv. Preprint posted online on Jun 12, 2024 URL: https://arxiv.org/abs/2406.19407v5 [accessed 2025-04-16] [doi: 10.48550/arXiv.2406.19407]
19. Ayyalasomayajula MMT, Chintala S, Ayyalasomayajula S. A cost-effective analysis of machine learning workloads in public clouds: is AutoML always worth using? Int J Comput Sci Trends Technol 2019(5):107-115 [FREE Full text]
20. Sun Y, Song Q, Gui X, Ma F, Wang T. AutoML in the wild: obstacles, workarounds, and expectations. 2023 Apr 19 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023; Hamburg, Germany p. 1-15. [doi: 10.1145/3544548.3581082]
21. What is Custom Vision? Microsoft. URL: https://learn.microsoft.com/en-us/azure/ai-services/custom-vision-service/overview [accessed 2025-04-17]
22. Vertex AI documentation. Google. URL: https://cloud.google.com/vertex-ai/docs [accessed 2025-04-24]
23. Interfaces for Vertex AI. Google. URL: https://cloud.google.com/vertex-ai/docs/start/introduction-interfaces [accessed 2025-04-24]
24. What is Amazon Rekognition Custom Labels? Amazon Web Services. URL: https://docs.aws.amazon.com/rekognition/latest/customlabels-dg/what-is.html [accessed 2025-04-24]
25. Ling S, Pastor A, Li J, et al. Few-shot pill recognition. Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020. [doi: 10.1109/CVPR42600.2020.00981]
26. Yaniv Z, Faruque J, Howe S, et al. The National Library of Medicine pill image recognition challenge: an initial report. IEEE Appl Imag Pattern Recognit Workshop 2016 Oct;2016:1-10. [doi: 10.1109/AIPR.2016.8010584] [Medline: 29854569]
27. What's new in Custom Vision? Microsoft. URL: https://learn.microsoft.com/en-us/azure/ai-services/custom-vision-service/whats-new [accessed 2025-04-17]
28. Prepare image training data for object detection. Google. URL: https://cloud.google.com/vertex-ai/docs/image-data/object-detection/prepare-data [accessed 2025-04-24]

XSL•FO
RenderX

29. About data splits for AutoML models. Google. URL: https://cloud.google.com/vertex-ai/docs/general/ml-use [accessed 2025-04-24]
30. Managing labels. Amazon Web Services. URL: https://docs.aws.amazon.com/rekognition/latest/customlabels-dg/md-labels.html [accessed 2025-04-24]
31. Creating training and test datasets with images. Amazon Web Services. URL: https://docs.aws.amazon.com/rekognition/latest/customlabels-dg/md-create-dataset.html [accessed 2025-04-24]
32. Pricing. Google. URL: https://cloud.google.com/vertex-ai/pricing [accessed 2025-06-14]
33. Amazon Rekognition pricing. Amazon Web Services. URL: https://aws.amazon.com/rekognition/pricing/ [accessed 2025-06-14]
34. Azure AI Custom Vision Service. Microsoft. URL: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/custom-vision-service/ [accessed 2025-06-14]

Abbreviations

AI: artificial intelligence
API: application programming interface
AutoML: automated machine learning
AWS: Amazon Web Services
FNR: false negative rate
GPU: graphics processing unit
IoU: intersection over union
mAP: mean average precision
ML: machine learning
OER: overall error rate
ONNX: Open Neural Network Exchange
SDK: software development kit
UCTS: uncontrolled clinical test set
YOLO: You Only Look Once

Edited by A Benis; submitted 16.Jun.2025; peer-reviewed by A Njoku, J Aarts, S Han; revised version received 12.Jan.2026; accepted 17.Feb.2026; published 10.Apr.2026.

Please cite as:
Ashraf AR, Rádli R, Vörösházi Z, Fittler A
Code-Based Versus AutoML Methods for Pill Recognition in Clinical Settings: Comparative Performance Study
JMIR Med Inform 2026;14:e79160
URL: https://medinform.jmir.org/2026/1/e79160
doi:10.2196/79160

© Amir Reza Ashraf, Richárd Rádli, Zsolt Vörösházi, András Fittler. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bridging Population Patterns and Individual Prediction: Framework for Prospective Multimorbidity Study

Qianyao Zhang[1], BBA; Runtong Zhang[1], Prof Dr; Weiguang Ma[1], MSc; Butian Zhao[2], PhD; Xiaomin Zhu[3], Prof Dr

[1]Department of Information Management, School of Economics and Management, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China

[2]School of Management, Beijing University of Chinese Medicine, Beijing, China

[3]School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing, China

Corresponding Author:
Runtong Zhang, Prof Dr
Department of Information Management, School of Economics and Management, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China

Abstract

Background: Multimorbidity has become a major global public health challenge. However, existing research primarily emphasizes the identification of disease patterns at the population level and lacks the capacity to provide predictive insights into individual future pattern membership. Bridging this gap is crucial for personalized prevention and management.

Objective: This study aims to propose an innovative framework that integrates population-level multimorbidity pattern recognition with individual-level predictive modeling, thus advancing multimorbidity research from descriptive analysis to prospective multimorbidity pattern prediction.

Methods: Using longitudinal health follow-up data, we first applied latent transition analysis (LTA) to identify temporally stable multimorbidity patterns. These patterns were subsequently transformed into predictive labels to construct a novel deep learning model, CLA-Net (Cross-Lag Attention Network). CLA-Net is designed to predict individual future multimorbidity patterns by leveraging the complementary strengths of Gated Recurrent Units (GRU) and transformer architectures. It introduces a bitemporal directed cross-attention mechanism to simultaneously capture temporal dependencies and complex feature interactions. We compared CLA-Net against several advanced baselines and conducted ablation studies to validate its architectural components.

Results: In terms of pattern recognition, the LTA identified 5 clinically meaningful multimorbidity patterns: Cardiometabolic-Multisystem, Hypertension-Arthritis, Respiratory-Musculoskeletal, Metabolic Syndrome, and Gastritis-Arthritis. In terms of prediction, experimental results demonstrated that CLA-Net significantly outperformed all baseline models. CLA-Net achieved an accuracy of 0.8352 (SD 0.0048), a precision of 0.8326 (SD 0.0053), a recall of 0.8312 (SD 0.0056), and an F_1-score of 0.8319 (SD 0.0051). Notably, it achieved an area under the curve of 0.9293, surpassing baseline models. Ablation studies confirmed the necessity of the dual-branch architecture and the directed cross-attention mechanism, as removing these components resulted in performance declines ranging from 0.93% to 2.50%.

Conclusions: This study extends the scope of LTA beyond descriptive statistical modeling and establishes the scientific value of multimorbidity pattern prediction as an independent research task. By bridging population-level insights with individual-level prediction, the proposed framework provides a data-driven tool for the prospective prediction of future multimorbidity pattern membership conditional on survival, thereby supporting stratified disease management and care planning, rather than general risk stratification for acute or end-stage deterioration. This offers new methodological and practical value for precision medicine and public health policymaking.

(JMIR Med Inform 2026;14:e84261) doi:10.2196/84261

KEYWORDS

multimorbidity; latent transition analysis; LTA; deep learning; population-level patterns; personalized medicine

Introduction

Multimorbidity, commonly defined as the coexistence of 2 or more chronic conditions within a single individual [1], has emerged as one of the most pressing public health challenges of the 21st century amid the accelerating global trend of population aging [2]. Worldwide, approximately one-third of adults have 2 or more chronic diseases, and the proportion exceeds 50% among individuals aged 60 years and older [3]. Multimorbidity not only substantially increases health care resource consumption and imposes a heavy burden on health systems, but also severely compromises patients' quality of life, leading to a markedly higher risk of functional disability and

XSL•FO
RenderX

premature mortality, thereby greatly increasing the complexity of chronic disease management and clinical decision-making [4,5]. Previous studies have demonstrated that the co-occurrence of chronic diseases is not a random phenomenon but tends to manifest as distinct multimorbidity patterns, in which specific combinations of diseases frequently occur in particular populations [6,7]. This suggests the potential existence of shared pathophysiological mechanisms, risk factors, or lifestyle determinants underlying these patterns [8]. Therefore, systematically identifying, understanding, and predicting these multimorbidity patterns holds strategic importance for the early detection of high-risk individuals, the advancement of precision medicine, and the optimization of health care resource allocation [9,10].

Although various methods for identifying multimorbidity patterns have been proposed, such as factor analysis, cluster analysis, latent class analysis (LCA), and network analysis [11-15], they share significant limitations and fall short of meeting the clinical demand for personalized and prospective management. First, most approaches group populations primarily based on disease prevalence or epidemiological characteristics, assuming homogeneity within groups while overlooking the inherent heterogeneity among individuals [14,16]. Such population-level stratification can facilitate the identification of common multimorbidity patterns at a macrolevel but fails to uncover microlevel differences within patterns, which is particularly inadequate for supporting fine-grained individualized management in clinical practice. In reality, even among patients belonging to the same multimorbidity pattern, disease trajectories may differ substantially. For example, some patients experience slow progression and remain stable over a long period, whereas others may rapidly develop multisystem dysfunction [17]. Neglecting such heterogeneity may directly result in biased risk assessments and inefficient resource allocation, thereby severely constraining the practical implementation of precision medicine. In addition, existing studies have mainly focused on static multimorbidity pattern mining [16,18], and the methods used often struggle to maintain consistency of pattern meanings in longitudinal data. This instability not only reduces the comparability of research findings but also weakens the ability to capture disease progression, thereby limiting the clinical interpretability and practical value of the conclusions.

To address the above challenges, latent transition analysis (LTA), a longitudinal latent variable model, has been introduced into health research. LTA can identify stable latent health states across multiple time points and ensure their consistency and comparability over time [19]. Compared to other longitudinal modeling approaches, LTA offers distinct advantages for identifying multimorbidity evolution. Unlike growth mixed models or latent growth curve models, which primarily characterize continuous trajectories of a single variable (eg, functional decline over time), LTA is specifically designed to model transitions between discrete, qualitative latent statuses, aligning perfectly with the categorical nature of multimorbidity patterns. Furthermore, in contrast to standard Markov chain models that rely on observed states, LTA separates measurement error from true structural transitions by inferring latent classes from multiple observed indicators, thereby ensuring more robust and clinically meaningful pattern identification. By modeling follow-up data, LTA can effectively capture heterogeneity within populations and provide new perspectives for understanding the dynamic evolution of multimorbidity patterns [20,21]. However, LTA is essentially a descriptive statistical method, with its primary focus on identifying and summarizing patterns from historical or current data, while lacking the capacity to predict individual future health states. This limitation hinders its ability to meet urgent clinical needs, such as providing precise early warnings for high-risk individuals, developing timely intervention strategies, and guiding personalized treatment planning. Therefore, constructing predictive models that can accurately forecast individuals' future multimorbidity patterns carries substantial clinical value and practical significance.

Predicting individual future multimorbidity patterns requires accounting for both the temporal evolution of health states and the complex interactions among multidimensional health features. On the one hand, the progression of chronic diseases exhibits marked and highly nonlinear temporal dependencies. A patient's current health status not only reflects their immediate physiological condition but also embodies the evolution from the previous health state, reflecting the staged characteristics of disease progression [22]. Ignoring such sequential state transitions weakens a model's ability to capture disease progression trends, thereby limiting its capacity to deliver reliable forward-looking predictions. On the other hand, multidimensional health features, including lifestyle factors, preexisting conditions, and functional indicators, are interconnected through intricate nonlinear couplings and synergistic effects. These dynamic interactions drive the formation and evolution of multimorbidity patterns [23]. Many conventional statistical methods, such as logistic regression, support vector machines, random forests, gradient boosting trees, and k-nearest neighbors (KNN), perform well in modeling static feature associations [24,25]. However, when confronted with high-dimensional, nonlinear, and temporally dependent medical data, their modeling capacity is limited, making it difficult to simultaneously capture nonlinear state transitions and deep couplings among features. Therefore, there is an urgent need to develop a model that overcomes these limitations by effectively integrating sequential health state changes and feature coupling effects in longitudinal data, thus enabling precise and prospective prediction of individual future multimorbidity patterns and providing robust data support for clinical decision-making.

Based on the above discussion, this study proposes an innovative framework that integrates population-level multimorbidity pattern recognition with precise prediction of individual future states, thereby advancing multimorbidity research from macrolevel description to prospective application. Specifically, the framework first uses LTA to identify population-level multimorbidity patterns with temporal consistency and clinical stability from longitudinal follow-up data, and then transforms these patterns into individual predictive labels. On this basis, we designed the Cross-Lag Attention Network (CLA-Net), which leverages immediate longitudinal context to model

XSL•FO
RenderX

short-term, nonlinear state transitions and complex feature interactions among health variables, thereby enabling accurate prediction of individual future multimorbidity patterns.

The contributions of this study can be summarized in 3 aspects. First, we propose a novel cross-paradigm framework that bridges a statistical latent-variable modeling approach with deep representation learning for longitudinal multimorbidity research. Specifically, population-level multimorbidity patterns identified through LTA are leveraged as clinically interpretable supervisory signals for individual-level prediction. By integrating latent structure discovery with prospective modeling, this framework overcomes the limitation of traditional multimorbidity studies that are largely restricted to retrospective pattern characterization and establishes multimorbidity pattern prediction as a distinct and clinically meaningful research task.

Second, we propose a novel hybrid neural network architecture, CLA-Net, for multimorbidity pattern prediction. CLA-Net integrates the inherent strength of Gated Recurrent Units (GRU) in efficiently modeling short-term, nonlinear state transitions with the powerful representational capacity of the transformer architecture in capturing complex feature interactions. This design surpasses the capacity of traditional linear models, enabling deep modeling of the intricate couplings within medical data. Furthermore, CLA-Net incorporates a bitemporal directed cross-attention mechanism, which establishes directional information channels between the preceding and current health states. This mechanism facilitates feature extraction across time steps and dynamic association modeling, allowing the network to simultaneously perceive and effectively integrate dynamic state changes and feature coupling effects within complex longitudinal medical data.

Third, the experimental results demonstrate that the proposed CLA-Net model significantly outperforms other mainstream baseline models. Further ablation studies confirm that the synergistic integration of GRU and transformer, together with the bitemporal directed cross-attention mechanism, is the key factor driving the performance improvement. These findings not only provide strong evidence for the superiority of CLA-Net and the soundness of its design but also offer new insights and paradigms for model development in the field of multimorbidity pattern prediction.

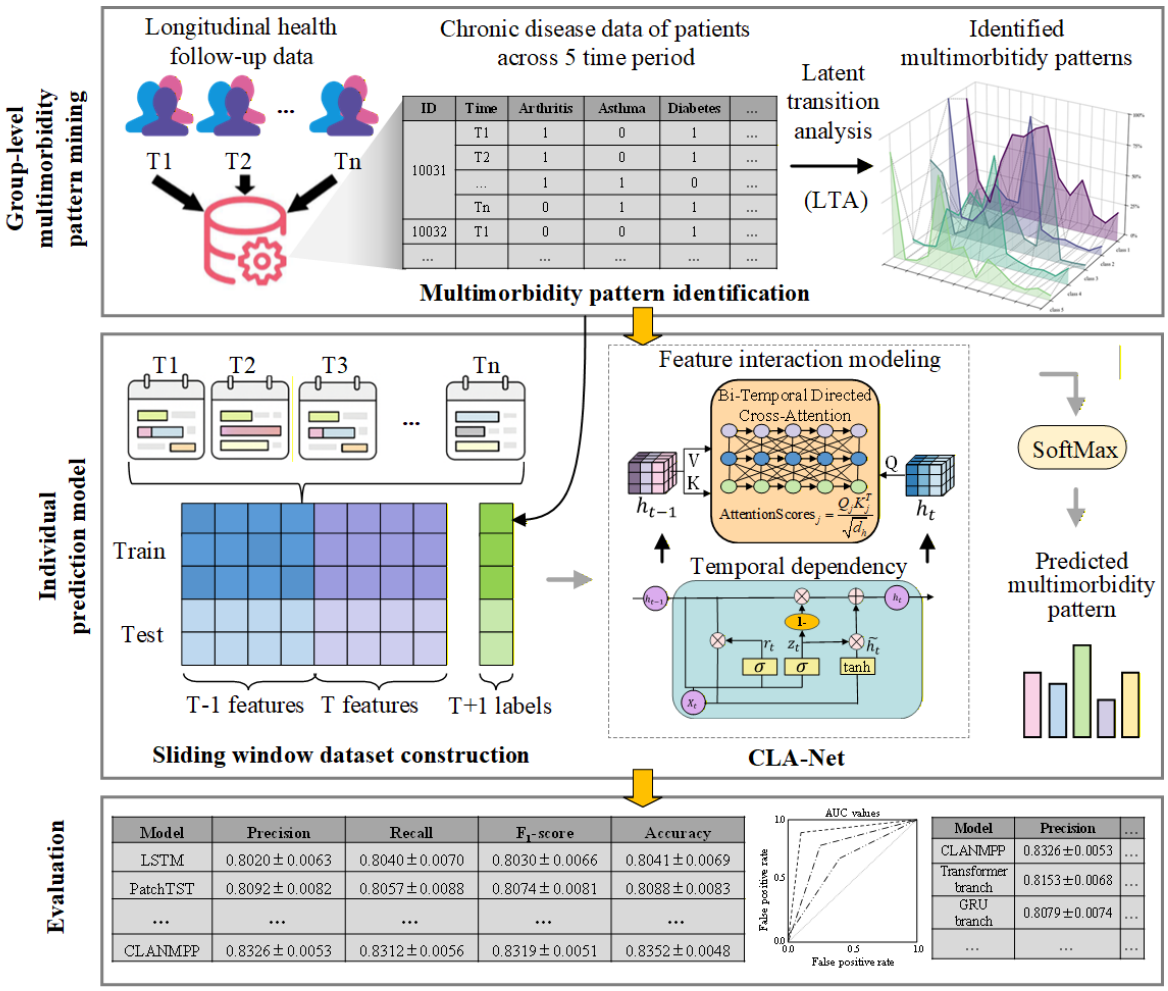
Methods

Overview

We designed a novel research framework that integrates multimorbidity pattern recognition with individualized prediction. The overarching idea is to first identify population-level latent multimorbidity patterns with temporal consistency from longitudinal health data, and then build a deep learning model to predict each individual's future pattern membership, thereby establishing a unified framework that bridges population-level recognition and individual-level prediction. As illustrated in Figure 1, the proposed framework consists of 3 stages. In the first stage, LTA is applied to longitudinal health records to identify latent multimorbidity patterns at the population level, ensuring that these patterns maintain consistent clinical meaning and comparability across different time points, thus providing a stable and reliable label basis for subsequent individual prediction tasks. In the second stage, based on the LTA-derived pattern labels, we design and construct the CLA-Net, which integrates temporal dependencies and complex feature interactions to enable accurate prediction of individual future pattern membership. In the third stage, we conduct a comprehensive performance evaluation and validation of the prediction model.

XSL•FO
RenderX

Figure 1. Proposed framework for multimorbidity pattern recognition and prediction. CLA-Net: Cross-Lag Attention Network; LTA: latent transition analysis.



Multimorbidity Pattern Identification

This study used SAS 9.4 (SAS Institute) to apply LTA for identifying the latent class structures of population-level multimorbidity patterns across 5 waves of China Health and Retirement Longitudinal Study (CHARLS) data. LTA is a longitudinal analytic method based on finite mixture models, which can detect unobserved health subgroups (latent statuses) within the sample across multiple time points. By imposing measurement invariance constraints, LTA ensures temporal consistency and stability of the latent classes [20], thereby guaranteeing that the identified classes maintain comparable and clinically meaningful interpretations across different time points. This provides stable and reliable classification labels for subsequent individual prediction modeling tasks. To control classification error and enhance model stability, we adopted a 2-step analytic strategy [26]. The detailed procedure is as follows: in step 1, we first conducted LCA separately at each time point to explore the optimal number of latent class partitions. In the model specification process, we considered different numbers of latent classes (ranging from 2 to 6) [26]. Model fit was evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), sample-size adjusted BIC (SaBIC), and entropy, to determine the optimal class solution.

Specifically, the AIC balances model fit and complexity, with smaller values indicating better fit [27]. The BIC adds a penalty for sample size, favoring more parsimonious models and thus being more conservative in class determination, particularly in large samples [28]. The SaBIC further adjusts BIC for sample size, offering greater robustness in large-sample contexts [29]. Entropy (ranging from 0 to 1) measures the certainty of individual classification, with values above 0.80 generally indicating high classification accuracy [30]. In addition, we used the Bootstrap Likelihood Ratio Test (BLRT) to compare log-likelihood differences between adjacent class models, in order to assess whether adding an extra class significantly improves model fit [31,32].

In step 2, after determining the optimal number of classes, we conducted chi-square tests on the log-likelihood values of each model [31] and performed invariance testing of class structures across different measurement time points, in order to verify whether the meanings of multimorbidity classes remained consistent over time [31]. This step is a critical prerequisite for ensuring that classes retain the same interpretive meaning across time points, thereby enhancing the interpretability and stability of the model [33].

In this study, LTA is used as a population-level structure discovery tool to identify clinically stable multimorbidity phenotypes, rather than as a component optimized for

XSL•FO
RenderX

downstream predictive performance. The primary objective of LTA in our framework is to define a consistent and interpretable latent outcome space that reflects the underlying disease co-occurrence structure at the population level. Accordingly, LTA was estimated using the full longitudinal dataset across all available survey waves to maximize phenotype stability and clinical interpretability by leveraging the complete temporal information within the observed follow-up window. Within this 5-wave longitudinal design, the use of all waves enables the identified latent classes and transition patterns to reflect the evolution of multimorbidity structures across the entire observed period, rather than being driven by partial or wave-specific information. This design choice allows the latent class definitions to be less sensitive to sampling variability across individual waves and ensures that the resulting multimorbidity patterns are representative of the population dynamics observed within the study timeframe, rather than tailored to a specific subsample.

Proposed CLA-Net Framework

Multimorbidity Pattern Prediction Dataset

This study constructed a task-oriented dataset for prospective prediction of multimorbidity patterns based on the CHARLS, a nationally representative longitudinal survey in China. To bridge the gap between population-level pattern discovery and individual-level prediction, we constructed a supervised learning task using LTA-derived latent classes as predictive targets. Since LTA is a probabilistic model, it outputs the posterior probability of an individual belonging to each latent class rather than a deterministic category. To transform these probabilistic outputs into deterministic prediction labels for CLA-Net, we used a maximum posterior probability assignment strategy. Specifically, for each individual n at time step t, the LTA model calculates a posterior probability vector $\boldsymbol{P}_{n,t} = [p_1, p_2,\ldots, p_K]$, where K is the number of latent patterns. The ground truth label $y_{n,t}$ is assigned to the class with the highest probability:

$$(1)\text{yn,t=arg maxk}\in\{1,...,K\}(\text{Pn,t(k)})$$

These deterministic labels ($y_{n,t}$) serve as the supervisory signals (ground truth) for training the CLA-Net.

We used 2 consecutive follow-up waves as the input window to predict the multimorbidity pattern at the subsequent wave. In CHARLS, the intervals between survey waves are 2‐3 years, which align with the slow evolution and long-term accumulation process of chronic diseases. Although the intervals vary slightly, the progression of chronic conditions is characterized by slow variables (ie, gradual progression, stable trends, and limited abrupt changes), making 2‐3 years sufficient to capture substantial changes in individual health status while reducing noise from short-term fluctuations. This temporal span is clinically reasonable for modeling chronic disease progression. According to the study objectives, we reorganized the raw data using a sliding-window approach to construct 3 temporally progressive subdatasets, each corresponding to a "past–current → future" window structure:

- Subdataset 1: 2011 features as early input ($\mathbf{x}_{t-1}$), 2013 features as current input ($\mathbf{x}_t$), and 2015 LTA-derived multimorbidity patterns as prediction labels (y_{t+1})
- Subdataset 2: 2013 features as ($\mathbf{x}_{t-1}$), 2015 features as ($\mathbf{x}_t$), and 2018 multimorbidity patterns as (y_{t+1})
- Subdataset 3: 2015 features as ($\mathbf{x}_{t-1}$), 2018 features as ($\mathbf{x}_t$), and 2020 multimorbidity patterns as (y_{t+1})

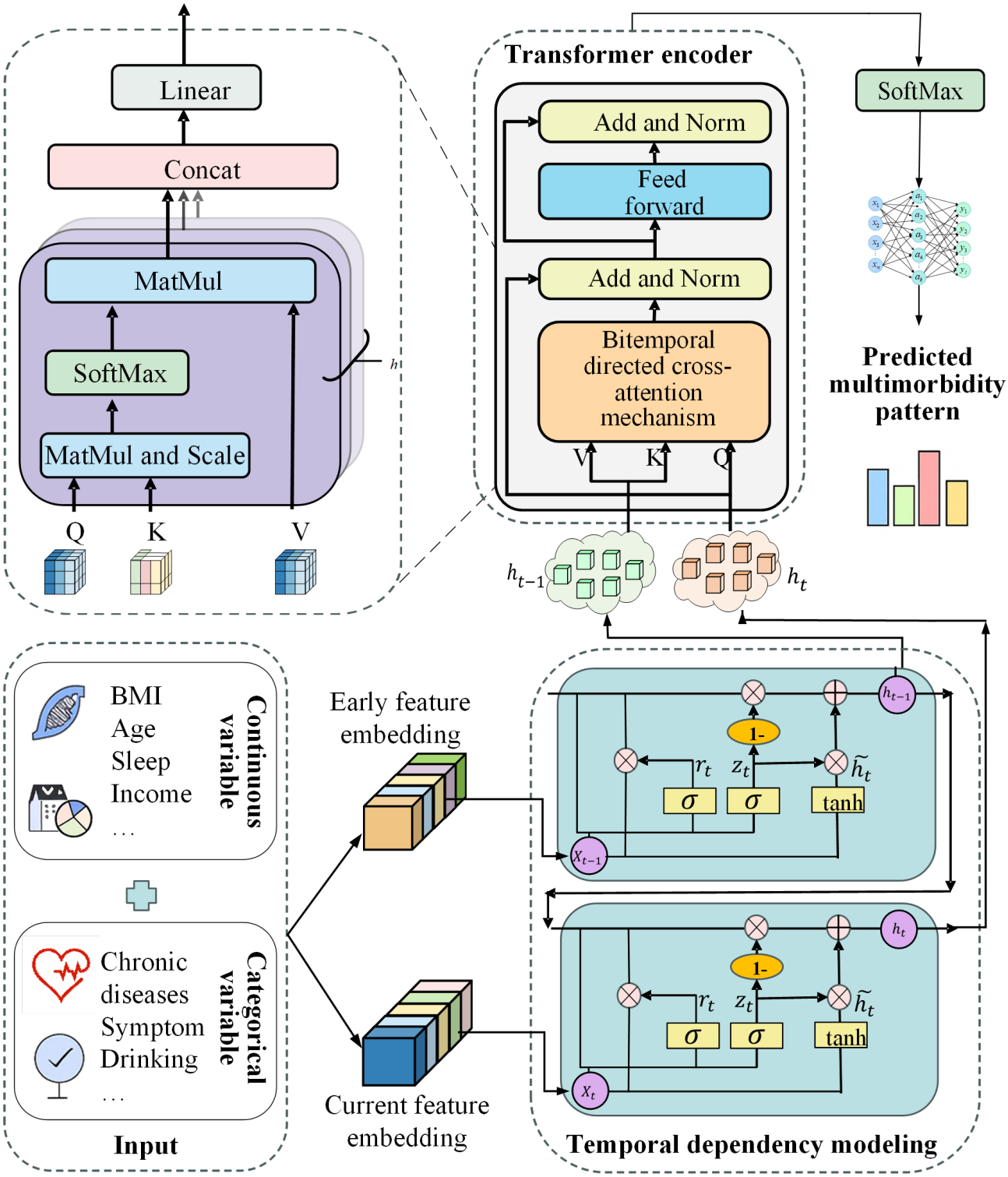
The 3 subdatasets were concatenated by rows to form a unified prediction dataset, structured as triplets ($\mathbf{x}_{t-1}$,$\mathbf{x}_t$, y_{t+1}), where $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ represent feature vectors from 2 consecutive time points, and y_{t+1} denotes the LTA-identified future multimorbidity pattern label. This construction preserves the natural temporal evolution of disease states without relying on long-sequence assumptions, while effectively augmenting the training data. It enables the model to learn multistage health state transitions within the same population, thereby providing a robust data foundation for subsequent individual-level prospective prediction.

Overview of the Model Architecture

This study proposes an innovative deep learning model, named the CLA-Net. CLA-Net integrates the sequential state encoding capability of GRU with the global feature interaction advantages of an improved transformer architecture, enabling fine-grained cross-time feature interactions through a bitemporal directed cross-attention mechanism. The model is composed of 4 core modules: an input embedding layer, a temporal GRU encoder, a transformer encoding layer based on the bitemporal directed cross-attention mechanism, and a classifier layer. An overview of the model architecture is shown in Figure 2.

XSL•FO
RenderX

Figure 2. The architecture of the Cross-Lag Attention Network.



Components of CLA-Net

Feature Embedding Layer

The input features include continuous variables (eg, biomarkers and cognitive scores) and binary indicator variables (eg, disease diagnoses and medication use). Socioeconomic variables, including total household income and per capita household consumption, were inspected for extreme values prior to normalization. Given the heavy-tailed distributions of these variables, a percentile-based outlier handling strategy was applied to mitigate the undue influence of extreme observations while preserving the original scale of the data. Min-max normalization was then used to rescale these variables to a common range, ensuring numerical stability and consistency with the scaling of other continuous inputs used in the model. Binary features were kept unchanged. Subsequently, heterogeneous features are mapped into a unified representation space through learnable linear transformations:

(2)Xt−1Wext−1+be∈ Rd

(3)Xt=WeXt+be∈ Rd

XSL•FO
RenderX

where xt∈ RD denotes the raw input features, Xt∈ Rddenotes the embedded representation, We∈ Rd×Dand be∈ Rdare shared parameters, *D* denotes the input dimension, and *d* represents the hidden space dimension.

Temporal GRU Encoder

After feature embedding, the representations of features from the earlier time point ($\mathbf{X}_{t-1}$) and the current time point ($\mathbf{X}_t$) are concatenated into a 2-step sequence, which is then fed into a GRU for temporal modeling. Through the update mechanism of its hidden states, the GRU efficiently captures the sequential evolution of the disease process, ensuring that the nonlinear dynamics of health state transitions are explicitly modeled. This design enables the network to capture the critical short-term progression from *t*–1 to *t* and to explicitly model the nonlinear dynamics underlying near-term health state transitions.

(4)ht−1=GRU(Xt−1,hinit)

(5)ht=GRU(Xt,ht−1)

where $\mathbf{h}_{init}$ denotes the initial hidden state, and $\mathbf{h}_{t-1}$ and $\mathbf{h}_t$ represent the hidden state outputs at the 2 time steps. The internal computational mechanism of the GRU is defined as follows:

(6)rt=σ(WrXt+Urht−1+br)

(7)zt=σ(WzXt+Uzht−1+bz)

(8)h~t=tanh (WhXt+Uh(rt⊙ht−1)+bh)

(9)ht=(1−zt)⊙ht−1+zt⊙h~t

Here, $\mathbf{X}_t$ denotes the input at time step *t*. $\mathbf{r}_t$ is the reset gate, which controls how much information from the previous time step is retained. $\mathbf{z}_t$ is the update gate, which determines the proportion of information to be updated at the current time step. ht~ represents the candidate's hidden state. σis the sigmoid activation function, and ⊙ denotes elementwise multiplication. $\mathbf{W}_*$ and $\mathbf{U}_*$ are learnable weight matrices, and $\mathbf{b}_*$ is the bias vector.

Transformer Encoding Layer

Bitemporal Directed Cross-Attention Mechanism

Although GRUs are effective at encoding immediate longitudinal context via gated state updates, they have limited capacity to explicitly characterize fine-grained interactions among heterogeneous features, especially when the predictive signal arises from short-term, nonlinear state transitions rather than long-range history. In our setting, an individual's future multimorbidity pattern is shaped not only by the recent state representation but also by complex cross-feature influences that operate across adjacent time points. To address this, we adopt an asymmetric attention paradigm, in which features from the current time point are used as the query vector **Q**, while features from the earlier time point serve as the keys (**K**) and values (**V**) in the attention computation, enabling the model to attend to historical representations when updating the current state. This design reflects the directional interactions between features across time. For each attention head $j \in \{1,...,H\}$, the GRU outputs are first projected into the **Q**, **K**, and **V** spaces:

(10){Qj=htWQ(j)Kj=ht−1WK(j)Vj=ht−1WV(j)

where Q_j, K_j, and V_j denote the query, key, and value representations of the *j*th attention head, which are vector-valued representations derived from the input hidden states. WQ(j), WK(j), WV(j)∈ Rd×dh denotes the projection matrix of the *j*th head, *d* is the input dimension, and $d_h=d/H$ is the dimension of each attention head. The attention scores are computed using scaled dot-product attention:

(11)Attention scoresj=QjKjTdh

The scaling factor 1/dh is critical for maintaining gradient stability. Without scaling, the dot-product values increase with feature dimensionality, which may push the SoftMax function into regions with extremely small gradients. The computation for each attention head is defined as

(12)headj=Attention(Qj,Kj,Vj)=SoftMax (QjKjTdh)Vj

After concatenating the outputs of all attention heads, a linear transformation is applied to obtain the multi-head cross-attention output:

(13)MultiHeadCrossAttn(ht−1,ht)=Concat(head1,…,headH)Wo

where $\mathbf{W}_O$ denotes the output projection matrix. Through the multihead mechanism, the model can learn distinct attention patterns in different representation subspaces, allowing it to attend to diverse types of feature relationships simultaneously and thereby enhancing its expressive capacity.

Stacked Transformer Encoder Architecture

To capture the hierarchical patterns of disease progression, we use a stack of *L* transformer encoder layers. Each layer builds upon the representations from the previous one, thereby progressively modeling more complex temporal dependencies. Each layer, $l \in \{1,..,L\}$ consists of 2 sublayers with residual connections. Specifically, Equation 14 corresponds to the bitemporal directed cross-attention mechanism sublayer, while Equation 15 represents the positionwise feed-forward network (FFN) sublayer:

(14)z^(l)=LayerNorm(z(l−1)+Dropout(MultiHeadCrossAttn(ht−1,z(l−1))))

(15)z(l)=LayerNorm (z^(l)+Dropout(FFN(z^(l))))

Here, the positionwise FFN uses the ReLU activation function and incorporates dropout to prevent overfitting, defined as

(16)FFN(x)=ReLU(xW1+b1)W2+b2

where $\mathbf{W}_1$ is the weight matrix of the first feed-forward layer, which expands the feature dimension by a factor of 4 to enhance nonlinear representational capacity. $\mathbf{W}_2$ is the weight matrix of the second feed-forward layer, which projects the expanded dimension back to the original size *d*. $\mathbf{b}_1$ and $\mathbf{b}_2$ denote the bias vectors for the first and second layers, respectively. After passing through *L* stacked layers, we obtain the final fused representation $\mathbf{z}^{(L)}$, which simultaneously encodes early and later temporal information of patients and is further enriched through direct cross-time feature interactions.

Classifier Layer

The fused features $\mathbf{z}^{(L)}$ output by the transformer encoder are fed into a 2-layer fully connected neural network as the classifier. This classifier consists of a linear layer, a ReLU

XSL•FO
RenderX

activation function, a dropout layer, and a final linear output layer, which maps the fused features into probability distributions over multimorbidity patterns.

First Layer: Linear Mapping and Nonlinear Activation

The first layer of the classifier consists of a linear mapping followed by a nonlinear activation function. This step is crucial for transforming the fused features into a more complex representation. The output of this layer is computed as follows:

(17)o=ReLU (z(L)W1(c)+b1(c))

where W1(c)∈ Rd×d and b1(c)∈ Rd are the weight matrix and bias vector of the first layer, respectively, and ReLU denotes the activation function. A dropout regularization is applied after this layer to reduce the risk of overfitting.

Second Layer: Output Mapping

The second layer of the classifier maps the transformed features to output logits, representing the class scores. This layer computes the logits as follows:

(18)ylogits=o W2(c)+b2(c)

where W2(c)∈ Rd×C and b2(c)∈ RC represent the weight matrix and bias vector of the second layer, respectively, and *C* denotes the number of output classes. The final prediction $\mathbf{y}_{\text{logits}}$ is obtained by applying the SoftMax function to transform the outputs into probabilities:

(19)y^=SoftMax(ylogits)=exp (ylogits, c)∑c′=1Cexp (ylogits, c′)

Loss Function

We trained the model using the standard cross-entropy loss function for multiclass, single-label classification. Based on the predicted probability distribution y^ output by the classifier, the cross-entropy loss is defined as

(20)L=−1N∑i=1N∑c=1Cyi(c)log (y^i(c))

where *C* denotes the number of classes, yi(c)is the one-hot encoded true label of the *i*th sample for class *c*, and yi(c)represents the predicted probability that the *i*th sample belongs to class *c*. The objective is to minimize the average loss across all training samples. Cross-entropy is particularly suitable for our multiclass single-label task because it encourages the network to assign high probability to the correct class while penalizing confident but incorrect predictions. Since the dataset has a relatively balanced class distribution, no class weighting was applied in the loss function. The use of SoftMax outputs with cross-entropy is equivalent to maximizing the likelihood of the correct class and is a standard choice for neural network classifiers.

Algorithm 1 in Textbox 1 presents the pseudocode of the CLA-Net model. The algorithm details how each component is sequentially executed on each batch of data.

XSL•FO
RenderX

Textbox 1. Algorithm 1: CLA-Net (Cross-Lag Attention Network) model.

Input:

xt: Features at time *t* (continuous features xtcont , discrete features xtdisc)

x_{t-1}: Features at time t–1 (continuous features xt−1cont , discrete features xt−1disc)

y: Label (5-class classification)

Output: Predicted probability distribution **p**, cross-entropy loss *L*

Process:

1: // *1. Feature preprocessing and embedding*

2: et−1←Embedding (xt−1disc)*// Embedding of discrete features*

3: et←Embedding (xtdisc)

4: ct−1←LayerNorm (xt−1cont)*// Normalization of continuous features*

5: ct←LayerNorm (xtcont)

6:xt−1←Linear (Concat(et−1, ct−1))*// Concatenation of features at time t−1*

7: xt←Linear (Concat(et, ct))*// Concatenation of features at time t*

8: Xt−1←We xt−1+be

9: Xt←We xt+be

10: *// 2. GRU-based temporal encoding*

11:ht−1←GRU (Xt−1, hinit)*// Hidden state at time t−1*

12: ht←GRU (Xt, ht−1)*// Hidden state at time t*

13: *// 3. Transformer encoder based on the bitemporal directed cross-attention mechanism*

14: for do

15: for head i=1,…,Ha do

16: Q(i)←WQ(i)ht, K(i)←WK(i)ht−1, V(i)←WV(head)ht−1

17: A(i)←Softmax (Q(i)(K(i)) dk)*// Attention weights*

18:head(i)←A(i)V(i)*// Context vector*

19: end for

20: Z←Concat (head(1),…,head(H))Wo

21: U←LayerNorm (ht+Dropout(Z))*// Residual connection and normalization*

22: ht←LayerNorm (U+Dropout (FFN(U)))

23: end for

24: Let z←ht*// final fused representation*

25: *// 4. Classification and prediction*

26: o←Flatten(z)

27: p←Softmax (Wfc o+bfc)*// Output probability distribution*

28: L←−∑i=15yilog pi*// Cross-entropy loss*

29: return **p**, *L*

Training Configuration and Implementation Details

All experiments were conducted using PyTorch v1.7.0 (Facebook AI Research, Meta Platforms) and Python 3.8 (Python Software Foundation) on Ubuntu 18.04 (Canonical Ltd), with an NVIDIA RTX 4090 GPU (24 GB memory) and 32 GB RAM (CUDA 11.0). The proposed model and all baselines were trained under identical configurations and hyperparameters (Table 1). The main settings were 60 training epochs, mini-batch size of 64, embedding and GRU hidden dimensions set to 128, Adam optimizer with a learning rate of 0.0001 (determined via grid search), cross-entropy loss, and a dropout rate of 0.3 applied to both transformer layers and the classifier. Model weights were initialized using PyTorch defaults. During data preprocessing, feature normalization was fitted only on the training set and then applied to the validation and test sets. The dataset was partitioned into 80% training and 20% testing sets using a subject-wise splitting strategy based on unique individual identifiers. This approach ensures that all

XSL•FO
RenderX

longitudinal observations from the same participant were assigned exclusively to a single subset, thereby preventing potential data leakage. From the training set, 20% (n/N) of the individuals were further held out as a validation set. During training, the macro-F_1-score on the validation set was used as the criterion for early stopping and model selection. Only the best-performing model on the validation set was retained, and a final evaluation was conducted on the independent test set.

Table . Training hyperparameters.

Parameters or hyperparameters	Value
Epoch	60
Learning rate	0.0001
Dropout	0.3
Random seed	42
Batch size	64
Hidden dimension	128

Evaluation Metrics

To evaluate the performance of the model in the multiclass classification task, we used commonly used metrics, including accuracy, precision, recall, and F_1-score, and reported their macroaveraged values. Macroaveraging assigns equal weight to each class, thereby avoiding the bias of microaveraging, which tends to favor majority classes in imbalanced datasets. This provides a more comprehensive and balanced assessment of the model's ability to identify all 5 multimorbidity patterns. The formulas for these evaluation metrics are as follows:

(21)Accuracy=∑c=1CTPc∑c=1C(TPc+FPc+FNc+TNc)

(22)Precisionc=TPcTPc+FPc

(23)Recallc=TPcTPc+FNc

(24)F1=2×Precisionc×RecallcPrecisionc+Recallc

where TP_c, FP_c, FN_c, and TN_c denote the true positives, false positives, false negatives, and true negatives for class c, respectively. Accuracy represents the proportion of correctly predicted samples to the total number of samples. Precision measures the proportion of samples predicted as a given class that actually belong to that class. Recall indicates the proportion of samples that are correctly identified among those that truly belong to a given class. The F_1-score, defined as the harmonic mean of precision and recall, provides a balanced measure of the model's classification performance for a given class. Macroaveraging is obtained by taking the arithmetic mean of the metric values across all classes. The formulas are defined as follows:

(25)Macro Precision=1C∑c=1CPrecisionc, Macro Recall=1C∑c=1CRecallc, Macro F1=1C∑c=1CF1c

In addition, we further introduced the receiver operating characteristic (ROC) curve and the area under the curve (AUC) as complementary evaluation metrics. While accuracy focuses on the correctness of class predictions, AUC assesses the ranking quality of predicted probabilities. Particularly in multiclass problems, a higher macroaveraged AUC together with higher accuracy provides stronger evidence that the model can achieve good discrimination across all classes, rather than merely benefiting from class size distributions. The ROC curve is plotted with the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis and is mathematically defined as:

(26)FPR=FPFP+TN, TPR=TPTP+FN

AUC measures the overall discriminative ability of a model, with values ranging from 0 to 1, where values closer to 1 indicate stronger separation between positive and negative classes. Commonly, AUC >0.9 is considered excellent, 0.8 - 0.9 good, 0.7 - 0.8 acceptable, and <0.7 poor. For multiclass prediction, we adopted a one-vs-rest strategy, treating each class as positive against all others to compute class-specific ROC curves and AUC values. To validate the effectiveness of the proposed model, we compared it with multiple baselines under identical test conditions and reported AUC as a key performance metric. The evaluation procedures followed recommended practices for multiclass classification to ensure reliable assessment of predictive performance.

In addition to AUC-ROC, we also reported AU-PRC (area under the precision-recall curve), which provides a complementary evaluation of model performance by emphasizing the trade-off between precision and recall. AU-PRC is particularly informative in settings with class imbalance, as it focuses on the model's ability to correctly identify positive instances without being overly influenced by true negatives. Similar to the ROC-based evaluation, AU-PRC was computed using a one-vs-rest strategy for multiclass prediction, and class-specific precision-recall (PR) curves were aggregated to assess overall performance.

Baseline Models

Logistic Regression

A linear classification model that estimates class probabilities through a logistic function, serving as a simple and interpretable baseline to assess the benefit of nonlinear feature interactions [34].

Support Vector Machine

A margin-based classifier that constructs an optimal separating hyperplane in the feature space, enabling the evaluation of nonlinear decision boundaries through kernel-based learning [34].

XSL•FO
RenderX

Random Forest

An ensemble learning method that aggregates multiple decision trees trained on bootstrapped samples, capturing nonlinear feature relationships and improving robustness against overfitting [35].

XGBoost

A gradient boosting framework that builds additive tree-based models in a stagewise manner, designed to optimize predictive performance by modeling complex nonlinear interactions among features [36].

Convolutional Neural Network

Extracts local temporal features via convolutional kernels without explicitly modeling temporal order, serving to test the importance of temporal structure [37].

Long Short-Term Memory

Captures long-term dependencies through gating mechanisms, used to compare different recurrent units for short-sequence modeling [38].

Transformer

Models relationships across time steps using self-attention, with concatenated features fed into a standard transformer encoder [39].

PatchTST

Reduces computational complexity through a "patching" strategy, improving efficiency for long-sequence prediction [40].

iTransformer

Applies attention along the feature dimension rather than the temporal dimension, better capturing inter-variable relationships in multivariate time series [41].

Mamba

Mamba uses a selective state space mechanism to model long-term dependencies with linear complexity [42].

MambaTS

An enhanced variant of Mamba that incorporates improved state space mechanisms to further boost long-sequence prediction performance [43].

Long Short-Term Memory+Transformer

A variant of CLA-Net that replaces the GRU encoder with a long short-term memory (LSTM), used to compare the effects of different recurrent units.

To ensure a fair comparison with the deep learning models, the input configuration for all traditional machine learning baselines (logistic regression, support vector machine, random forest, and XGBoost) was aligned with the dual-time-point setup of CLA-Net. Specifically, the feature vectors from the historical time point (Xt-1) and the current time point (Xt) were concatenated to form a single flattened input vector [Xt-1, Xt]. This ensures that all models use the exact same longitudinal information, preventing any performance bias due to information asymmetry.

Ethical Considerations

Ethical approval for the original data collection in the CHARLS project was granted by the Biomedical Ethics Review Committee of Peking University (approval number: IRB00001052-11015). All participants provided written informed consent at the time of enrollment. As this study involves a secondary analysis of publicly available, deidentified data, the requirement for additional informed consent was waived. To ensure privacy and confidentiality, all direct identifiers were removed from the dataset by the CHARLS team prior to public release. Furthermore, this paper and its supplementary materials do not contain any images or information that could identify individual participants.

Results

Dataset

This study adopted a longitudinal cohort design using publicly available data from the CHARLS. CHARLS is a nationally representative longitudinal survey covering 28 provinces, 150 counties and districts, and 450 communities, and provides extensive demographic, chronic disease, and other health-related information. Using 2011 as the baseline, we included respondents with at least 2 chronic conditions and complete demographic data who participated in all 5 waves, yielding 3644 individuals and 18,220 person-wave observations. The sample selection process is shown in Figure 3. Attrition during follow-up occurred due to both mortality and nonmortality loss to follow-up. Specifically, the without-follow-up observations (30,483 person-wave observations) corresponded to 1617 person-wave observations attributable to death and 28,866 person-wave observations attributable to other reasons for nonparticipation.

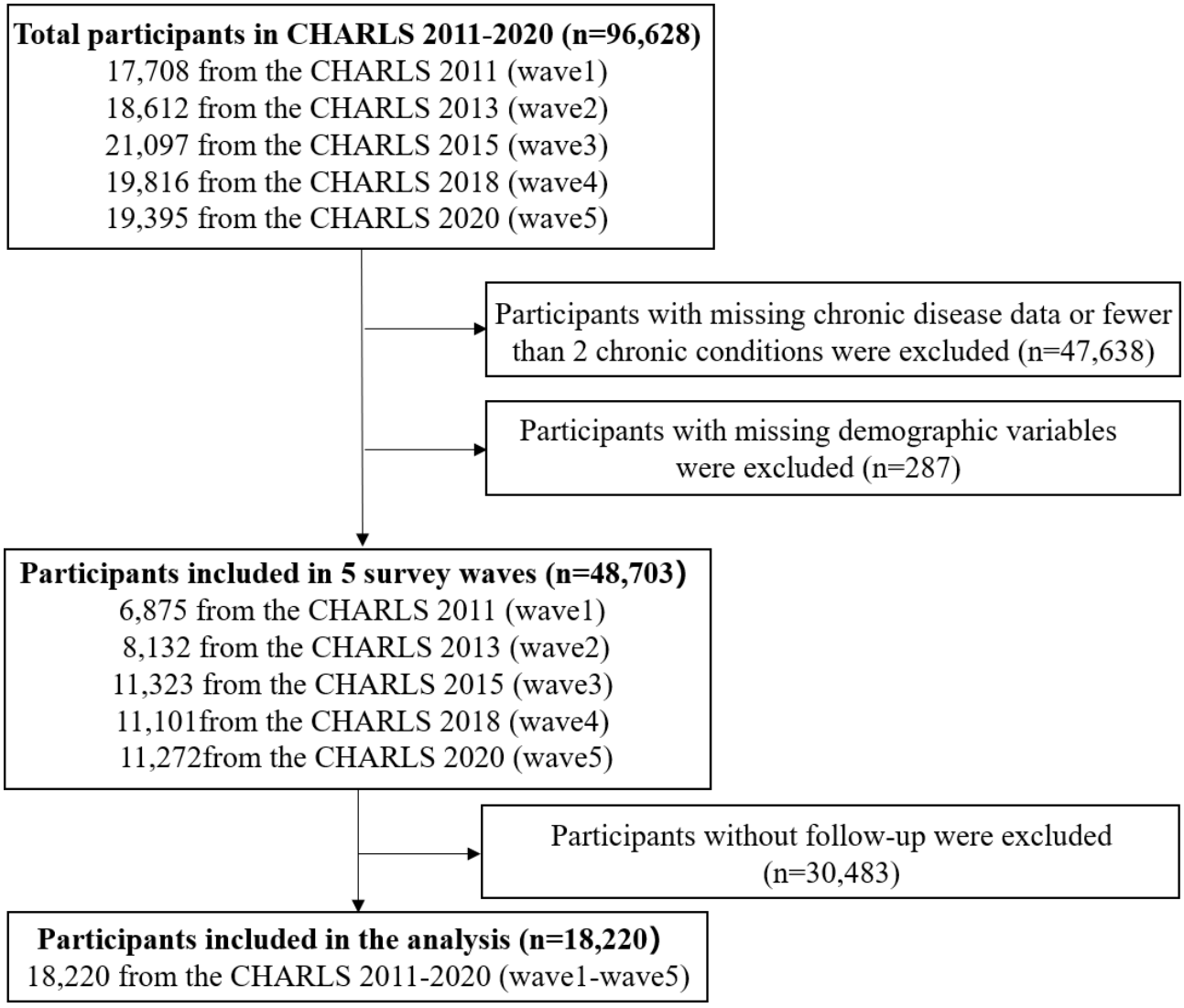
Data preprocessing included logical error correction and outlier removal. Regarding missing value treatment, to strictly prevent data leakage, we adhered to a rigorous "split-then-impute" strategy. The dataset was first partitioned into training and independent test sets. Subsequently, missing values were handled using the MiceForest method for multiple imputation [44]. Crucially, the imputation model was fitted exclusively on the training set. The MiceForest method is based on random forests and predictive mean matching, allowing it to capture nonlinear relationships and improve imputation accuracy [45,46]. To ensure data quality, variables with more than 50% missingness were excluded, and only those with relatively low missingness were retained [45].

The final dataset variables are summarized in Table S1 in Multimedia Appendix 1 and include (1) individual identifiers and survey time (ID and year); (2) 14 chronic disease variables used for LTA-based identification and transition modeling of multimorbidity patterns, which also served as target variables in the prediction task; and (3) symptom variables along with demographic, behavioral, psychological, socioeconomic, and macropolicy features, which were used as input features for the prediction model. A final cohort of 3644 individuals with multimorbidity who maintained participation across all 5 survey waves was included in the analysis. Table 2 presents the baseline sociodemographic characteristics of the participants (recorded

XSL•FO
RenderX

in 2011). The cohort was predominantly female (n=2132, 58.51%) and comprised mainly middle-aged and older adults, with 93.33% (n=3401) of participants aged between 45 and 74 years. In terms of socioeconomic status, the sample was characterized by relatively low educational attainment, as 90.12% (n=3284) of participants had an education level below junior high school. The majority were married (n=3283, 90.09%) and resided in rural areas (n=2306, 63.28%), while less than half (n=1550, 42.54%) were covered by pension insurance.

Figure 3. Participant selection flowchart. CHARLS: China Health and Retirement Longitudinal Study.



XSL•FO
RenderX

Table . Demographic characteristics.

Variables	Values
Gender, n (%)	
Men	1512 (41.49)
Women	2132 (58.51)
Age (years), n (%)	
<18	1 (0.03)
18 - 44	65 (1.78)
45 - 59	1881 (51.62)
60 - 74	1520 (41.71)
≥75	177 (4.86）
Education level, n (%)	
Below junior high school	3284 (90.12)
High school or vocational training	313 (8.59)
Higher education	47 (1.29)
Marital status, n (%)	
Unmarried	361 (9.91)
Married	3283 (90.09)
Residence, n (%)	
Rural	2306 (63.28)
Urban	1338 (36.72)
Pension coverage, n (%)	
No	2094 (57.46)
Yes	1550 (42.54)

LTA-Based Identification of Multimorbidity Patterns

To clarify the consistency and heterogeneity of multimorbidity pattern structures across different time points, we first conducted LCA at each wave to provide the basis for subsequent LTA. The model fit results (Table 3) indicated that the 5-class solution performed best across multiple criteria: at T1, T2, T3, and T5, it achieved the lowest AIC and SaBIC values with good classification quality (entropy=0.838 - 0.926); at T4, although the 4-class and 5-class solutions showed comparable entropy, the 5-class model outperformed in AIC, BIC, and SaBIC. The BLRT consistently supported the 5-class solution across all time points (all P<.001). Moreover, the class distributions of the 5-class solution were relatively stable and clinically plausible across waves. Taken together, the 5-class model was selected as the optimal solution for the LTA.

After establishing the 5-class structure at each time point, we tested measurement invariance to ensure the validity of cross-time comparisons. As shown in Table 4, constraining item-response probabilities to be equal across time did not significantly worsen model fit compared to the freely estimated model ($\Delta\chi^2_{75}$=88.7; P=.14). This nonsignificant likelihood ratio test confirmed measurement invariance, indicating that the latent classes maintained consistent meanings across all 5 waves.

XSL•FO
RenderX

Table . Latent class analysis results for multimorbidity patterns at 5 time points based on the CHARLS (China Health and Retirement Longitudinal Study) database.

Time and class		Percent (%)	AIC[a]	BIC[b]	SaBIC[c]	Entropy	BLRT[d]
T1							
	2	40.5/59.5	4196.03	4375.85	4283.70	0.720	—[e]
	3	47.8/16.5/35.8	3598.93	3871.77	3731.96	0.802	<0.001
	4	21.4/15.6/34.8/28.2	3378.95	3744.80	3557.32	0.839	<0.001
	5	11.6/15.3/39.8/9.4/23.9	3230.96	3782.83	3500.03	0.926	<0.001
	6	17.1/15.0/32.5/19.4/14.4/1.6	3293.62	3752.49	3517.35	0.856	<0.001
T2							
	2	37.7/62.3	4083.43	4263.25	4171.11	0.711	—
	3	46.0/15.3/38.6	3498.78	3771.62	3631.81	0.803	<0.001
	4	10.7/15.1/40.3/33.8	3285.65	3651.50	3464.03	0.860	<0.001
	5	11.1/13.6/39.5/2.1/33.7	3197.59	3656.45	3421.32	0.889	<0.001
	6	14.7/12.8/34.9/7.4/27.6/2.6	3167.49	3719.36	3436.56	0.878	<0.001
T3							
	2	36.6/63.4	4199.14	4378.97	4286.82	0.701	—
	3	39.5/12.7/47.8	3603.58	3876.42	3736.61	0.800	<0.001
	4	20.5/12.0/35.6/32.0	3446.95	3812.80	3625.33	0.802	<0.001
	5	5.0/9.6/33.0/24.1/28.3	3238.82	3697.68	3462.54	0.838	<0.001
	6	2.4/8.1/0.2/30.1/16.8/13.6/4.6	3218.26	3770.13	3487.33	0.884	<0.001
T4							
	2	41.3/58.7	5174.42	5354.24	5262.09	0.687	—
	3	18.4/42.2/39.4	4448.84	4721.68	4581.87	0.798	<0.001
	4	11.2/44.0/35.8/9.0	4034.00	4399.85	4212.37	0.849	<0.001
	5	8.7/9.8/29.9/26.8/24.7	3869.07	4327.93	4092.79	0.842	<0.001
	6	8.0/9.9/30.8/19.4/27.6/4.2	3832.77	4384.64	4101.84	0.834	<0.001
T5							
	2	49.0/51.0	5576.80	5756.62	5664.47	0.694	—
	3	22.7/35.5/41.8	4771.20	5044.04	4904.23	0.812	<0.001
	4	11.8/34.9/43.4/9.9	4272.68	4638.53	4451.05	0.848	<0.001
	5	9.8/12.5/32.2/37.2/8.2	4087.68	4639.55	4356.75	0.849	<0.001
	6	13.2/9.0/30.6/3.4/8.8/36	4135.51	4594.37	4359.23	0.839	<0.001

[a]AIC: Akaike Information Criterion.

[b]BIC: Bayesian Information Criterion.

[c]SaBIC: sample-size adjusted Bayesian Information Criterion.

[d]BLRT: Bootstrap Likelihood Ratio Test.

[e]Not applicable.

XSL•FO
RenderX

Table . Likelihood ratio tests for measurement invariance of the latent transition analysis model.

Model	LogLik[a]	−2LL[b]	Chi-square[c] (*df*)[d]	*P* value
M_free[e]	−97100	194,200	—[f]	—
M_invariant[g]	−97144.4	194,288.7	88.7 (75)	.14

[a] LogLik: log-likelihood.

[b]−2LL: −2 log-likelihood.

[c]$\Delta\chi^2$: chi-square difference test.

[d]Δdf: difference in degrees of freedom.

[e]M_free: model with all parameters freely estimated across time.

[f]Not applicable.

[g]M_invariant: model with item-response probabilities constrained to be equal across time points.

The LTA clustering results for the 5-class model are shown in Figure 4. Class 1 (19.8%) had the highest overall disease burden, with nearly universal hypertension (82.7%), accompanied by high prevalence of heart disease (78.1%), dyslipidemia (69.3%), gastritis (68.9%), diabetes (43.9%), arthritis (90.4%), lung disease (44.3%), and stroke (24.1%). This class represented severe multisystem involvement [47] and was labeled the "Severe Cardiometabolic-Multisystem Pattern."

Class 2 (22.3%) was characterized by extremely high probabilities of hypertension (99.7%) and arthritis (99.6%). Gastritis showed a moderate prevalence (37.5%), while dyslipidemia (27.3%) and heart disease (26.6%) were present in about one-quarter of individuals. Other conditions were rare (<13%), indicating a pattern primarily dominated by hypertension and joint disease. This class was labeled the "Hypertension-Arthritis Pattern."

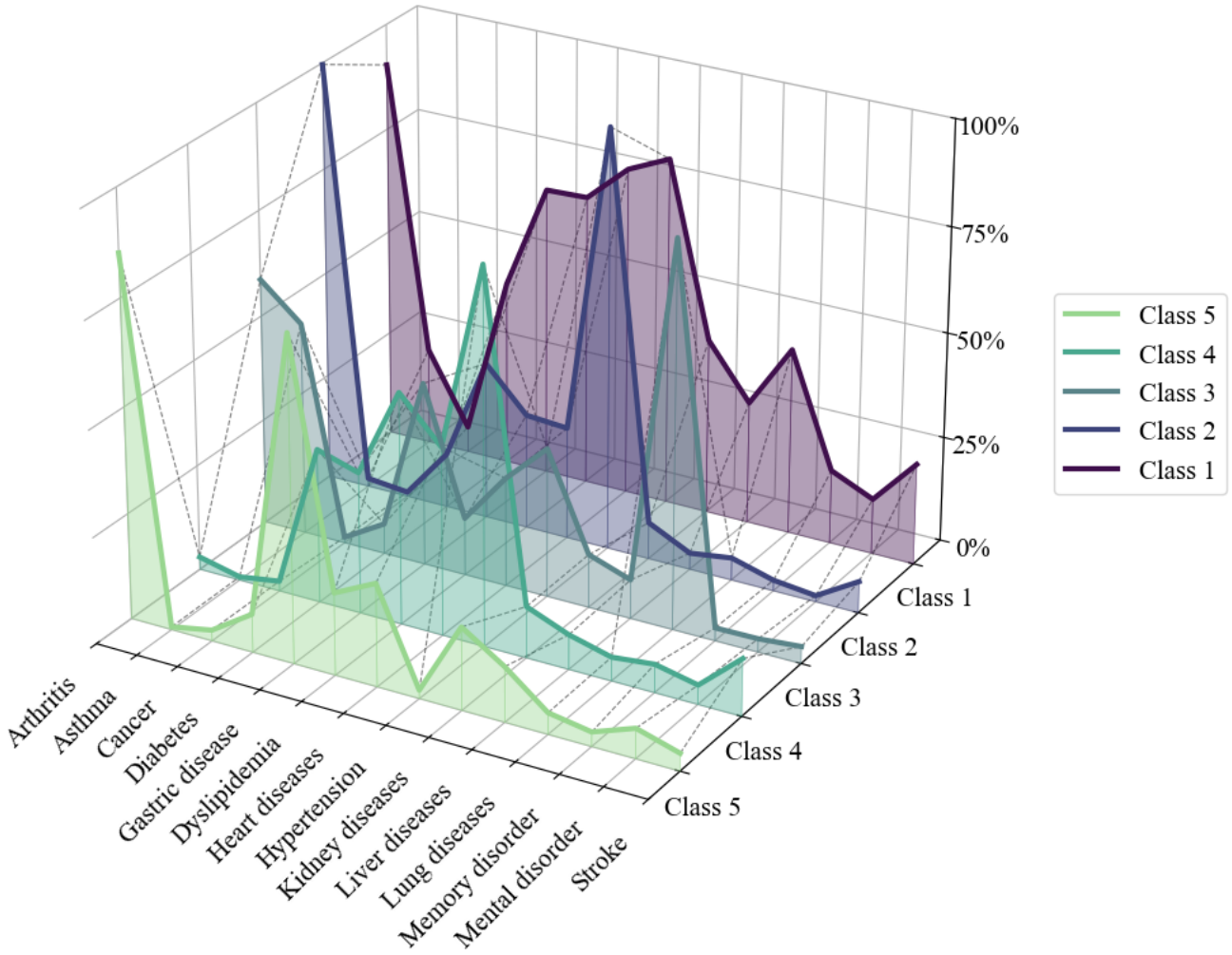
Class 3 (15.6%) was dominated by lung disease (90.8%), with co-occurrence of arthritis (58.7%), asthma (50.4%), gastritis (43.4%), and hypertension (35.3%). This class was labeled the "Respiratory-Musculoskeletal Pattern."

Class 4 (24.1%) exhibited typical metabolic syndrome features, including high prevalence of hypertension (88.3%), dyslipidemia (54.5%), and heart disease (44.4%), with moderate diabetes (36.3%) and gastritis (33.4%). Other diseases were less common (<14%), indicating a primarily cardiometabolic profile without extensive multisystem involvement. This class was labeled the "Metabolic Syndrome Pattern."

Class 5 (18.2%) had the lowest overall disease burden, dominated by arthritis (85.2%) and gastritis (76.4%), with all other diseases below 25%. This class was labeled the "Gastritis-Arthritis Pattern." The naming of latent multimorbidity patterns was guided by the dominant disease combinations and their clinical interpretation, following conventions widely used in epidemiologic and multimorbidity-cluster studies [19,48-52].

XSL•FO
RenderX

Figure 4. Disease probabilities in the 5 multimorbidity patterns.



Performance Evaluation Results of CLA-Net

In the experiments, the proposed method was trained on the training set and evaluated on an independent test set. To ensure robustness, the entire process was repeated 50 times, and the mean and SD of all evaluation metrics across these runs were reported as the final performance results. The comparative performance of all models on the multimorbidity pattern prediction task is summarized in Table 5, the ROC curves are presented in Figures 5 and 6, and the detailed results of the ablation study are shown in Table 6.

As shown in Table 5, CLA-Net consistently outperformed all baseline models in multimorbidity pattern prediction. Traditional machine learning approaches perform relatively poorly, while deep learning models designed for temporal modeling yield substantial improvements. Among the baselines, Mamba performed best overall, while convolutional neural network (CNN) performed the worst, underscoring the importance of temporal feature modeling in longitudinal health data. The hybrid LSTM+transformer model achieved relatively high accuracy, highlighting the benefits of combining sequential modeling with attention mechanisms. Nevertheless, CLA-Net surpassed all baselines across accuracy, precision, recall, and F_1-score, with low SDs confirming its stability and robustness.

We further conducted formal statistical significance testing to assess whether the observed performance gains of CLA-Net over baseline models were statistically robust. Specifically, 2-sided Wilcoxon signed-rank tests were applied to paired performance scores obtained from 50 repeated experimental runs for each evaluation metric. To account for multiple comparisons across baseline models and metrics, *P* values were adjusted using the Holm step-down procedure. The Wilcoxon signed-rank test results are given in Table 7. The results indicate that CLA-Net achieves statistically significant performance improvements over all baseline models. The consistent significance across repeated experiments confirms that the observed gains are robust rather than driven by random fluctuations.

XSL•FO
RenderX

Table . The results of the performance comparison.[a]

Types of models and models	Precision	Recall	F_1-score	Accuracy
Machine learning models, mean (SD)				
Logistic regression	0.7385 (0.0075)	0.7350 (0.0080)	0.7320 (0.0082)	0.7335 (0.0079)
SVM[b]	0.7540 (0.0069)	0.7510 (0.0074)	0.7480 (0.0076)	0.7495 (0.0072)
Random forest	0.7650 (0.0062)	0.7620 (0.0068)	0.7590 (0.0070)	0.7605 (0.0066)
XGBoost	0.7856 (0.0059)	0.7820 (0.0062)	0.7835 (0.0060)	0.7892 (0.0052)
Deep learning models, mean (SD)				
CNN[c]	0.7835 (0.0090)	0.7779 (0.0084)	0.7807 (0.0086)	0.7860 (0.0082)
LSTM[d]	0.8020 (0.0063)	0.8040 (0.0070)	0.8030 (0.0066)	0.8041 (0.0069)
Transformer	0.8138 (0.0079)	0.8074 (0.0076)	0.8106 (0.0072)	0.8120 (0.0069)
iTransformer	0.8167 (0.0074)	0.8189 (0.0078)	0.8178 (0.0071)	0.8184 (0.0070)
PatchTST	0.8092 (0.0082)	0.8057 (0.0088)	0.8074 (0.0081)	0.8088 (0.0083)
MambaTS	0.8191 (0.0061)	0.8214 (0.0067)	0.8202 (0.0062)	0.8210 (0.0064)
Mamba	0.8226 (0.0062)	0.8258 (0.0068)	0.8242 (0.0061)	0.8242 (0.0064)
LSTM + Transformer	0.8234 (0.0060)	0.8256 (0.0065)	0.8245 (0.0062)	0.8247 (0.0058)
CLA-Net (our model)	*0.8326 (0.0053)*	*0.8312 (0.0056)*	*0.8319 (0.0051)*	*0.8352 (0.0048)*

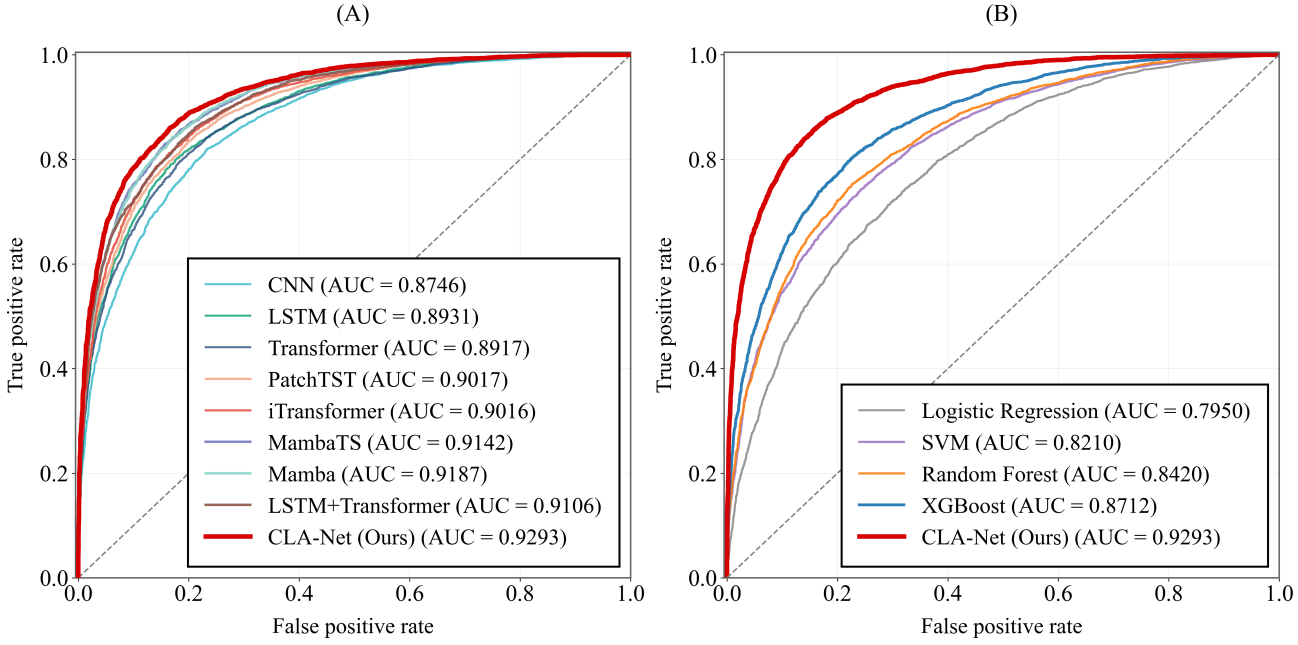
[a]The italicized values represent the best performance of each data handling strategy on the evaluation metrics. These italicized values are used to highlight the most outstanding results among the different strategies.

[b]SVM: support vector machine.

[c]CNN: convolutional neural network.

[d]LSTM: long short-term memory.

Figure 5. The area under the receiver operating characteristic curve of the Cross-Lag Attention Network and baseline models. AUC: area under the curve; CLA-Net: Cross-Lag Attention Network; CNN: convolutional neural network; LSTM: long short-term memory; SVM: support vector machine.



XSL•FO
RenderX

Figure 6. The area under the precision-recall curve of the Cross-Lag Attention Network and baseline models. AU -PRC: area under the precision-recall curve; CLA-Net: Cross-Lag Attention Network; CNN: convolutional neural network; LSTM: long short-term memory; SVM: support vector machine.

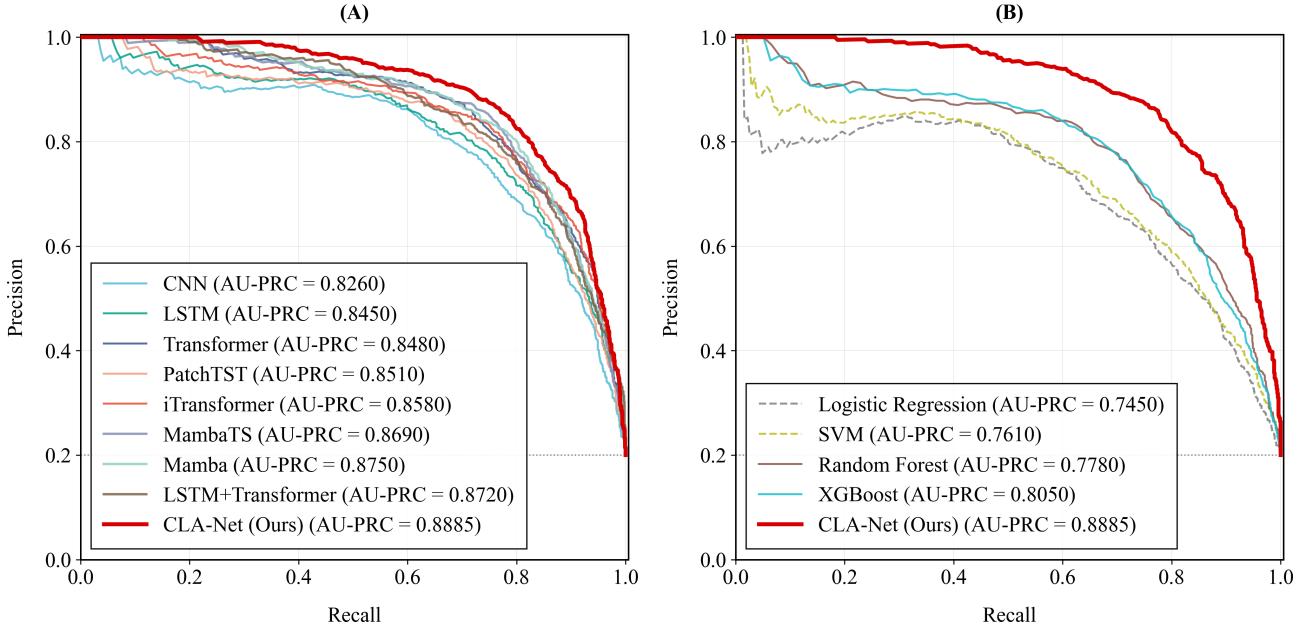


Table . Results of the ablation study.[a]

Type and ablation setup		Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)	Accuracy, mean (SD)
Full model					
	CLA-Net[b] (our model)	*0.8326 (0.0053)*	*0.8312 (0.0056)*	*0.8319 (0.0051)*	*0.8352 (0.0048)*
Module ablation					
	Transformer branch	0.8153 (0.0068)	0.8116 (0.0071)	0.8134 (0.0069)	0.8166 (0.0065)
	GRU[c] branch	0.8079 (0.0074)	0.8067 (0.0078)	0.8073 (0.0075)	0.8102 (0.0072)
	Replacing the bitemporal directed cross-attention mechanism with self-attention	0.8218 (0.0059)	0.8196 (0.0062)	0.8207 (0.0060)	0.8244 (0.0057)
Input configuration ablation					
	Single-time-point input	0.7982 (0.0083)	0.7964 (0.0087)	0.7973 (0.0085)	0.7985 (0.0082)
	Three-time-point input	*0.8338 (0.0055)*	0.8265 (0.0058)	0.8301 (0.0056)	0.8311 (0.0053)
	Four-time-point input	0.8251 (0.0057)	*0.8316 (0.0061)*	0.8283 (0.0059)	0.8287 (0.0056)
Architecture design					
	Reversing the order of transformer and GRU components	0.8230 (0.0061)	0.8208 (0.0064)	0.8219 (0.0062)	0.8235 (0.0059)

[a]The italicized values represent the best performance of each data handling strategy on the evaluation metrics.

[b]CLA-Net: Cross-Lag Attention Network.

[c]GRU: Gated Recurrent Unit.

XSL•FO
RenderX

Table . Statistical test results.[a]

Baseline	Precision *W*	Recall *W*	F_1-score *W*	Accuracy *W*
Logistic regression	0***	0***	0***	0***
SVM[b]	0***	0***	0***	0***
Random forest	0***	0***	0***	0***
XGBoost	0***	0***	0***	0***
CNN[c]	0***	0***	0***	0***
LSTM[d]	0***	0***	0***	0***
Transformer	1***	0***	0***	1***
iTransformer	13***	68***	20***	15***
PatchTST	0***	0***	0***	0***
MambaTS	27***	82***	39***	0***
Mamba	52***	258***	101***	11***
LSTM+transformer	80***	227***	126***	11***

[a]Two-sided Wilcoxon signed-rank tests were conducted on paired metric scores obtained from 50 repeated runs (n=50) to compare Cross-Lag Attention Network against each baseline model; Reported W corresponds to the Wilcoxon test statistic; To control for multiple comparisons across 12 baselines and 4 metrics, *P* values were adjusted using the Holm procedure; ***: $P<.001$.

[b]SVM: support vector machine.

[c]CNN: convolutional neural network.

[d]LSTM: long short-term memory.

Figure 5 presents the AUC-ROC curves comparing CLA-Net with a range of baseline models. As shown in Figure 5A, among deep learning approaches, CLA-Net achieves the highest AUC of 0.9293, demonstrating superior discriminative ability in multimorbidity pattern prediction. Models explicitly designed for temporal modeling, such as Mamba (AUC=0.9187), MambaTS (AUC=0.9142), PatchTST (AUC=0.9017), and iTransformer (AUC=0.9016), outperform conventional CNN-based architectures, highlighting the importance of capturing long-range temporal dependencies in longitudinal health data. The hybrid LSTM+transformer model attains an AUC of 0.9106, ranking among the strongest deep learning baselines, while CNN yields the lowest AUC (0.8746) within this group.

To further assess performance under class imbalance, Figure 6 reports the PR curves and corresponding AUC-PRC values. As shown in Figure 6, CLA-Net achieves the highest AUC-PRC among deep learning models (0.8885), indicating a superior PR trade-off across recall levels. Temporal models such as Mamba, iTransformer, and LSTM+transformer also perform competitively, whereas CNN shows relatively weaker performance. Figure 6 demonstrates that CLA-Net substantially outperforms traditional machine learning baselines, including logistic regression, support vector machine, random forest, and XGBoost. Overall, the PRC results complement the ROC analysis and further confirm the robustness of CLA-Net in identifying future multimorbidity patterns under imbalanced conditions.

As summarized in Table 6, the full CLA-Net consistently outperformed all ablation variants, underscoring the importance of integrating both GRU and the bitemporal directed cross-attention mechanism. Removing either component or replacing cross-attention with self-attention led to clear performance drops, confirming the effectiveness of the original design.

Regarding input configurations, single-time-point input resulted in the largest decline, while 3-time-point and 4-time-point settings provided only marginal gains but reduced overall accuracy, indicating that the 2-time-point design achieves the optimal balance, efficiently capturing the most relevant temporal signals without introducing the complexity or noise associated with longer historical windows. These results indicate that, under the current architecture, incorporating immediate temporal context is critical for prediction, whereas extending the input window beyond 2 time points yields diminishing or even negative returns. In addition, reversing the order of transformer and GRU also impaired performance, supporting the strategy of “temporal encoding before interaction.”

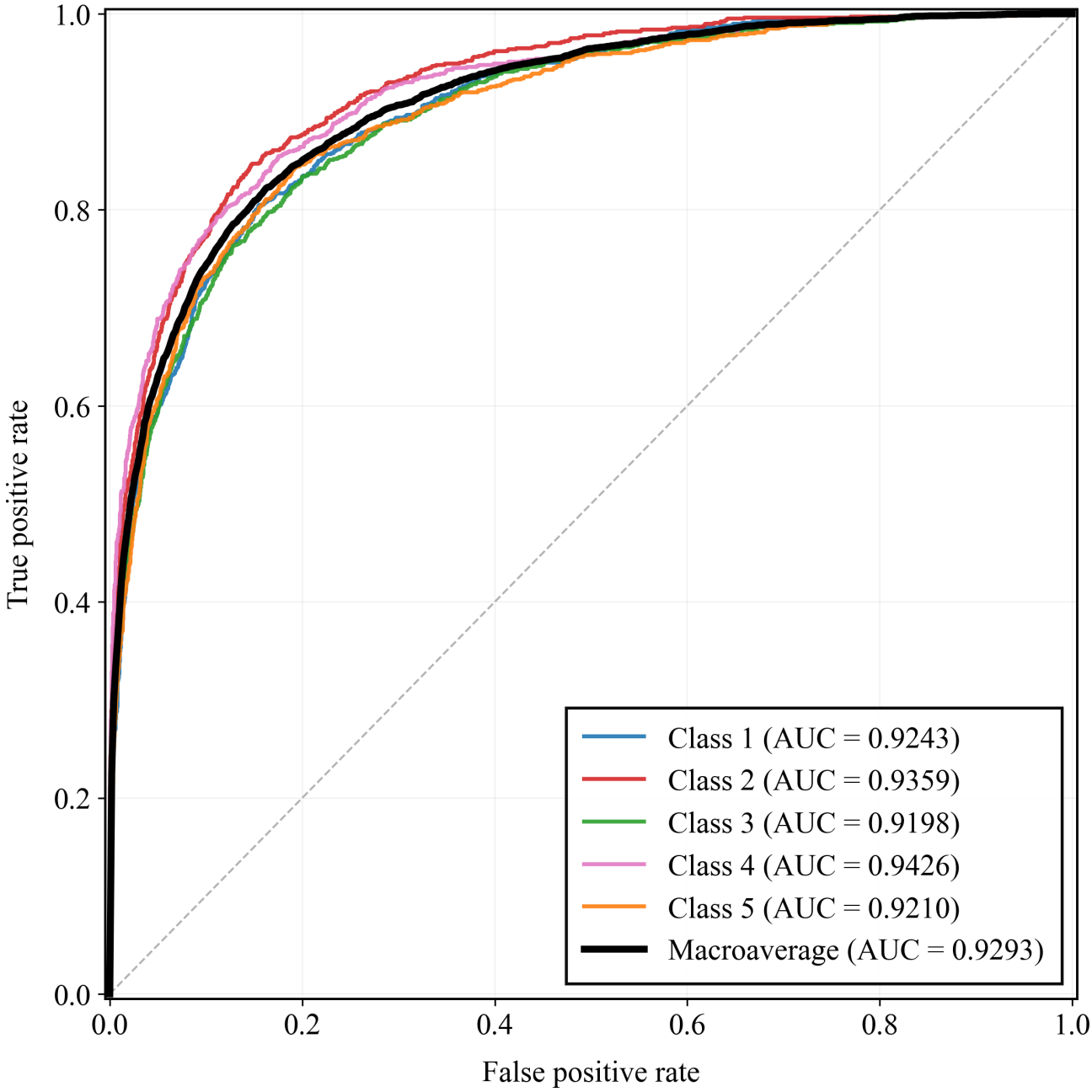
Figures 7 and 8 show the class-specific ROC and PR curves for the 5 multimorbidity patterns. All classes achieved consistently high ROC performance, with AUCs ranging from 0.9198 to 0.9426 and a macroaverage of 0.9293. The PRC analysis yielded AUC-PRC values between 0.8310 and 0.8620 (macroaverage=0.8429), indicating stable precision-recall trade-offs across classes. Overall, CLA-Net demonstrates robust and balanced predictive performance for all multimorbidity patterns.

To complement this threshold-free evaluation, Table 8 reports classwise precision, recall, and F_1-scores, indicating consistently good predictive performance across patterns, with F_1-scores ranging from 0.7906 to 0.8508. As further illustrated by the confusion matrix in Figure 9, classification errors are primarily

XSL•FO
RenderX

concentrated among clinically related multimorbidity patterns, suggesting that residual misclassifications arise from intrinsic overlap between multimorbidity profiles rather than systematic model failure.

Figure 7. Multiclass area under the receiver operating characteristic curve of the Cross-Lag Attention Network. AUC: area under the curve.



XSL•FO
RenderX

Figure 8. Multiclass area under the precision-recall curve of Cross-Lag Attention Network. AU-PRC: area under the precision-recall curve.

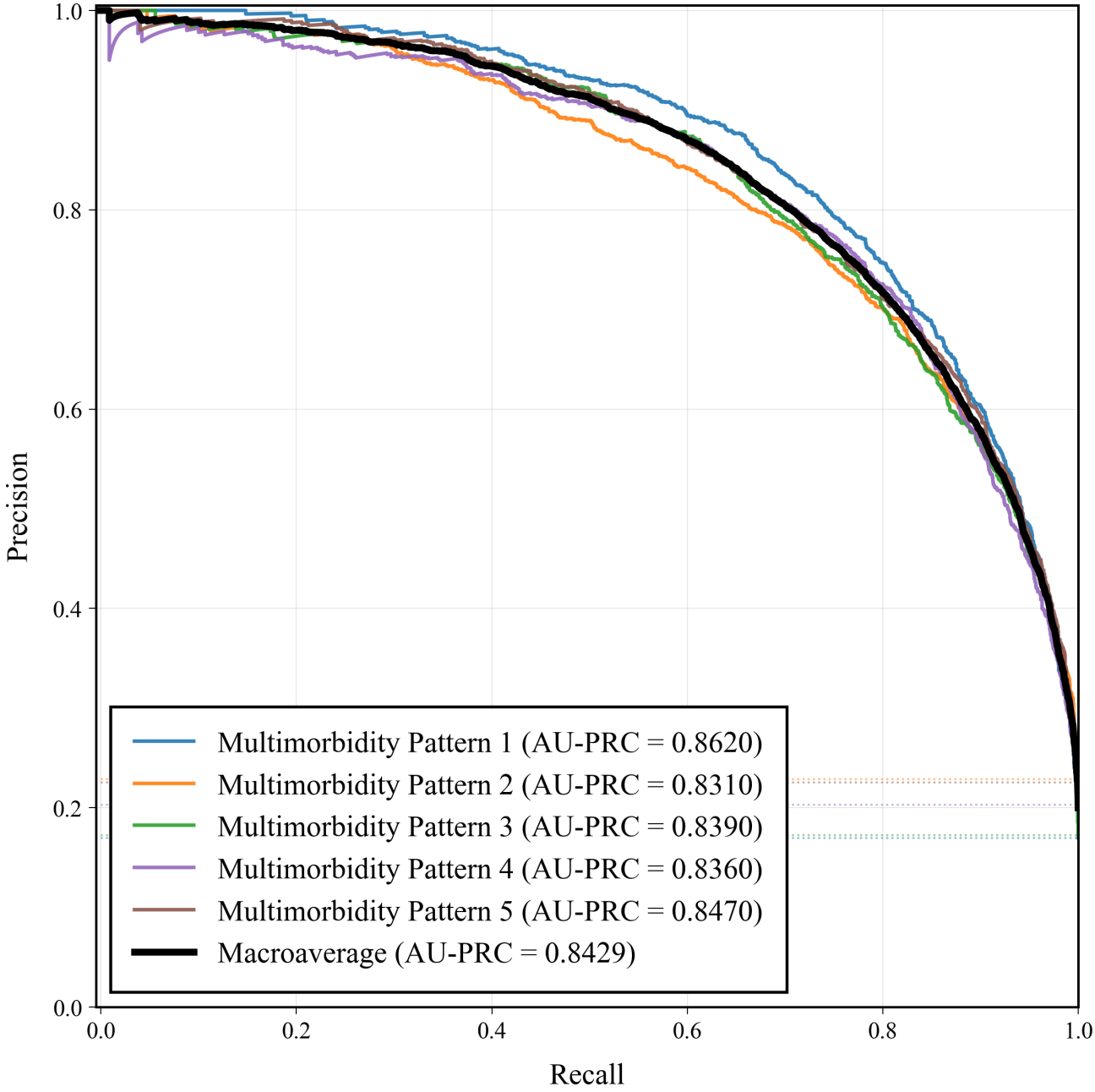


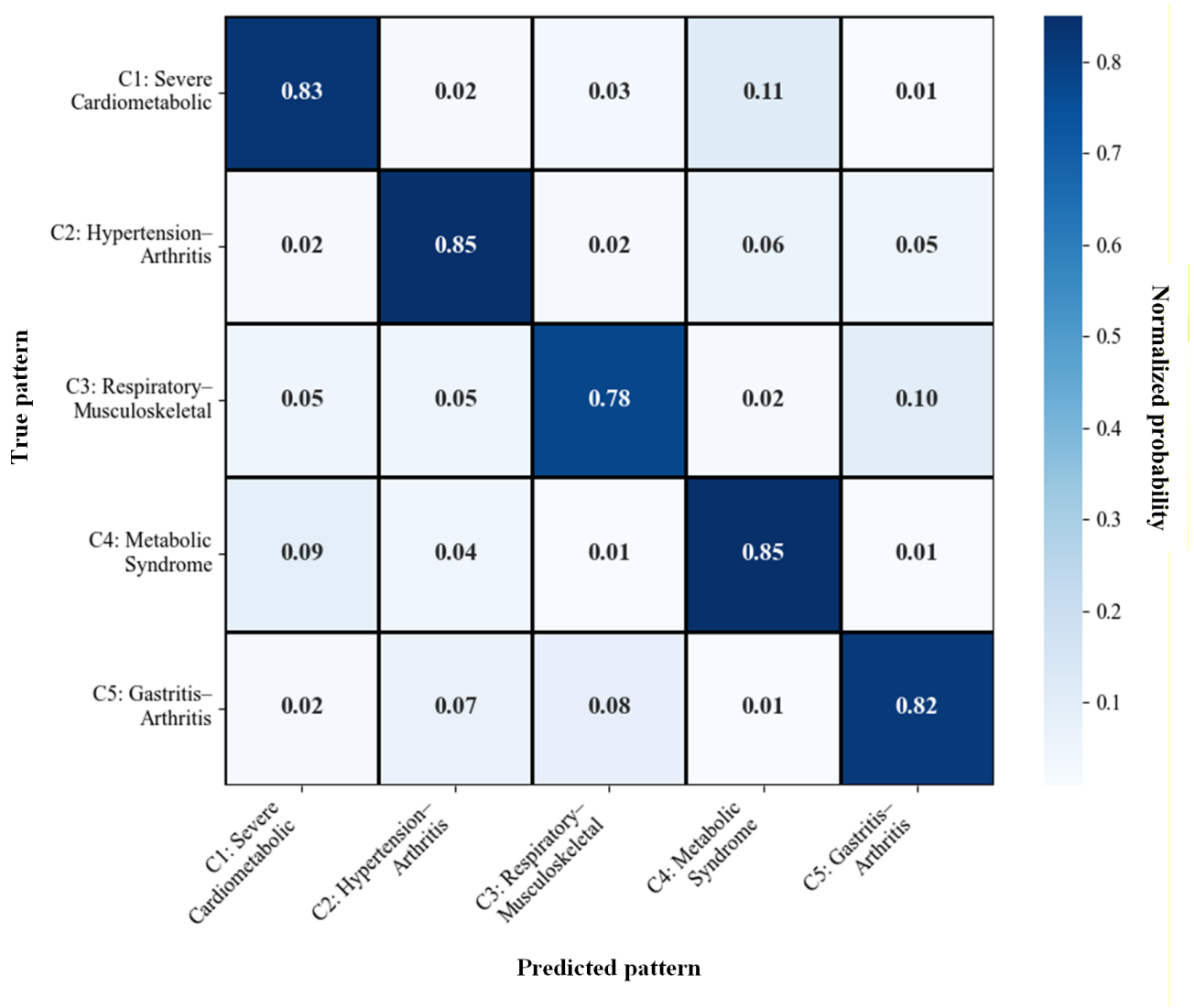
Table . Detailed classwise performance metrics of Cross-Lag Attention Network.

Multimorbidity pattern (class)	Prevalence (%)	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)
C1: Severe Cardiometabolic-Multisystem	19.80	0.8403 (0.006)	0.8314 (0.007)	0.8359 (0.006)
C2: Hypertension-Arthritis	22.30	0.8402 (0.005)	0.8515 (0.006)	0.8458 (0.005)
C3: Respiratory-Musculoskeletal	15.60	0.8002 (0.009)	0.7813 (0.010)	0.7906 (0.009)
C4: Metabolic Syndrome	24.10	0.8480 (0.006)	0.8535 (0.006)	0.8508 (0.006)
C5: Gastritis-Arthritis	18.20	0.8221 (0.007)	0.8194 (0.008)	0.8207 (0.007)
Macroaverage	—[a]	0.8303 (0.007)	0.8274 (0.008)	0.8288 (0.007)
Weighted average (overall)	100	0.8326 (0.006)	0.8312 (0.007)	0.8319 (0.006)

[a]Not applicable.

XSL•FO
RenderX

Figure 9. Confusion matrix across the 5 multimorbidity patterns.



Sensitivity Analysis: Robustness of Class Stability

To address concerns regarding potential information leakage and the temporal robustness of latent class definitions, we conducted a comprehensive sensitivity analysis in which the LTA was reestimated using only the training-period waves (2011 - 2015), thereby excluding all future observations used as prediction targets in the main analysis. This analysis was designed to evaluate whether the multimorbidity pattern structure identified in the full longitudinal dataset was inherently stable, rather than being driven by information from later waves.

We first reassessed the optimal number of latent classes by fitting LTA models with 2 to 6 classes under a free-parameter specification using the training-period data only. As shown in Table 9, information criteria (AIC, BIC, and SaBIC) consistently improved as the number of classes increased from 2 to 5, while a marginal deterioration in fit was observed when moving from 5 to 6 classes. The BLRT strongly supported the 5-class solution over the 4-class solution (P<.001), whereas the improvement from 5 to 6 classes was only marginal (P=.03), suggesting diminishing returns with increased model complexity.

XSL•FO
RenderX

Table . Model fit of latent transition analysis based on training-period waves (2011 - 2015).

Models (waves used) and number of classes	AIC[a]	BIC[b]	SaBIC[c]	Entropy	BLRT[d] (*P* value)
LTA[e] (free parameters; 2011-2015)					
2	6128.4	6316.2	6201.7	0.72	—[f]
3	5486.9	5759.8	5594.3	0.8	<.001
4	5142.6	5500.7	5284.9	0.84	<.001
5	4926.8	5369.9	5104.2	0.87	<.001
6	4954.1	5482.4	5166.5	0.85	.03
Selected model (2011-2015)					
5	4926.8	5369.9	5104.2	0.87	—
LTA (invariant)					
5	4968.3	5389.6	5136.8	0.86	—

[a]AIC: Akaike Information Criterion.

[b]BIC: Bayesian Information Criterion.

[c]SaBIC: sample-size adjusted Bayesian Information Criterion.

[d]BLRT: Bootstrap Likelihood Ratio Test.

[e]LTA: latent transition analysis.

[f]Not applicable.

In addition, classification quality, as reflected by entropy, reached its highest value under the 5-class solution (entropy=0.87), indicating satisfactory class separation despite the reduced number of waves and observations compared with the full-sample model. Taken together, these results demonstrate that the 5-class structure remained the optimal and most parsimonious representation of multimorbidity patterns even when the analysis was restricted to the training-period data.

To ensure consistency of class interpretation across time, measurement invariance constraints were imposed on the selected 5-class model. Although imposing invariance resulted in a modest increase in information criteria and a slight reduction in entropy (from 0.87 to 0.86), the overall classification quality remained high, indicating that the core latent structure was robust to parameter constraints and not dependent on future observations.

Beyond overall model fit, we directly examined the stability of class definitions by comparing disease-specific item-response probabilities between the full-sample LTA model and the training-period-only LTA model. For each latent class, the top 4 representative chronic conditions were retained, and their conditional probabilities were contrasted across the 2 modeling strategies (Table 10).

XSL•FO
RenderX

Table . Concordance of class definitions between full-sample latent transition analysis and training-only latent transition analysis.

Multimorbidity pattern (class) and top 4 representative conditions[a]	Probability (full model)[b]	Probability (training period only)[c]	Difference (bias)	Pearson correlation coefficient (r)[d]
Class 1: Cardiometabolic-Multisystem				0.954
Arthritis	0.904	0.774	−0.130	
Hypertension	0.827	0.854	0.027	
Heart disease	0.781	0.781	0.000	
Dyslipidemia	0.693	0.662	−0.031	
Class 2: Hypertension-Arthritis				0.983
Hypertension	0.997	0.917	−0.080	
Arthritis	0.996	0.958	−0.038	
Gastric disease	0.374	0.446	0.072	
Dyslipidemia	0.273	0.331	0.058	
Class 3: Respiratory-Musculoskeletal				0.970
Lung disease	0.908	0.783	−0.125	
Arthritis	0.587	0.604	0.017	
Asthma	0.504	0.540	0.036	
Gastric disease	0.434	0.466	0.032	
Class 4: Metabolic Syndrome				0.973
Hypertension	0.883	0.841	−0.042	
Dyslipidemia	0.545	0.646	0.101	
Heart disease	0.444	0.346	−0.098	
Diabetes	0.363	0.308	−0.055	
Class 5: Gastritis-Arthritis				0.965
Arthritis	0.852	0.892	0.040	
Gastric disease	0.764	0.818	0.054	
Heart disease	0.242	0.302	0.060	
Kidney disease	0.194	0.253	0.059	
Overall stability				
Average across all classes	—[g]	—	0.048[e]	0.969[f]

[a]To ensure consistency, the top 4 conditions with the highest prevalence in the full model were selected for comparison for each pattern.

[b]Derived from the complete dataset, including Waves 1 through 5 (2011 - 2020).

[c]Derived exclusively from the training dataset (Waves 1 through 3, 2011 - 2015), blinding the model to future test data.

[d]Pearson correlation coefficient calculated based on the probability vectors of all 14 chronic conditions for the corresponding class.

[e]Mean absolute deviation value.

[f]Mean value.

[g]Not applicable.

Across all 5 classes, the composition and relative ordering of dominant conditions remained highly consistent. The cardiometabolic-multisystem class continued to be characterized by a high prevalence of arthritis, hypertension, heart disease, and dyslipidemia; the hypertension-arthritis class preserved its defining dual dominance of hypertension and arthritis; the respiratory-musculoskeletal class remained anchored by lung disease and asthma; the metabolic syndrome class retained its core metabolic features; and the gastritis-arthritis class continued to reflect co-occurring gastrointestinal and musculoskeletal conditions. Importantly, no class exhibited a change in its defining disease constellation when future waves were excluded from model estimation.

Quantitatively, the similarity between item-response probability profiles was high across all classes, with Pearson correlation coefficients ranging from 0.954 to 0.983 and a mean correlation

XSL•FO
RenderX

of 0.969. These correlations were computed using the full set of disease-specific probabilities, indicating strong structural concordance rather than superficial agreement limited to a subset of conditions [53]. Differences in absolute probabilities were generally modest, with a mean absolute deviation of 0.048 across classes, reflecting only minor shifts in disease prevalence rather than substantive changes in class meaning.

Taken together, these findings provide strong evidence that the multimorbidity class structure identified in the main analysis is stable and not driven by information from future observations. Even when latent classes were defined exclusively using training-period data, both the optimal number of classes and their clinical interpretations remained highly consistent with those derived from the full-sample model. The observed differences were limited to modest shifts in absolute disease probabilities, while the underlying pattern structure and relative disease importance were preserved.

We acknowledge that, under a strictly unbiased evaluation framework, latent class definitions could alternatively be derived exclusively from training-period data and then applied to the test set, for example, by projecting test observations onto a training-derived latent space. In this study, however, we adopted a population-level LTA fitted on all available waves to define stable multimorbidity patterns, which were subsequently treated as fixed targets for prediction. Importantly, as demonstrated in Table 9, the latent class structures derived from training-period data and from the full dataset exhibited a high degree of concordance (mean r=0.969), indicating that the class definitions were not driven by any single wave or subset of the data. Therefore, the use of the full-sample LTA model to generate predictive targets does not introduce meaningful bias into the evaluation. The "ground truth" labels for the test set would remain virtually unchanged even if they were generated by projecting test data onto a latent space fitted strictly to training parameters. The reported predictive performance of CLA-Net reflects a genuine ability to forecast future multimorbidity states rather than an artifact of information leakage in the target generation process.

To explicitly assess whether the use of full-sample LTA targets introduces any optimistic bias in predictive evaluation, we conducted an additional experiment in which the LTA model was estimated exclusively on training-period data, with all model parameters fixed thereafter. The downstream prediction pipeline was then fully repeated using these training-only LTA-derived targets, and the CLA-Net model was retrained and reevaluated accordingly.

As summarized in Table 11, the predictive performance obtained under training-only LTA targets remains highly consistent with the original results based on full-sample LTA targets. Specifically, the differences in accuracy, precision, recall, and F_1-score are all below 0.3%, and all metrics remain well within 1 SD of the original estimates. The slight performance reductions observed are expected and can be attributed to minor numerical variations in latent class assignment, rather than to any systematic bias arising from information leakage.

Table . Comparison of predictive performance under full-sample and training-only LTA[a] targets.

Metric	Full-sample LTA targets, mean (SD)	Training-only LTA targets, mean (SD)	Difference
Accuracy	0.8352 (0.0048)	0.8338 (0.0050)	−0.0014 (−0.17%)
Precision	0.8326 (0.0053)	0.8315 (0.0055)	−0.0011 (−0.13%)
Recall	0.8312 (0.0056)	0.8290 (0.0060)	−0.0022 (−0.26%)
F_1-score	0.8319 (0.0051)	0.8302 (0.0054)	−0.0017 (−0.20%)

[a]LTA: latent transition analysis.

Transition-Focused Evaluation of Multimorbidity Class Prediction

To disentangle the prediction of genuine multimorbidity progression from trivial class persistence, we conducted a stratified evaluation by separating individuals into those whose multimorbidity class remained unchanged between consecutive waves (stayer-only) and those who experienced a true class transition (transition-only). Performance was assessed using macroaveraged precision, recall, F_1-score, and overall accuracy for each subset.

To further account for chance agreement and the high prevalence of class persistence, we additionally report chance-adjusted performance metrics. Across the full sample, CLA-Net achieved a Cohen κ of 0.7935 (SD 0.0055) and a Matthews correlation coefficient of 0.7968 (SD 0.0058), reflecting substantial chance-adjusted concordance between predicted and observed class labels. These values are consistent with the maintained performance on the transition-only subset, confirming that CLA-Net possesses robust discriminative power to identify genuine multimorbidity evolution rather than merely replicating static baselines.

As shown in Table 12, the model retained substantial discriminative ability on the transition-only subset, where individuals experienced a true change in multimorbidity class between tand t+1. On this subset, macroaveraged precision, recall, and F_1-score reached 0.7615 (SD 0.0106), 0.7632 (SD 0.0112), and 0.7623 (SD 0.0102), respectively, with an overall accuracy of 0.7654 (SD 0.0096). Although performance on the transition-only subset was lower than that observed in the full-sample and stayer-only evaluations, this reduction is expected given the increased complexity of predicting off-diagonal class transitions. Because multimorbidity patterns tend to remain relatively stable across adjacent waves in populations with largely irreversible chronic conditions, the transition-only subset represents a smaller and more

XSL•FO
RenderX

heterogeneous group, which inherently increases prediction difficulty and performance variability, as reflected by the larger SDs observed.

Table . Performance evaluation on different dynamic subsets.

Evaluation subset	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)	Accuracy, mean (SD)
Full sample	0.8326 (0.0053)	0.8312 (0.0056)	0.8319 (0.0051)	0.8352 (0.0048)
Transition-only ($Y_t \neq Y_{t+1}$; 23.47%)	0.7615 (0.0106)	0.7632 (0.0112)	0.7623 (0.0102)	0.7654 (0.0096)
Stayer-only ($Y_t = Y_{t+1}$; 76.53%)	0.8544 (0.0049)	0.8521 (0.0052)	0.8533 (0.0047)	0.8566 (0.0044)

As an important clinical benchmark, we further evaluated a naive persistence model that simply predicts class($t+1$)=class(t). On the test set, the naive model obtained an overall accuracy of 0.7653, substantially below CLA-Net (0.8352, SD 0.0048). Notably, on the transition-only subset, the naive model achieves 0% accuracy by definition, whereas CLA-Net maintains 0.7654 accuracy, demonstrating its added value in identifying genuine transitions rather than merely reproducing stable states. Together, these results indicate that the overall performance gain is not driven by trivial persistence, but by the model's capacity to capture progression-related signals associated with changes in multimorbidity structure.

To further account for chance agreement and the high prevalence of class persistence, we additionally report chance-adjusted performance metrics. Across the full sample, CLA-Net achieved a Cohen κ of 0.7935 (SD 0.0055) and a Matthews correlation coefficient of 0.7968 (SD 0.0058), reflecting substantial chance-adjusted concordance between predicted and observed class labels. These values are consistent with the maintained performance on the transition-only subset, confirming that CLA-Net possesses robust discriminative power to identify genuine multimorbidity evolution rather than merely replicating static baselines.

Sensitivity Analysis of Data Imputation Methods

To strictly verify the validity and robustness of the imputation strategy used in this study (ie, MiceForest), we conducted a comparative experiment to assess how different missing data handling methods affect the final predictive performance of CLA-Net. In this experiment, the standard MiceForest algorithm was replaced by 4 alternative strategies: complete case analysis (CCA; directly discarding samples with missing values), random imputation (filling missing values with randomly selected observed values), mean imputation (filling with the variable mean), and KNN imputation. The comparative results are presented in Table 13.

Table . Performance comparison of Cross-Lag Attention Network using different missing data handling strategies.[a]

Imputation strategy	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)	Accuracy, mean (SD)
Complete case analysis	0.8015 (0.0072)	0.7980 (0.0078)	0.7995 (0.0075)	0.8050 (0.0070)
Random imputation	0.8080 (0.0085)	0.8050 (0.0090)	0.8065 (0.0088)	0.8110 (0.0082)
Mean imputation	0.8150 (0.0065)	0.8120 (0.0069)	0.8135 (0.0066)	0.8180 (0.0062)
KNN[b] imputation	0.8265 (0.0058)	0.8240 (0.0062)	0.8252 (0.0060)	0.8290 (0.0055)
MiceForest (ours)	*0.8326 (0.0053)*	*0.8312 (0.0056)*	*0.8319 (0.0051)*	*0.8352 (0.0048)*

[a]The italicized values represent the best performance of each data handling strategy on the evaluation metrics. These italicized values are used to highlight the most outstanding results among the different strategies.

[b]KNN: k-nearest neighbors.

As shown in Table 13, the model using MiceForest for data imputation achieved the best performance across all metrics (accuracy=0.8352, F_1-score=0.8319). In contrast, CCA resulted in the lowest performance (accuracy=0.8050). It is important to note that the CCA metrics were evaluated on a reduced test set (subset of individuals with complete data only), whereas imputation methods were evaluated on the full test set. Despite being tested on this potentially "cleaner" subset, CCA still underperformed. This confirms that the substantial reduction in training sample size caused by discarding data severely compromises the model's ability to learn robust patterns, thereby limiting its generalizability. Random imputation also performed poorly (accuracy=0.8110), primarily because randomly assigned values introduce significant noise and disrupt the true correlation structure between variables. While mean imputation (accuracy=0.8180) and KNN imputation (accuracy=0.8290) showed improvements over random methods, they still lagged behind MiceForest.

These results indicate that MiceForest, which leverages iterative random forests to model non-linear interactions, effectively preserves the underlying data structure better than stochastic (random) or distance-based (KNN) methods, confirming it as the optimal choice for our framework.

XSL•FO
RenderX

Discussion

Principal Results

This study proposes an innovative research framework that integrates population-level multimorbidity pattern recognition with individual-level future prediction, achieving a key transition in multimorbidity research from descriptive statistics to prospective prediction.

In terms of multimorbidity pattern recognition, this study identified 5 clinically meaningful classes using LTA: Cardiometabolic-Multisystem, Hypertension-Arthritis, Respiratory-Musculoskeletal, Metabolic Syndrome, and Gastritis-Arthritis. Patients in the Cardiometabolic-Multisystem pattern bore the heaviest overall disease burden, with widespread hypertension, heart disease, dyslipidemia, diabetes, gastritis, arthritis, lung disease, and kidney disease, while the coexistence of stroke further indicated advanced vascular damage. The interactions and cascading effects among these chronic conditions have been confirmed in multiple studies. For instance, cardiometabolic diseases such as hypertension, diabetes, dyslipidemia, and heart disease frequently co-occur and significantly increase the risk of cardiovascular events and mortality [50]. Inflammatory arthritis (eg, rheumatoid arthritis) has also been associated with elevated cardiovascular risk [54]. This multisystem involvement highlights the importance of comprehensive and integrated management for such patients. The Hypertension-Arthritis pattern was characterized by the strong co-occurrence of hypertension and arthritis. This pattern is well supported by prior studies showing that hypertension is a common multimorbidity of osteoarthritis, potentially linked through shared mechanisms such as inflammation, oxidative stress, and vascular dysfunction, with arthritis treatments possibly influencing blood pressure [51,55]. The Respiratory-Musculoskeletal pattern was defined by a high prevalence of lung disease, arthritis, and asthma, often accompanied by gastritis and hypertension. This finding is consistent with previous clinical evidence [52,56]. The Metabolic Syndrome class exhibited high rates of hypertension, dyslipidemia, heart disease, and diabetes, without significant multisystem involvement. This aligns with the classical definition of metabolic syndrome and its strong association with increased risks of cardiovascular disease and type 2 diabetes [57]. Finally, the Gastritis-Arthritis pattern showed the lowest overall disease burden, primarily involving arthritis and gastritis. This finding is supported by prior medical studies reporting comorbidities such as enteropathic arthritis in patients with inflammatory bowel disease or gastrointestinal complications related to arthritis medications [58].

In multimorbidity pattern prediction, this study introduces CLA-Net, a novel deep learning framework specifically designed to forecast multimorbidity patterns from longitudinal health data with short temporal lags. Our comprehensive evaluation demonstrated that CLA-Net consistently outperformed all baseline methods, achieving an accuracy of 0.8352 and an AUC of 0.9293. Several strong deep learning baselines included in the comparison, such as Mamba, PatchTST, and iTransformer, were originally developed for long-sequence time series forecasting and are optimized to exploit long-range temporal dependencies through mechanisms such as selective state-space modeling or patch-based representations. In the present task, the input sequence consists of 2 adjacent time points $t-1$ and t, which may limit the extent to which these architectures can fully leverage their design strengths. In contrast, CLA-Net explicitly focuses on modeling short-lag transitions through a dual-branch architecture and a bitemporal directed cross-attention mechanism, enabling more effective capture of immediate disease-state dependencies. Compared with the best-performing baseline Mamba model, CLA-Net improved accuracy by 1.10% (0.8352 vs 0.8242) and by 1.05% over the hybrid LSTM+transformer architecture (0.8352 vs 0.8247). More notably, CLA-Net increased AUC by 1.87% relative to LSTM+transformer (0.9293 vs 0.9106), indicating superior discriminative ability across decision thresholds, which is particularly important for clinical applications requiring flexible risk stratification strategies.

Our ablation experiments revealed the critical contributions of individual components and validated the soundness of the architectural design. Only the transformer branch reduced accuracy to 0.8166 (−1.86%), highlighting its essential role in capturing global cross-temporal dependencies. Retaining only the GRU branch further decreased accuracy to 0.8102 (−2.50%), indicating that while GRU effectively models temporal dynamics, the absence of global feature interactions severely limits its ability to capture complex multimorbidity patterns. The performance gap between the 2 branches (0.8166 vs 0.8102) also suggests that global interactions are slightly more important than local temporal dependencies, yet their integration is necessary to achieve optimal performance. Replacing the bitemporal directed cross-attention with standard self-attention reduced accuracy to 0.8244 (−0.93%), with precision and F_1-scores declining by 1.08% and 1.12%, respectively. This underscores the importance of the directed mechanism in maintaining predictive stability and balance. Unlike standard self-attention, which may cause “future information leakage,” the dual-window design enforces a “history-to-current” information flow that better aligns with the clinical logic of predicting future risk based on historical data.

In terms of input configuration, single-time-point input resulted in a sharp drop in accuracy to 0.7985 (−3.67%) and recall to 0.7964 (−3.48%), indicating that the absence of historical information causes the model to miss many true multimorbidity cases, likely because early signals of certain patterns can only be detected through temporal evolution. Interestingly, adding more historical time points (3 or 4) did not improve overall performance: with 3 time points, precision slightly increased (0.8355 vs 0.8334) but accuracy dropped to 0.8228 and recall to 0.8185; with 4 time points, recall reached its highest (0.8360), but accuracy (0.8268) and F_1-score (0.8262) fell below the 2-time-point configuration. This empirical finding suggests that for this specific multimorbidity prediction task, immediate history is significantly more predictive than long-term history. Including more distant waves (eg, from 6‐8 years prior) appears to introduce noise or weaker relevance that obscures the strong signals from the most recent health state transition. Therefore, the 2-time-point configuration effectively captures

XSL•FO
RenderX

the critical "immediate" evolution while minimizing interference from outdated information, thus striking the best balance between information richness and model complexity. Importantly, this temporal design also has clear implications for practical clinical use. By leveraging a "past–current" window spanning approximately 2 - 3 years to predict multimorbidity patterns 2 - 3 years into the future, CLA-Net is not intended for short-term clinical decision-making (eg, acute care), but rather for midterm risk stratification and early intervention planning. In real-world settings, such predictions could be used to identify individuals who are likely to transition into more complex or higher-burden multimorbidity patterns within the next few years, thereby providing a clinically meaningful window for intervention. Finally, reversing the order of GRU and transformer (GRU-transformer) lowered accuracy to 0.8235 (−1.17%), confirming the effectiveness of the "temporal encoding before interaction" design. GRU provides temporally aware representations that serve as more suitable inputs for subsequent attention mechanisms, enabling more effective cross-temporal interactions, whereas applying attention without temporal context may lead to suboptimal feature modeling.

Considering the chronic and partially irreversible nature of many included conditions, class persistence is expected to be high. Consistent with this, approximately 76.53% of test instances were stayers. While the naive model achieves high accuracy on stable patients, it has zero clinical utility for risk warning, as it fails to identify any patient whose health state is deteriorating or changing (0% accuracy on the transition subset). In contrast, CLA-Net successfully identifies 76.54% of the patients who undergo pattern transitions. From a clinical perspective, the primary value of a predictive model lies in its ability to provide early warnings for high-risk transitions rather than merely confirming stability. Therefore, the complexity of the deep learning approach is justified by its ability to capture these critical, nonlinear disease progressions that a simple persistence rule completely misses.

Regarding generalizability, the proposed framework is designed to be transferable to other longitudinal electronic health record (EHR) datasets. The LTA component relies on routinely collected chronic disease diagnoses, which are widely available across EHR systems, allowing population-level pattern structures to be reestimated or adapted to different health care contexts. Meanwhile, the CLA-Net architecture operates on generic temporal feature representations and does not depend on dataset-specific coding schemes. This modular design facilitates scalability to larger cohorts and longer observation periods and enables potential transfer learning scenarios in which a pretrained representation model can be fine-tuned using locally derived multimorbidity structures. These properties suggest that the framework has the potential for broader application beyond the present dataset. However, due to the inherent survivorship bias of longitudinal follow-up cohorts, the current model is primarily applicable to chronic disease management in surviving, community-dwelling populations. Importantly, it should not be applied to high-acuity clinical settings (eg, emergency or Intensive Care Unit populations) where mortality risk is a competing outcome and may dominate near-term trajectories. In such contexts, mortality-aware endpoints and competing-risk modeling would be required prior to any clinical deployment.

Practical Implications

The ability of CLA-Net to anticipate future multimorbidity pattern membership offers important implications for the longitudinal management of patients with established chronic conditions. Rather than responding to isolated disease events, clinicians can leverage pattern-based forecasts to organize care around expected configurations of coexisting conditions, thereby supporting a more coordinated and forward-looking approach to chronic disease management.

First, knowledge of an individual's likely future multimorbidity pattern can inform proactive care planning within patients with multimorbidity. For example, individuals predicted to enter the metabolic syndrome pattern may benefit from earlier emphasis on lifestyle modification, metabolic regulation, and adherence support, with the aim of slowing progression toward more complex cardiometabolic profiles. Similarly, those expected to belong to the cardiometabolic-multisystem pattern can be prioritized for closer surveillance and integrated management across cardiovascular, metabolic, and musculoskeletal domains, helping to mitigate cumulative disease burden over time.

Second, pattern-level prediction provides a practical framework for structuring multidisciplinary collaboration. Instead of relying on uniform or reactive referral pathways, health care providers may tailor multidisciplinary team composition according to the anticipated constellation of conditions. For instance, the respiratory-musculoskeletal pattern underscores the value of coordinated input from pulmonology, rheumatology, and rehabilitation services to jointly address respiratory impairment and functional limitation. Likewise, the gastritis-arthritis pattern highlights the need for alignment between gastroenterology and pain management to balance long-term anti-inflammatory therapy with gastrointestinal protection.

At the organizational level, pattern-based predictions can facilitate more efficient allocation of follow-up intensity, monitoring priorities, and supportive services for patients with multimorbidity within long-term chronic care systems. Consistent with its design scope, the proposed framework does not seek to model acute deterioration, end-stage progression, or mortality-related outcomes. Rather, it functions as a scalable decision-support layer to identify patients who may benefit from anticipatory, pattern-oriented management strategies in settings focused on sustained multimorbidity care.

Limitations and Future Direction

Although this study has made significant progress in multimorbidity pattern prediction, several areas remain for improvement. First, although CLA-Net demonstrated strong predictive performance on a longitudinal cohort, the current evaluation was conducted using a single data source. Differences in population characteristics, disease coding systems, and follow-up structures across EHR datasets may affect model behavior and warrant further investigation. Future work will focus on validating the framework across diverse EHR datasets, exploring transfer learning strategies, and extending the model to other chronic disease domains and health care systems.

XSL•FO
RenderX

Second, this study does not explicitly incorporate medication-related information into the modeling framework. Chronic conditions were therefore represented as binary indicators of presence or absence, without distinguishing between actively treated and untreated disease states. This abstraction is appropriate for predicting future multimorbidity pattern membership at the population level, but it may obscure clinically relevant differences in disease control achieved through pharmacological management. Consequently, the predicted pattern assignments should be interpreted as reflecting disease co-occurrence status rather than treatment-adjusted clinical responses. Future work integrating longitudinal medication data may further enhance the clinical granularity of multimorbidity pattern prediction.

Third, the model's predictive performance reflects a combined but functionally distinct reliance on objective disease history and subjective self-assessment measures, such as self-rated health and self-rated memory. While objective disease variables anchor predictions to diagnosed multimorbidity status, subjective assessments provide higher-level summaries of perceived health and functional burden and may contribute more strongly for individuals with borderline or heterogeneous disease profiles. Although the proposed framework demonstrates strong overall predictive performance, the relative contributions of subjective versus objective inputs cannot be fully disentangled within the current framework without dedicated interpretability analyses, such as feature attribution or attention visualization. Addressing this limitation and providing greater transparency regarding the sources of predictive power will be a key focus of future work.

Fourth, although the proposed framework demonstrates strong predictive performance, incorporating detailed interpretability analyses, such as SHAP (Shapley Additive Explanations) values or attention weight visualizations, may further enhance its practical utility. Owing to the complexity of the model architecture, such analyses are beyond the scope of this study. Future research will focus on developing tailored interpretability strategies to better elucidate the learned multimorbidity representations and support clinical interpretation.

Fifth, regarding temporal granularity, the survey waves in the CHARLS dataset are separated by nonuniform intervals, ranging from approximately 2 to 3 years. In the current CLA-Net architecture, transitions between adjacent waves are modeled as equidistant steps, without explicitly encoding the varying time intervals. This simplification may introduce temporal noise into the estimation of transition dynamics, particularly when interpreting changes across waves with heterogeneous follow-up durations. Future work could address this limitation by incorporating time-aware representations or modeling strategies explicitly designed for irregular temporal intervals.

Another methodological consideration concerns the transformation of probabilistic latent class assignments into deterministic labels for supervised prediction. In this study, posterior class probabilities obtained from the LTA were converted into hard labels using a maximum posterior probability assignment strategy. While this approach provides a clear and stable supervisory signal for training the prediction model, it inevitably discards information regarding classification uncertainty. In particular, individuals with ambiguous or borderline posterior distributions are treated as belonging entirely to a single class, which may encourage overconfident predictions. Future work could address this limitation by leveraging soft targets, such as full posterior probability vectors, within the learning objective. Incorporating probabilistic supervision through distribution-based loss functions or uncertainty-aware architectures may enable a more nuanced representation of multimorbidity structure and further improve predictive robustness.

Finally, to ensure the integrity of longitudinal trajectory modeling and the stability of LTA patterns, this study included participants present in all 5 waves (2011‐2020). We acknowledge that excluding individuals who died or dropped out due to severe illness may introduce survivorship bias, potentially skewing the dataset toward a "healthier survivor" cohort. However, this inclusion criterion was a necessary methodological trade-off to accurately capture the continuous evolution of multimorbidity patterns. It is worth noting that the primary goal of this framework is to support chronic disease management and secondary prevention in the general population, rather than predicting acute mortality risk in end-stage patients. To address this limitation, future research could incorporate advanced strategies such as joint longitudinal-survival modeling, inverse probability-of-censoring weighting, or competing-risk frameworks to explicitly account for nonrandom attrition and mortality, thereby extending the model's generalizability to high-risk clinical populations.

Conclusion

This study addresses the complex challenges of multimorbidity management in the context of population aging by proposing an innovative framework that bridges population-level pattern recognition and individual risk prediction. Using LTA, we identified multimorbidity patterns with temporal consistency and clinical stability (ie, Cardiometabolic-Multisystem, Hypertension-Arthritis, Respiratory-Musculoskeletal, Metabolic Syndrome, and Gastritis-Arthritis), which were then transformed into predictive labels to develop the CLA-Net deep learning model. CLA-Net integrates the sequential modeling capacity of GRU with the feature interaction advantages of a bitemporal directed cross-attention mechanism, achieving superior performance in capturing both temporal dependencies and complex feature interactions in chronic disease progression. Experimental results demonstrated that CLA-Net significantly outperformed existing methods across accuracy, recall, precision, F_1-score, and AUC.

Acknowledgments

This work is supported by the Beijing Logistics Informatics Research Base. We used ChatGPT-5 for language polishing.

XSL•FO
RenderX

Funding

This work was supported by the National Social Science Foundation of China (grant number 18ZDA086), the National Natural Science Foundation of China (grant number 62173025), the Fundamental Research Funds for the Central Universities (grant number 2025-JYB-XJSJJ-022), and Beijing Logistics Informatics Research Base.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QZ, WM
Data curation: QZ, WM
Formal analysis: QZ
Funding acquisition: RZ
Investigation: QZ
Methodology: QZ, WM, BZ
Project administration: RZ, WM, XZ
Resources: RZ, WM
Software: QZ
Validation: QZ, WM, BZ
Writing – original draft: QZ
Writing – review & editing: QZ, RZ, WM, BZ, XZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final dataset variables.
[DOCX File, 22 KB - medinform_v14i1e84261_app1.docx]

References

1. MacMahon S, Calverley P, Chaturvedi N, et al. Multimorbidity: a priority for global health research. : The Academy of Medical Sciences; 2018 URL: https://acmedsci.ac.uk/file-download/82222577 [accessed 2026-02-20]
2. Skou ST, Mair FS, Fortin M, et al. Multimorbidity. Nat Rev Dis Primers 2022 Jul 14;8(1):48. [doi: 10.1038/s41572-022-00376-4] [Medline: 35835758]
3. Du Y, de Bock GH, Vonk JM, et al. Lifestyle factors and incident multimorbidity related to chronic disease: a population-based cohort study. Eur J Ageing 2024 Nov 28;21(1):37. [doi: 10.1007/s10433-024-00833-x] [Medline: 39609306]
4. Nunes BP, Flores TR, Mielke GI, Thumé E, Facchini LA. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2016;67:130-138. [doi: 10.1016/j.archger.2016.07.008] [Medline: 27500661]
5. Fortin M, Lapointe L, Hudon C, Vanasse A, Ntetu AL, Maltais D. Multimorbidity and quality of life in primary care: a systematic review. Health Qual Life Outcomes 2004 Sep 20;2(1):51. [doi: 10.1186/1477-7525-2-51] [Medline: 15380021]
6. Violan C, Foguet-Boreu Q, Flores-Mateo G, et al. Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. PLoS One 2014;9(7):e102149. [doi: 10.1371/journal.pone.0102149] [Medline: 25048354]
7. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. J Clin Epidemiol 2014 Mar;67(3):254-266. [doi: 10.1016/j.jclinepi.2013.09.021] [Medline: 24472295]
8. Dong X, Ma Y, Zhang H, Wang P. Individual-level transitions between chronic disease multimorbidity clusters and the risk of five-year mortality in longitudinal cohort of Chinese middle-aged and older adults. Aging Clin Exp Res 2025 Jul 9;37(1):216. [doi: 10.1007/s40520-025-03078-5] [Medline: 40632375]
9. Muth C, Blom JW, Smith SM, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. J Intern Med 2019 Mar;285(3):272-288. [doi: 10.1111/joim.12842] [Medline: 30357955]
10. Zhang J, Zhou J, Zhou L, Ba Z. Extracting multifaceted characteristics of patients with chronic disease comorbidity: framework development using large language models. JMIR Med Inform 2025 May 15;13(1):e70096. [doi: 10.2196/70096] [Medline: 40373298]
11. Held FP, Blyth F, Gnjidic D, et al. Association rules analysis of comorbidity and multimorbidity: the concord health and aging in men project. J Gerontol A Biol Sci Med Sci 2016 May;71(5):625-631. [doi: 10.1093/gerona/glv181]

XSL•FO
RenderX

12. Hu Y, He H, Ou Q, et al. Prevalence of common chronic disease and multimorbidity patterns in Guangdong province with three typical cultures: analysis of data from the Diverse Life-Course Cohort study. Front Public Health 2023 May 4;11:1163791. [doi: 10.3389/fpubh.2023.1163791]
13. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018 Oct 1;47(5):1687-1704. [doi: 10.1093/ije/dyy134] [Medline: 30016472]
14. Zhang Z, He P, Yao H, et al. A network-based study reveals multimorbidity patterns in people with type 2 diabetes. iScience 2023 Oct;26(10):107979. [doi: 10.1016/j.isci.2023.107979]
15. Nicolet A, Assouline D, Le Pogam MA, et al. Exploring patient multimorbidity and complexity using health insurance claims data: a cluster analysis approach. JMIR Med Inform 2022 Apr 4;10(4):e34274. [doi: 10.2196/34274] [Medline: 35377334]
16. Wang X, Zhang R, Zhu X. What can we learn from multimorbidity? A deep dive from its risk patterns to the corresponding patient profiles. Decis Support Syst 2024 Nov;186:114313. [doi: 10.1016/j.dss.2024.114313]
17. Zhu J, Wu Y, Lin S, Duan S, Wang X, Fang Y. Identifying and predicting physical limitation and cognitive decline trajectory group of older adults in China: a data-driven machine learning analysis. J Affect Disord 2024 Apr 1;350:590-599. [doi: 10.1016/j.jad.2024.01.095] [Medline: 38218258]
18. Strauss VY, Jones PW, Kadam UT, Jordan KP. Distinct trajectories of multimorbidity in primary care were identified using latent class growth analysis. J Clin Epidemiol 2014 Oct;67(10):1163-1171. [doi: 10.1016/j.jclinepi.2014.06.003] [Medline: 25063556]
19. Zacarías-Pons L, Vilalta-Franch J, Turró-Garriga O, Saez M, Garre-Olmo J. Multimorbidity patterns and their related characteristics in European older adults: a longitudinal perspective. Arch Gerontol Geriatr 2021;95:104428. [doi: 10.1016/j.archger.2021.104428] [Medline: 33991948]
20. Collins LM, Lanza ST. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences: John Wiley & Sons; 2009.
21. Wang Y, Xu X, Liu J, et al. Latent transition analysis of instrumental activities of daily living in Chinese elderly: based on the 2014–2018 wave of the Chinese Longitudinal Healthy Longevity Survey. BMC Geriatr 2024 Jan 22;24(1):83. [doi: 10.1186/s12877-023-04631-5]
22. Paukner M, Ladner DP, CAPriCORN Team, Zhao L. Measuring disease burden with individual cumulative incidence in patients with cirrhosis. J Biomed Inform 2025 Sep;169:104883. [doi: 10.1016/j.jbi.2025.104883] [Medline: 40752671]
23. Zeng H, Miao C, Wang R, et al. Influence of comorbidity of chronic diseases on basic activities of daily living among older adults in China: a propensity score-matched study. Front Public Health 2024;12:1292289. [doi: 10.3389/fpubh.2024.1292289] [Medline: 38638478]
24. Uddin S, Wang S, Lu H, Khan A, Hajati F, Khushi M. Comorbidity and multimorbidity prediction of major chronic diseases using machine learning and network analytics. Expert Syst Appl 2022 Nov;205:117761. [doi: 10.1016/j.eswa.2022.117761]
25. Yang D, Kim J, Yoo J, Cha WC, Paik H. Identifying the risk of sepsis in patients with cancer using digital health care records: machine learning-based approach. JMIR Med Inform 2022 Jun 15;10(6):e37689. [doi: 10.2196/37689] [Medline: 35704364]
26. Nylund-Gibson K, Grimm R, Quirk M, Furlong M. A latent transition mixture model using the three-step specification. Struct Equ Modeling 2014 Jul 3;21(3):439-454. [doi: 10.1080/10705511.2014.915375]
27. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods 2012 Jun;17(2):228-243. [doi: 10.1037/a0027127] [Medline: 22309957]
28. Lorah J, Womack A. Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. Behav Res 2019 Feb;51(1):440-450. [doi: 10.3758/s13428-018-1188-3]
29. Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 1987 Sep;52(3):333-343 [FREE Full text] [doi: 10.1007/BF02294360]
30. Larose C, Harel O, Kordas K, Dey DK. Latent class analysis of incomplete data via an entropy-based criterion. Stat Methodol 2016 Sep;32:107-121. [doi: 10.1016/j.stamet.2016.04.004] [Medline: 27695391]
31. Park J, Kim HJ, Duffy RD. Using latent transition analysis to explore changes in decent work across time. J Couns Psychol 2025 Jan;72(1):80-91. [doi: 10.1037/cou0000773] [Medline: 39680017]
32. Dziak JJ, Lanza ST, Tan X. Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. Struct Equ Modeling 2014 Jan 1;21(4):534-552. [doi: 10.1080/10705511.2014.919819] [Medline: 25328371]
33. Rinne LF, Ye A, Jordan NC. Development of fraction comparison strategies: a latent transition analysis. Dev Psychol 2017 Apr;53(4):713-730. [doi: 10.1037/dev0000275] [Medline: 28221051]
34. Kalantar B, Pradhan B, Naghibi SA, Motevalli A, Mansor S. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomat Nat Hazards Risk 2018 Jan 1;9(1):49-69. [doi: 10.1080/19475705.2017.1407368]

XSL•FO
RenderX

35. Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 2016 Apr;114:24-31. [doi: 10.1016/j.isprsjprs.2016.01.011]
36. Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform 2020;17(6):2131-2140. [doi: 10.1109/TCBB.2019.2911071] [Medline: 30998478]
37. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278-2324. [doi: 10.1109/5.726791]
38. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
39. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017 URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2025-09-09]
40. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: long-term forecasting with transformers. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2211.14730]
41. Liu Y, Hu T, Zhang H, et al. ITransformer: inverted transformers are effective for time series forecasting. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2310.06625]
42. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2312.00752]
43. Cai X, Zhu Y, Wang X, Yao Y. MambaTS: improved selective state space models for long-term time series forecasting. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2405.16440] [Medline: 38800661]
44. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 Jan 1;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
45. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychol Methods 2021 Aug;26(4):466-485. [doi: 10.1037/met0000381] [Medline: 33507765]
46. Zhao B, Zhang R, Chen D, et al. A machine-learning-based approach for identifying diagnostic errors in electronic medical records. IEEE Trans Rel 2024 Jun;73(2):1172-1186. [doi: 10.1109/TR.2023.3330733]
47. Wang S, Li Z, Wang X, et al. Bioinformatics analysis of comorbid mechanisms between ischemic stroke and end stage renal disease. Sci Rep 2025 May 16;15(1):17060. [doi: 10.1038/s41598-025-01049-4]
48. Xi J, Miao M, Li PWC, Yu DSF. Prognostic effects of multimorbidity clusters on health outcomes in adults: a systematic review and meta-analysis. Ageing Res Rev 2025 Dec;112:102897. [doi: 10.1016/j.arr.2025.102897] [Medline: 40934974]
49. Garin N, Koyanagi A, Chatterji S, et al. Global multimorbidity patterns: a cross-sectional, population-based, multi-country study. J Gerontol A Biol Sci Med Sci 2016 Feb;71(2):205-214. [doi: 10.1093/gerona/glv128] [Medline: 26419978]
50. Canoy D, Tran J, Zottoli M, et al. Association between cardiometabolic disease multimorbidity and all-cause mortality in 2 million women and men registered in UK general practices. BMC Med 2021 Oct 28;19(1):258. [doi: 10.1186/s12916-021-02126-x] [Medline: 34706724]
51. Ching K, Houard X, Berenbaum F, Wen C. Hypertension meets osteoarthritis - revisiting the vascular aetiology hypothesis. Nat Rev Rheumatol 2021 Sep;17(9):533-549. [doi: 10.1038/s41584-021-00650-x] [Medline: 34316066]
52. Katz J, Bartels CM. Multimorbidity in rheumatoid arthritis: literature review and future directions. Curr Rheumatol Rep 2024 Jan;26(1):24-35. [doi: 10.1007/s11926-023-01121-w] [Medline: 37995046]
53. Adler J, Parmryd I. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. Cytometry A 2010 Aug;77(8):733-742. [doi: 10.1002/cyto.a.20896] [Medline: 20653013]
54. Nurmohamed MT, Heslinga M, Kitas GD. Cardiovascular comorbidity in rheumatic diseases. Nat Rev Rheumatol 2015 Dec;11(12):693-704. [doi: 10.1038/nrrheum.2015.112] [Medline: 26282082]
55. Lu K, Ma F, Yi D, Yu H, Tong L, Chen D. Molecular signaling in temporomandibular joint osteoarthritis. J Orthop Translat 2022 Jan;32:21-27. [doi: 10.1016/j.jot.2021.07.001] [Medline: 35591935]
56. Cielen N, Maes K, Gayan-Ramirez G. Musculoskeletal disorders in chronic obstructive pulmonary disease. Biomed Res Int 2014;2014:965764. [doi: 10.1155/2014/965764] [Medline: 24783225]
57. Bozkurt B, Aguilar D, Deswal A, et al. Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the American Heart Association. Circulation 2016 Dec 6;134(23):e535-e578. [doi: 10.1161/CIR.0000000000000450] [Medline: 27799274]
58. Zohar A, Cohen AD, Bitterman H, et al. Gastrointestinal comorbidities in patients with psoriatic arthritis. Clin Rheumatol 2016 Nov;35(11):2679-2684. [doi: 10.1007/s10067-016-3374-y] [Medline: 27530409]

Abbreviations

AIC: Akaike Information Criterion
AU-PRC: area under the precision-recall curve
AUC: area under the curve
BIC: Bayesian Information Criterion
BLRT: Bootstrap Likelihood Ratio Test

XSL•FO
RenderX

CCA: complete case analysis
CHARLS: China Health and Retirement Longitudinal Study
CLA-Net: Cross-Lag Attention Network
CNN: convolutional neural network
EHR: electronic health record
FFN: feed-forward network
FPR: false positive rate
GRU: Gated Recurrent Unit
KNN: k-nearest neighbors
LCA: latent class analysis
LSTM: long short-term memory
LTA: latent transition analysis
PR: precision-recall
ROC: receiver operating characteristic
SaBIc: sample-size adjusted Bayesian Information Criterion
SHAP: Shapley Additive Explanations
TPR: true positive rate

Edited by A Coristine; submitted 16.Sep.2025; peer-reviewed by J Wang, L Dong, MZ Hossain, Q Lu, W Niu, X Shi; accepted 29.Jan.2026; published 10.Mar.2026.

Please cite as:
Zhang Q, Zhang R, Ma W, Zhao B, Zhu X
Bridging Population Patterns and Individual Prediction: Framework for Prospective Multimorbidity Study
JMIR Med Inform 2026;14:e84261
URL: https://medinform.jmir.org/2026/1/e84261
doi:10.2196/84261

© Qianyao Zhang, Runtong Zhang, Weiguang Ma, Butian Zhao, Xiaomin Zhu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhanced Prediction of Atrial Fibrillation in Patients With Ischemic Stroke Through Electronic Medical Records and Text Mining: Algorithm Development and Validation

Yu-Wei Chen[1,2,3*], MD, PhD; Sheng-Feng Sung[4,5*], MD, PhD; Ya-Han Hu[6,7], PhD; Yu-Hsuan Yang[6], MBA

[1]Department of Neurology, Landseed International Hospital, Taoyuan City, Taiwan

[2]Department of Neurology, National Taiwan University Hospital, Taipei, Taiwan

[3]Center for General Education, National Central University, Taoyuan City, Taiwan

[4]Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi, Taiwan

[5]Department of Nursing, Fooyin University, Kaohsiung, Taiwan

[6]Department of Information Management, National Central University, No. 300, Zhongda Rd, Zhongli District, Taoyuan City, Taiwan

[7]Asian Institute for Impact Measurement and Management, National Central University, Taoyuan City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Ya-Han Hu, PhD
Department of Information Management, National Central University, No. 300, Zhongda Rd, Zhongli District, Taoyuan City, Taiwan

Abstract

Background: Stroke remains one of the leading causes of mortality and long-term disability worldwide. Atrial fibrillation (AF) is a major and often underdiagnosed risk factor for ischemic stroke as it is frequently asymptomatic and may remain undetected until a catastrophic cerebrovascular event occurs. The lack of timely identification and preventive treatment for AF substantially increases stroke risk. Although previous studies have proposed various predictive models for AF detection, many rely primarily on structured clinical variables and are developed using data from a single institution, which limits their generalizability and real-world applicability across different health care settings.

Objective: The objective of this study was to develop a robust and generalizable AF risk prediction model for patients with stroke using electronic medical records. By integrating structured clinical variables with features derived from unstructured clinical text, this study aimed to construct a more comprehensive representation of patient health status. Furthermore, this study emphasized systematic internal and external validation, along with calibration assessment, to evaluate model stability and generalizability across multiple hospital datasets, thereby supporting its potential use in routine clinical practice.

Methods: This study analyzed datasets from 2 hospitals in Taiwan: Landseed International Hospital (LIH), with 3988 patients, and Chia-Yi Christian Hospital (CYCH), with 5821 patients. We applied 5 feature engineering techniques to extract features from unstructured electronic medical record data, addressed data imbalance using 6 distinct resampling methods, and used 9 classification algorithms to compare model performance across both internal and external validation sets. This study identified the top 20 most important features from the best-performing models for both the LIH and CYCH datasets.

Results: The optimal predictive model for LIH was based solely on structured variables, whereas the model for CYCH achieved superior results by integrating structured variables with text-derived variables obtained from unstructured clinical notes using term frequency–inverse document frequency. Notably, feature importance analysis consistently identified the ratio of E- to A-wave velocities, left atrial size, and age as the top 3 predictive factors across both datasets, underscoring their critical role in AF risk assessment among patients with stroke.

Conclusions: This study demonstrated the development of predictive models for AF in patients with ischemic stroke. Notably, the integration of structured variables with variables derived from unstructured clinical text improved predictive performance in selected model configurations. Rigorous internal and external validation processes confirmed the superior performance of ensemble learning–based machine learning models compared with alternative algorithms, underscoring the potential of this approach for AF risk prediction.

(JMIR Med Inform 2026;14:e78117) doi:10.2196/78117

KEYWORDS

atrial fibrillation; stroke; electronic medical records; text mining; clinical decision support

XSL•FO
RenderX

Introduction

Stroke ranks as the second leading cause of death and the third leading cause of disability worldwide, with both related mortality rates and the global disease burden continuing to rise [1,2]. The most common types of recurrent stroke are acute ischemic stroke (AIS) and transient ischemic attack (TIA) [3]. Within 1 year following an ischemic stroke, the risk of recurrence exceeds 7% [4-6]. Moreover, patients who experience a recurrent stroke within 10 years face more than twice the risk of death or disability compared with those who experience a first stroke [7].

Atrial fibrillation (AF) is a common sustained or paroxysmal cardiac arrhythmia [8,9]. Treatment may often be delayed as AF can be asymptomatic or episodic, making it challenging to detect [10]. AF is closely associated with cerebrovascular disease: irregular heartbeats over time can impair efficient blood contraction and clearance, potentially leading to blood clot formation. These clots may then travel to the brain and cause a stroke if AF remains untreated, increasing the risk of acute complications [11,12]. Patients with AF have twice the risk of dying from cardiovascular disease and up to a 5-fold increased risk of stroke [13]. Consequently, AF impacts patients with stroke significantly in terms of risk, treatment, and daily care requirements [14-17]. Regarding medication, for patients with stroke and AF, oral anticoagulants (eg, factor IIa and Xa inhibitors and warfarin) have been shown to reduce mortality by one-quarter and the risk of recurrent stroke by two-thirds [12]. Comparatively, antiplatelet drugs can reduce the risk of recurrent stroke by three-fifths [18]. Newer direct oral anticoagulants such as apixaban, rivaroxaban, edoxaban, and dabigatran are considered relatively safer and more effective, with a reduced risk of major bleeding compared with warfarin [19,20]. Physicians should find more underdiagnosed patients with AF, mostly paroxysmal AF, among those with AIS and TIA to administer appropriate safe and effective treatment for stroke prevention [21].

Risk assessment scales such as CHA2DS2-VASc [22] and its 2024 updated CHA2DS2-VA version [23] are widely used as clinical tools for evaluating AF risk based on variables collected from patient data. However, these assessments often require additional time or supplementary screening items to thoroughly identify associated risks. Long-term screening studies reveal that the diagnosis rate of AF, particularly paroxysmal AF, tends to increase with extended electrocardiographic (electrocardiogram; ECG) monitoring, although this also raises costs. Furthermore, collecting ECG data is time-intensive and can be cumbersome. Some researchers advocate for the use of continuous ECG monitoring (eg, Holter monitors) for AF detection, noting that recordings exceeding 24 hours may improve detection rates. However, given limited medical resources [7], the presence or absence of AF in patients with stroke significantly influences both treatment outcomes and subsequent risk levels. Thus, it is crucial to develop an efficient screening method that enables physicians to more accurately identify high-risk patients [10].

Electronic medical records (EMRs) contain rich structured and unstructured information that can be leveraged to support risk stratification and early detection of AF among patients with ischemic stroke [24]. While prior studies have demonstrated the potential of machine learning models for AF prediction, many have relied primarily on structured variables and data from a single institution, which may limit model generalizability and real-world applicability across different clinical settings [25-27]. In addition, the incremental value of incorporating unstructured clinical text into predictive models has not been systematically evaluated using rigorous external validation.

The aim of this study was to develop and validate a robust and generalizable machine learning model for predicting AF during hospitalization among patients with ischemic stroke using EMRs. Specifically, we sought to (1) integrate structured clinical variables with features derived from unstructured clinical text; (2) systematically evaluate multiple feature extraction methods, resampling strategies, and classification algorithms; and (3) assess model performance and calibration through rigorous internal and external validation across 2 independent hospital datasets. By emphasizing generalizability and real-world applicability, this study aimed to provide an effective decision support tool to facilitate early identification of high-risk patients and support clinical screening strategies for AF after stroke.

Methods

Study Population

This retrospective study was conducted using EMR datasets from 2 hospitals in Taiwan. The study population included all inpatients who were diagnosed with or suspected of having ischemic stroke by a physician during hospitalization, as well as those who presented with stroke-related symptoms. The dataset from Landseed International Hospital (LIH) covered the period from 2018 to 2022 and included a total of 3988 patients, while the dataset from Chia-Yi Christian Hospital (CYCH) spanned 2007 to 2020 and comprised 5821 patients. In total, 9809 patients were initially identified across the 2 hospitals.

Because the objective of this study was to develop predictive models for AF among patients with ischemic stroke without a prior history of AF, several exclusion criteria were applied. Consistent with prior studies, patients with documented AF prior to the index stroke were classified as having known AF before stroke and were excluded from model development. Known AF before stroke was defined as an AF diagnosis recorded in the medical history at the time of admission, indicating that AF had been identified through outpatient visits or cardiac monitoring before the current hospitalization. In addition, patients with AF detected on admission were excluded. AF detected on admission was identified based on the initial ECG, admission records, or documentation in the present illness or past medical history.

After applying the inclusion and exclusion criteria, the LIH dataset consisted of 1969 inpatients with unstructured clinical records, 2032 inpatients with structured clinical data, and 1226 inpatients with both structured and unstructured data. Similarly, the CYCH dataset included 3319 inpatients with unstructured

XSL•FO
RenderX

data, 1441 inpatients with structured data, and 1072 inpatients with both data types.

Patients with both structured and unstructured data were defined as the intersecting cohort, representing overlapping patients across the 2 data sources. This intersecting cohort constituted the final analytic population used for model development and evaluation to ensure consistent feature availability across all included patients. Patients with data available in only 1 data source were not included in the final modeling analyses.

In the LIH intersecting cohort, 7% (86/1226) of the patients were identified as having AF during hospitalization following ischemic stroke. In contrast, in the CYCH intersecting cohort, 32.6% (350/1072) of the patients met the criteria for AF. Therefore, both cohorts exhibited class imbalance with notably different proportions of AF cases, which motivated the use of resampling strategies in subsequent model development.

Ethical Considerations

The study protocol was approved by the institutional review boards of LIH (IRB-22-049) and Ditmanson Medical Foundation Chia-Yi Christian Hospital (IRB2022071). As this was a retrospective study using de-identified medical records, the requirement for informed consent was waived by both institutional review boards. To ensure patient confidentiality, all personal identifiers were removed and replaced with unique study identification numbers prior to analysis. No participants received any financial compensation, as no direct contact or intervention was involved.

Outcome Variable

The primary outcome variable in this study was the occurrence of AF during hospitalization following ischemic stroke among patients in the intersecting cohort. AF status was determined using a predefined set of criteria that integrated multiple data sources from the EMRs, as summarized in Table 1. Specifically, AF was ascertained based on evidence from structured diagnostic information, unstructured clinical text, and medication records. Structured data included *International Classification of Diseases* diagnosis codes indicative of AF, whereas unstructured sources comprised textual mentions of AF in clinical narratives and examination reports, such as Holter monitoring reports; cardiac ultrasound reports; ECG reports, excluding the first ECG on admission; and general medical records, including discharge summaries. Keyword-based text searches were conducted using commonly used terms referring to AF. In addition, prescriptions of antiarrhythmic agents and oral anticoagulants were incorporated as supportive evidence for AF identification [28]. A patient was classified as having AF if any one of the criteria listed in Table 1 was satisfied.

Table . Criteria used to ascertain atrial fibrillation (AF) status in the intersecting cohort.

Criterion	Conditions
AF-1	AF documented in the stroke registry during hospitalization
AF-2	AF-related keywords identified in Holter monitoring report text
AF-3	AF-related keywords identified in cardiac ultrasound report text
AF-4	AF-related keywords identified in ECG[a] report text, excluding the first ECG on admission
AF-5	AF-related keywords identified in clinical narratives, including hospital discharge summaries
AF-6	Presence of an AF-related *ICD*[b] diagnosis code (*ICD-9-CM*[c] 427.31 or *ICD-10-CM*[d] I48.91)
AF-7	Prescription of antiarrhythmic medications, including amiodarone, propafenone, or dronedarone
AF-8	Prescription of oral anticoagulants, including warfarin, apixaban, edoxaban, rivaroxaban, or dabigatran

[a]ECG: electrocardiogram.

[b]*ICD*: *International Classification of Diseases.*

[c]*ICD-9-CM*: *ICD, Ninth Revision, Clinical Modification.*

[d]*ICD-10-CM*: *ICD, 10th Revision, Clinical Modification.*

Predictor Variables and Text Preprocessing

Predictor variables comprised data obtained within 72 hours of admission. After excluding patients with a history of AF or diagnosed with AF upon admission, all variables were standardized across both hospitals.

The structured variables, detailed in Multimedia Appendix 1, comprised a comprehensive set of clinical predictors that are routinely available at hospital admission and during early hospitalization. These structured predictors included demographic characteristics (sex, age, height, and weight); vascular imaging findings derived from carotid duplex sonography; echocardiographic parameters reflecting cardiac structure and function (eg, aortic diameter, left atrial size, ventricular dimensions, ejection fraction, and diastolic function indexes); vital signs (body temperature, heart rate, respiratory rate, and blood pressure); and laboratory measurements covering hepatic function, renal function, coagulation profiles, inflammatory markers, lipid profiles, and hematological indexes. In addition, neurological severity was captured using individual

XSL•FO
RenderX

National Institutes of Health Stroke Scale (NIHSS) subitems, as well as the overall NIHSS score, enabling fine-grained representation of stroke-related neurological deficits. Family medical history variables related to hypertension, diabetes mellitus, stroke, and ischemic heart disease were also included as nominal predictors. Together, these structured variables provided a multidimensional characterization of patients' demographic profiles, cardiovascular status, metabolic conditions, neurological severity, and familial risk factors.

Unstructured data encompassed magnetic resonance imaging results, chief concerns, and present illness. Text preprocessing of variables derived from unstructured clinical text involved Chinese-English translation, spell-checking, abbreviation expansion, nonword symbol removal, AF-suggestive word deletion, lowercase conversion, word form reduction, negation marking as "_NEG" using additional negation words [29,30] and the Natural Language Toolkit's (Team NLTK) *mark_negation* function, and stop word removal.

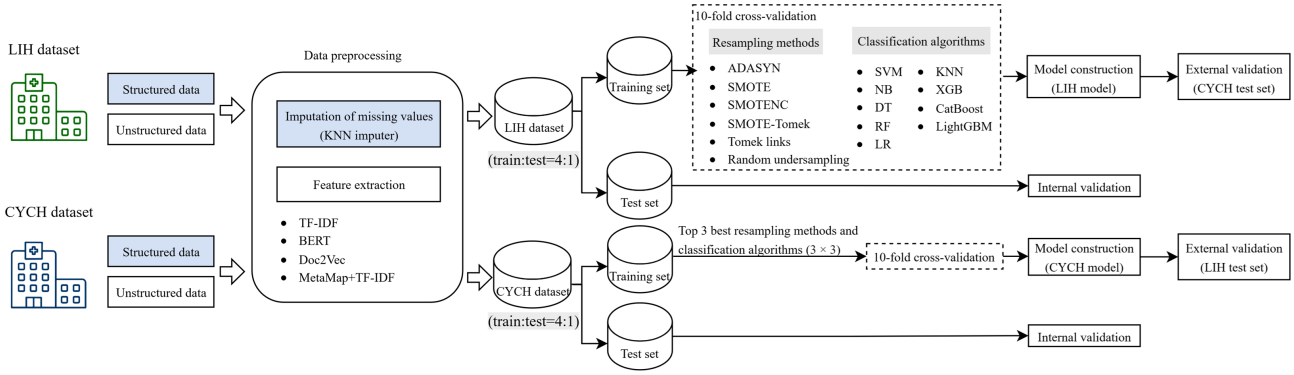
For text feature extraction, we used bidirectional encoder representations from transformers (BERT) [31,32], Doc2Vec [33,34], term frequency–inverse document frequency (TF-IDF) [35,36], and MetaMap [37,38]. These 4 methods were selected to represent complementary paradigms in clinical natural language processing, ranging from deep contextualized language models to traditional statistical and knowledge-based approaches. BERT has demonstrated strong performance in modeling contextual semantics and capturing complex linguistic patterns in clinical narratives and has been widely adopted in biomedical and clinical natural language processing tasks. Doc2Vec was included as a representative neural embedding method that captures document-level semantics with lower computational complexity, making it suitable for large-scale clinical text analysis. TF-IDF remains a widely used and robust baseline in clinical text mining due to its interpretability and effectiveness in high-dimensional sparse representations. MetaMap was incorporated to leverage domain-specific medical knowledge by mapping clinical text to standardized concepts in the Unified Medical Language System, which has been shown to enhance semantic normalization and downstream clinical prediction tasks.

Medical concepts (concept unique identifiers) were identified via MetaMap and subsequently transformed into feature vectors using TF-IDF applying a term frequency threshold with maximum and minimum values set to 0.9 and 0.1, respectively. In addition, Chinese text translation and spell-checking were conducted using *googletrans* (version 3.1.0a0) and *pyspellchecker* (version 0.8.1) to improve text consistency prior to feature extraction.

Prediction Model Development and Evaluation

Figure 1 delineates a comprehensive machine learning model validation workflow encompassing both internal and external validation phases. The internal validation phase uses the LIH dataset and a 10-fold cross-validation strategy. During this process, various combinations of resampling methods and modeling techniques are evaluated, with the top 3 best-performing combinations (totaling 9 configurations) selected for further analysis. This step is crucial for identifying the most appropriate models and parameter settings for the given data.

Figure 1. The process of machine learning model construction. ADASYN: adaptive synthetic sampling; BERT: bidirectional encoder representations from transformers; CatBoost: categorical boosting; CYCH: Chia-Yi Christian Hospital; DT: decision tree; KNN: k-nearest neighbor; LightGBM: light gradient-boosting machine; LIH: Landseed International Hospital; LR: logistic regression; NB: naive Bayes; RF: random forest; SMOTE: synthetic minority oversampling technique; SMOTENC: SMOTE for nominal and continuous features; SVM: support vector machine; TF-IDF: term frequency–inverse document frequency; XGB: extreme gradient boosting.



The external validation phase leverages an independent dataset from CYCH. In this stage, the optimal model and resampling method combinations identified during internal validation are used to train models, which are subsequently evaluated on the external dataset. This phase is essential for assessing the models' generalizability, ensuring their robustness and efficacy when applied to novel, unseen data. This rigorous 2-stage validation approach enhances the reliability and external validity of the developed predictive models, providing a robust framework for clinical application in AF risk assessment among patients with stroke. For both the internal and external validation phases, we used k-nearest neighbor imputation to address missing values within structured data. Additionally, we applied multiple feature extraction techniques to the unstructured text, including TF-IDF, BERT, Doc2Vec, and MetaMap combined with TF-IDF, followed by subsequent evaluations to compare model performance. To comprehensively evaluate the predictive power of different data types, we constructed and compared three distinct model configurations: (1) models using only structured variables, (2) models using only variables derived from unstructured clinical text (ie, text-derived variables), and (3) models combining both structured and text-derived variables.

XSL•FO
RenderX

To address the class imbalance present in the datasets from both hospitals, we implemented various resampling techniques during the model training phase, including adaptive synthetic sampling, synthetic minority oversampling technique (SMOTE), SMOTE for nominal and continuous features, SMOTE-Tomek, Tomek links, and random undersampling. We stratified and split the data into training and test sets with a 4:1 ratio using the parameter *stratify=df[['AF']]* to maintain consistent class ratios across sets.

We evaluated multiple classifiers, including support vector machine, Gaussian naïve Bayes, k-nearest neighbor, decision tree, random forest, logistic regression, Extreme Gradient Boosting (XGBoost), categorical boosting (CatBoost), and light gradient-boosting machine (LightGBM). The training process incorporated 10-fold cross-validation with stratification (*stratify=df[['AF']]*) to obtain robust performance estimates. We assessed model performance using multiple evaluation metrics, including the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and expected calibration error. Following the identification of optimal model combinations, we conducted both internal and external validation. The final results reflected the outcomes of the test set.

To enhance model interpretability, we used Shapley values (using *shap* version 0.45.0) to explain the best-performing models in both the LIH and CYCH datasets. This approach provides insights into the relative importance of different features in the prediction task.

Baseline Models

The baseline model was compared with machine learning models using traditional risk-scoring models (AS5F [age, stroke severity NIHSS>5 to find atrial fibrillation] [39-41] and CHASE-LESS [coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke] [15,42,43]). CHASE-LESS was developed using data derived from a Taiwanese population, which is consistent with the population of this study, and its variables include age, NIHSS score, coronary heart disease, congestive heart failure, hyperlipidemia, diabetes mellitus, and prior stroke or TIA. The threshold value of CHASE-LESS was set to 6, and the scoring items were based on the variables in the EMR at the time of admission. The AS5F score is designed to determine the need for long-term ECG monitoring in patients with AIS or TIA.

Results

We conducted modeling and prediction using both structured and text-derived variables from the two hospitals. In the internal validation for LIH (Table 2), the combination of structured variables with logistic regression yielded the best performance, achieving an AUC of 0.896. For text-derived variables, the model using Doc2Vec, random undersampling, and LightGBM achieved an AUC of 0.674. When integrating both structured and text-derived variables, the model incorporating Doc2Vec, SMOTE, and CatBoost performed well, with an AUC of 0.885. Some models showed enhanced performance when integrating both types of variables. For instance, in the LightGBM model implementation, the integration of structured and text-derived variables, coupled with the adaptive synthetic sampling technique, demonstrated superior performance, with an AUC of 0.876. This feature-integrated approach substantially outperformed both models using only structured variables (AUC=0.850) and those using only text-derived variables (AUC=0.468). However, not all combined approaches demonstrated such improvement.

XSL•FO
RenderX

Table . Comparative analysis of Landseed International Hospital (LIH) models for atrial fibrillation prediction for the area under the receiver operating characteristic curve (AUC) metric: internal validation using the LIH dataset.

Model	Structured variables, AUC	Text-derived variables, AUC				Structured+text-derived variables, AUC			
		BERT[a]	Doc2Vec	MetaMap	TF-IDF[b]	BERT	Doc2Vec	MetaMap	TF-IDF
AS5F[c]	0.597	—[d]	—	—	—	—	—	—	—
CHASE-LESS[e]	0.657	—	—	—	—	—	—	—	—
SVM[f]									
Imbalanced	0.835	0.403	0.496	0.499	0.64	0.654	0.837	0.809	0.862
Undersampling	0.778	0.579	0.61	0.48	0.595	0.482	0.76	0.779	0.841
ADASYN[g]	0.812	0.548	0.514	0.466	0.602	0.504	0.846	0.81	0.868
SMOTE[h]	0.803	0.54	0.511	0.51	0.672	0.504	0.824	0.783	0.796
Gaussian NB[i]									
Imbalanced	0.52	0.439	0.604	0.572	0.466	0.404	0.535	0.566	0.468
Undersampling	0.763	0.495	0.671	0.483	0.574	0.616	0.764	0.551	0.599
ADASYN	0.455	0.402	0.533	0.534	0.643	0.452	0.456	0.545	0.52
SMOTE	0.458	0.507	0.532	0.55	0.525	0.477	0.42	0.547	0.609
KNN[j]									
Imbalanced	0.618	0.463	0.47	0.519	0.666	0.486	0.642	0.58	0.567
Undersampling	0.646	0.398	0.451	0.491	0.527	0.526	0.576	0.628	0.648
ADASYN	0.627	0.529	0.45	0.483	0.506	0.421	0.603	0.691	0.625
SMOTE	0.594	0.446	0.427	0.511	0.548	0.405	0.577	0.582	0.608
DT[k]									
Imbalanced	0.542	0.529	0.512	0.535	0.503	0.613	0.576	0.531	0.553
Undersampling	0.621	0.433	0.509	0.49	0.539	0.673	0.634	0.501	0.531
ADASYN	0.666	0.439	0.584	0.521	0.448	0.723	0.631	0.567	0.675
SMOTE	0.545	0.45	0.599	0.574	0.439	0.713	0.673	0.604	0.548
RF[l]									
Imbalanced	0.815	0.548	0.535	0.533	0.586	0.627	0.758	0.757	0.756
Undersampling	0.846	0.376	0.658	0.553	0.523	0.609	0.838	0.738	0.829
ADASYN	0.82	0.523	0.532	0.487	0.476	0.6	0.832	0.79	0.821
SMOTE	0.86	0.502	0.539	0.522	0.565	0.621	0.845	0.815	0.851
LR[m]									
Imbalanced	0.896	0.523	0.659	0.504	0.641	0.798	0.862	0.819	0.805

XSL•FO
RenderX

Model	Structured variables, AUC	Text-derived variables, AUC				Structured+text-derived variables, AUC			
		BERT[a]	Doc2Vec	MetaMap	TF-IDF[b]	BERT	Doc2Vec	MetaMap	TF-IDF
Undersampling	0.84	0.462	0.664	0.492	0.558	0.769	0.783	0.805	0.827
ADASYN	0.865	0.463	0.635	0.537	0.621	0.752	0.847	0.789	0.804
SMOTE	0.866	0.471	0.624	0.543	0.644	0.757	0.833	0.803	0.787
XGB[n]									
Imbalanced	0.845	0.402	0.541	0.571	0.653	0.813	0.812	0.822	0.822
Undersampling	0.829	0.408	0.617	0.537	0.547	0.797	0.787	0.824	0.826
ADASYN	0.849	0.41	0.5	0.46	0.504	0.865	0.868	0.86	0.874
SMOTE	0.856	0.449	0.591	0.474	0.581	0.882	0.868	0.857	0.819
CatBoost[o]									
Imbalanced	0.869	0.41	0.609	0.517	0.652	0.774	0.85	0.778	0.849
Undersampling	0.868	0.407	0.657	0.536	0.573	0.772	0.854	0.806	0.835
ADASYN	0.864	0.524	0.522	0.481	0.558	0.809	0.875	0.857	0.864
SMOTE	0.855	0.46	0.567	0.529	0.564	0.833	0.885	0.84	0.866
LightGBM[p]									
Imbalanced	0.829	0.524	0.46	0.619	0.538	0.801	0.75	0.837	0.798
Undersampling	0.783	0.361	0.674	0.435	0.513	0.777	0.79	0.783	0.81
ADASYN	0.85	0.471	0.642	0.604	0.468	0.852	0.843	0.773	0.876
SMOTE	0.858	0.488	0.665	0.616	0.525	0.811	0.873	0.845	0.837

[a]BERT: bidirectional encoder representations from transformers.

[b]TF-IDF: term frequency–inverse document frequency.

[c]AS5F: age, stroke severity National Institutes of Health Stroke Scale >5 to find atrial fibrillation.

[d]Not applicable.

[e]CHASE-LESS: coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke.

[f]SVM: support vector machine.

[g]ADASYN: adaptive synthetic sampling.

[h]SMOTE: synthetic minority oversampling technique.

[i]NB: naïve Bayes.

[j]KNN: k-nearest neighbor.

[k]DT: decision tree.

[l]RF: random forest.

[m]LR: logistic regression.

[n]XGB: extreme gradient boosting.

[o]CatBoost: categorical boosting.

[p]LightGBM: light gradient-boosting machine.

In the external validation of the LIH model using the CYCH dataset (Table 3), the best performance was achieved using a combination of BERT and SMOTE, with an AUC of 0.795, which exceeded the AUC for models using only structured variables (AUC=0.781) and those using only text-derived variables. In the internal validation for CYCH (Table 4), the best-performing model was a combination of BERT with undersampling and XGBoost, achieving an AUC of 0.861. In the external validation, the optimal combination involved an imbalanced technique with CatBoost, resulting in an AUC of

XSL•FO
RenderX

0.832. These findings suggest that combining structured and text-derived variables can improve predictive performance in some models, highlighting the importance of selecting appropriate models and resampling techniques to enhance model outcomes.

Figure 2 presents the external validation results from the 2 hospitals, indicating that the combination of CatBoost and SMOTE ranked third in both cases as determined by the AUC metric. Overall, ensemble learning demonstrated superior performance throughout the validation process. While the incorporation of text-derived variables did not consistently yield the highest AUC values in both internal and external validation, it is noteworthy that their inclusion generally resulted in improved outcomes based on the expected calibration error during external validation.

Table . Performance of the top 3 Landseed International Hospital models incorporating structured and unstructured feature combinations: external validation using the Chia-Yi Christian Hospital dataset.

Model	AUC[a]	Sensitivity	Specificity	ECE[b]
XGB[c]				
Structured	0.781	0.651	0.776	0.187
Structured+MRI[d]	0.771	0.549	0.843	0.174
Structured+CC[e]	0.776	0.623	0.796	0.171
Structured+PI[f]	0.773	0.597	0.798	0.178
Structured+MRI+CC+PI	0.775	0.571	0.821	0.168
CatBoost[g]				
Structured	0.789	0.463	0.892	0.09
Structured+MRI	0.781	0.311	0.942	0.147
Structured+CC	0.788	0.483	0.895	0.093
Structured+PI	0.79	0.46	0.91	0.104
Structured+MRI+CC+PI	0.788	0.443	0.911	0.107
LightGBM[h]				
Structured	0.781	0.611	0.81	0.161
Structured+MRI	0.794	0.509	0.889	0.111
Structured+CC	0.772	0.5	0.864	0.124
Structured+PI	0.784	0.509	0.873	0.124
Structured+MRI+CC+PI	0.795	0.574	0.855	0.133

[a]AUC: area under the receiver operating characteristic curve.

[b]ECE: expected calibration error.

[c]XGB: extreme gradient boosting.

[d]MRI: magnetic resonance imaging.

[e]CC: chief concern.

[f]PI: present illness.

[g]CatBoost: categorical boosting.

[h]LightGBM: light gradient-boosting machine.

XSL•FO
RenderX

Table . Performance of the top 3 Chia-Yi Christian Hospital (CYCH) models incorporating structured and unstructured (magnetic resonance imaging+chief concerns+present illness) feature combinations.

Model	Internal validation (CYCH dataset), AUC[a]	External validation (LIH[b] dataset)	
		AUC	ECE[c]
XGB[d]			
Imbalanced	0.858	0.816	0.08
Undersampling	0.861	0.776	0.176
ADASYN[e]	0.852	0.73	0.338
SMOTE[f]	0.845	0.692	0.334
CatBoost[g]			
Imbalanced	0.851	0.832	0.129
Undersampling	0.859	0.816	0.253
ADASYN	0.843	0.76	0.363
SMOTE	0.843	0.776	0.305
LightGBM[h]			
Imbalanced	0.84	0.764	0.225
Undersampling	0.816	0.769	0.366
ADASYN	0.789	0.726	0.37
SMOTE	0.778	0.662	0.409

[a]AUC: area under the receiver operating characteristic curve.

[b]LIH: Landseed International Hospital.

[c]ECE: expected calibration error.

[d]XGB: extreme gradient boosting.

[e]ADASYN: adaptive synthetic sampling.

[f]SMOTE: synthetic minority oversampling technique.

[g]CatBoost: categorical boosting.

[h]LightGBM: light gradient-boosting machine.

XSL•FO
RenderX

Figure 2. Ranking of the top 3 model and resampling method combinations for external validation: validation of the Landseed International Hospital (LIH) model using the Chia-Yi Christian Hospital (CYCH) dataset (A) and validation of the CYCH model using the LIH dataset (B). ADASYN: adaptive synthetic sampling; AUC: area under the receiver operating characteristic curve; CatBoost: categorical boosting; ECE: expected calibration error; LightGBM: light gradient-boosting machine; SMOTE: synthetic minority oversampling technique; XGB: extreme gradient boosting.

XSL•FO
RenderX

Discussion

Summary of Main Findings

This study developed and validated machine learning–based prediction models for identifying AF during hospitalization among patients with ischemic stroke by leveraging EMRs from 2 independent hospitals. The main findings can be summarized in 3 points. First, models trained on structured clinical variables achieved strong predictive performance, and the integration of unstructured clinical text further improved discrimination and calibration in selected model configurations, particularly in the external validation setting. Second, ensemble learning methods, including CatBoost, XGBoost, and LightGBM, consistently outperformed traditional risk scores and conventional classifiers across internal and external validation. Third, feature importance analysis using SHAP revealed that echocardiographic parameters, including the ratio of E- to A-wave velocities, left atrial size, and age, were among the most influential predictors across both datasets. Together, these findings demonstrate the feasibility and generalizability of incorporating structured and unstructured EMR data to support early risk stratification for AF after ischemic stroke.

Implications of Integrating Structured and Text-Derived Variables

Although the integration of structured variables with variables derived from unstructured clinical text improved predictive performance in several model configurations, the results in Table 2 indicate that such integration did not uniformly outperform models based solely on structured variables. This finding warrants careful interpretation rather than being viewed as a limitation of the proposed framework. In this study, structured clinical variables already encompassed well-established and highly discriminative predictors of AF, including echocardiographic parameters such as the ratio of E- to A-wave velocities, left atrial size, and age. When such strong predictors are present, models trained exclusively on structured variables may achieve near-optimal performance, leaving limited room for additional gains from text-derived variables.

Moreover, variables derived from unstructured clinical text are inherently high-dimensional and potentially noisy, particularly when generated using representation techniques such as TF-IDF or Doc2Vec. For certain classifiers that are sensitive to feature dimensionality or redundancy, the inclusion of text-derived variables may dilute the contribution of strong structured predictors, resulting in marginal or inconsistent performance gains. This phenomenon is consistent with prior observations in studies integrating clinical text with structured EMR data.

Importantly, our results suggest that the benefit of feature-level integration is highly model dependent. Ensemble learning–based algorithms, including LightGBM, XGBoost, and CatBoost, demonstrated a greater capacity to exploit complementary information from heterogeneous feature sources, particularly when combined with appropriate resampling strategies. In contrast, simpler or less flexible classifiers did not consistently benefit from the inclusion of text-derived variables. These findings indicate that integrating unstructured clinical text should be viewed as a complementary strategy whose effectiveness depends on the interaction among feature representation, classifier architecture, and imbalance-handling methods.

In addition, Table 2 indicates that resampling did not consistently improve performance for all classifiers as some models achieved comparable results when trained on the original imbalanced data. This finding suggests that the necessity of resampling is model dependent. Ensemble learning–based classifiers may inherently mitigate class imbalance effects, whereas other models benefit more substantially from resampling techniques. Therefore, resampling should be considered a robustness-oriented strategy rather than a mandatory preprocessing step.

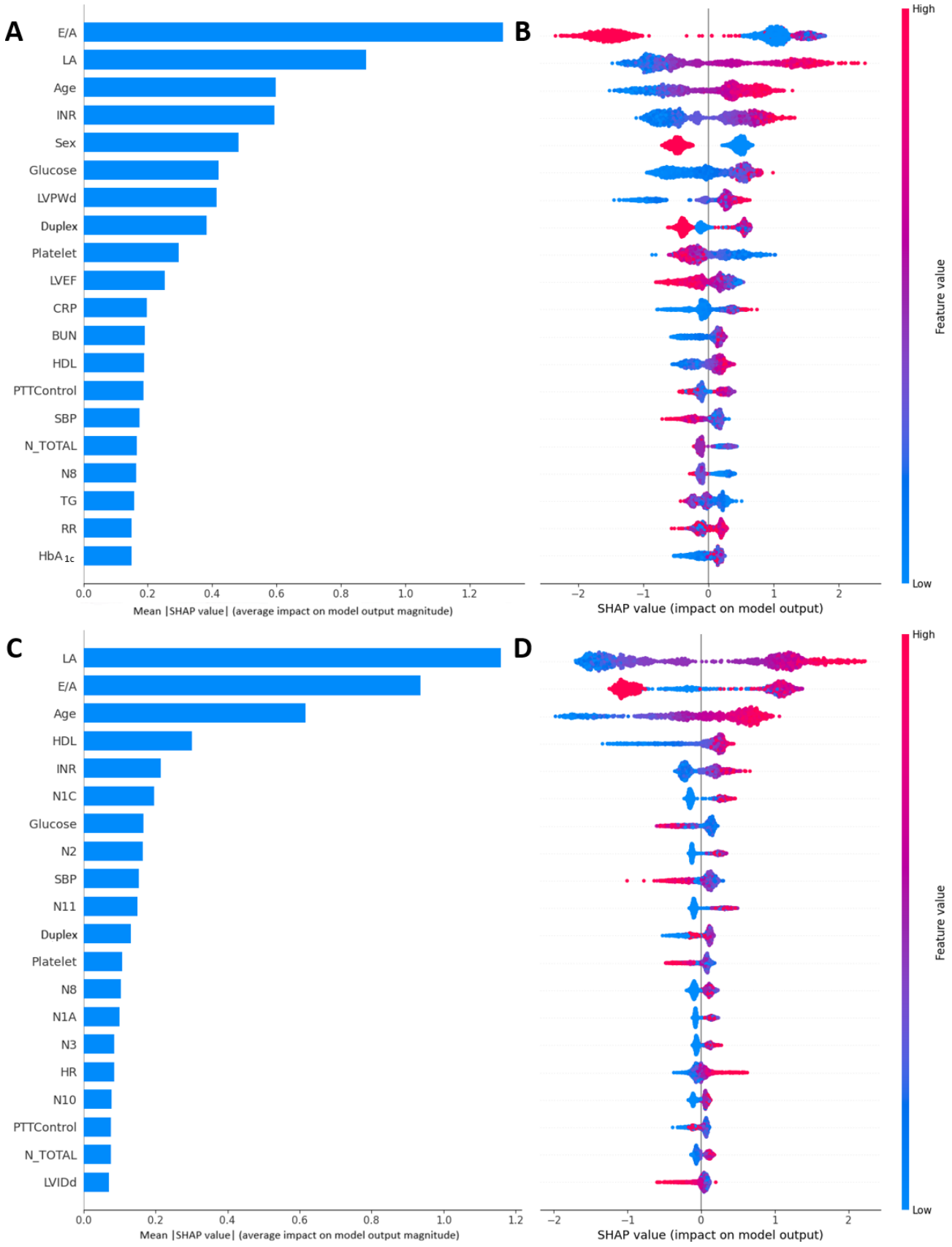
Significant Features

Figure 3 illustrates the 20 most significant features identified for the LIH and CYCH datasets using shap (version 0.45.0). In Figure 3, features are ranked according to their mean absolute Shapley values, which quantify the global importance of each feature in relation to the model's output. Figure 3A presents the feature ordering for LIH, whereas Figure 3C depicts the ordering for CYCH. Figure 3 shows beeswarm plots that visualize the Shapley values for each patient concerning these features. The beeswarm plots convey the contribution of each feature to the model's output based on the magnitude and direction of the Shapley values. Figure 3B represents the contributions of LIH patients, whereas Figure 3D illustrates those of CYCH patients. Notably, the graphs reveal that the top 3 features consistently identified across both datasets were ratio of E- to A-wave velocities, left atrial size, and age.

XSL•FO
RenderX

Figure 3. The top 20 most important features identified by the model based on structured variables. Panels A and C display bar charts representing the mean absolute Shapley values, which indicate the average contribution of each feature to the model's output. Panels B and D present beeswarm plots of individual Shapley values for each feature across patients, where the position of each dot on the x-axis reflects that feature's contribution to the prediction for a given patient. The color of each dot corresponds to the relative value of the associated feature. BUN: blood urea nitrogen; CRP: C-reactive protein; E/A: ratio of E- to A-wave velocities; HbA_{1c}: hemoglobin A_{1c}; HDL: high-density lipoprotein; HR: heart rate; INR: international normalized ratio; LA: left atrial size; LVEF: left ventricular ejection fraction; LVIDd: left ventricular internal dimension in diastole; LVPWd: left ventricular posterior wall in diastole; N1A: level of consciousness responsiveness; N1C: level of consciousness commands; N2: best gaze; N3: visual fields; N8: sensory; N10: dysarthria; N11: extinction and inattention; PTTControl: partial thromboplastin time control; RR: respiratory rate; SBP: systolic blood pressure; SHAP: Shapley additive explanations; TG: triglyceride.

XSL•FO
RenderX



Clinical Applications and Significance

Predicting AF following a stroke is critical for effective secondary stroke prevention as it directly influences medical decision-making and patient outcomes. Given the constraints of health care resources, it is essential to optimize their use by incorporating efficient screening protocols. In this context, machine learning offers a highly effective approach. This study identified the optimal combination of variables derived from unstructured clinical text through rigorous external validation

XSL•FO
RenderX

supplemented by internal validation using the LIH dataset and further validated using EMR data from CYCH. Although models built on internal data are often preferred for clinical applications, our experiments revealed that models developed using the CYCH data occasionally outperformed those based on LIH data. This discrepancy may be attributed to missing information and the relatively small sample size of the LIH dataset. The models developed in this study aim to support clinicians by identifying high-risk patients who would benefit from additional screening measures, such as ECG, ultimately reducing the risk of missed diagnoses and ensuring timely and appropriate treatment.

Limitations

This study has several limitations that should be acknowledged. First, the data source poses a potential issue as the data were derived from EMRs within the hospitals. It is possible that patients were diagnosed with AF outside these institutions, introducing bias, particularly given the older age demographic of the hospital population. Second, regarding the data themselves, while generalizability was tested across 2 different hospitals, the variability in clerical styles between the hospitals may affect the consistency of the variables derived from unstructured clinical text. For structured variables, a high proportion of missing values in factors known to be associated with stroke or AF (eg, smoking history), uncertainties about the accuracy of the records, and the small and variable sample size could all impact the overall efficacy of the model. Finally, it should be noted that the predictors contributing to the model in this study do not imply causality.

Conclusions

This study demonstrates the feasibility of developing machine learning–based predictive models to identify AF in patients with ischemic stroke using EMR data. The integration of structured variables with variables derived from unstructured clinical text, including medical history and magnetic resonance imaging findings extracted from free-text sources, showed potential to improve predictive performance in selected model configurations. Internal and external validation results indicate that ensemble learning–based models performed favorably compared with other algorithmic approaches, supporting their potential utility for AF risk prediction in clinical settings.

Funding

This research was supported in part by the National Central University–Landseed Hospital Research and Development Office (NCU-LSH-111-B-008) and the Ministry of Science and Technology of the People's Republic of China (grant MOST 111-2410-H-008-026-MY2).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient privacy and institutional data protection regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YWC, YHH, SFS
Data curation: YWC, SFS
Formal analysis: YWC, YHH, SFS, YHY
Investigation: YWC, SFS
Methodology: YWC, YHH, SFS
Supervision: YHH
Validation: YWC, YHH, SFS, YHY
Writing—original draft: YWC, YHH, SFS, YHY
Writing—review and editing: YWC, YHH, SFS, YHY
All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed definitions and summary statistics of structured clinical variables extracted from electronic medical records at Landseed International Hospital and Chia-Yi Christian Hospital, including variable descriptions, data types, distributions, and proportions of missing values.
[DOCX File, 31 KB - medinform_v14i1e78117_app1.docx]

References

1. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. Int J Stroke 2022 Jan;17(1):18-29. [doi: 10.1177/17474930211065917] [Medline: 34986727]

XSL•FO
RenderX

2. Feigin VL, Forouzanfar MH, Krishnamurthi R, et al. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. Lancet 2014 Jan 18;383(9913):245-254. [doi: 10.1016/s0140-6736(13)61953-4] [Medline: 24449944]
3. Pennlert J, Eriksson M, Carlberg B, Wiklund PG. Long-term risk and predictors of recurrent stroke beyond the acute phase. Stroke 2014 Jun;45(6):1839-1841. [doi: 10.1161/STROKEAHA.114.005060] [Medline: 24788972]
4. Bergström L, Irewall AL, Söderström L, Ögren J, Laurell K, Mooe T. One-year incidence, time trends, and predictors of recurrent ischemic stroke in Sweden from 1998 to 2010: an observational study. Stroke 2017 Aug;48(8):2046-2051. [doi: 10.1161/STROKEAHA.117.016815] [Medline: 28706114]
5. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res 2017 Feb 3;120(3):439-448. [doi: 10.1161/CIRCRESAHA.116.308413] [Medline: 28154096]
6. Lee M, Wu YL, Ovbiagele B. Trends in incident and recurrent rates of first-ever ischemic stroke in Taiwan between 2000 and 2011. J Stroke 2016 Jan;18(1):60-65. [doi: 10.5853/jos.2015.01326] [Medline: 26687123]
7. Lip GYH, Hunter TD, Quiroz ME, Ziegler PD, Turakhia MP. Atrial fibrillation diagnosis timing, ambulatory ECG monitoring utilization, and risk of recurrent stroke. Circ Cardiovasc Qual Outcomes 2017 Jan;10(1):e002864. [doi: 10.1161/CIRCOUTCOMES.116.002864] [Medline: 28096204]
8. Lee SR, Choi EK, Han KD, Cha MJ, Oh S. Trends in the incidence and prevalence of atrial fibrillation and estimated thromboembolic risk using the CHA2DS2-VASc score in the entire Korean population. Int J Cardiol 2017 Jun 1;236:226-231. [doi: 10.1016/j.ijcard.2017.02.039] [Medline: 28233629]
9. Ball J, Carrington MJ, McMurray JJ, Stewart S. Atrial fibrillation: profile and burden of an evolving epidemic in the 21st century. Int J Cardiol 2013 Sep 1;167(5):1807-1824. [doi: 10.1016/j.ijcard.2012.12.093] [Medline: 23380698]
10. Jones NR, Taylor CJ, Hobbs FD, Bowman L, Casadei B. Screening for atrial fibrillation: a call for evidence. Eur Heart J 2020 Mar 7;41(10):1075-1085. [doi: 10.1093/eurheartj/ehz834] [Medline: 31811716]
11. Huang CK, Wang JC, Chung CH, Chen SJ, Liao WI, Chien WC. The risk and timing of acute ischemic stroke after electrical cardioversion for atrial fibrillation in Taiwan: a nationwide population-based cohort study. Int J Cardiol 2022 Mar 15;351:55-60. [doi: 10.1016/j.ijcard.2021.12.035] [Medline: 34954280]
12. Mtwesi V, Amit G. Stroke prevention in atrial fibrillation: the role of oral anticoagulation. Med Clin North Am 2019 Sep;103(5):847-862. [doi: 10.1016/j.mcna.2019.05.006] [Medline: 31378330]
13. Lee E, Choi EK, Han KD, et al. Mortality and causes of death in patients with atrial fibrillation: a nationwide population-based study. PLoS ONE 2018 Dec 26;13(12):e0209687. [doi: 10.1371/journal.pone.0209687] [Medline: 30586468]
14. Chao TF, Liu CJ, Tuan TC, et al. Lifetime risks, projected numbers, and adverse outcomes in Asian patients with atrial fibrillation: a report from the Taiwan nationwide AF cohort study. Chest 2018 Feb;153(2):453-466. [doi: 10.1016/j.chest.2017.10.001] [Medline: 29017957]
15. Hsieh CY, Kao HM, Sung KL, Sposato LA, Sung SF, Lin SJ. Validation of risk scores for predicting atrial fibrillation detected after stroke based on an electronic medical record algorithm: a registry-claims-electronic medical record linked data study. Front Cardiovasc Med 2022 Apr 29;9:888240. [doi: 10.3389/fcvm.2022.888240] [Medline: 35571191]
16. Hsieh FI, Lien LM, Chen ST, et al. Get with the guidelines-stroke performance indicators: surveillance of stroke care in the Taiwan stroke registry: get with the guidelines-stroke in Taiwan. Circulation 2010 Sep 14;122(11):1116-1123. [doi: 10.1161/CIRCULATIONAHA.110.936526] [Medline: 20805428]
17. Hsieh FI, Chiou HY. Stroke: morbidity, risk factors, and care in Taiwan. J Stroke 2014 May;16(2):59-64. [doi: 10.5853/jos.2014.16.2.59] [Medline: 24949310]
18. Garg N, Kumar A, Flaker GC. Antiplatelet therapy for stroke prevention in atrial fibrillation. Mo Med 2010;107(1):44-47. [Medline: 20222295]
19. Själander S, Sjögren V, Renlund H, Norrving B, Själander A. Dabigatran, rivaroxaban and apixaban vs. high TTR warfarin in atrial fibrillation. Thromb Res 2018 Jul;167:113-118. [doi: 10.1016/j.thromres.2018.05.022] [Medline: 29803981]
20. Yang L, Brooks MM, Glynn NW, Zhang Y, Saba S, Hernandez I. Real-world direct comparison of the effectiveness and safety of apixaban, dabigatran, rivaroxaban, and warfarin in Medicare beneficiaries with atrial fibrillation. Am J Cardiol 2020 Jul 1;126:29-36. [doi: 10.1016/j.amjcard.2020.03.034] [Medline: 32359718]
21. Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011 Sep 15;365(11):981-992. [doi: 10.1056/NEJMoa1107039] [Medline: 21870978]
22. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on atrial fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
23. Van Gelder IC, Rienstra M, Bunting KV, et al. 2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2024 Sep 29;45(36):3314-3414. [doi: 10.1093/eurheartj/ehae176] [Medline: 39210723]
24. Karnik S, Tan SL, Berg B, et al. Predicting atrial fibrillation and flutter using electronic health records. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:5562-5565. [doi: 10.1109/EMBC.2012.6347254] [Medline: 23367189]
25. Chamberlain AM, Roger VL, Noseworthy PA, et al. Identification of incident atrial fibrillation from electronic medical records. J Am Heart Assoc 2022 Apr 5;11(7):e023237. [doi: 10.1161/JAHA.121.023237] [Medline: 35348008]

XSL•FO
RenderX

26. Hulme OL, Khurshid S, Weng LC, et al. Development and validation of a prediction model for atrial fibrillation using electronic health records. JACC Clin Electrophysiol 2019 Nov;5(11):1331-1341. [doi: 10.1016/j.jacep.2019.07.016] [Medline: 31753441]
27. Verhaeghe J, De Corte T, Sauer CM, et al. Generalizable calibrated machine learning models for real-time atrial fibrillation risk prediction in ICU patients. Int J Med Inform 2023 Jul;175:105086. [doi: 10.1016/j.ijmedinf.2023.105086] [Medline: 37148868]
28. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. Am J Cardiol 2016 Jan 15;117(2):221-225. [doi: 10.1016/j.amjcard.2015.10.031] [Medline: 26684516]
29. Hu M, Liu B. Mining and summarizing customer reviews. : ACM Presented at: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 22-25, 2004; Seattle, WA. [doi: 10.1145/1014052.1014073]
30. Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web. Presented at: Proceedings of the 14th International Conference on World Wide Web; May 10-14, 2005. [doi: 10.1145/1060745.1060797]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019. [doi: 10.18653/v1/N19-1423]
32. Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. BERT applications in natural language processing: a review. Artif Intell Rev 2025;58(6):166. [doi: 10.1007/s10462-025-11162-5]
33. Le Q, Mikolov T. Distributed representations of sentences and documents. Proc Mach Learn Res 2014;32(2):1188-1196 [FREE Full text]
34. Getzen E, Ruan Y, Ungar L, Long Q. Mining for health: a comparison of word embedding methods for analysis of EHRs data. In: Zhao Y, Chen DG, editors. Statistics in Precision Health: Theory, Methods and Applications: Springer International Publishing; 2024:313-338. [doi: 10.1007/978-3-031-50690-1_13]
35. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972 Jan;28(1):11-21. [doi: 10.1108/eb026526]
36. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
37. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236. [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
38. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
39. Uphaus T, Weber-Krüger M, Grond M, et al. Development and validation of a score to detect paroxysmal atrial fibrillation after stroke. Neurology 2019 Jan 8;92(2):e115-e124. [doi: 10.1212/WNL.0000000000006727] [Medline: 30530796]
40. Zheng X, Wang F, Zhang J, et al. Using machine learning to predict atrial fibrillation diagnosed after ischemic stroke. Int J Cardiol 2022 Jan 15;347:21-27. [doi: 10.1016/j.ijcard.2021.11.005] [Medline: 34774886]
41. Sung SF, Sung KL, Pan RC, Lee PJ, Hu YH. Automated risk assessment of newly detected atrial fibrillation poststroke from electronic health record data using machine learning and natural language processing. Front Cardiovasc Med 2022 Jul 29;9:941237. [doi: 10.3389/fcvm.2022.941237] [Medline: 35966534]
42. Hsieh CY, Lee CH, Sung SF. Development of a novel score to predict newly diagnosed atrial fibrillation after ischemic stroke: the CHASE-LESS score. Atherosclerosis 2020 Feb;295:1-7. [doi: 10.1016/j.atherosclerosis.2020.01.003] [Medline: 31972497]
43. Ratajczak-Tretel B, Lambert AT, Al-Ani R, et al. Prediction of underlying atrial fibrillation in patients with a cryptogenic stroke: results from the NOR-FIB Study. J Neurol 2023 Aug;270(8):4049-4059. [doi: 10.1007/s00415-023-11680-8] [Medline: 37162578]

Abbreviations

AF: atrial fibrillation
AIS: acute ischemic stroke
AS5F: age, stroke severity National Institutes of Health Stroke Scale>5 to find atrial fibrillation
AUC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
CatBoost: categorical boosting
CHASE-LESS: coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke
CYCH: Chia-Yi Christian Hospital
ECG: electrocardiogram
EMR: electronic medical record
LightGBM: light gradient-boosting machine

XSL•FO
RenderX

LIH: Landseed International Hospital
NIHSS: National Institutes of Health Stroke Scale
SHAP: Shapley additive explanations
SMOTE: synthetic minority oversampling technique
TF-IDF: term frequency–inverse document frequency
TIA: transient ischemic attack
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 02.Jun.2025; peer-reviewed by PJ Lee, SY Lin; revised version received 10.Jan.2026; accepted 19.Jan.2026; published 10.Mar.2026.

Please cite as:
Chen YW, Sung SF, Hu YH, Yang YH
Enhanced Prediction of Atrial Fibrillation in Patients With Ischemic Stroke Through Electronic Medical Records and Text Mining: Algorithm Development and Validation
JMIR Med Inform 2026;14:e78117
URL: https://medinform.jmir.org/2026/1/e78117
doi:10.2196/78117

© Yu-Wei Chen, Sheng-Feng Sung, Ya-Han Hu, Yu-Hsuan Yang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning–Based Risk Prediction for Coronary Heart Disease Complicated by Hyperhomocysteinemia: Retrospective Study

Ming-Yuan Du[1,2,3], MMD; Meng-Ke Lyu[3], MMD; Hai-long Liu[1,2,3], MMD; Yi-zhuo Li[1,2,3], MD; Hai-feng Yan[1,2,3], MD; Xiao-hui Li[1,2,3], MD

[1]Heart Center, The First Affiliated Hospital of Henan University of Chinese Medicine, National Regional (TCM) Cardiovascular Diagnosis and Treatment Center, Zhengzhou, China

[2]Collaborative Innovation Center of Prevention and Treatment of Major Diseases by Chinese and Western Medicine, Zhengzhou, China

[3]The First Affiliated Hospital of Henan University of Traditional Chinese Medicine, No. 19 Renmin Road, Jinshui District, Henan Province, Zhengzhou, China

Corresponding Author:
Meng-Ke Lyu, MMD
The First Affiliated Hospital of Henan University of Traditional Chinese Medicine, No. 19 Renmin Road, Jinshui District, Henan Province, Zhengzhou, China

Abstract

Background: Hyperhomocysteinemia (HHcy) is recognized as an independent risk factor for coronary heart disease (CHD), yet accurately predicting CHD risk in patients with HHcy remains a challenge. This study aimed to develop and validate multiple machine learning models for predicting CHD risk in patients with HHcy and elucidate key predictors using Shapley Additive Explanation (SHAP) algorithms.

Objective: This study aims to develop and validate machine learning models for predicting the risk of coronary heart disease in individuals with normal homocysteine levels, aiming to improve early risk stratification and clinical decision-making.

Methods: This single-center retrospective study collected data from patients who were diagnosed with HHcy through electronic medical records, which were randomly divided into training (n=364, 70%), validation (n=78, 15%), and test (n=78, 15%) sets. Seven machine learning models were constructed, including logistic regression, k-nearest neighbor, decision tree, random forest, extreme gradient boost, light gradient boosting machine (LightGBM), and stacking. Six core variables (age, weight, hypertension, continuous drinking history, activated partial thromboplastin time, and carotid plaque) were utilized as inputs, with performance evaluation metrics encompassing area under the receiver operating characteristic curve, accuracy, F_1-score, calibration curve, Brier score, and decision curve analysis. Additionally, SHAP interpretation was conducted on the optimal LightGBM model.

Results: The LightGBM model exhibited superior performance in the test set (area under the receiver operating characteristic curve=0.807, F_1-score=0.606), demonstrated good calibration (Brier score=0.2415), and yielded high clinical net benefit. SHAP analysis revealed age and activated partial thromboplastin time as the most influential predictors, followed by hypertension, weight, carotid plaque, and continuous drinking history. The correlation heat map illustrated low collinearity among variables, ensuring model stability.

Conclusions: The LightGBM model demonstrated high accuracy and interpretability in forecasting CHD risk among patients with HHcy. The integration of machine learning and interpretable artificial intelligence methods holds promise for delivering personalized early risk assessment and intervention strategies in clinical settings.

(JMIR Med Inform 2026;14:e80809) doi:10.2196/80809

KEYWORDS

retrospective study; hyperhomocysteinemia; coronary heart disease; machine learning; predictive model

Introduction

Hyperhomocysteinemia (HHcy) predisposes individuals to heightened and premature susceptibility to coronary heart disease (CHD) [1,2]. Existing research has substantiated that HHcy expedites atherosclerosis progression via various mechanisms, including vascular endothelial impairment, pro-oxidative stress, and disrupted coagulation pathways, establishing it as a stand-alone CHD risk factor [3-5]. Nevertheless, the CHD risk profile in HHcy cohorts is heterogeneous, potentially modulated by a confluence of factors

XSL•FO
RenderX

encompassing age, biochemical milieu, lifestyle choices, and vascular morphology.

The current assessment tools for CHD risk in patients with HHcy are limited [6]. Traditional models, such as Cox or logistic regression, are typically based on a small number of variables and struggle to capture the intricate nonlinear interactions among variables in real clinical settings. Furthermore, these models often rely on predictors that necessitate specialized knowledge or image interpretation, hindering their widespread adoption among noncardiovascular specialist clinicians [7]. Given that HHcy is a common condition and many patients are initially diagnosed in primary or nonspecialist clinics, there is a pressing need for a predictive tool that offers predictability, universality, and interpretability to facilitate early risk assessment and intervention in clinical practice.

Machine learning technology holds potential for predicting the likelihood of diabetes, arrhythmia, and various cardiovascular diseases by effectively capturing intricate multidimensional patterns [8]. Ensemble models, such as light gradient boosting machine (LightGBM) and extreme gradient boosting (XGBoost), excel in automatically discerning key features and enhancing predictive accuracy [9]. Nevertheless, there is a scarcity of dedicated modeling research on CHD risk among individuals with HHcy. Moreover, the clinical interpretability of numerous machine learning models remains inadequate, constraining their reliability and utility in medical settings [10,11].

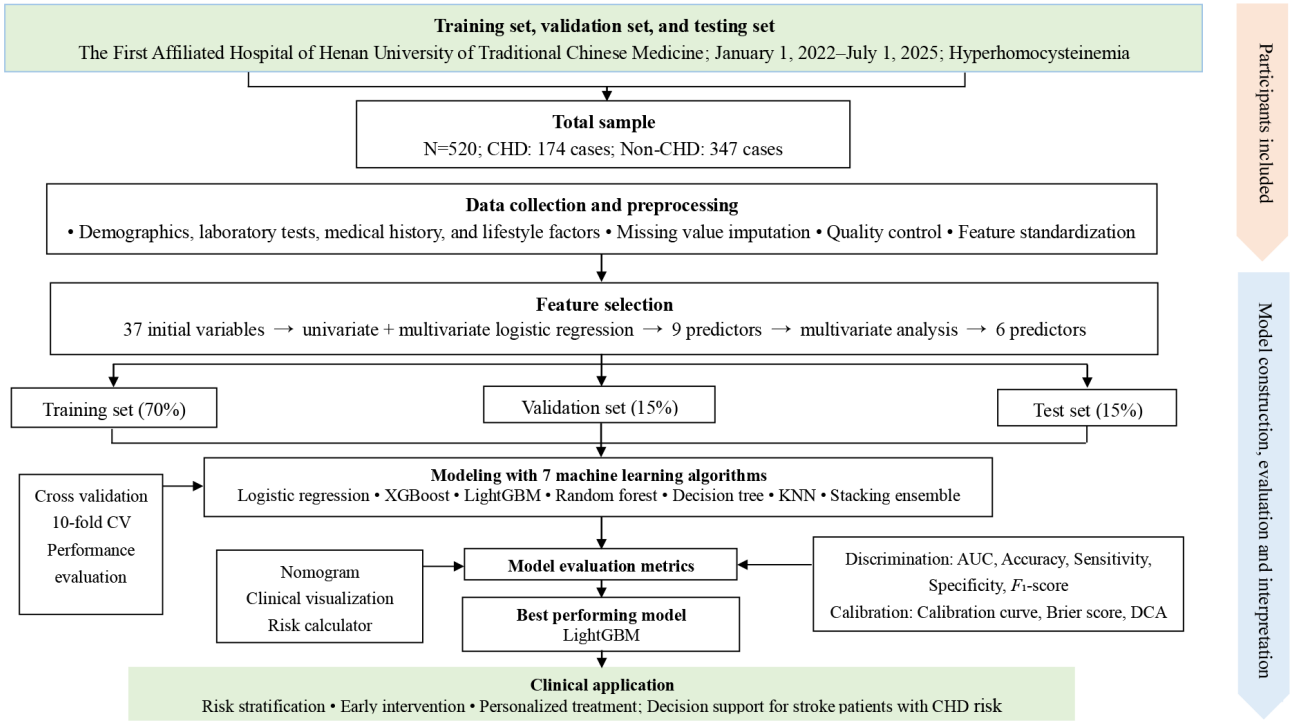
This study seeks to develop CHD risk prediction models for individuals with HHcy using real-world electronic health record data from a single center [12]. The study aims to compare the performance of 7 mainstream machine learning algorithms in terms of discriminative power, calibration, and clinical net benefit. To improve clinical interpretability, the Shapley Additive Explanation (SHAP) method was used to assess model decision-making and identify key variables. Emphasis was placed on the generalizability of variable selection and form, prioritizing standardized biochemical and lifestyle indicators that are routinely accessible to avoid reliance on specific specialties or complex imaging techniques. The objective is to create a practical model that can be easily implemented across various clinical settings, balancing predictability and interpretability. This model aims to serve as an early risk assessment tool for high-risk HHcy populations and establish a basis for future multicenter external validation studies.

Methods

Research Materials

A comprehensive workflow of the study, including data collection, feature selection, model development, evaluation, and interpretation, is illustrated in Figure 1.

Figure 1. Flowchart of developing and validating machine learning–based models for predicting coronary heart disease (CHD) risk in patients with hyperhomocysteinemia. This flowchart outlines the entire process of constructing and evaluating risk prediction models for CHD in patients with hyperhomocysteinemia. AUC: area under the receiver operating characteristic curve; CV: cross-validation, DCA: decision curve analysis; KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; XGBoost: extreme gradient boosting.



General Information

This retrospective registry study examines patients with HHcy treated at the First Affiliated Hospital of Henan University of Traditional Chinese Medicine between January 1, 2022, and July 1, 2025.

Patients were categorized into either the CHD group or the non-CHD group based on clinical diagnoses. Subsequent to

XSL•FO
RenderX

admission, demographic features, laboratory values, medical history, familial history, and additional indicators were gathered for statistical assessment.

Ethical Considerations

Ethical approval was obtained from the Ethics Committee of the First Affiliated Hospital of Henan University of Chinese Medicine (approval number: 2025HL-202‐01). The data used were fully deidentified, and no personally identifiable information was accessible to the researchers. The requirement for informed consent was waived by the ethics committee in accordance with national regulations, citing the retrospective nature of the study. No compensation was provided to participants.

Diagnostic Criteria

The patient was diagnosed with stable CHD in accordance with the 2025 American College of Cardiology/American Heart Association (ACC/AHA) guideline for chronic coronary syndromes [13]. Acute coronary syndromes were identified following the criteria outlined in the 2023 ESC and 2021 ACC/AHA guidelines for non–ST-elevation acute coronary syndrome [14,15].

Inclusion and Exclusion Criteria

Inclusion Criteria

Participants were required to meet the diagnostic criteria for HHcy, be aged between 18 and 80 years, and have complete baseline clinical data available for analysis.

Exclusion Criteria

Participants excluded from the study encompass those with comorbid severe organic conditions (eg, advanced tumors, hepatorenal insufficiency), nonatherosclerotic heart ailments (eg, congenital heart disease, myocarditis), individuals with mental illness or cognitive deficits hindering study cooperation, pregnant or lactating women, and those engaged in other clinical trials within the preceding 3 months.

Observation Indicators

This study utilized data collected through a standardized case report form, encompassing five key categories of core indicators: (1) demographics and basic information, such as age, gender, height, weight, BMI, educational level, and marital status; (2) medical history and lifestyle factors, including a history of diseases, such as hypertension, diabetes, transient ischemic attack, as well as lifestyle habits, such as persistent smoking and alcohol consumption; (3) laboratory examination indicators, comprising blood routine parameters (red blood cell, white blood cell, and platelet counts), biochemical markers (total cholesterol, triglycerides, and creatinine), and coagulation function tests (activated partial thromboplastin time [APTT], prothrombin time, and fibrinogen); (4) imaging examination indicators, specifically the presence or absence of carotid plaques as determined by ultrasound or computed tomography angiography; and (5) quality control measures involving the uniform extraction of all data by trained researchers through the electronic medical record system and laboratory information system. A double-person review system was implemented to ensure data accuracy. Following the integrity assessment, no missing values were identified in the variables included, thus obviating the need for imputation or data elimination procedures. The detailed data extraction workflow and quality control procedures are provided in Multimedia Appendix 1.

Machine Learning Modeling and Related Statistical Methods

Statistical analysis was performed using SPSS Statistics 27. Machine learning modeling utilized R 4.5 (caret package) and Python 3.13 (scikit-learn, xgboost, lightgbm) libraries. Quantitative data were reported as mean (SD). Group differences were assessed using independent-sample 2-tailed *t* tests (for normal distribution) or Mann-Whitney *U* tests (for nonnormal distribution), based on data distribution. Categorical data were presented as n (%), with group differences analyzed using the chi-square test or Fisher exact probability method.

Variable Screening and Dataset Division

Variable screening involved a 2-step dimensionality reduction approach utilizing “univariate+multivariate logistic regression.” Initially, 9 variables exhibiting significant differences ($P<.05$) between the CHD and non-CHD groups were selected for the candidate pool. Subsequently, multivariate analysis was conducted using stepwise regression (backward method, inclusion criterion $\alpha=.05$, exclusion criterion $\alpha=.10$), resulting in the retention of 6 independent predictors: age, body weight, hypertension, history of continuous alcohol consumption, APTT, and carotid artery plaque. The variance inflation factors for all variables were less than 5, indicating the absence of significant multicollinearity.

The data were stratified based on the presence of CHD comorbidity, resulting in division into a training set (n=364), a validation set (n=78), and a test set (n=78) at a ratio of 7:1.5:1.5. To ensure result reproducibility, a random seed (random_state=2024) was set. The training set was used for model training and initial parameter optimization, the validation set for fine-tuning model parameters and classification threshold optimization, and the test set as an independent dataset for the final model performance assessment.

Machine Learning Model Construction

Seven classification models were developed to forecast the risk of CHD, categorized into traditional models (logistic regression, decision tree, and k-nearest neighbors [KNNs]) and ensemble models (random forest, XGBoost, LightGBM, and stacking). Hyperparameter optimization was performed on all models using grid search (GridSearchCV) on the training set to enhance the area under the receiver operating characteristic curve (AUC) on the validation set.

Logistic regression employs L2 regularization and class weight balancing (class_weight='balanced') with a regularization strength of C=0.1. The decision tree and random forest models optimize the max_depth, which is ultimately determined to be 7, with min_samples_split set to 2 for the decision tree and n_estimators set to 50 for the random forest. KNN utilizes the Euclidean distance metric, with the number of neighbors ultimately set to 9.

XSL•FO
RenderX

In the context of boosting models, both XGBoost and LightGBM were configured with a scale_pos_weight of 2 to address class imbalance. Parameter optimization yielded the following settings: XGBoost (max_depth=3 and learning_rate=0.3) and LightGBM (max_depth=3 and learning_rate=0.01). The stacking model used the aforementioned 6 models as base learners, with logistic regression serving as the meta-learner. The integration of outputs was achieved through 5-fold cross-validation.

All parameter tuning details are listed in Table S2 in Multimedia Appendix 2 to enhance reproducibility.

Model Evaluation and Validation

A multidimensional index is used to thoroughly assess the model's performance, with the subsequent metrics computed on the training, validation, and test sets, respectively:

Discrimination is primarily assessed using the AUC, computed through "sklearn.metrics.roc_auc_score." Additionally, metrics, such as accuracy, recall, specificity, and F_1-score, are documented. To provide a comprehensive evaluation of classification performance, the precision-recall (PR) curve is generated, and the PR-AUC is determined.

Calibration is assessed through a calibration curve generated using the sklearn.calibration.calibration_curve function. An optimal calibration curve closely aligns with the diagonal line. The Brier score quantifies the accuracy of predicted probabilities by measuring the mean squared error between predicted probabilities and true labels. A lower Brier score signifies more precise predicted probabilities.

We used a 10-fold cross-validation method on the training set to assess model stability, recording the mean AUC and standard deviation. To evaluate generalization capability, we considered the metrics of the test set as the reference standard. A significant decrease in the test set's AUC compared to the training set (eg, ΔAUC<0.1) suggested potential overfitting.

Sample Imbalance Processing and Threshold Optimization

To address the 1:2 sample imbalance between the CHD and non-CHD groups, the following strategies were implemented to enhance the model's ability to identify the minority class (CHD group):

For models supporting weight parameters (eg, logistic regression, XGBoost, LightGBM, and random forest), class_weight='balanced' or scale_pos_weight=2 was set. Specifically, class_weight='balanced' automatically calculates class weights inversely proportional to class frequencies, amplifying the contribution of CHD samples in the loss function. Meanwhile, scale_pos_weight=2 directly sets the weight of positive samples (CHD group) to twice that of negative samples (non-CHD group), aligning with the 1:2 sample ratio and mitigating the model's bias toward the majority class caused by imbalance.

Although class weight adjustment balances class importance during training, the default classification threshold (0.5) may not meet the clinical demand for high sensitivity in CHD screening. Therefore, further threshold optimization was performed on the best-performing LightGBM model. Based on predicted probabilities from the validation set, candidate thresholds ranging from 0.1 to 0.6 (stepping at 0.05) were evaluated, with sensitivity, specificity, and F_1-score calculated for each threshold. The optimal threshold was determined by balancing these metrics—prioritizing high sensitivity (to identify as many true CHD patients as possible) while maintaining acceptable specificity. This approach breaks the limitation of a fixed threshold, enabling the model to better align with the practical goal of "reducing missed diagnoses" in imbalanced scenarios (see Figure S1 in Multimedia Appendix 3).

Analysis of Model Interpretability

The LightGBM model's optimal performance was assessed using the SHAP algorithm for both global and local interpretability analyses. Globally, the SHAP summary plot illustrated the average contribution of each variable to the model prediction, identifying key factors influencing the risk of CHD. Locally, the SHAP dependence plot depicted the functional relationship between each feature and the model output, highlighting the positive and negative effects of variables, such as age and APTT, on the predicted value. Additionally, a Pearson correlation heatmap was generated to assess collinearity among features, confirming that no strong correlations (r<0.5) existed among the model variables. This analysis ensured the model's stability and bolstered the credibility of its interpretation.

Sample Size Estimation

Based on the recommendations by van Smeden et al [16,17] and Riley et al [18,19], we estimated the minimum sample size required to develop a reliable multivariable prediction model using 5 candidate predictors. Assuming an outcome prevalence of approximately 30%, a shrinkage factor of 0.9, and an expected Cox-Snell R^2 of 0.15, the minimum required sample size was calculated to be approximately 218. This study developed models with a larger sample size, thus mitigating the potential for major overestimation bias based on these parameters.

External Benchmark Models

For comparative purposes, Framingham risk score [20] and the ACC/AHA pooled cohort equations (ASCVD score) [21] were calculated for all patients using baseline clinical parameters. These scores were included as benchmark models to evaluate the incremental value of the model.

Results

Comparison of Clinical Data Between the 2 Groups

This study comprised 520 patients with HHcy, with 174 in the CHD group and 347 in the non-CHD group, resulting in a CHD detection rate of 33.46%. The analysis of baseline data revealed significant differences in 9 variables between the 2 groups, all with P<.05, as detailed in Table 1.

XSL•FO
RenderX

Table . Comparison of baseline characteristics between patients with hyperhomocysteinemia with and without coronary heart disease (CHD)[a].

Characteristic	Control group (n=347)	Case group (n=174)	*P* value
Demographic and general information			
Age (y), mean (SD)	65.30 (10.36)	70.49 (8.76)	<.001[b]
Sex, n (%)			<.001[b]
Male	263 (0.71)	107 (0.29)	
Female	83 (0.55)	67 (0.45)	
Height (cm), mean (SD)	167.89 (7.29)	167.62 (7.30)	.78
Weight (kg), mean (SD)	66.96 (10.60)	71.52 (9.91)	.04[b]
BMI, mean (SD)	24.73 (3.22)	25.12 (3.06)	.07
Education level, n (%)			.86
Junior high school or below	224 (0.66)	114 (0.34)	
High school and above	122 (0.67)	60 (0.33)	
Household annual income (RMB 10,000 yuan[c]), n (%)			.13
<8	280 (0.67)	150 (0.33)	
≤8	66 (0.73)	24 (0.27)	
Occupation in the 6 months prior to onset, n (%)			.66
Mental workers	43 (0.64)	24 (0.36)	
Physical workers	303 (0.74)	105 (0.26)	
Payment type, n (%)			.12
Basic medical care	232 (0.69)	102 (0.31)	
Urban medical care	113 (0.61)	72 (0.39)	
Marital status, n (%)			.65
Married	333 (0.67)	166 (0.33)	
Separated	13 (0.62)	8 (0.32)	
Body temperature	36.46 (0.29)	36.45 (0.22)	.85
Respiratory rate	18.95 (4.67)	18.68 (1.76)	.82
Heart rate	77.14 (12.62)	77.41 (12.64)	.88
Systolic blood pressure	143.87 (21.69)	141.49 (21.09)	.16
Diastolic blood pressure	84.19 (13.92)	81.67 (14.37)	.05
Medical history, n (%)			
Hypertension	209 (0.61)	134 (0.39)	<.001[b]
Diabetes mellitus	176 (0.70)	76 (0.30)	.14
Hyperlipidemia	79 (0.70)	36 (0.30)	.58
History of TIA[d]	20 (0.56)	16 (0.44)	.15
Continuous smoking history	135 (0.75)	44 (0.25)	<.001[b]
Continuous drinking history	82 (0.78)	20 (0.22)	<.001[b]
Carotid plaque	275 (0.64)	154 (0.36)	.01[b]
Blood biochemical indicators, mean (SD)			
Red blood cell count	4.46 (0.48)	4.43 (0.52)	.32
White blood cell count	6.90 (4.28)	6.61 (1.93)	.80
Platelet count	207.82 (60.41)	211.49 (53.30)	.32

XSL•FO
RenderX

Characteristic	Control group (n=347)	Case group (n=174)	*P* value
Hemoglobin concentration	141.76 (29.06)	143.21 (38.77)	.15
Total cholesterol	4.42 (1.27)	4.39 (1.24)	.63
Triglycerides	1.73 (1.59)	1.65 (1.10)	.39
Low-density lipoprotein	2.63 (0.87)	2.53 (0.98)	.11
High-density lipoprotein	1.14 (0.29)	1.20 (0.30)	.06
Prothrombin time	11.37 (2.76)	10.86 (2.96)	.84
Fibrinogen content	3.15 (0.95)	3.19 (0.75)	.11
Activated partial thromboplastin time	30.52 (4.95)	29.44 (3.82)	.01[b]
Thrombin time	18.67 (64.39)	15.23 (1.89)	.34
Glycated hemoglobin	6.50 (1.63)	6.44 (1.28)	.17
Creatinine	74.44 (70.07)	65.23 (21.57)	.04[b]
Uric acid	313.56 (92.76)	310.13 (90.29)	.64

[a]The table includes demographic information, clinical history, lifestyle factors, and laboratory indicators for a total of 520 patients. Quantitative variables are presented as mean (SD), while categorical variables are shown as counts and percentages. Group differences were assessed using independent-sample *t* tests for normally distributed data, Mann-Whitney *U* tests for nonnormally distributed data, and chi-square tests for categorical data.

[b]P<.05.

[c]Conversion based on an exchange rate of 1 US $≈7 RMB.

[d]TIA: transient ischemic attack.

Univariate Regression Analysis of Risk Factors for CHD in HHcy Patients

Nine statistically significant factors were selected as independent variables for logistic regression analysis to predict the occurrence of CHD. The results revealed that 6 variables—age, weight, hypertension, continuous drinking history, APTT, and carotid plaque—significantly influenced the occurrence of CHD (all *P*<.05), as shown in Table 2.

Table . Logistic regression analysis of risk factors for coronary heart disease (CHD) in patients with hyperhomocysteinemia.[a]

Factor	OR[b]	SE	Wald statistics	*P* value	95% CI
Age	1.06	0.01	22.68	<.001	1.08-1.03
Gender	0.69	0.26	2.12	.15	1.14-0.42
Weight	1.04	0.01	13.90	<.001	1.06-1.02
Hypertension	1.99	0.23	9.26	<.001	3.09-1.28
Continuous smoking history	0.90	0.27	0.14	.71	1.54-0.53
Continuous drinking history	0.51	0.33	4.03	.045	0.99-0.27
APTT[c]	0.95	0.02	5.82	.02	0.99-0.90
Cr[d]	0.99	0.01	1.96	.16	1.00-0.98
Carotid plaque	1.79	0.29	3.99	.046	3.18-1.01

[a]The table presents the outcomes of univariate logistic regression analysis that pinpoint factors linked to the incidence of coronary heart disease in patients with hyperhomocysteinemia. It encompasses odds ratios, SE, Wald statistics, and 95% CI for each factor.

[b]OR: odds ratio.

[c]APTT: activated partial thromboplastin time.

[d]Cr: creatinine.

XSL•FO
RenderX

Establishment and Evaluation of 7 Machine Learning Models for CHD Risk Prediction in Patients With HHcy

This study utilized 7 machine learning algorithms, namely logistic regression, XGBoost, LightGBM, random forest, decision tree, KNNs, and stacking ensemble, to develop risk prediction models for CHD in patients with HHcy based on 6 core predictors: age, weight, hypertension, continuous drinking history, APTT, and carotid plaque. In the training set assessment, each model exhibited distinct performance characteristics. Notably, the decision tree, random forest, and XGBoost models achieved perfect discrimination (AUC=1.000) and classification metrics (accuracy, sensitivity, specificity, and F_1-score=1.000), indicating potential overfitting, unlike the logistic regression and KNN models. The LightGBM model displayed strong performance with an AUC of 0.987, accuracy of 0.854, and F_1-score of 0.818. Conversely, the stacking ensemble model yielded an AUC of 0.933, accuracy of 0.810, and F_1-score of 0.660, demonstrating inferior performance compared to the aforementioned potentially overfitting models.

During the validation phase, model performance varied notably among different algorithms. The XGBoost model exhibited the highest AUC of 0.802 and an F_1-score of 0.689. The stacking model outperformed others with an AUC of 0.800, the highest accuracy of 0.769, and an F_1-score of 0.625. The LightGBM model achieved an AUC of 0.780 and an F_1-score of 0.648. In comparison, the logistic model yielded an AUC of 0.755 and an F1-score of 0.639. The random forest model attained an AUC of 0.777; however, it demonstrated low sensitivity at 0.355 and an F_1-score of 0.489. The decision tree model, with an AUC of 0.679 and an F_1-score of 0.613, displayed limited generalization ability and was notably impacted by overfitting during training.

During testing, the LightGBM model demonstrated the highest AUC of 0.807, a sensitivity of 0.913, and an F_1-score of 0.636, indicating promising performance. The logistic model yielded an AUC of 0.796, an F_1-score of 0.603, a sensitivity of 0.826, and a specificity of 0.618, striking a balance between prediction accuracy and clinical relevance. In contrast, the stacking model achieved an AUC of 0.802 but exhibited lower sensitivity at 0.478 and an F_1-score of 0.564, suggesting limited generalizability. Although the random forest model achieved the highest accuracy at 0.769, its F_1-score was 0.571. The XGBoost model's performance notably decreased, with an AUC of 0.757 and an F_1-score of 0.523 (Table 3).

XSL•FO
RenderX

Table . Performance metrics of the 7 machine learning models for predicting coronary heart disease (CHD) risk in patients with hyperhomocysteinemia across different datasets.

Model and stage	AUC[a]	Accuracy	Sensitivity	Specificity	F_1-score
Training					
Decision tree	1.000	1.000	1.000	1.000	1.000
KNN[b]	0.795	0.734	0.475	0.861	0.540
LightGBM[c]	0.987	0.854	0.992	0.787	0.818
Logistic	0.701	0.646	0.683	0.627	0.560
Random forest	1.000	1.000	1.000	1.000	1.000
Stacking	0.933	0.810	0.558	0.934	0.660
XGBoost[d]	1.000	1.000	1.000	1.000	1.000
Validation					
Decision tree	0.679	0.692	0.613	0.745	0.613
KNN	0.751	0.731	0.452	0.915	0.571
LightGBM	0.780	0.679	0.742	0.638	0.648
Logistic	0.755	0.667	0.742	0.617	0.639
Random forest	0.777	0.705	0.355	0.936	0.489
Stacking	0.800	0.769	0.484	0.957	0.625
XGBoost	0.802	0.756	0.677	0.809	0.689
Testing					
Decision Tree	0.656	0.615	0.696	0.582	0.516
KNN	0.768	0.744	0.565	0.818	0.565
LightGBM	0.807	0.692	0.913	0.600	0.636
Logistic	0.796	0.679	0.826	0.618	0.603
Random Forest	0.765	0.769	0.522	0.873	0.571
Stacking	0.802	0.782	0.478	0.909	0.564
XGBoost	0.757	0.603	0.739	0.545	0.523

[a]AUC: area under the receiver operating characteristic curve.

[b]KNN: k-nearest neighbors.

[c]LightGBM: light gradient boosting machine.

[d]XGBoost: extreme gradient boosting.

In conclusion, both the LightGBM and logistic regression models demonstrated consistent performance across the validation and testing phases, indicating their heightened clinical utility.

The performance of the 7 machine learning models (decision tree, KNN, LightGBM, logistic regression, random forest, stacking, and XGBoost) in predicting CHD in patients with hyperhomocysteinemia was evaluated. Performance metrics, such as area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, and F_1-score, were assessed on the training, validation, and test sets. The results indicate consistent discriminative ability and stability across different data partitions, with LightGBM demonstrating superior performance in the test set (AUC=0.807, F_1-score=0.636).

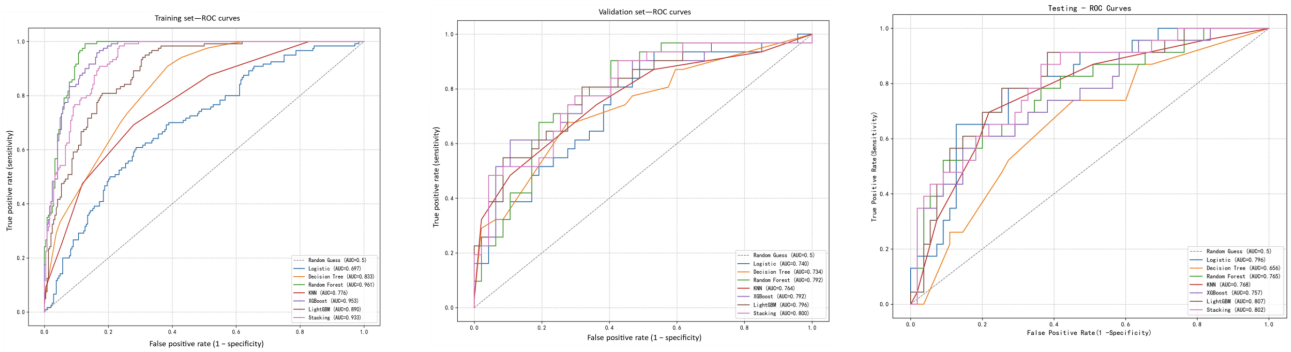
Model Performance Evaluation

Upon analyzing the ROC curves, it was evident that within the training set, the random forest (AUC=0.961), XGBoost (AUC=0.953), stacking (AUC=0.933), and LightGBM (AUC=0.890) models displayed exceptional discriminatory capabilities, indicating potential overfitting. Conversely, the logistic regression model exhibited a comparatively lower AUC of 0.697, while the KNN and decision tree models achieved AUC values of 0.776 and 0.833, respectively (Figure 2A). In the validation set, the stacking model maintained superior performance (AUC=0.800), closely followed by LightGBM (AUC=0.796), XGBoost (AUC=0.792), and random forest (AUC=0.792). The logistic model also demonstrated satisfactory performance (AUC=0.740), surpassing the decision tree (AUC=0.734) and KNN (AUC=0.764) models (Figure 2B). Within the testing set, the LightGBM (AUC=0.807) and stacking (AUC=0.802) models achieved the highest AUC values, trailed

XSL•FO
RenderX

by KNN (AUC=0.768), random forest (AUC=0.765), and XGBoost (AUC=0.757). The logistic regression model also displayed robust discriminatory ability (AUC=0.796), while the decision tree model exhibited the lowest AUC (0.656; Figure 2C).

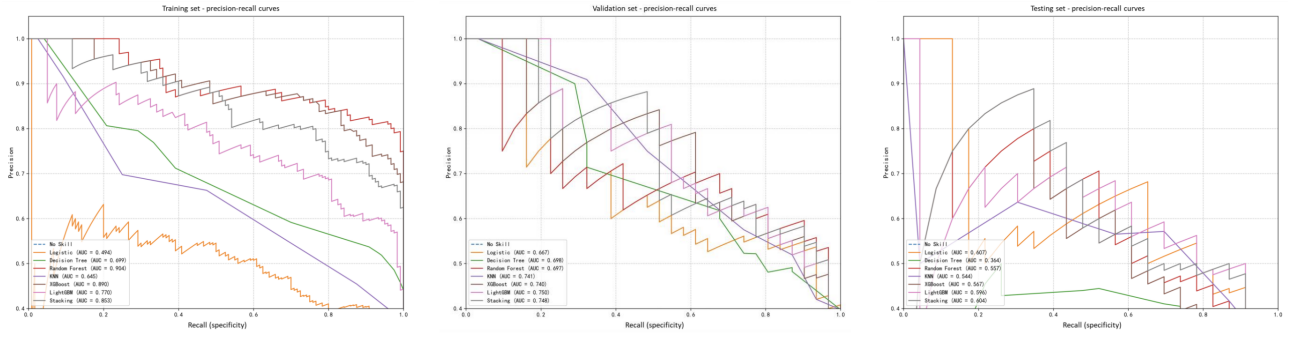
Figure 2. Receiver operating characteristic (ROC) curves of coronary heart disease (CHD) risk prediction models constructed by 7 machine learning algorithms in Hyperhomocysteinemia patients ROC) curves were utilized to assess the discriminative performance of 7 distinct machine learning models, including logistic regression, k-nearest neighbors (KNN), decision tree, random forest, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and stacking, on various datasets: the training set, validation set, and test set denoted as (A), (B), and (C) respectively. The area under the receiver operating characteristic curve (AUC) was employed as a metric to quantify the models' performance, where increased values are indicative of enhanced capability in discriminating between individuals with CHD and those without.



A B C

Upon analyzing the PR curves, distinct variations in model efficacy were evident across the three datasets. In the training dataset, the random forest model exhibited the highest PR-AUC value of 0.904, outperforming XGBoost (0.890), stacking (0.853), and LightGBM (0.770), underscoring its robust classification capability on the internal data. Conversely, the Logistic model displayed limited discriminatory power with an AUC of 0.494, while the decision tree and KNN models achieved PR-AUCs of 0.699 and 0.645, respectively (Figure 3A).

Figure 3. Precision-recall (PR) curves of coronary heart disease (CHD) risk prediction models constructed by the 7 machine learning algorithms. PR curves evaluate model performance in imbalanced datasets, emphasizing precision (positive predictive value) and recall (sensitivity). The results for the training set are displayed in (A), the validation set in (B), and the test set in (C). The PR-AUC (area under the precision-recall curve) is reported for each model, with greater values denoting better classification performance for the positive class (CHD). KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; XGBoost: extreme gradient boosting.



A B C

In the validation dataset, the LightGBM model demonstrated the highest PR-AUC of 0.750, followed closely by stacking (0.748) and KNN (0.741). The XGBoost and random forest models performed comparably with AUCs of 0.740 and 0.697, respectively, whereas the logistic model yielded a PR-AUC of 0.667, and the decision tree model slightly lagged behind with an AUC of 0.698 (Figure 3B).

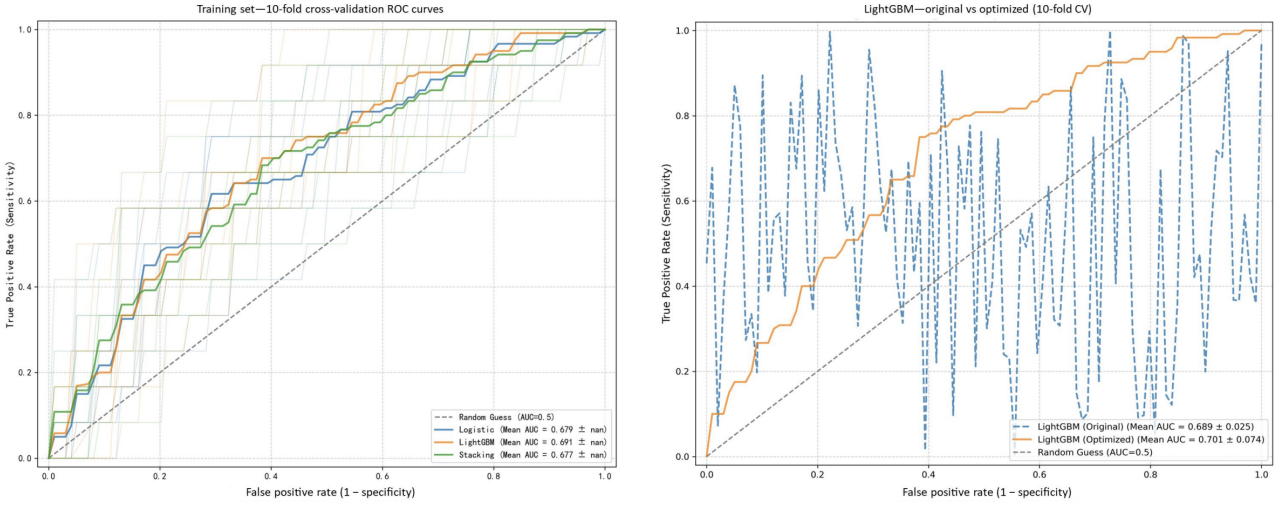
In the testing dataset, the LightGBM (AUC=0.596) and stacking (AUC=0.604) models exhibited moderate PR performance. The logistic model followed with an AUC of 0.607. Conversely, XGBoost (0.567), random forest (0.557), and KNN (0.544) demonstrated slightly lower PR-AUCs, with the decision tree model displaying the weakest performance (AUC=0.364; Figure 3C).

ROC Curve Analysis of Risk Prediction Models Based on 10-Fold Cross-Validation

ROC analysis using 10-fold cross-validation was used to assess the efficacy of 3 leading models: logistic regression, LightGBM, and stacking. As depicted in Figure 4A, during the training phase, the LightGBM model demonstrated the highest average AUC of 0.691, marginally surpassing logistic regression (AUC=0.679) and stacking (AUC=0.677), underscoring its superior discriminatory capability. Additionally, as illustrated in Figure 4B, following hyperparameter optimization, LightGBM's performance was further bolstered, with its mean AUC escalating from 0.689 (0.025) to 0.701 (0.074). This enhancement substantiates LightGBM's resilience across iterations. Consequently, LightGBM was ultimately designated as the preferred model for subsequent analyses.

XSL•FO
RenderX

Figure 4. Receiver operating characteristic (ROC) curve analysis of risk prediction models based on 10-fold cross-validation. (A) ROC curves were generated for the 3 best-performing models, namely logistic regression, light gradient boosting machine (LightGBM), and stacking, through 10-fold cross-validation on the training dataset. LightGBM demonstrated the highest average area under the receiver operating characteristic curve (AUC) of 0.691, with logistic regression and stacking following closely at 0.679 and 0.677, respectively. (B) A comparison of LightGBM performance prehyperparameter and posthyperparameter optimization revealed enhancements. The optimized model exhibited a higher mean AUC of 0.701 (SD 0.074), surpassing the original model's performance of 0.689 (SD 0.025), thus validating its increased stability.

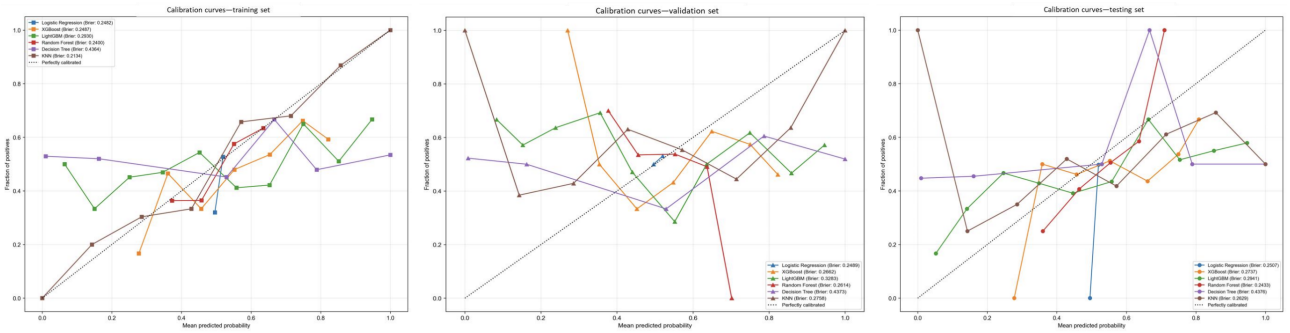


A B

Calibration Curves and Brier Score Analysis

Calibration curves and Brier scores were utilized to assess the alignment between predicted probabilities and observed outcomes across various datasets. In the training set (Figure 5A), all models exhibited moderate calibration performance. The KNN model displayed the lowest Brier score (0.2134), closely trailed by random forest (0.240), logistic regression (0.2482), and XGBoost (0.2487). Within the validation set (Figure 5B), random forest exhibited favorable calibration (Brier score=0.2614), with XGBoost (0.2662) and KNN (0.2758) also demonstrating relatively strong performance. In the test set (Figure 5C), logistic regression demonstrated the most accurate calibration (Brier score=0.2507), followed by XGBoost (0.2737) and KNN (0.2629). While the Brier scores indicated acceptable calibration for multiple models, no single model consistently outperformed others across all 3 datasets. Nevertheless, XGBoost and KNN showcased consistent and favorable calibration performance.

Figure 5. Calibration curves and Brier scores of 6 machine learning models. Calibration curves were utilized to evaluate the concordance between predicted probabilities and observed outcomes of coronary heart disease (CHD) across 3 sets: the training set, validation set, and test set. The diagonal dotted line in the graphs symbolizes ideal calibration, with Brier scores (inset) serving as a measure of prediction accuracy (lower scores indicating improved calibration). Both extreme gradient boosting (XGBoost) and k-nearest neighbors (KNN) models demonstrated consistent calibration performance across all datasets.



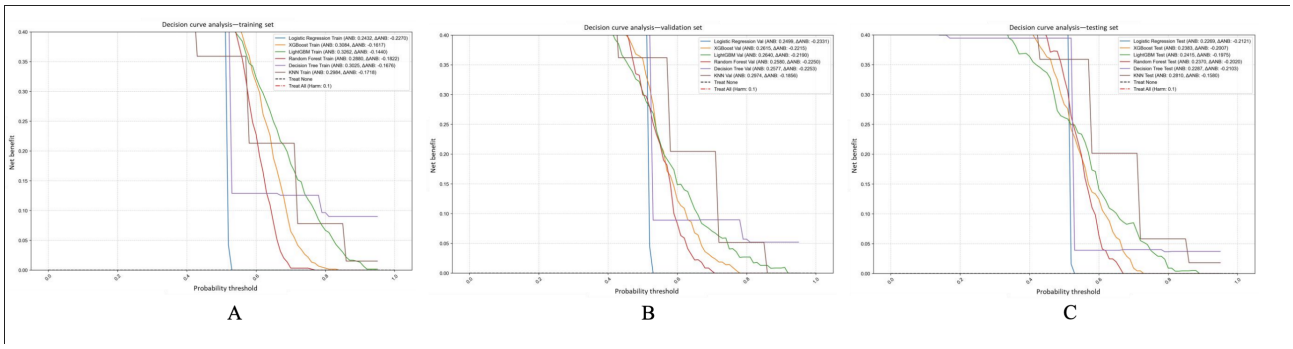
A B C

Clinical Utility Evaluation Based on Decision Curve Analysis

Decision curve analysis was used to assess the clinical efficacy of various machine learning models on the training, validation, and test sets. Figure 6 illustrates that the KNN model yielded the highest average net benefit across all 3 datasets (training: 0.2984, validation: 0.2974, and test: 0.281), closely trailed by LightGBM (training: 0.3262, validation: 0.264, and test: 0.2415) and XGBoost (train: 0.3084, validation: 0.2615, and test: 0.2383). These models consistently outperformed the "Treat None" and "Treat All" strategies in terms of net benefit over a broad range of probability thresholds. Notably, LightGBM and KNN consistently demonstrated superior clinical utility, particularly within the 0.3 - 0.7 threshold range, suggesting their potential for guiding personalized preventive interventions in practical healthcare settings.

XSL•FO
RenderX

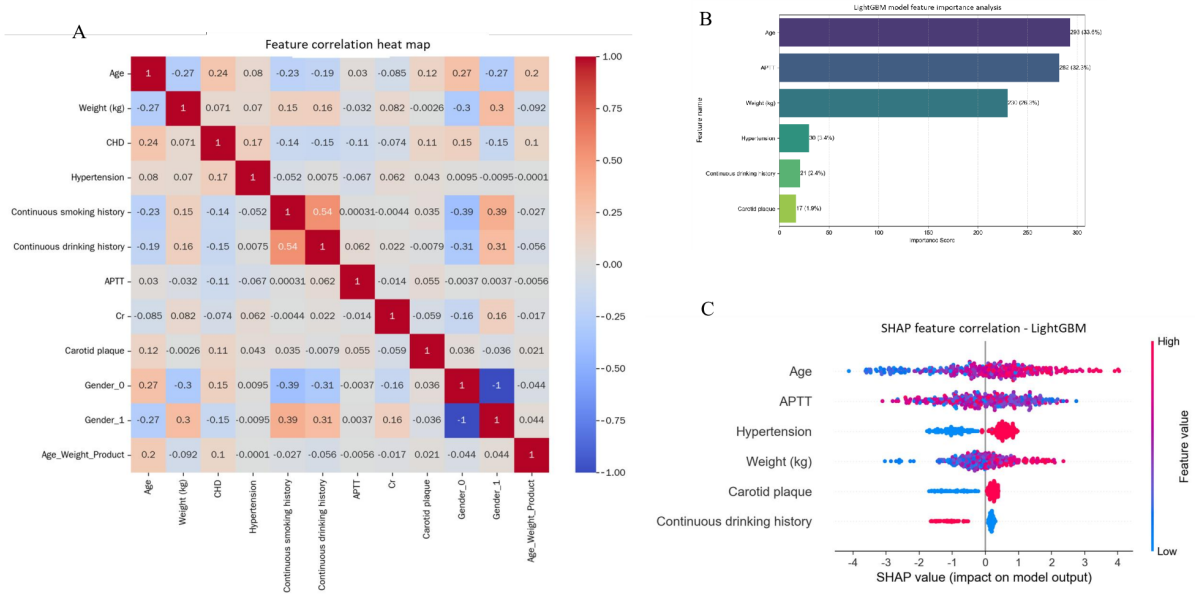
Figure 6. Decision curve analysis (DCA) of 6 machine learning models. The clinical utility of models is assessed by DCA through the calculation of net benefit at different probability thresholds in the training, validation, and test sets. Net benefit is determined as the difference between the proportion of correctly identified cases of coronary heart disease (CHD) and the proportion of unnecessary interventions. K-nearest neighbors (KNN), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost) consistently exhibited superior net benefit compared to strategies of treating all or treating none, especially within the 0.3 - 0.7 threshold range.



SHAP-Based Interpretation of the Optimal LightGBM Model

To enhance the interpretability of the LightGBM model, SHAP values were utilized to assess the significance of individual features on the model's predictions. The feature importance summary plot, derived from the mean absolute SHAP values, identified age, APTT, hypertension, weight, carotid plaque, and continuous drinking history as the primary predictors influencing the risk of CHD in patients with HHcy (Figure 7A). Notably, age and APTT exhibited the most substantial average impact, underscoring their pivotal roles in the model's predictive capabilities.

Figure 7. Shapley Additive Explanation (SHAP)–based interpretation and feature correlation analysis of the light gradient boosting machine (LightGBM) model. (A) A SHAP summary plot displays the average absolute impact of individual features on the prediction of coronary heart disease (CHD) risk, highlighting age and activated partial thromboplastin time (APTT) as the most influential factors. (B) A SHAP dependence plot illustrates the direction (positive or negative) and strength of feature effects on the model's output, indicating that higher age and APTT levels are associated with an increased risk of CHD. (C) A Pearson correlation heatmap of key features demonstrates minimal multicollinearity ($|r|$<0.5 for all features except continuous drinking or smoking history, where r=0.54), thereby ensuring the stability of the model. Cr: creatinine.



The SHAP dependence plot (Figure 7B) illustrates the impact of each feature on the model output in terms of both direction and magnitude. The elevated values of age and APTT were consistently linked to a higher predicted risk of CHD, with hypertension and carotid plaque presence also notably increasing the SHAP value. Notably, the influence of continuous drinking history exhibited variability across individual patient profiles, suggesting potential interactions with other variables.

A feature correlation heatmap (Figure 7C) was generated to evaluate multicollinearity and the interplay among essential features. The heatmap reveals a moderate positive correlation between continuous drinking history and continuous smoking history (r=0.54) and a slight negative correlation between age

XSL•FO
RenderX

and weight (r=–0.27). These findings endorse the incorporation of these variables into the LightGBM model without introducing unnecessary redundancy.

These results demonstrate that the LightGBM model offers both superior predictive accuracy and the ability to reveal the significance and interplay of clinical features. This capability can enhance personalized risk evaluation and targeted preventive measures for coronary heart disease in individuals with HHcy.

Discussion

This research conducted a comparative analysis of 7 machine learning algorithms to forecast the likelihood of CHD in patients with HHcy using authentic single-center electronic health records. The models were developed utilizing 6 fundamental variables (age, weight, hypertension, history of sustained alcohol consumption, APTT, and carotid artery plaque), and their discriminatory capacity, calibration, and clinical applicability were thoroughly assessed across training, validation, and test datasets. The findings revealed that the LightGBM model exhibited superior overall performance, demonstrating the highest AUC (0.807) and a relatively elevated F_1-score (0.636) in the test set, underscoring its promising clinical predictive capability.

A significant discovery is that the model does not depend on variables specific to CHD or HHcy. Instead, it opts for indicators that are commonly available and extensively standardized in routine clinical settings, including coagulation function (activated partial thromboplastin time), arterial structure markers (presence of carotid artery plaques), basal metabolism (body weight), and behavioral aspects (alcohol consumption history). These variables do not necessitate expert image analysis or sophisticated molecular assays, thereby augmenting the model's practicality in primary health care facilities or noncardiovascular disciplines.

Upon applying the SHAP algorithm to interpret the model, we validated that age and APTT are the most impactful variables within the model. Notably, the predictive significance of APTT warrants special consideration. Serving as a marker for the intrinsic coagulation pathway, it could potentially signify a hypercoagulable state within the HHcy population. This pathological characteristic has been identified as a key factor contributing to accelerated atherosclerosis [22]. The SHAP dependence plot and correlation heatmap provided additional confirmation of the minimal collinearity among variables, thereby augmenting the interpretive clarity and reliability of the model.

The stacking model exhibits a high AUC of 0.800 in the validation set; however, its F_1-score decreases to 0.33 in the test set, with a sensitivity of only 0.26, indicating limited generalization ability and a predisposition toward overfitting. Conversely, the LightGBM model demonstrates consistent performance across the training, validation, and test sets, underscoring its robustness in diverse sample scenarios. While the logistic regression model shows a slightly lower AUC, its balanced F_1-score, sensitivity, and specificity underscore its reliability in terms of generalization stability.

In this study, we aim to streamline variable processing by avoiding artificial binning, segmentation, high-dimensional mapping transformations, and the creation of interaction terms. This approach is intended to enhance the model's generalizability across different datasets by minimizing reliance on specific data structures. Additionally, we refrain from using oversampling methods, such as synthetic minority over-sampling technique, to mitigate issues related to feature distribution drift resulting from synthetic sample generation. Instead, we tackle data imbalance by adjusting category weights and optimizing thresholds.

However, this study is subject to several limitations. First, the data were sourced from a singular medical institution, potentially introducing regional, health care–seeking motivation, or detection biases. For instance, individuals with high homocysteine levels who consented to coagulation function tests may exhibit distinct health awareness and disease profiles compared to those who declined the tests. Second, the retrospective design of the study resulted in the exclusion of crucial variables that could impact CHD risk, such as dietary patterns, nutritional status, and renal function. Consequently, the interpretation of changes in serum markers remains inconclusive. Furthermore, the study solely predicted the present CHD status of patients documented in electronic health records. The absence of longitudinal follow-up data precluded the modeling of event timing or disease progression. To address these limitations, future research should integrate prospective cohorts and long-term follow-up data to validate the model's reliability and establish a dynamic risk assessment framework.

In conclusion, the LightGBM model developed in this study exhibits exceptional predictive accuracy and interpretability in individuals with HHcy, offering a robust basis for real-world implementation. We propose its consideration as a potential tool for early CHD risk stratification to aid clinicians in promptly identifying high-risk patients. Subsequent investigations may concentrate on external validation and model refinement in a multicenter context.

Funding

This research was supported by the National Natural Youth Science Foundation of China (grant numbers 82205021 and 82405290) and the Science and Technology Innovation Guidance Plan Project in the Medical and Health Field of Zhengzhou City (grant number 2024YLZDJH075).

Data Availability

The datasets produced and/or scrutinized in this study are not publicly accessible to safeguard patient confidentiality and comply with institutional guidelines. However, interested parties may obtain the data from the corresponding author upon request.

XSL•FO
RenderX

Authors' Contributions

The study was designed by and data analysis was supervised by MKL and MYD. Data collection was conducted by HL, HY, XL, and YL. MKL and MYD contributed to manuscript writing and revision. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed data extraction and quality-control protocols.
[DOCX File, 18 KB - medinform_v14i1e80809_app1.docx]

Multimedia Appendix 2
Grid search ranges and final hyperparameters for all machine learning models.
[DOCX File, 18 KB - medinform_v14i1e80809_app2.docx]

Multimedia Appendix 3
Threshold optimization curve for the light gradient boosting machine (LightGBM) model.
[PDF File, 223 KB - medinform_v14i1e80809_app3.pdf]

References

1. Guéant JL, Guéant-Rodriguez RM, Oussalah A, Zuily S, Rosenberg I. Hyperhomocysteinemia in cardiovascular diseases: revisiting observational studies and clinical trials. Thromb Haemost 2023 Mar;123(3):270-282. [doi: 10.1055/a-1952-1946] [Medline: 36170884]
2. Tinelli C, Di Pino A, Ficulle E, Marcelli S, Feligioni M. Hyperhomocysteinemia as a risk factor and potential nutraceutical target for certain pathologies. Front Nutr 2019;6:49. [doi: 10.3389/fnut.2019.00049] [Medline: 31069230]
3. Balint B, Jepchumba VK, Guéant JL, Guéant-Rodriguez RM. Mechanisms of homocysteine-induced damage to the endothelial, medial and adventitial layers of the arterial wall. Biochimie 2020 Jun;173:100-106. [doi: 10.1016/j.biochi.2020.02.012] [Medline: 32105811]
4. Herrero-Fernandez B, Gomez-Bris R, Somovilla-Crespo B, Gonzalez-Granado JM. Immunobiology of atherosclerosis: a complex net of interactions. Int J Mol Sci 2019 Oct 24;20(21):5293. [doi: 10.3390/ijms20215293] [Medline: 31653058]
5. Cetin E, Raby AC. Understanding atherosclerotic plaque cellular composition: recent advances driven by single cell omics. Cells 2025 May 23;14(11):770. [doi: 10.3390/cells14110770] [Medline: 40497946]
6. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019;14(5):e0213653. [doi: 10.1371/journal.pone.0213653] [Medline: 31091238]
7. Wysocki A, Fułek M, Macek P, et al. Ultrasound carotid plaque score and severity of coronary artery disease assessed by computed tomography angiography in patients with arterial hypertension. Diagnostics (Basel) 2024 May 25;14(11):1101. [doi: 10.3390/diagnostics14111101] [Medline: 38893628]
8. Pal P, Singh HV, Grover V, Manikandan R, Karimi R, Khishe M. Interactive cardiovascular disease prediction system using learning techniques: Insights from extensive experiments. Results Control Optim 2025 Jun;19:100560. [doi: 10.1016/j.rico.2025.100560]
9. Hajihosseinlou M, Maghsoudi A, Ghezelbash R. A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Nat Resour Res 2023 Dec;32(6):2417-2438. [doi: 10.1007/s11053-023-10249-6]
10. Gulum MA, Trombley CM, Ozen M, Esen E, Aksamoglu M, Kantardzic M. Why are explainable AI methods for prostate lesion detection rated poorly by radiologists? Appl Sci (Basel) 2024;14(11):4654. [doi: 10.3390/app14114654]
11. Obeagu EI, Ezeanya CU, Ogenyi FC, Ifu DD. Big data analytics and machine learning in hematology: transformative insights, applications and challenges. Medicine (Baltimore) 2025;104(10):e41766. [doi: 10.1097/MD.0000000000041766] [Medline: 40068020]
12. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12(4):e0174944. [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
13. Williams MS, Levine GN, Kalra D, et al. Correction to: 2025 AHA/ACC clinical performance and quality measures for patients with chronic coronary disease: a report of the American College of Cardiology/American Heart Association joint committee on performance measures. Circ Cardiovasc Qual Outcomes 2025 Jul;18(7):e000141. [doi: 10.1161/HCQ.0000000000000141] [Medline: 40663623]

XSL•FO
RenderX

4. Byrne RA, Rossello X, Coughlan JJ, et al. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J 2023 Oct 12;44(38):3720-3826. [doi: 10.1093/eurheartj/ehad191] [Medline: 37622654]
15. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur J Prev Cardiol 2022 Feb 19;29(1):5-115. [doi: 10.1093/eurjpc/zwab154] [Medline: 34558602]
16. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016 Nov 24;16(1):163. [doi: 10.1186/s12874-016-0267-3] [Medline: 27881078]
17. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019 Aug;28(8):2455-2474. [doi: 10.1177/0962280218784726] [Medline: 29966490]
18. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II—binary and time-to-event outcomes. Stat Med 2019 Mar 30;38(7):1276-1296. [doi: 10.1002/sim.7992] [Medline: 30357870]
19. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I—continuous outcomes. Stat Med 2019 Mar 30;38(7):1262-1275. [doi: 10.1002/sim.7993] [Medline: 30347470]
20. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart study. Circulation 2008 Feb 12;117(6):743-753. [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]
21. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014 Jun 24;129(25 Suppl 2):S49-S73. [doi: 10.1161/01.cir.0000437741.48606.98] [Medline: 24222018]
22. Yin H, Cheng X, Liang Y, et al. High perceived stress may shorten activated partial thromboplastin time and lead to worse clinical outcomes in patients with coronary heart disease. Front Cardiovasc Med 2021;8:769857. [doi: 10.3389/fcvm.2021.769857] [Medline: 34912866]

Abbreviations

ACC/AHA: American College of Cardiology/American Heart Association
APTT: activated partial thromboplastin time
AUC: area under the receiver operating characteristic curve
CHD: coronary heart disease
HHcy: hyperhomocysteinemia
KNN: k-nearest neighbor
LightGBM: light gradient boosting machine
PR: precision-recall
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanation
XGBoost: extreme gradient boosting

Edited by J Klann; submitted 17.Jul.2025; peer-reviewed by L Hou, Y Zhou; revised version received 17.Jan.2026; accepted 20.Jan.2026; published 19.Mar.2026.

Please cite as:
Du MY, Lyu MK, Liu HL, Li YZ, Yan HF, Li XH
Machine Learning–Based Risk Prediction for Coronary Heart Disease Complicated by Hyperhomocysteinemia: Retrospective Study
JMIR Med Inform 2026;14:e80809
URL: https://medinform.jmir.org/2026/1/e80809
doi:10.2196/80809

© Ming-Yuan Du, Meng-Ke Lyu, Hai-long Liu, Yi-zhuo Li, Hai-feng Yan, Xiao-hui Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research on the Prediction of Coal Workers' Pneumoconiosis Based on Easily Detectable Clinical Data: Machine Learning Model Development and Validation Study

Haiquan Li[1,2], PhD; Jiaqi Jia[3], ME; Xu Shi[2], MM; Yudie Dong[3], ME; Songquan Wang[3], PhD; Yuming Cui[3], PhD; Wenlu Hang[2], PhD; Dekun Zhang[1], PhD

[1]School of Chemical Engineering & Technology, China University of Mining and Technology, Xuzhou, Jiangsu, China

[2]Department of Respiratory and Critical Care Medicine, Second Affiliated Hospital of Xuzhou Medical University, 32 Meijian Road, Xuzhou, Jiangsu, China

[3]School of Mechatronic Engineering, Jiangsu Normal University, Xuzhou, Jiangsu, China

Corresponding Author:
Wenlu Hang, PhD
Department of Respiratory and Critical Care Medicine, Second Affiliated Hospital of Xuzhou Medical University, 32 Meijian Road, Xuzhou, Jiangsu, China

Abstract

Background: Coal workers' pneumoconiosis (CWP) is the most prevalent occupational disease that causes irreversible lung damage. Early prediction of CWP is the key to blocking the irreversible process of pulmonary fibrosis. The prediction of CWP based on imaging data and biomarker detection is constrained due to high cost and poor convenience.

Objective: The study aimed to use easily detectable clinical data to construct a prediction model for CWP through machine learning (ML) methods.

Methods: A prediction framework was established using a moderate-sized dataset and multidimensional clinical features, including occupational information, lung function parameters, and blood indicators. Six ML algorithms (light gradient boosting machine, random forest, extreme gradient boosting, categorical boosting, support vector machine, and logistic regression) were trained and evaluated using a stratified 5-fold cross-validation and a held-out test set. Hyperparameter optimization was performed using a unified Optuna-based strategy to ensure fair comparison across models. Model interpretability was assessed using Shapley Additive Explanation on top-performing models. In addition, an ablation analysis was conducted by retraining models after excluding job type to assess the independent predictive value of clinical biomarkers.

Results: All 6 models achieved consistently high predictive performance, and the differences among the top-performing models were small on the test set. After Optuna-based optimization, light gradient boosting machine and categorical boosting achieved high test-set area under curve values (0.974 and 0.975, respectively), while extreme gradient boosting achieved the highest recall (0.926) and F_1-score (0.952). Compared with the baseline models, hyperparameter optimization resulted in only minor performance changes, indicating robust prediction under the current feature set and evaluation protocol. Shapley Additive Explanation analysis consistently identified age, forced expiratory volume/forced vital capacity, and platelet count as key contributors to CWP risk prediction. The ablation analysis further showed that model performance remained strong after removing job type, supporting the independent predictive value of clinical features beyond occupational history.

Conclusions: The research results have confirmed the potential of combining simple multidimensional features with ML algorithms for predicting CWP and provided new ideas for early diagnosis and intervention of patients with CWP.

(JMIR Med Inform 2026;14:e80156) doi:10.2196/80156

KEYWORDS

coal workers' pneumoconiosis; disease prediction; machine learning; clinical data; job-type

Introduction

As a traditional fossil energy source, coal has long held an important position in the global energy system. A large amount of respiratory coal dust can be generated during coal mining, processing, loading, and transportation and enter the human lungs through diffusion and sedimentation, inducing the occurrence of coal workers' pneumoconiosis (CWP) [1]. The pathogenesis of CWP is complex, characterized by strong concealment in the early stage, high mortality rate in the later stage, and poor quality of life for patients. Once diagnosed, the course of the disease is irreversible and there is currently no

XSL•FO
RenderX

effective cure in clinical practice [2,3]. Early identification of CWP can delay the deterioration of the condition and prevent it from developing into progressive mass fibrosis or respiratory failure.

High kilovoltage X-ray examination is the gold standard for CWP diagnosis. To avoid the problem of overlapping and occlusion of tissue and organ images, computed tomography detection technology has also been used for CWP diagnosis [4,5]. However, the imaging differences among early patients with CWP are not significant, and there are also issues such as high cost, high radiation risk, and convenient equipment use, which collectively constrain the early identification of patients with CWP. At present, the development of biomarker detection technology has significantly improved the clinical feasibility of CWP early screening [6,7]. As a measurable biological indicator, biomarkers can objectively reflect the physiological and pathological status of the body, such as proteins, genes, and metabolites, etc. The research on the expression levels of biomarkers in the serum of patients with CWP is the most extensive, including transforming growth factor-α [8], interleukin-8 [9], noncoding RNA (such as microRNA) [10], and common lipid metabolites such as phosphatidylethanolamines and free fatty acids [11], all of which have been proven to have important guiding significance for early identification of CWP. On the other hand, the occurrence and development of lung diseases usually have an impact on changes in lung microbiota and respiratory flora. MicroRNA expression profiles [12], surfactant-associated protein A and surfactant-associated protein D [13] in bronchoalveolar lavage fluid, and transforming growth factor-β, interleukin-1β, and matrix metalloproteinase-9 in sputum [14], as well as benzene and aldehydes in volatile organic compounds of exhaled breath [15], are also commonly used for early identification of CWP. However, metabolic processes are regulated by multiple factors. The lower specificity and sensitivity reduce the reliability of early screening of CWP through a single biomarker. Meanwhile, the high cost of detecting specific biomarkers also limits the early identification of patients with CWP.

Previous studies have shown that blood routine examination, as an economical, efficient, and easy-to-operate screening method in clinical practice, has important guiding significance for early identification and risk assessment of diseases, especially in mining areas where medical resources are relatively scarce [16,17]. CWP usually leads to lung infections or the occurrence of inflammatory diseases, which are often reflected in lung function, coagulation function, inflammatory markers, etc. This provides the possibility for early identification of patients with CWP [18]. At present, there is a relative lack of research on CWP prediction based on routine clinical blood data. This study aims to develop a low-cost CWP early screening tool based on machine learning (ML) models. By establishing a 3D feature space of occupational exposure history, lung function parameters, and routine blood indicators, and combining 6 algorithms including light gradient boosting machine (LightGBM), random forest (RF), extreme gradient boosting (XGBoost), categorical boosting (CatBoost), support vector machine (SVM), and logistic regression (LR) for comparative analysis of predictive performance. In addition, an Optuna-based hyperparameter optimization strategy was applied to tune the models under a unified evaluation protocol. Finally, the Shapley Additive Explanation (SHAP) method was used to interpret model predictions and analyze the contributions of key parameters such as lung function indicators and blood indicators. The high-precision and interpretable prediction model constructed can provide theoretical basis for early screening of CWP.

Methods

Ethical Considerations

This study was approved by the Second Affiliated Hospital of Xuzhou Medical University ([2024] 082701). Due to the retrospective nature of the study and the use of deidentified data, the requirement for informed consent was waived by the institutional review board. To ensure privacy and confidentiality, all personal identifiers, such as names and national identification numbers, were removed and replaced with unique study IDs before data analysis. No financial compensation was provided to the participants as the data were extracted from routine clinical and physical examination records. Furthermore, we confirm that no identifiable information or images of individual participants are included in this manuscript or its supplementary materials.

Data Sources

Two hundred eighty-seven patients with CWP were admitted to a large tertiary hospital from June 28, 2022, to September 20, 2024. Dust-exposed workers undergoing annual occupational health examinations at the same hospital from 2022 to 2024 were considered as controls. Considering some workers attended examinations in multiple years, records were deduplicated using a unique personal identifier, and only the most recent examination record per worker was retained, yielding 2446 unique controls. These data were retrospectively extracted from the hospital's electronic medical records system and physical examination database. All participants were male, aged between 22 and 90 years, and were employees of a certain mining group. The testing report included common demographic information, job types, and routine biochemical indicators.

The clinical test data of dust-exposed workers and patients with CWP were intersected, and 17 indicators were selected as candidate features. These indicators included job type, age, forced expiratory volume/forced vital capacity (FEV1/FVC), white blood cell count (WBC), absolute neutrophil count (ANC), absolute lymphocyte count (ALC), absolute monocyte count (AMC), absolute eosinophil count (AEC), red blood cell count, hemoglobin, platelet count (PLT), alanine aminotransferase (ALT), glucose, triglycerides, cholesterol, high-density lipoprotein, and low-density lipoprotein. The raw dataset initially comprised 36 job types. However, when categorized by disease status, a highly significant class imbalance was observed that the number of healthy individuals exposed to dust was approximately 8 times that of patients with CWP. This severe imbalance phenomenon can cause the model to lean toward the majority class during training, thereby reducing its ability to recognize diseased samples and affecting the model's generalization performance [19]. Therefore, the original 36 job

XSL•FO
RenderX

titles were first merged into 11 broader job categories based on similarity in work environment and job tasks. And then examined the distribution of CWP cases across these categories and found substantial imbalance (eg, only 1 CWP case among vehicle drivers vs 218 cases among mixed excavation and coal mining workers). To avoid unstable estimates driven by rare categories and to ensure adequate case representation for modeling, we restricted the analytic cohort to 5 job categories with sufficient CWP case counts, including mixed excavation and coal mining workers, excavation workers, coal miners, winch operators, and conveyor operators. After this restriction, the final dataset included 1085 dust-exposed healthy individuals and 271 participants with CWP.

Data Preprocessing

After verification, it was found that FEV1/FVC and low-density lipoprotein had missing values, accounting for 3.68% (50/1356) and 1.48% (20/1356), respectively. In order to avoid the impact of missing values on subsequent analysis and model training, the *k*-nearest neighbor (KNN) algorithm was used to fill in the missing value variables. The KNN imputation was performed within the training data for each fold in cross-validation, ensuring that the test data remained unseen during preprocessing. The specific calculation method is shown in equation (1).

(1)d(xi,xj)∑k=1p(xik−xjk)2

Among them, x_i and x_j, respectively, represent the feature vectors of 2 samples, and x_{ik} and x_{jk}, respectively, represent the *k*th feature of these 2 samples.

The categorical variable job type was processed using one-hot encoding, which converts each category into a binary feature column. Categories include mixed excavation and coal mining, excavation workers, winch operators, conveyor operators, and coal miners. The remaining 15 continuous feature variables were standardized, and each feature was transformed into a distribution with a mean 0 (SD 1) for the model to analyze. Standardization was performed only on the training data, with the same scaling applied to the validation and test sets. The calculation method for the standard score of each feature is shown in equation (2).

(2)Z=x−μσ

Among them, x is a certain value of the feature in the original data, μ is the mean of the feature, and σ is the SD of the feature.

Statistical Analysis and Feature Selection

Statistical analysis was conducted using SPSS Statistics (version 26.0; IBM Corp). The normality of quantitative data was tested using the *K-S* test, and the homogeneity of variance was tested using the Levene test. Data with normal distribution were represented by mean (SD), and intergroup comparison was tested using 2 independent samples *t* test. The data with nonnormal distribution were represented by median (P25-P75), and Mann-Whitney *U* test was used for intergroup comparison. Categorical variables were represented by the number of examples (%), and comparison between groups was conducted using the chi-square test. The difference was statistically significant with $P<.05$. The Least Absolute Shrinkage and Selection Operator (LASSO) algorithm, an embedded feature selection method, was used for regression analysis to identify key feature variables associated with CWP, thereby reducing model complexity and enhancing generalization capability. LASSO regression achieved feature selection by performing α regularization on coefficients, shrinking the coefficients of less important features to 0 [20]. To ensure no information leakage, LASSO feature selection was performed within each fold's training data, and the same selected features were applied to the validation data within each fold. This method was combined with correlation analysis in filtering methods to comprehensively select features.

Construction and Evaluation of ML Models

Six representative ML models including tree-based ensemble learning models (LightGBM, XGBoost, RF, and CatBoost) and traditional classification algorithms (LR and SVM) were used for constructing CWP prediction models. A brief overview of each model's key characteristics and its relevance to this study is provided below.

For tree-based ensemble learning models, XGBoost uses second-order Taylor expansion for high accuracy and speed, incorporates regularization to prevent overfitting, and supports parallel computing for efficient training [21]. LightGBM uses a leaf-wise growth strategy and histogram-based feature discretization for efficiency, with built-in class weight adjustments beneficial for imbalanced datasets [22]. CatBoost uses an ordered boosting strategy for better generalization, directly handles categorical features, and uses a symmetric tree structure to reduce overfitting [23]. RF builds multiple decision trees from bootstrapped samples, randomly selects features at each split, and aggregates predictions through voting for robust classification [24].

To compare model performance on imbalanced datasets, 2 traditional models were also selected for comparison with the ensemble models. LR as a generalized linear model predicts probabilities using a sigmoid function [25]. SVM finds an optimal hyperplane to separate classes, using slack variables and a radial basis function kernel for inseparable data. Class weights were also incorporated into its objective function for imbalance handling [26]. In this study, class weights were applied in the loss functions of both traditional models to effectively handle the class imbalance issue.

Python (version 3.8.0; Python Software Foundation) software was used for model training and evaluation, randomly dividing the dataset into training and test set in an 8:2 ratio. To further assess model stability, 5-fold stratified cross-validation within the training data was used. In this procedure, the dataset was randomly divided into 5 nonoverlapping subsets. In each fold, 4 subsets were used as the training set, and the remaining 1 subset was used for validation. This process was repeated 5 times, and the average performance across all folds was taken as the evaluation metric. The test set was only used for the final model evaluation, ensuring it remained unseen during model training and hyperparameter tuning. In the process of model training and evaluation, based on the confusion matrix, the performance of the model was comprehensively judged through accuracy, precision, recall, F_1-score value, and the area under

XSL•FO
RenderX

curve (AUC) of the subjects. The corresponding calculation formulas are as follows:

(3)Accuracy=TN+FNTP+FP

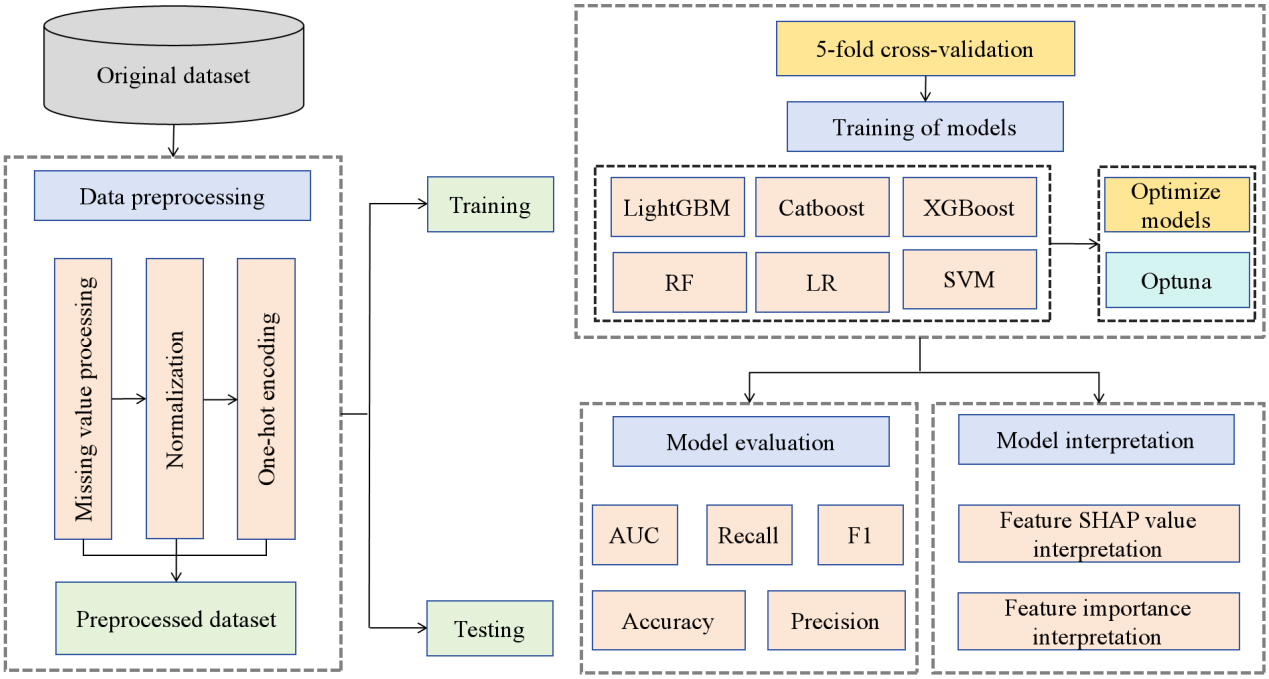
(4)Precision=TPTP+FP

(5)Recall=TPTP+FN

(6)F1=2∗precision∗recallprecision+recall

Among them, *TP* represents true positive, *TN* represents true negative, *FP* represents false positive, and *FN* represents false negative. Figure 1 shows the technical roadmap of this study.

Figure 1. The technical roadmap of this study. AUC: area under the curve; CatBoost: categorical boosting; LightGBM: light gradient boosting machine; LR: logistic regression; RF: random forest; SHAP: Shapley Additive Explanation; SVM: support vector machine; XGBoost: extreme gradient boosting.



Results

Results of Statistical Analysis and Feature Selection

Table 1 shows the comparison of basic characteristics between the dust-exposed workers and the patients with CWP. It is found that 13 indicators, including job type, age, FEV1/FVC, WBC, ANC, ALC, AMC, AEC, hemoglobin, PLT, ALT, cholesterol, and glucose, have statistical significance (P<.05) compared between the 2 groups.

XSL•FO
RenderX

Table . Comparison of basic characteristics between the dust-exposed workers and the patients with coal workers' pneumoconiosis.

	Healthy	Disease	Test	Statistic	*P* value
Job type/Case	1085	271	Pearson χ^2	—[a]	<.001
Mixed workers for excavation and coal mining, n (%)	160 (14.7)	219 (80.8)	—	—	—
Excavator workers, n (%)	253 (23.3)	35 (12.9)	—	—	—
Winch operator, n (%)	260 (24)	3 (1.1)	—	—	—
Conveyor operator, n (%)	234 (21.6)	6 (2.2)	—	—	—
Coal miners, n (%)	178 (16.4)	8 (3)	—	—	—
Age, median (IQR)	48 (39-52)	69 (62-77)	Mann-Whitney *U*	*U*=11,900, *Z*=−23.43	<.001
FEV1/FVC[b] (%), median (IQR)	90 (86-97)	75.86 (68-89)	Mann-Whitney *U*	*U*=177,046, *Z*=12.02	<.001
WBC[c] ($\times 10^9$ /L), median (IQR)	6.36 (5.42-7.60)	5.84 (4.92-6.96)	Mann-Whitney *U*	*U*=177,670, *Z*=5.32	<.001
ANC[d] ($\times 10^9$ /L), median (IQR)	3.53 (2.86-4.39)	2.97 (2.21-3.98)	Mann-Whitney *U*	*U*=185,504, *Z*=6.67	<.001
ALC[e] ($\times 10^9$ /L), median (IQR)	2.21 (1.82-2.69)	2.01 (1.58-2.72)	Mann-Whitney *U*	*U*=163,138, *Z*=2.80	.005
AMC[f] ($\times 10^9$ /L), median (IQR)	0.39 (0.33-0.48)	0.47 (0.38-0.68)	Mann-Whitney *U*	*U*=94,124, *Z*=−9.17	<.001
AEC[g] ($\times 10^9$ /L), median (IQR)	0.16 (0.10-0.25)	0.13 (0.07-0.19)	Mann-Whitney *U*	*U*=174,854, *Z*=4.83	<.001
RBC[h] ($\times 10^{12}$ /L), median (IQR)	4.89 (4.64-5.12)	4.85 (4.46-5.33)	Mann–Whitney *U*	*U*=150,982, *Z*=0.69	.49
HB[i] (g/L), median (IQR)	151 (144-158)	148 (136-160)	Mann-Whitney *U*	*U*=161,103, *Z*=2.44	.01
PLT[j] ($\times 10^9$ /L), median (IQR)	244 (211-276)	200 (158.50-240)	Welch t	*U*=208,796, *Z*=10.71	<.001
ALT[k] (U/L), median (IQR)	20 (15-27)	17 (12-24)	Mann-Whitney *U*	*U*=180,185, *Z*=5.75	<.001
GLU[l] (mmol/L), median (IQR)	5.25 (4.85-5.75)	4.91 (4.41-5.62)	Mann-Whitney *U*	*U*=180,125, *Z*=6.14	<.001
TG[m] (mmol/L), median (IQR)	1.42 (0.99-2.33)	1.28 (0.96-2.01)	Mann-Whitney *U*	*U*=157,897, *Z*=1.89	.06
CHOL[n] (mmol/L), median (IQR)	4.86 (4.29-5.52)	4.48 (3.67-5.22)	Mann-Whitney *U*	*U*=181,528, *Z*=5.98	<.001
HDL[o] (mmol/L), median (IQR)	1.27 (1.12-1.47)	1.23 (1.05-1.54)	Mann-Whitney *U*	*U*=150,479, *Z*=0.60	.55
LDL[p] (mmol/L), median (IQR)	2.67 (2.29-3.11)	2.69 (2.13-3.21)	Mann-Whitney *U*	*U*=129,398, *Z*=0.67	.50

[a]Not available.

[b]FEV1/FVC: forced expiratory volume/forced vital capacity.

[c]WBC: white blood cell.

[d]ANC: absolute neutrophil count.

[e]ALC: absolute lymphocyte count.

[f]AMC: absolute monocyte count.

XSL•FO
RenderX

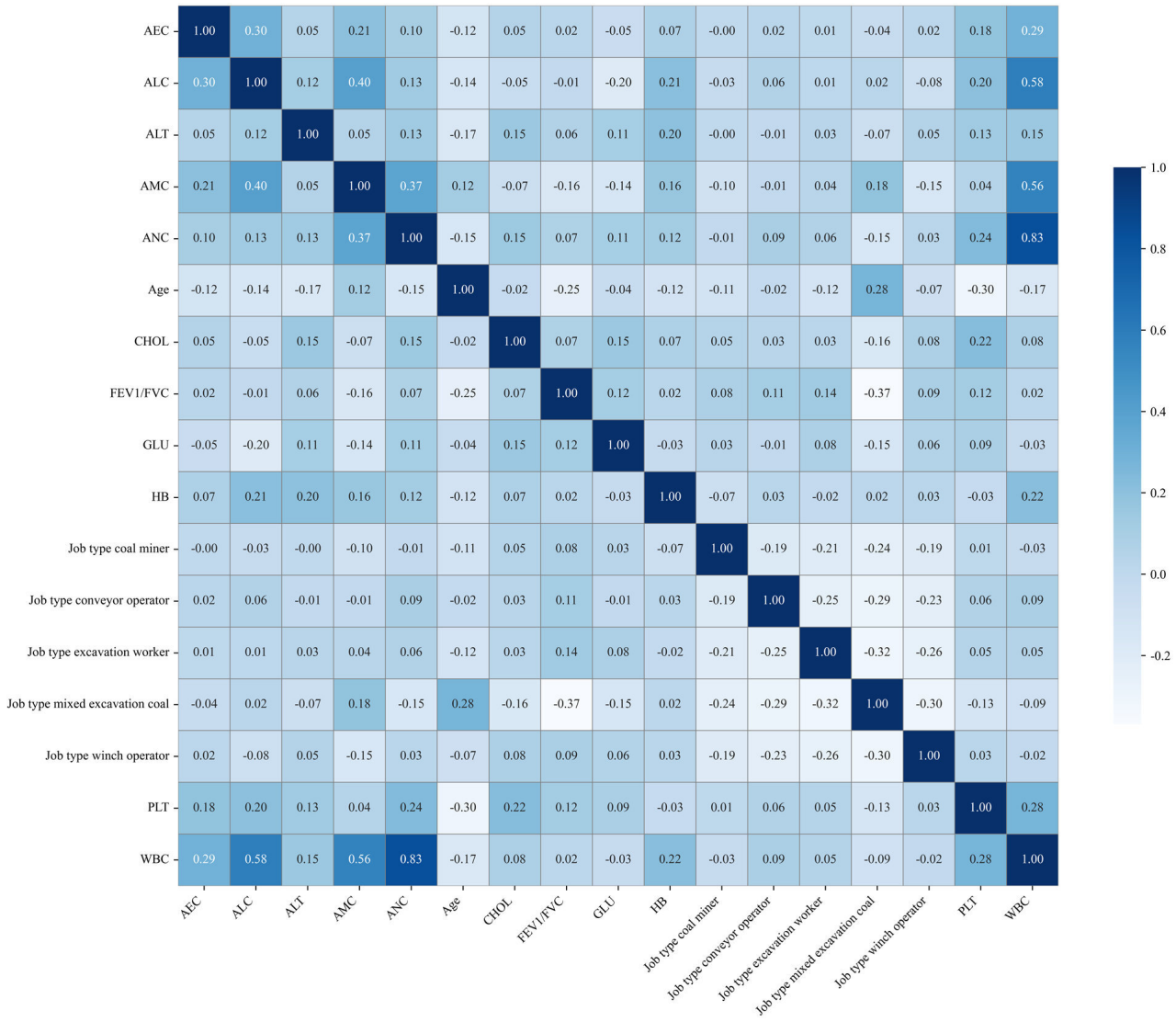
[g]AEC: absolute eosinophil count.
[h]RBC: red blood cell count.
[i]HB: hemoglobin.
[j]PLT: platelet count.
[k]ALT: alanine aminotransferase.
[l]GLU: glucose.
[m]TG: triglycerides.
[n]CHOL: cholesterol.
[o]HDL: high-density lipoprotein.
[p]LDL: low-density lipoprotein.

Figure S1 in Multimedia Appendix 1 showed the cross-validation curve of LASSO regression. When the α value is low, the model may contain too many irrelevant features, resulting in significant errors (overfitting). When the α value is large, the model may remove too many important features, which also leads to an increase in error (underfitting). At the optimal α value, the cross-validation error is minimized. The LASSO coefficient plot was shown in Figure S2 in Multimedia Appendix 1, which showed that the coefficients of 9 features, including job-type mixed excavation coal, job-type excavation worker, job-type conveyor operator, job-type winch operator, age, FEV1/FVC, AMC, PLT, and ANC, were not 0 at the optimal α value. This indicated the criticality of these features and their significant explanatory power for the target variable; therefore, they should be retained in the final model for prediction.

In order to avoid the problem of multicollinearity caused by strong correlation between features, this study used correlation analysis in a filtering method to comprehensively select features based on the 17 features selected through statistical analysis and LASSO regression screening in the early stage. By calculating the Spearman correlation coefficient matrix (Figure 2), the threshold was set to an absolute value of *r* greater than 0.8, and highly correlated terms in the feature pairs were selected. Based on the absolute value of LASSO regression weights, the features that contribute more to the target were retained, thereby eliminating redundant variables. Specifically, a Spearman correlation matrix was computed on the training data, and pairs with an absolute value of *r* greater than 0.8 were considered highly correlated. For each highly correlated pair, the feature with the larger absolute LASSO coefficient was retained and the other feature was removed. In the current dataset, WBC and ANC showed high correlation (r=0.833); thus, WBC was removed and ANC was retained. After redundancy filtering, 16 nonredundant features were used as inputs for subsequent model development, including job-type mixed excavation coal, job-type excavation worker, job-type conveyor operator, job-type winch operator, job-type coal miner, age, FEV1/FVC, ANC, ALC, AMC, AEC, hemoglobin, PLT, ALT, cholesterol, and glucose.

XSL•FO
RenderX

Figure 2. Spearman correlation coefficient matrix. AEC: absolute eosinophil count; ALC: absolute lymphocyte count; ALT: alanine aminotransferase; AMC: absolute monocyte count; ANC: absolute neutrophil count; CHOL: cholesterol; FEV1/FVC: forced expiratory volume/forced vital capacity; GLU: glucose; HB: hemoglobin; PLT: platelet count; WBC: white blood cell count.



Evaluation of the CWP Prediction Model

The selected 16 clinical features were input as feature variables into 6 ML models, and 5-fold cross-validation was performed for each model during training. The performance of each model was comprehensively evaluated based on the test-set data, and the output results of each model were organized and are summarized in Table 2. Meanwhile, the visualized results of the data in Table 2 are shown in Figure S3 in Multimedia Appendix 1. From the figure, it can be seen that the AUC of each fold of the 6 models in cross-validation was consistently high (all folds >0.90), indicating that the generalization ability of the 6 models is strong and stable. The AUC values of the 6 models were ranked in descending order as CatBoost (0.979), LightGBM (0.978), XGBoost (0.976), RF (0.972), SVM (0.968), and LR (0.967). This result indicated that the performance of ensemble learning models on imbalanced datasets was superior to traditional models, verifying the advantages of ensemble learning models in dealing with such problems [27]. Taking into account accuracy, precision, recall, and F_1-score, both the LightGBM and RF models achieved the highest accuracy (0.982), precision (1), and F_1-score (0.951), while the LR model had the highest recall (0.926). Therefore, based on the overall performance across multiple evaluation indicators, LightGBM and RF were preliminarily considered as top-performing predictive models.

XSL•FO
RenderX

Table . Evaluation of predictive performance of different models[b].

Model	Accuracy	Precision	Recall	F_1-score	AUC[a]
LightGBM[c]	0.982[b]	1.000[b]	0.907	0.951[b]	0.978
CatBoost[d]	0.978	0.980	0.907	0.942	0.979[b]
XGBoost[e]	0.978	0.980	0.907	0.942	0.976
RF[f]	0.982[b]	1.000[b]	0.907	0.951[b]	0.972
LR[g]	0.956	0.862	0.926[b]	0.893	0.967
SVM[h]	0.963	0.907	0.907	0.907	0.968

[a]AUC: area under the curve.

[b] Significant values.

[c]LightGBM: light gradient boosting machine.

[d]CatBoost: categorical boosting.

[e]XGBoost: extreme gradient boosting.

[f]RF: random forest.

[g]LR: logistic regression.

[h]SVM: support vector machine.

In order to further improve the predictive performance and to ensure a fair comparison among candidate models, this study used the Optuna algorithm to optimize the hyperparameters of all 6 ML models (LightGBM, CatBoost, XGBoost, RF, LR, and SVM) under the same optimization budget, and also conducted 5-fold cross-validation during training. The output results of each optimized model were summarized in Table 3, and the visualized results were shown in Figure S4 in Multimedia Appendix 1. The results showed that the AUC of each fold in cross-validation was consistently high (all folds >0.90), indicating that the generalization ability of the 6 models was strong and stable. After applying an equivalent hyperparameter optimization strategy, the overall performance of the 6 models remains high and the differences among the top-performing models were small. Specifically, CatBoost and LightGBM achieved high test-set AUC values (0.975 and 0.974, respectively). In addition, XGBoost achieved the highest recall (0.926) and F_1-score (0.952) on the test set. Compared with the baseline results, hyperparameter optimization led to only small changes in performance. Overall, the 6 models maintained consistently high performance under the current evaluation protocol.

Table . Performance evaluation of optimized models.

Model	Accuracy	Precision	Recall	F_1-score	AUC[a]
LightGBM[b]-Optuna	0.982	1	0.907	0.951	0.974
CatBoost[c]-Optuna	0.982	1	0.907	0.951	0.975
XGBoost[d]-Optuna	0.982	0.98	0.926	0.952	0.969
RF[e]-Optuna	0.974	0.961	0.907	0.933	0.968
LR[f]-Optuna	0.952	0.847	0.926	0.885	0.968
SVM[g]-Optuna	0.967	0.925	0.907	0.916	0.962

[a]AUC: area under the curve.

[b]LightGBM: light gradient boosting machine.

[c]CatBoost: categorical boosting.

[d]XGBoost: extreme gradient boosting.

[e]RF: random forest.

[f]LR: logistic regression.

[g]SVM: support vector machine.

Model Interpretability

In order to gain a deeper understanding of the impact of various clinical features on the model's prediction results, this study used the SHAP method to conduct interpretability analysis on the 2 representative top-performing models LightGBM-Optuna and CatBoost-Optuna. The calculation method was shown in Equation 7.

$$(7)\text{SHAP}(y)= \text{SHAP}(\text{base})+\sum_{i=1}^{n}\text{SHAP}(x_i)$$

XSL•FO
RenderX

Among them, *SHAP (base)* is the baseline value of the entire model, and *SHAP* (x_i) is the contribution of each sample to the final prediction result.

The summary results of SHAP values are shown in Figure 3, which displayed the distribution of SHAP values for 17 input feature variables. Each point in the figure represented a feature, and the position of the point represented the SHAP value of the feature, which was the contribution of the feature to the model output. If the SHAP value is positive, it indicates that the feature increases the risk of disease and has a positive impact on the output results. Conversely, if it is negative, it indicates that the feature reduces the risk of disease and has a negative impact on the output results. In addition, the color range from blue to red reflects the actual value of the feature, with red indicating high values and blue indicating low values. The darker the color, the stronger the impact of the feature on the target variable. Overall, both models showed consistent patterns in feature effects. Age is the most influential variable, and higher age values were mainly associated with positive SHAP values, suggesting that older individuals tended to have a higher predicted disease risk. In contrast, higher values of FEV1/FVC were mostly distributed on the negative side, indicating that better lung function (higher FEV1/FVC) was related to a lower predicted risk. PLT showed a similar tendency, with higher values generally corresponding to negative SHAP values. On the other hand, higher AMC values tended to correspond to positive SHAP values, indicating a positive association with increased predicted risk. These results suggested that the model predictions were largely driven by age-related factors and lung function indicators, together with selected hematological and biochemical variables.

Figure 3. Summary chart of SHAP values. (a) LightGBM-Optuna; (b) CatBoost-Optuna. AEC: absolute eosinophil count; ALC: absolute lymphocyte count; ALT: alanine aminotransferase; AMC: absolute monocyte count; ANC: absolute neutrophil count; CHOL: cholesterol; FEV1/FVC: forced expiratory volume/forced vital capacity; GLU: glucose; HB: hemoglobin; PLT: platelet count; SHAP: Shapley Additive Explanations.

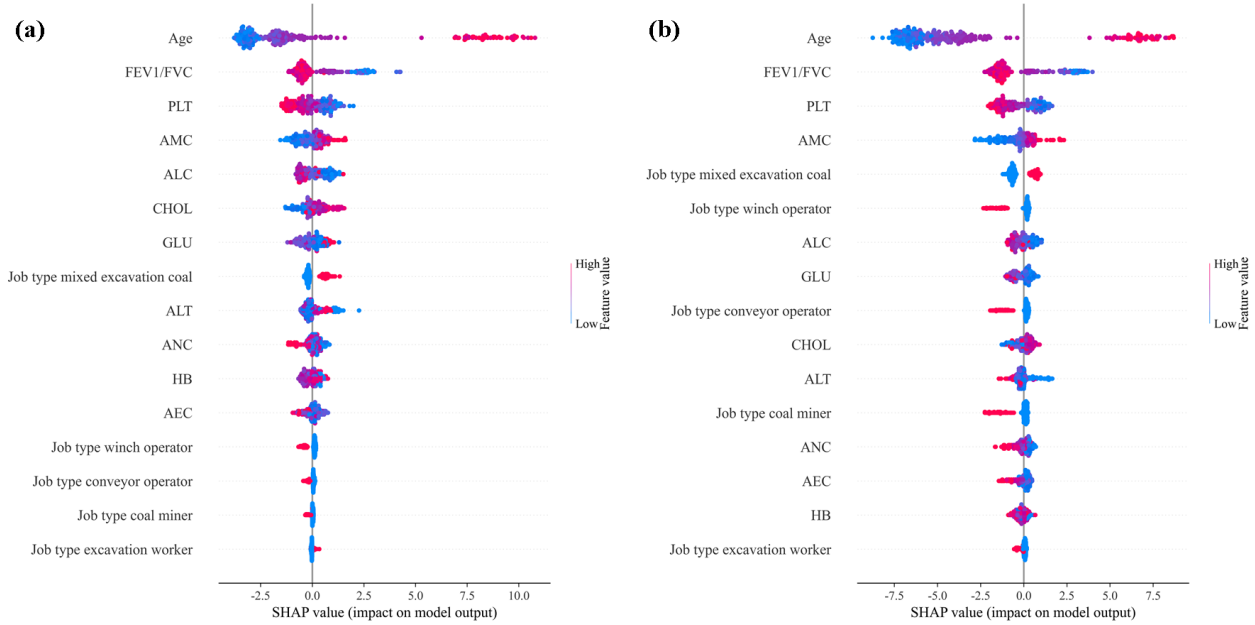
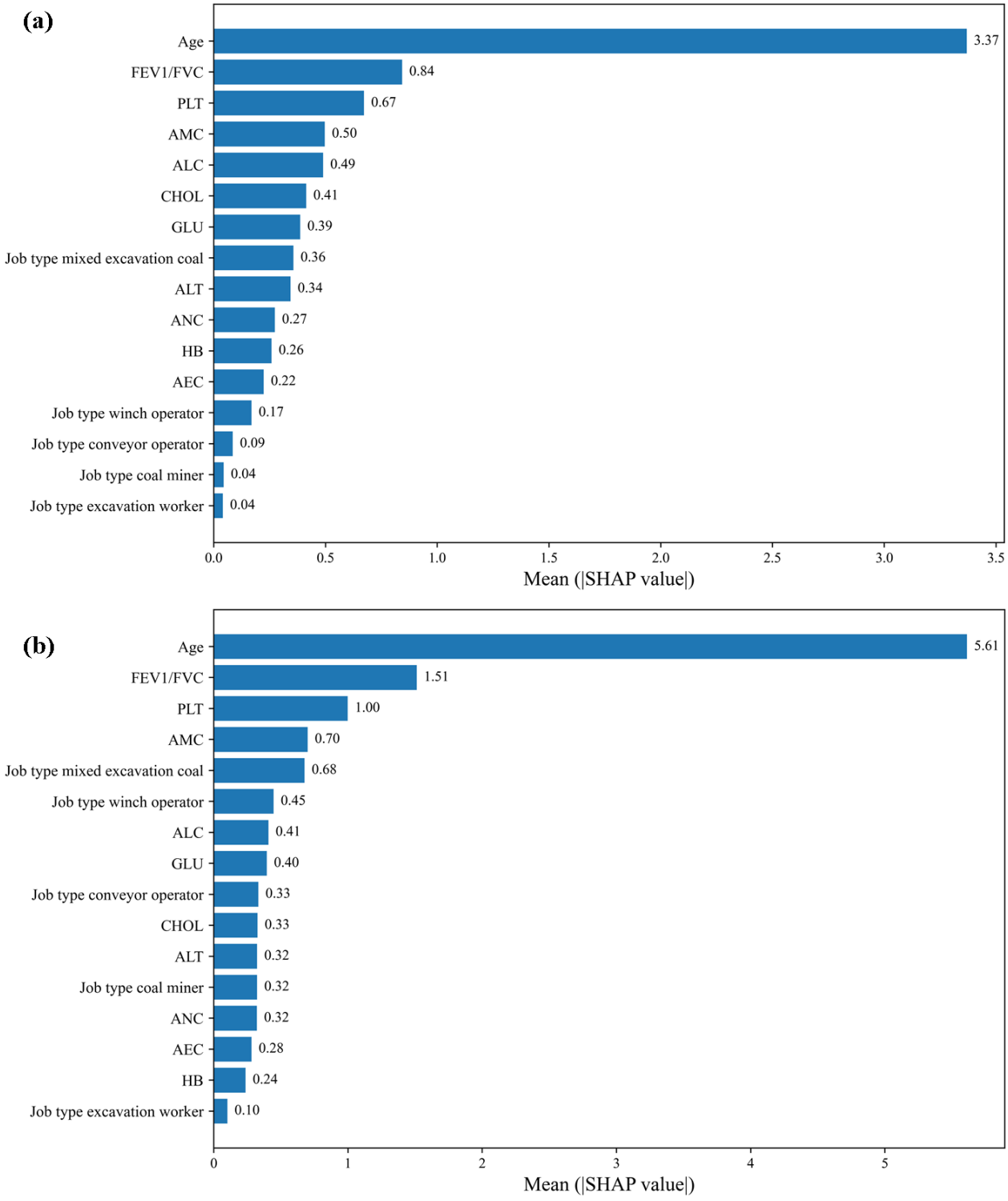


Figure 4 showed the SHAP feature importance matrix, which arranged the average SHAP absolute value of each feature from high to low. The horizontal axis represented the contribution value, and the larger the value, the greater the contribution to the model results. In both the Optuna-tuned LightGBM and CatBoost models, age showed the highest contribution, followed by FEV1/FVC and PLT, which indicates that these variables play the most important roles in the prediction of CWP risk.

XSL•FO
RenderX

Figure 4. Matrix diagram of SHAP feature importance. (a) LightGBM-Optuna; (b) CatBoost-Optuna. AEC: absolute eosinophil count; ALC: absolute lymphocyte count; ALT: alanine aminotransferase; AMC: absolute monocyte count; ANC: absolute neutrophil count; CHOL: cholesterol; FEV1/FVC: forced expiratory volume/forced vital capacity; GLU: glucose; HB: hemoglobin; PLT: platelet count; SHAP: Shapley Additive Explanation.



To further examine how the top 3 influential features affect the model output, SHAP dependence plots were generated (Figure S5 in Multimedia Appendix 1). As shown in the dependence plot for age, the SHAP values generally increase with age, and the risk contribution became more pronounced after approximately 55 - 60 years, suggesting that older age was associated with a higher predicted disease risk. For FEV1/FVC, lower values correspond to higher SHAP values, indicating an increased risk, whereas higher values (approximately 85% - 100%) were associated with SHAP values close to zero or negative, suggesting a lower predicted risk. Similarly, the PLT dependence plot showed that lower PLT levels tended to contribute positively to disease risk, while higher PLT values (approximately 250 - 350×10^9/L) were more often associated with negative SHAP values, indicating a reduced predicted risk.

XSL•FO
RenderX

Discussion

Principal Findings

In clinical research, it is very common for medical records to have missing values in a certain examination due to the complexity of data collection and individual differences among different patients. By calculating the Euclidean distance between samples, KNN interpolation can identify *K*-neighboring samples that are most similar to missing values and then use the average feature information of neighboring samples to fill in the missing values, effectively restoring the integrity of the data [28]. As a categorical variable, the values of job type do not have a sequential relationship, and categories with larger values do not necessarily have greater weights than categories with smaller values. In order to enable ML models to better capture the relationship between feature variables and target variables, single hot encoding was applied to categorical features in this study [29]. The remaining 15 characteristic variables are all continuous features, but their units and ranges of values vary greatly. For example, PLT is measured in unit ×10^9/L and has a wide range of variable values, while glucose is measured in unit mmol/L and has a smaller range of values. This inconsistent scale may lead to the model being more sensitive to certain features with larger numerical ranges during training and ignoring other features with smaller scales, thereby affecting the training effectiveness of the model. Standardizing irregular continuous variables is often the key to solving such problems [30].

Feature selection is a crucial step in ML applications, aimed at selecting the most relevant features to the target variable in order to improve model performance and interpretability [31]. The common feature selection methods mainly include embedded, wrapped, and filtered methods [32]. As an embedded feature selection method, LASSO uses regression analysis to screen out key feature variables related to CWP, reducing model complexity and improving model generalization ability. In this study, LASSO regression analysis is applied to compress the coefficients of some unimportant features to 0 by performing α regularization on the coefficients, thereby achieving feature selection [20]. The preprocessed dataset still suffers from class imbalance, with the number of dust-exposed workers without CWP being about 4 times that of patients with CWP. Regarding the issue of class imbalance, commonly used methods in model construction include data sampling and ensemble learning [33]. In order to preserve the distribution characteristics of the original data as much as possible and avoid the bias and noise that may be introduced by data oversampling methods [34], ensemble learning models were used to handle imbalanced data in this study. Specifically, we selected decision tree-based ensemble learning models such as LightGBM, RF, CatBoost, and XGBoost and compared them with traditional LR and SVM models.

To ensure a fair comparison among candidate models, this study adopted a unified hyperparameter optimization strategy based on Optuna. Optuna is a Bayesian optimization framework that uses a tree-structured Parzen estimator to efficiently explore the hyperparameter space by prioritizing promising regions. Under the same optimization budget and the same stratified *k*-fold cross-validation protocol, all 6 models were tuned and evaluated consistently. The results show that after optimization, all models achieved consistently high cross-validated performance, while the differences among the top-performing models remained small on the held-out test set. This finding suggests that the current feature set and evaluation setting already provide strong predictive ability, and further improvements are more likely to depend on feature refinement or decision strategy rather than extensive hyperparameter tuning. In addition, the top-performing models show comparable overall performance, but each presents advantages under different evaluation priorities. Specifically, models such as LightGBM and CatBoost demonstrate stronger overall discrimination, whereas XGBoost tends to perform better when recall- or F_1-related sensitivity is emphasized. Therefore, LightGBM and CatBoost were both retained as top-performing models for subsequent interpretability analysis. Job type reflects different dust exposure scenarios in coal mining and therefore contributes to CWP risk prediction, which is consistent with epidemiological evidence [35]. The concentration, particle size, and composition of coal dust have a significant impact on the pathogenesis and prevalence of CWP [36,37]. The different working scenarios in coal mines in the same region also have a significant impact on the prevalence of CWP. The excavator workers are mainly responsible for developing tunnels, and the cut rocks are rich in free silica. The pathogenicity of silica dust is much higher than that of coal dust, which can lead to more severe pulmonary fibrosis (silicosis) and a shorter onset period [38,39]. Coal miners mainly come into contact with coal dust (carbon-based dust), which has relatively weaker pathogenicity compared to silica dust and slower disease progression. At the same time, the excavation face is a temporary work site, and the ventilation and dust removal facilities are usually not as complete as those in the coal mining face, resulting in greater difficulty in dust control [40]. Mixed workers for excavation and coal mining are exposed to silica dust and coal dust simultaneously, and the synergistic effect of the 2 types of dust may accelerate lung damage [41,42].

To examine whether the clinical variables provide predictive value beyond occupational history, we conducted an ablation analysis by retraining the baseline models after removing the job-type variable. Because hyperparameter tuning led to only minor changes in performance, using baseline models for this analysis was sufficient to evaluate the independent contribution of clinical variables. The results of each model are summarized in Table 4, which showed that model performance remained highly robust. For example, the AUC of LightGBM only slightly changed from 0.978 (with job type) to 0.973 (without job type), indicating that the physiological signals captured by clinical features and biomarkers are major contributors to the model's predictive capability.

XSL•FO
RenderX

Table . Ablation analysis of model performance without job type.

Model	Accuracy	Precision	Recall	F_1-score	AUC[a]
LightGBM[b]	0.982	1	0.907	0.951	0.973
CatBoost[c]	0.978	0.98	0.907	0.942	0.977
XGBoost[d]	0.982	1	0.907	0.951	0.978
RF[e]	0.982	1	0.907	0.951	0.976
LR[f]	0.963	0.893	0.926	0.909	0.969
SVM[g]	0.974	0.98	0.889	0.932	0.971

[a]AUC: area under curve.

[b]LightGBM: light gradient boosting machine.

[c]CatBoost: categorical boosting.

[d]XGBoost: extreme gradient boosting.

[e]RF: random forest.

[f]LR: logistic regression.

[g]SVM: support vector machine.

This finding is also supported by the SHAP interpretation results. The decrease in FEV1/FVC reflects impaired lung function, while the corresponding increase in SHAP value suggests that ventilation function may be an important feature of CWP. This is consistent with the results of existing studies suggesting that alveolar-arterial oxygen gradient in lung function can be used as a predictor of CWP [43]. Platelets, as important cells for hemostasis and coagulation, play a role by participating in systemic inflammatory and immune responses, providing new therapeutic targets for inflammatory diseases [44]. For example, a previous study found that lower PLT levels were associated with a higher risk of developing severe *mycoplasma pneumoniae* pneumonia [45]. Due to the important role of platelets in inflammation and tissue repair, this phenomenon may be related to inflammation or weakened immune function leading to lung damage. These findings emphasize the importance of clinical features in CWP risk assessment and provide new perspectives for a deeper understanding of the pathogenesis of CWP.

Despite its contributions, this study has several limitations. First, the cohort was derived from a single center and a specific occupational group, which may introduce regional or selection bias and limit generalizability to other settings. Second, although model performance was evaluated rigorously, interpretability remains limited and warrants further investigation. Most importantly, smoking history was not available in the retrospective physical examination records. Because smoking is a major confounder for both lung function and inflammatory biomarkers, part of the observed discrimination may reflect unmeasured differences in smoking behavior rather than CWP status alone. Furthermore, differences in physical demands and lifestyle factors associated with distinct job roles could potentially influence certain biomarkers. While our analysis indicates strong independent predictive value for the biomarkers, future studies should consider more granular lifestyle adjustments.

Conclusions

This study developed a ML-based model for CWP prediction using multidimensional clinical features. The 6 candidate models achieved consistently high performance, and Optuna-based tuning resulted in only small changes, suggesting robust prediction under the current protocol. SHAP analysis identified age, FEV1/FVC, and PLT as key contributors to CWP risk prediction. Moreover, ablation analysis showed that the models remained highly accurate even without job type, indicating that clinical biomarkers provide strong predictive signals beyond occupational information. These results support the potential of routine clinical data for early CWP screening and intervention.

Funding

This study was supported by the National Natural Science Foundation of China (82405130) and the Natural Science Foundation of Jiangsu Province (BK20220236).

Data Availability

The datasets analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

HL, YC, WH, and DZ were responsible for the conceptualization and design of the study. JJ, XS, and YD performed the data extraction and the data analysis. HL and WH provided clinical advice and recommendations on usability and clinical relevance. HL, JJ, and SW drafted the original manuscript. All authors critically reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cross-validation curve of least absolute shrinkage and selection operator regression, least absolute shrinkage and selection operator coefficient chart, cross-validation of 6 models and receiver operating characteristic curves of the final model, receiver operating characteristic curves of optimized model, and Shapley Additive Explanation dependency graph.
[DOCX File, 17478 KB - medinform_v14i1e80156_app1.docx]

References

1. Vanka KS, Shukla S, Gomez HM, et al. Understanding the pathogenesis of occupational coal and silica dust-associated lung disease. Eur Respir Rev 2022 Sep 30;31(165):210250. [doi: 10.1183/16000617.0250-2021] [Medline: 35831008]
2. Blackley DJ, Hall NB, Flattery J, Harris DA, Cummings KJ, Laney AS. Rise in lung transplants for coal workers' pneumoconiosis and silicosis. Am J Respir Crit Care Med 2025 Apr;211(4):642-644. [doi: 10.1164/rccm.202409-1767RL] [Medline: 39805092]
3. Wang Z, Zhang J, Yang Y, et al. Current status, trends, and predictions in the burden of coal worker's pneumoconiosis in 204 countries and territories from 1990 to 2019. Heliyon 2024 Oct 15;10(19):e37940. [doi: 10.1016/j.heliyon.2024.e37940] [Medline: 39381106]
4. Akira M, Suganuma N. Imaging diagnosis of pneumoconiosis with predominant nodular pattern: HRCT and pathologic findings. Clin Imaging 2023 May;97:28-33. [doi: 10.1016/j.clinimag.2023.02.010] [Medline: 36878176]
5. Mandrioli D, Schlünssen V, Ádám B, et al. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of occupational exposure to dusts and/or fibres and of the effect of occupational exposure to dusts and/or fibres on pneumoconiosis. Environ Int 2018 Oct;119:174-185. [doi: 10.1016/j.envint.2018.06.005] [Medline: 29958118]
6. Deng H, Chen Y, Wu M, Zhang T. ERCC1 which affects lipids metabolism and actin dynamics in coal workers' pneumoconiosis is a candidate biomarker for early warning and diagnosis. PLoS One 2024;19(9):e0308082. [doi: 10.1371/journal.pone.0308082] [Medline: 39283905]
7. Yu X, Xiong T, Yu L, et al. Gut microbiome and metabolome profiling in coal workers' pneumoconiosis: potential links to pulmonary function. Microbiol Spectr 2024 Nov 5;12(11):e0004924. [doi: 10.1128/spectrum.00049-24] [Medline: 39283109]
8. Lee JS, Shin JH, Lee KM, et al. Serum levels of TGF-β1 and MCP-1 as biomarkers for progressive coal workers' pneumoconiosis in retired coal workers: a three-year follow-up study. Ind Health 2014;52(2):129-136. [doi: 10.2486/indhealth.2013-0112] [Medline: 24464026]
9. Lee JS, Shin JH, Choi BS. Serum levels of IL-8 and ICAM-1 as biomarkers for progressive massive fibrosis in coal workers' pneumoconiosis. J Korean Med Sci 2015 Feb;30(2):140-144. [doi: 10.3346/jkms.2015.30.2.140] [Medline: 25653483]
10. Huang R, Yu T, Li Y, Hu J. Upregulated has-miR-4516 as a potential biomarker for early diagnosis of dust-induced pulmonary fibrosis in patients with pneumoconiosis. Toxicol Res 2018;7(3):415-422. [doi: 10.1039/C8TX00031J]
11. Chen Z, Shi J, Zhang Y, et al. Lipidomics profiles and lipid metabolite biomarkers in serum of coal workers' pneumoconiosis. Toxics 2022 Aug 26;10(9):496. [doi: 10.3390/toxics10090496] [Medline: 36136461]
12. Song J, Xu M, Wang T, et al. Exosomal miRNAs contribute to coal dust particle-induced pulmonary fibrosis in rats. Ecotoxicol Environ Saf 2023 Jan 1;249:114454. [doi: 10.1016/j.ecoenv.2022.114454] [Medline: 38321673]
13. Zhou Y, Wang H, Xing J, et al. Expression levels of surfactant-associated proteins and inflammation cytokines in serum and bronchoalveolar lavage fluid among coal miners: a case-control study. J Occup Environ Med 2014 May;56(5):484-488. [doi: 10.1097/JOM.0000000000000169] [Medline: 24806560]
14. Prince P, Boulay ME, Pagé N, Desmeules M, Boulet LP. Induced sputum markers of fibrosis and decline in pulmonary function in asbestosis and silicosis: a pilot study. Int J Tuberc Lung Dis 2008 Jul;12(7):813-819. [Medline: 18544209]
15. Xuan W, Zheng L, Bunes BR, Crane N, Zhou F, Zang L. Engineering solutions to breath tests based on an e-nose system for silicosis screening and early detection in miners. J Breath Res 2022 Apr 7;16(3). [doi: 10.1088/1752-7163/ac5f13] [Medline: 35303733]
16. Shi H, You M, Li X, et al. Evaluation of factors associated with adult skeletal fluorosis in coal-burning type of endemic fluorosis and initial screening model based on machine learning in Guizhou, Southwest China. Ecotoxicol Environ Saf 2025 Mar 15;293:118018. [doi: 10.1016/j.ecoenv.2025.118018] [Medline: 40073783]

XSL•FO
RenderX

17. Qian X, Li Y, Zhang X, et al. A cardiovascular disease prediction model based on routine physical examination indicators using machine learning methods: a cohort study. Front Cardiovasc Med 2022;9:854287. [doi: 10.3389/fcvm.2022.854287] [Medline: 35783868]
18. Hang W, Bu C, Cui Y, et al. Research progress on the pathogenesis and prediction of pneumoconiosis among coal miners. Environ Geochem Health 2024 Jul 16;46(9):319. [doi: 10.1007/s10653-024-02114-z] [Medline: 39012521]
19. Mrad AB, Lahiani A, Mefteh-Wali S, Mselmi N. Correction: a comparative analysis of machine learning techniques for imbalanced data. Ann Oper Res 2024 Oct;341(2-3):1349-1349. [doi: 10.1007/s10479-024-06079-1]
20. Zheng H, Zhu W, Niu Z, et al. A novel nutrition-based nomogram to predict prognosis after curative resection of gastric cancer. Front Nutr 2021;8:664620. [doi: 10.3389/fnut.2021.664620] [Medline: 34760907]
21. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021 Mar;54(3):1937-1967. [doi: 10.1007/s10462-020-09896-5]
22. Hajihosseinlou M, Maghsoudi A, Ghezelbash R. A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Nat Resour Res 2023 Dec;32(6):2417-2438. [doi: 10.1007/s11053-023-10249-6]
23. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94. [doi: 10.1186/s40537-020-00369-8] [Medline: 33169094]
24. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
25. Mirbagheri B, Alimohammadi A. Improving urban cellular automata performance by integrating global and geographically weighted logistic regression models. Trans GIS 2017 Dec;21(6):1280-1297 [FREE Full text] [doi: 10.1111/tgis.12278]
26. Chen W, Pourghasemi HR, Naghibi SA. A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China. Bull Eng Geol Environ 2018 May;77(2):647-664. [doi: 10.1007/s10064-017-1010-y]
27. Rahmatinejad Z, Dehghani T, Hoseini B, et al. A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department. Sci Rep 2024 Feb 10;14(1):3406. [doi: 10.1038/s41598-024-54038-4] [Medline: 38337000]
28. Xiao JL. SVM and KNN ensemble learning for traffic incident detection. Physica A Stat Mech Its Appl 2019 Mar;517:29-35. [doi: 10.1016/j.physa.2018.10.060]
29. Budholiya K, Shrivastava SK, Sharma V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. J King Saud Univ Comput Inf Sci 2022 Jul;34(7):4514-4523. [doi: 10.1016/j.jksuci.2020.10.013]
30. Thorjussen CBH, Liland KH, Måge I, Solberg LE. Computational test for conditional independence. Algorithms ;17(8):323. [doi: 10.3390/a17080323]
31. Khan F, Tarimer I, Alwageed HS, et al. Effect of feature selection on the accuracy of music popularity classification using machine learning algorithms. Electronics (Basel) 2022;11(21):3518. [doi: 10.3390/electronics11213518]
32. Senbagamalar L, Logeswari S. Genetic clustering algorithm-based feature selection and divergent random forest for multiclass cancer classification using gene expression data. Int J Comput Intell Syst 2024;17(1):23. [doi: 10.1007/s44196-024-00416-9]
33. Yuan X, Sun C, Chen S. A clustering-based adaptive undersampling ensemble method for highly unbalanced data classification. Appl Soft Comput 2024 Jul;159:111659. [doi: 10.1016/j.asoc.2024.111659]
34. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc 2022 Aug 16;29(9):1525-1534. [doi: 10.1093/jamia/ocac093] [Medline: 35686364]
35. Cool CD, Murray J, Vorajee NI, et al. Pathologic findings in severe coal workers' pneumoconiosis in contemporary US coal miners. Arch Pathol Lab Med 2024 Jul 1;148(7):805-817. [doi: 10.5858/arpa.2022-0491-OA] [Medline: 37852172]
36. Peng H, Tang A, Cheng Y, et al. Study of dust deposition pattern in the respiratory tract of dust particles less than 10 μm in size. Powder Technol 2024 Aug;444:120033. [doi: 10.1016/j.powtec.2024.120033]
37. Sun Y, Kinsela AS, Cen X, et al. Impact of reactive iron in coal mine dust on oxidant generation and epithelial lung cell viability. Sci Total Environ 2022 Mar 1;810:152277. [doi: 10.1016/j.scitotenv.2021.152277] [Medline: 34902414]
38. Yang B, Liu X, Peng C, Meng X, Jia Q. Silicosis: from pathogenesis to therapeutics. Front Pharmacol 2025;16:1516200. [doi: 10.3389/fphar.2025.1516200] [Medline: 39944632]
39. Keles C, Sarver E. A study of respirable silica in underground coal mines: particle characteristics. Minerals 2022;12(12):1555. [doi: 10.3390/min12121555]
40. Wang Z, Li S, Ren T, Wu J, Lin H, Shuang H. Respirable dust pollution characteristics within an underground heading face driven with continuous miner: a CFD modelling approach. J Clean Prod 2019 Apr;217:267-283. [doi: 10.1016/j.jclepro.2019.01.273]
41. Zhang Y, Yan J, Ren Y, et al. Coal dust particles can upregulate the expression of NLRP3 inflammasome components in rat alveolar macrophages through phagocytosis. Sci Rep 2025 Mar 15;15(1):8989. [doi: 10.1038/s41598-025-93946-x] [Medline: 40089559]

XSL•FO
RenderX

42. Sun Y, Kinsela AS, Waite TD. Elucidation of alveolar macrophage cell response to coal dusts: role of ferroptosis in pathogenesis of coal workers' pneumoconiosis. Sci Total Environ 2022 Jun 1;823:153727. [doi: 10.1016/j.scitotenv.2022.153727] [Medline: 35149061]
43. Dong H, Zhu B, Kong X, Zhang X. Efficient clinical data analysis for prediction of coal workers' pneumoconiosis using machine learning algorithms. Clin Respir J 2023 Jul;17(7):684-693. [doi: 10.1111/crj.13657] [Medline: 37380332]
44. Bo Y, Lu Q, Li B, Sha R, Yu H, Miao C. The role of platelets in central hubs of inflammation: a literature review. Medicine (Baltimore) 2024 May 10;103(19):e38115. [doi: 10.1097/MD.0000000000038115] [Medline: 38728509]
45. Ye Y, Gao Z, Zhang Z, Chen J, Chu C, Zhou W. A machine learning model for predicting severe mycoplasma pneumoniae pneumonia in school-aged children. BMC Infect Dis 2025 Apr 21;25(1):570. [doi: 10.1186/s12879-025-10958-8] [Medline: 40259232]

Abbreviations

AEC: absolute eosinophil count
ALC: absolute lymphocyte count
ALT: alanine aminotransferase
AMC: absolute monocyte count
ANC: absolute neutrophil count
AUC: area under curve
CatBoost: categorical boosting
CWP: coal workers' pneumoconiosis
FEV1/FVC: forced expiratory volume/forced vital capacity
KNN: k-nearest neighbor
LASSO: Least Absolute Shrinkage and Selection Operator
LightGBM: light gradient boosting machine
LR: logistic regression
ML: machine learning
PLT: platelet count
RF: random forest
SHAP: Shapley Additive Explanation
SVM: support vector machine
WBC: white blood cell count
XGBoost: extreme gradient boosting

Edited by A Coristine, A Benis; submitted 05.Jul.2025; peer-reviewed by XC Wang, X Liang, Z Yang; revised version received 04.Jan.2026; accepted 08.Jan.2026; published 13.Feb.2026.

Please cite as:
Li H, Jia J, Shi X, Dong Y, Wang S, Cui Y, Hang W, Zhang D
Research on the Prediction of Coal Workers' Pneumoconiosis Based on Easily Detectable Clinical Data: Machine Learning Model Development and Validation Study
JMIR Med Inform 2026;14:e80156
URL: https://medinform.jmir.org/2026/1/e80156
doi:10.2196/80156

© Haiquan Li, Jiaqi Jia, Xu Shi, Yudie Dong, Songquan Wang, Yuming Cui, Wenlu Hang, Dekun Zhang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Impact of Sociodemographic Factors on Artificial Intelligence Models in Predicting Dementia: Retrospective Cohort Study

Xingyi Liu[1], PhD; Muskan Garg[1], PhD; Maria Vassilaki[2], MD, PhD; Jennifer St Sauver[2], PhD; Ronald C Petersen[3], MD, PhD; Momin M Malik[4], PhD; Chung Il Wi[4], MD; Young J Juhn[4], MD, PhD; Sunghwan Sohn[1], PhD

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Harwick Building, 7th Fl., 205 3rd Ave SW, Rochester, MN, United States

[2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

[3]Department of Neurology, Mayo Clinic, Rochester, MN, United States

[4]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Sunghwan Sohn, PhD
Department of Artificial Intelligence and Informatics, Mayo Clinic, Harwick Building, 7th Fl., 205 3rd Ave SW, Rochester, MN, United States

Abstract

Background: Artificial intelligence (AI) is increasingly applied to health care, yet concerns about fairness persist, particularly in relation to sociodemographic disparities. Previous studies suggest that socioeconomic status (SES) and sex may influence AI model performance, potentially affecting groups that are historically underserved or understudied.

Objective: This study aimed to (1) assess algorithmic bias in AI-driven dementia prediction models based on SES and sex (biological sex), (2) compare the utility of an individual-level SES measure (the Housing-Based Socioeconomic Status [HOUSES] Index) versus an area-level measure (the Area Deprivation Index) for bias detection, and (3) evaluate the effectiveness of an oversampling technique (the Synthetic Minority Oversampling Technique for Nominal and Continuous features) for bias mitigation.

Methods: This study used data from two population-based cohorts: the Mayo Clinic Study on Aging (n=3041) and the Rochester Epidemiology Project (n=19,572). Four AI models (random forest, logistic regression, support vector machine, and Naïve Bayes) were trained using a 5-year observation window of structured electronic health record data to predict dementia onset within the subsequent 1-year window. Fairness and model performance were assessed using the balanced error rate (BER) across intersecting SES-sex subgroups. The Synthetic Minority Oversampling Technique for Nominal and Continuous features algorithm was applied to the training data to balance the representation of SES groups.

Results: Across both cohorts, individuals with lower SES generally exhibited higher BERs (worse performance) than high SES groups, confirming the presence of bias. In the MCSA cohort, males with high SES, as indicated by the HOUSES Index, consistently exhibited the lowest BERs across all evaluated models. Balancing the training data based on a specific SES measure showed a trend toward reducing the BER disparity when evaluated using that same measure. However, this targeted improvement demonstrated nonuniversal benefits; in some cases, it exacerbated disparities when evaluated using other, unbalanced SES measures. This pattern suggests that fairness interventions are not universally beneficial across different definitions of the protected attribute. While the balancing approach improved fairness in model performance for lower SES groups, it often came at the cost of a slight reduction in overall model performance. However, an exception was observed in the MCSA cohort when balancing based on the HOUSES Index using logistic regression, support vector machine, and Naïve Bayes, where the performances of both the high and low SES groups improved.

Conclusions: This research highlights the importance of incorporating sociodemographic context into AI modeling in health care. The choice of SES measure may lead to different assessments of algorithmic bias. The HOUSES Index, as a validated individual-level SES measure, may be more effective for bias mitigation than area-level measures. Future AI development should integrate bias mitigation strategies to ensure models do not reinforce existing disparities in health outcomes.

(JMIR Med Inform 2026;14:e80405) doi:10.2196/80405

KEYWORDS

artificial intelligence; AI bias; dementia; Area Deprivation Index; socioeconomic status; machine learning

XSL•FO
RenderX

Introduction

In 2024, Alzheimer dementia affected approximately 6.9 million people in the United States, and this number is projected to escalate to 13.8 million by 2060. This growth places an immense burden on patients and their families, society, and the health care system, manifesting in an estimated 18.4 billion hours of unpaid care annually, valued at US $346.6 billion [1]. Despite this considerable prevalence, cognitive impairment, including mild cognitive impairment (MCI) and dementia, often goes undetected in clinical settings. Diagnoses frequently come late in the cognitive decline process, reducing opportunities to enhance patients' quality of life and exacerbating existing health disparities. Ongoing research in health care seeks to leverage artificial intelligence (AI) to enhance dementia prediction.

The rapid advancements in computing power, data storage, and predictive analytics have significantly accelerated the integration of AI tools into the US health care system [2]. By 2014, 97% of nonfederal acute care hospitals had adopted a certified electronic health record (EHR) [3]. As of January 2025, the US Food and Drug Administration has approved more than 1000 clinical AI applications [4]. This foundational shift has enabled the widespread adoption of AI. AI-powered solutions have already demonstrated transformative potential in health care.

Beyond diagnostics, AI research continues to explore ways to address health disparities, such as reducing unexplained pain disparities through image-based algorithms in underserved populations [5]. However, alongside these advancements lies the critical challenge of mitigating biases inherent in AI systems. Studies have revealed that AI algorithms trained on biased data can perpetuate health disparities, particularly among underresourced and understudied populations [6-9]. For example, a race-based algorithm for estimating kidney function produced higher estimates for Black patients than for White patients, resulting in delayed referrals for organ transplants among Black individuals [10]. Such findings underscore the ethical imperative to scrutinize differential model performance by socioeconomic status (SES) and other social determinants of health (SDOH), which have profound implications for underserved populations and care delivery. SES, as a core component of SDOH, plays a pivotal role in shaping health outcomes through biological, behavioral, and environmental pathways [11-20].

Recent studies have highlighted the pervasive issue of bias in health care AI systems, particularly SES. For instance, a 2019 study [21] defines algorithmic bias in health care as the application of an algorithm that compounds the challenges affecting underserved groups across SES, race, background, religion, sex, or disability, thereby amplifying disparities in health systems. The authors discuss how AI can inadvertently perpetuate and exacerbate challenges faced by underserved populations if not carefully designed and implemented. Another study outlines various elements of potential bias in the development and implementation of AI algorithms in health care. It emphasizes that such biases can propagate stereotypes and discrimination, further marginalizing underserved populations and contributing to socioeconomic health care disparities [22]. Similarly, research indicates that lower SES is associated with poorer predictive model performance, potentially due to incomplete or inaccurate EHR data [23].

In addition to SES disparities, sex differences have emerged as a critical factor influencing AI model performance in dementia prediction. Research indicates that females often exhibit more severe cognitive impairment and experience a faster rate of cognitive decline compared to males at the onset of Alzheimer disease [24]. Sociodemographic factors, such as lower education levels and SES, disproportionately affect older female individuals, increasing their risk for dementia [25-27]. Predictive models for cognitive decline, such as Cox regression, demonstrate varying performances across sexes, indicating potential biases in predictive accuracy [28]. Additionally, transfer learning approaches have revealed that cognitive deficits in female individuals, once detected at the MCI stage, tend to deteriorate more consistently over time, whereas male participants exhibit a wider variety of declines across multiple cognitive functions [29].

Despite these findings, there remains a significant gap in strategies to mitigate biases in AI models. While some studies have proposed methods to detect and quantify such biases, comprehensive approaches to address and correct them are limited [30]. Our study seeks to fill this gap by not only examining the relationship between SES, sex, and AI model performance in dementia prediction but also implementing oversampling techniques to ensure adequate representation of socioeconomically disadvantaged groups. By doing so, we aim to reduce disparities in model performance across different SES groups, thereby enhancing fairness in AI-driven health care solutions.

Methods

Study Population

In our study, we used 2 cohorts. First, we studied individuals participating in the Mayo Clinic Study on Aging (MCSA) [31]. The MCSA, which commenced in 2004, is a population-based research study focused on investigating the epidemiology of MCI, dementia, and related biomarkers (N=5890; dementia, n=682, 11.6%; and non-dementia, n=5208, 88.4%). Participants for the study were chosen randomly from among persons living in Olmsted County, Minnesota, at the time of recruitment. These participants completed detailed cognitive assessments, including the Clinical Dementia Rating scale, a full-scale neurological evaluation, and various neuropsychological tests. Using established criteria, a consensus committee made diagnoses, classifying participants as cognitively unimpaired, having MCI, or having dementia. MCI diagnostic criteria have been disclosed in previous publications [32], and dementia diagnoses adhered to the criteria laid out in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* [33].

Second, we studied a larger cohort of persons living in Olmsted County, MN, between 2004 and 2020, derived from the Rochester Epidemiology Project (REP) medical records linkage system (N=290,528; dementia, n=8,205, 2.8%; and non-dementia, n=282,323, 97.2%) [34]. The onset of dementia

XSL•FO
RenderX

was identified by the date of the first diagnosis code (*International Classification of Diseases [ICD], Ninth or Tenth Revision code*) received between 2004 and 2020.

Socioeconomic Measures

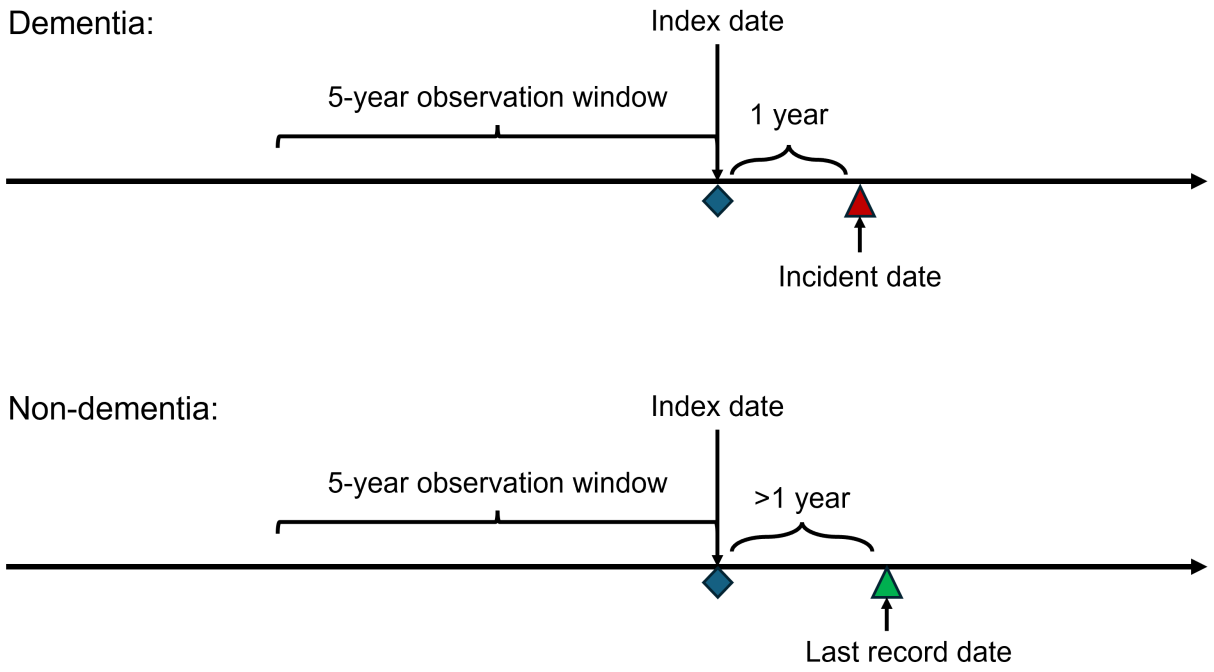
We assessed potential biases in model performance using two residence-based SES measures: the Area Deprivation Index (ADI) [35] and the Housing-Based Socioeconomic Status (HOUSES) Index [36]. The ADI is a multidimensional evaluation of a person's socioeconomic conditions at the census block group level. We used both national-level and state-level ADI rankings. The national-level ADI rankings were used to divide participants into two groups: those in the higher ADI quartile group (76-100) were considered to reside in an area with high socioeconomic deprivation (hereafter, low SES), and those in the ADI group (0 - 75) were considered the referent group (hereafter, high SES). Similarly, the state-level ADI ranking was categorized into two groups: ADI (8-10) for low SES and ADI (0 - 7) for high SES. The HOUSES Index is a housing-based SES indicator, derived from an analysis of 4 specific variables related to a housing unit. The bottom quartile was considered low SES, and the remaining quartiles were considered high SES. The HOUSES Index was only used in the MCSA cohort because it is not available for the REP cohort. By incorporating both the HOUSES Index and ADI measures, our study aimed to evaluate and compare their effectiveness in identifying SES-related biases in the performance of AI models. In addition to quantifying bias in model performance based on SES, our study also considered sex.

Prediction Timeline Window

Our goal was to use records from the past 5 years (observation window) to forecast whether a patient would be diagnosed with dementia in the following year (prediction window).

For both cohorts, we applied consistent inclusion and exclusion criteria. For patients with dementia (case), the incident date was the date of dementia onset. We excluded patients whose age at the incident date was less than 50 years. The index date (point for prediction) is set to 1 year before the dementia incident date. We performed frequency matching for age between participants with dementia and participants without dementia. For patients without dementia (control), we assigned index dates to match age with cases—that is, patients without dementia at the index date have a similar age distribution to patients with dementia at their index dates. As shown in Figure 1, the 5-year period preceding the index date serves as the observation window. The 1-year period following the index date is the prediction window for both groups. We exclude patients with no record within the observation window. We also exclude patients without dementia who do not have at least 1 year of follow-up after their index date (whose last record date is not at least 1 year after the index date).

Figure 1. Timeline overview for dementia and non-dementia patients. For patients with dementia (cases), the index date is set 1 year before their diagnosis. For patients without dementia (controls), the index date is assigned to match the age distribution of the cases. The model uses a 5-year observation window preceding the index date to predict the onset of dementia within the subsequent 1-year prediction window.



Assessing AI Bias Across Different SES and Sex Categories

Four machine learning algorithms were selected to capture a variety of model classes commonly used in clinical prediction tasks, allowing for an assessment of whether observed bias patterns were robust across different algorithmic assumptions and complexities. The selected models were (1) logistic regression (LR), a well-established, interpretable linear model that serves as a standard baseline in clinical research; (2) support vector machine (SVM), a powerful kernel-based method capable of learning complex, nonlinear decision boundaries; (3) random

XSL•FO
RenderX

forest (RF), a nonlinear ensemble of decision trees that is robust to overfitting and can implicitly capture complex interactions between features; and (4) Naïve Bayes (NB), a simple yet efficient probabilistic classifier based on Bayes' theorem with strong independence assumptions. This selection enables a more comprehensive evaluation of how algorithmic bias manifests across linear, nonlinear, ensemble, and probabilistic modeling paradigms. We trained and tested the 4 AI models to predict the incidence of dementia in the following year.

The experimental process begins with preprocessing the data. The 5-year observation period is divided into five 1-year windows. During each 1-year window, patients may have multiple visits. For each visit, we extracted ICD-9 and ICD-10 codes, categorized the codes using the Clinical Classifications Software (CCS) [37], which groups diagnoses and procedures into clinically relevant categories and then generated a one-hot–encoded vector representing the CCS categories for the visit. We computed the element-wise sum of all visit vectors within each 1-year window. To robustly distinguish between years with no visit and years with visits but no relevant diagnoses, we used a missing indicator method. For 1-year windows where a patient had no visits, the corresponding annual vector was filled with zeros. Concurrently, we introduced 5 binary features, one for each year of the observation window, setting the feature to "1" to explicitly flag a year with no visits and "0" otherwise. Finally, we formed the final input vector by concatenating the 5 annual vectors, sex, age at the index date, and the 5 created binary visit-indicator features. Our primary objective was to assess the baseline performance and bias of models built only on the most common, structured data elements that are widely available for nearly all patients within a standard EHR system.

In this study, we developed and evaluated 4 AI models for each cohort. We randomly partitioned the dataset into two subsets: 70% for training and 30% for testing. Within the training set, we used a 5-fold cross-validation strategy to identify the optimal parameters for each model, thereby ensuring robust performance estimates and minimizing the risk of overfitting. After determining the best-performing configuration for each model, we retrained the model on the entire training set and subsequently evaluated its performance using the 30% test set. All model development and evaluation procedures were conducted independently within each of the 2 cohorts. We did not perform cross-cohort validation.

Within each cohort, and for each model, we then conducted subgroup analyses by comparing performance across different sex groups and among various SES-sex subgroups. This approach was used to identify potential differences in predictive bias.

To assess the fairness of these models, we used the balanced error rate (BER) [23], defined as the unweighted mean of the false-positive rate (FPR) and the false-negative rate (FNR), as follows:

$$BER=\frac{FPR+FNR}{2}$$

The FPR, also referred to as 1−specificity, represents the proportion of negative instances that are incorrectly classified as positive. Formally,

$$FPR=\frac{\text{False positive (FP)}}{\text{True negative (TN)}+\text{False positive (FP)}}$$

The FNR, also referred to as 1−sensitivity, represents the proportion of positive instances that are incorrectly classified as negative. Formally,

$$FNR=\frac{\text{False negative (FN)}}{\text{True positive (TP)}+\text{False negative (FN)}}$$

By incorporating both the FPR and FNR, BER captures imbalances in how models treat positively and negatively labeled individuals, allowing us to identify potential disparities or biases in the predictive performance.

To assess the variability and reliability of the BER estimates, we calculated bootstrap CIs for each SES-sex group:

1. Bootstrapping procedure: We generated 1000 bootstrap samples by randomly resampling with replacement from the original test datasets.
2. BER calculation: For each bootstrap sample, we calculated the BER.
3. CI estimation: The distribution of the bootstrap BERs was used to compute the 95% CI, providing an estimate of the uncertainty around the BER.

To assess statistical differences in model performance (BER) between SES-sex groups, we conducted a series of pairwise 2-tailed *t* tests with a significance threshold of 0.05. The Levene test was used to assess variance homogeneity, with the Welch *t* test applied for unequal variances and the standard *t* test for equal variances. Each test was performed independently to evaluate a specific contrast (eg, low vs high SES within a sex group). To correct for the increased risk of type 1 errors (false positives) due to multiple comparisons, we applied the Benjamini-Hochberg procedure. For each of the 4 AI models, we collected the *P* values from all pairwise *t* tests performed for that model and applied the Benjamini-Hochberg correction to this complete set of *P* values. This method controls the false discovery rate at our chosen significance level of 0.05. All reported statistical significances for pairwise comparisons are based on the results of this false discovery rate correction.

In addition to these pairwise comparisons, a separate baseline analysis was conducted to contextualize the performance of all subgroups relative to a single reference point. For this analysis, the "male in high SES" subgroup was designated as the baseline category. We then calculated the difference in BER for all other sociodemographic subgroups by comparing their BERs to the BER of this baseline group. This method provides a consistent benchmark for evaluating the performance disparities relative to the reference group.

Reducing SES-Related Bias

To address disparities in AI model performance related to SES, we implemented the Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC) [38]. SMOTE-NC is a variant of an oversampling technique designed to generate synthetic examples in datasets that include both continuous and categorical variables. This method is

XSL•FO
RenderX

particularly useful in addressing imbalances in datasets, which can lead to biased model performance.

In the context of our study, we applied SMOTE-NC to balance the representation of understudied SES groups within our training dataset. This process preserved the ratio of dementia cases and controls in the cohort (Tables 1 and 2). We set the number of neighbors parameter to 5 and used a fixed random seed of 42. This process ensured that the distribution of SES groups in the training dataset was balanced, without introducing synthetic examples into the test set. Following the oversampling, the AI models were retrained on the balanced dataset using the same cross-validation strategy. To assess the effectiveness of SES balancing, we compared the performance (BER) of models trained on the oversampled datasets with those trained on the original, imbalanced datasets.

Table . Participant characteristics in Mayo Clinic Study on Aging (MCSA; N=3041), stratified by dementia cases (n=679, 22.3%) and non-dementia controls (n=2362, 77.7%).

MCSA	Case	Control
Age (y; at index date), n (%)		
55 - 64	1 (0)	3 (0)
65 - 74	32 (5)	112 (5)
75 - 84	259 (38)	906 (38)
85 - 94	364 (54)	1261 (53)
≥95	23 (3)	80 (4)
Sex, n (%)		
Male	353 (52)	1192 (51)
Female	326 (48)	1170 (49)
Race, n (%)		
Native American or Alaska Native	0 (0)	0 (0)
Asian	8 (1)	11 (1)
Hawaiian or Pacific Islander	0 (0)	1 (0)
Black or African American	1 (0)	5 (0)
White	666 (98)	2337 (99)
More than one	3 (1)	5 (0)
Missing	1 (0)	3 (0)
Background, n (%)		
Hispanic	2 (0)	6 (0)
Non-Hispanic	674 (99)	2350 (100)
Missing	3 (1)	6 (0)
HOUSES[a] Index, n (%)		
Q1 (low SES[b])	146 (22)	506 (21)
Q2-Q4 (high SES)	533 (78)	1856 (79)
National-level ADI[c], n (%)		
76 - 100 (low SES)	35 (5)	94 (4)
0 - 75 (high SES)	644 (95)	2268 (96)
State-level ADI, n (%)		
8 - 10 (low SES)	202 (30)	572 (24)
0 - 7 (high SES)	477 (70)	1790 (76)
Number of visits, mean (SD)	93.3 (65.0)	79.5 (62.7)

[a]HOUSES: Housing-Based Socioeconomic Status.

[b]SES: socioeconomic status.

[c]ADI: Area Deprivation Index.

XSL•FO
RenderX

Table . Participant characteristics in Rochester Epidemiology Project (REP; N=19,572), stratified by dementia cases (n=7335, 37.5%) and nondementia controls (n=12,237, 62.5%).

REP	Case	Control
Age (y; at index date), n (%)		
50 - 54	97 (1)	194 (1)
55 - 64	405 (6)	810 (7)
65 - 74	1199 (16)	2398 (20)
75 - 84	2830 (39)	5660 (46)
85 - 94	2485 (34)	3016 (25)
≥95	319 (4)	159 (1)
Sex, n (%)		
Male	3040 (41)	5298 (43)
Female	4295 (59)	6939 (57)
Race, n (%)		
Black	80 (1)	123 (1)
Asian	118 (2)	229 (2)
Hawaiian or Pacific Islander	3 (0)	4 (0)
American Indian	3 (0)	11 (0)
Other or mixed	58 (1)	121 (1)
White	7049 (96)	11,681 (96)
Refusal	11 (0)	24 (0)
Unknown	13 (0)	44 (0)
Background, n (%)		
Hispanic	105 (1)	238 (2)
Non-Hispanic	6504 (89)	10,536 (86)
Missing	726 (10)	1463 (12)
National-level ADI[a], n (%)		
76 - 100 (low SES[b])	444 (6)	651 (5)
0 - 75 (high SES)	6891 (94)	11,586 (95)
State-level ADI, n (%)		
8 - 10 (low SES)	2098 (29)	3176 (26)
0 - 7 (high SES)	5237 (71)	9061 (74)
Number of visits, mean (SD)	76.2 (70.4)	55.8 (52.2)

[a]ADI: Area Deprivation Index.

[b]SES: socioeconomic status.

Ethical Considerations

The study was approved by the institutional review boards (IRBs) of the Mayo Clinic (20 - 004498) and the Olmsted Medical Center (028-OMC-20). The need for informed consent was waived for the study. Approval for the MCSA (IRB number 14 - 004401) was also obtained from the IRBs of the Mayo Clinic and Olmsted Medical Center in Rochester, Minnesota. Participants of the MCSA provided written informed consent before participation. The MCSA informed consent and the study's IRB protocol allow secondary analysis without additional consent. All data were accessed and analyzed within secure and institutionally approved computing environments. No individual identifiers were presented in the manuscript. The results are reported only in aggregate form, ensuring that patients cannot be identified. No compensation was provided to participants, as this study involved the retrospective analysis of medical records.

XSL•FO
RenderX

Results

Cohort Characteristics

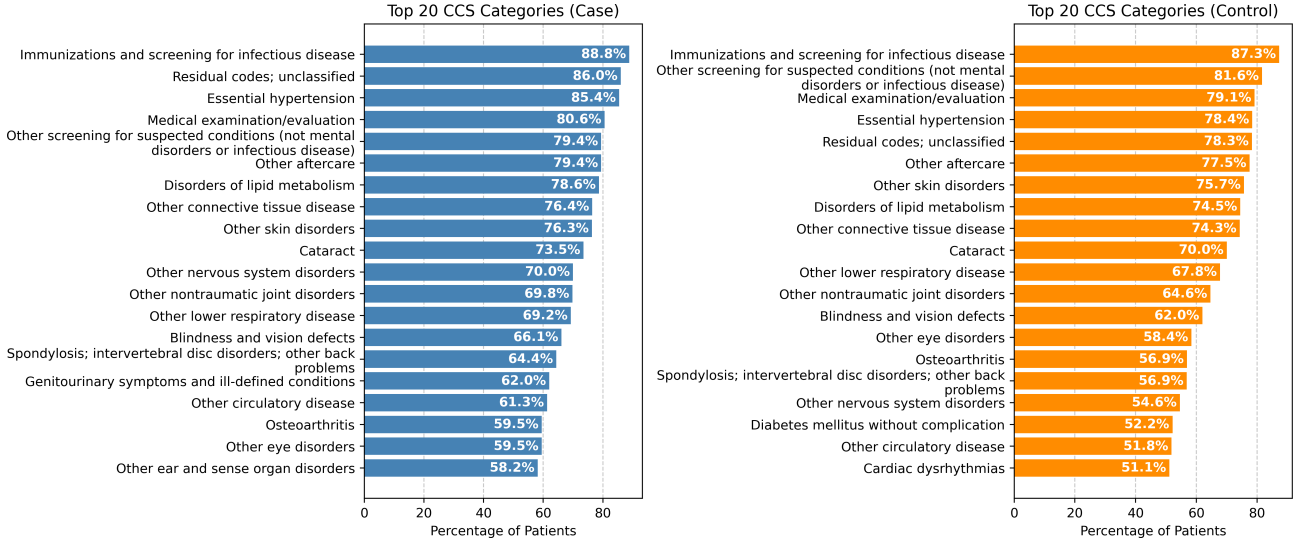
Table 1 presents the participant characteristics for the MCSA cohort. Approximately 21% (652/3041) of cases were categorized as low SES based on the HOUSES Index, 4% (129/3041) were categorized based on the national-level ADI, and 25% (774/3041) were categorized based on the state-level ADI. To understand the relationship between SES measures, we assessed the concordance between the HOUSES Index and both ADI measures in the MCSA cohort. There was limited overlap between the classifications. Among those identified as low SES by the HOUSES Index, only 12.0% (78/652) were concurrently classified as low SES by the national-level ADI, and 40.0% (261/652) were concurrently classified as low SES by the state-level ADI. This lack of concordance indicates that while these metrics all aim to measure disadvantage, they identify distinct subpopulations.

Table 2 summarizes the participant characteristics for the REP cohort. SES analyses indicate that approximately 6% (1095/19,572) of cases were classified as low SES using the national-level ADI, and 27% (5274/19,572) of cases were classified as low SES using the state-level ADI (the HOUSES Index is not available in the REP cohort). SES measure distributions were similar for controls in both populations.

The average number of visits during the 5-year observation window was 93.3 (SD 65.0) for patients with dementia and 79.5 (SD 62.7) for patients without dementia in the MCSA cohort, and 76.2 (SD 70.4) and 55.8 (SD 52.2), respectively, in the REP cohort. We also assessed data completeness by calculating the average number of years missed, defined as the count of years within the 5-year observation window where a patient had no recorded visits. When analyzing the combined population of patients with dementia and patients without dementia, most SES measures indicated that the low SES group exhibited a higher average number of years missed compared to the high SES group. In the MCSA cohort, the high SES group averaged 0.12 missed years compared to 0.17 missed years for the low SES group based on the HOUSES Index. Similarly, the national-level ADI showed 0.13 missed years for the high SES group and 0.16 missed years for the low SES group, while the state-level ADI showed a divergent pattern, with 0.14 missed years for the high SES group compared to 0.11 missed years for the low SES group. A consistent pattern emerged in the REP cohort, where the national-level ADI showed averages of 0.54 missed years for the high SES group and 0.71 for the low SES group, and the state-level ADI showed 0.54 for the high SES group compared to 0.59 for the low SES group.

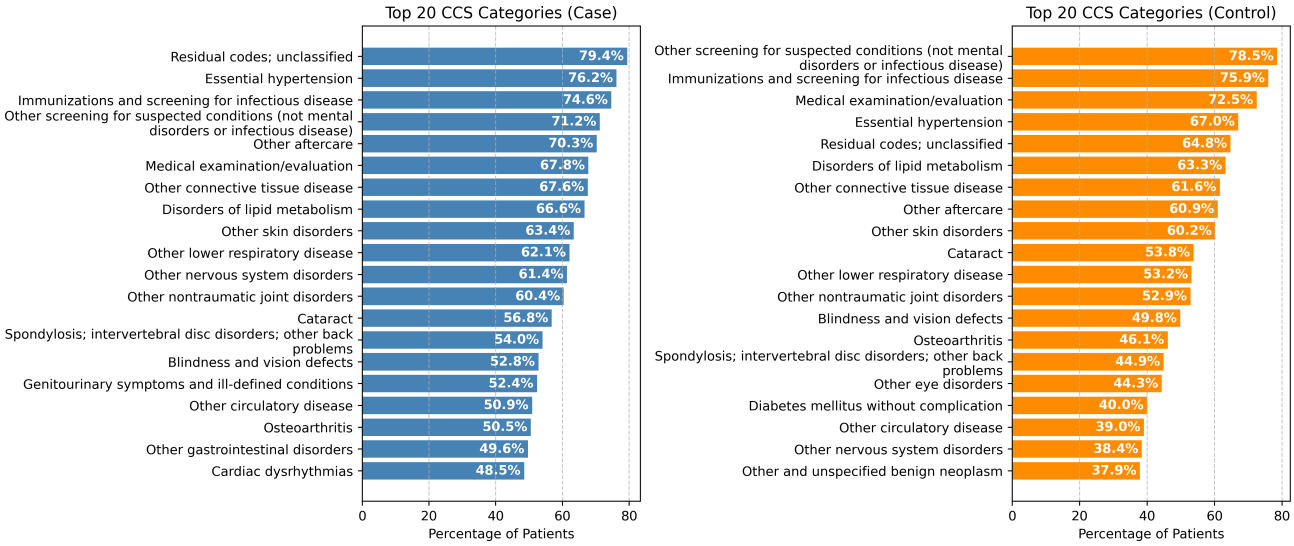
Figures 2 and 3 show the top 20 CCS categories that appear among cases versus controls, ordered by the percentage of patients who had those CCS categories in the MCSA cohort and the REP cohort, respectively. In both cohorts, more patients with dementia visited for “essential hypertension” than patients without dementia during the 5-year observation period (MCSA: 85.4% vs 78.4%; and REP: 76.2% vs 67.0%). Similarly, more patients with dementia visited for “other nervous system disorders” than patients without dementia during the same period (MCSA: 70.0% vs 54.6%; and REP: 61.3% vs 38.4%).

Figure 2. Top 20 Clinical Classifications Software (CCS) categories that appear among cases versus controls during the 5-year observation period, ordered by the percentage of patients who had those CCS categories in the Mayo Clinic Study on Aging cohort.



XSL•FO
RenderX

Figure 3. Top 20 Clinical Classifications Software (CCS) categories that appear among cases versus controls during the 5-year observation period, ordered by the percentage of patients who had those CCS categories in the Rochester Epidemiology Project cohort.



Bias Measurement

Table 3 presents a comparison of BERs for AI-based dementia prediction models stratified by SES, as indicated by both the ADIs and the HOUSES Index in the MCSA cohort. The first 2 rows compare BERs among males, the next 2 rows compare BERs among females, and the final 2 rows compare BERs for both males and females combined. Between each pair of rows, superscript letters denote statistically significant differences in BER for the low SES group compared to the high SES group (based on Benjamini-Hochberg–corrected $P<.05$): “h” indicates a significantly lower BER, “i” indicates a significantly higher BER, and “j” indicates no statistically significant difference. Across all evaluated models, individuals (males and females combined) with low SES, as measured by both the ADIs and the HOUSES Index, consistently exhibited higher BERs. When focusing on males only, those with low SES similarly demonstrated higher or equal BERs across models. However, for females, the patterns differed: in nearly all evaluated models, females with low SES, as indicated by the national-level ADI, consistently showed higher BERs, while female participants with low SES, as indicated by the state-level ADI, showed comparable BERs. In contrast, those with low SES as measured by the HOUSES Index consistently exhibited lower or equal BERs.

Table 4 compares BERs within each AI-based dementia prediction model among SES-sex subgroups in the MCSA cohort. We selected “male in high SES” as the baseline category. For each model, the table displays how the BERs for “male in low SES,” “female in high SES,” and “female in low SES” compare to this baseline. “Lower” indicates a lower BER than the baseline, “higher” indicates a higher BER, and empty cells signify no statistically significant difference. From the table, it is evident that males with high SES, as indicated by the HOUSES Index, consistently exhibited the lowest BERs across all evaluated models. Furthermore, males with high SES measured by the 2 ADIs showed the lowest BERs in almost all models, with the exception of the RF when using the national-level ADI and the NB model when using either the state-level or national-level ADI.

Table 5 presents a comparison of BERs for AI-based dementia prediction models stratified by SES—as indicated by both ADIs—in the REP cohort (the HOUSES Index was not available in the REP cohort). In every evaluated model except NB, individuals (combining both males and females) with low SES, as determined by both ADIs, consistently exhibited higher BERs. When focusing on females only, those with low SES similarly demonstrated higher BERs across almost all models, with the exception of the NB model when using the state-level ADI. Furthermore, males with high SES showed lower BERs in almost all models, with the exception of the NB model when using either the state-level or national-level ADI.

XSL•FO
RenderX

Table . Comparison of balanced error rates (BERs) for artificial intelligence–based dementia prediction models in the Mayo Clinic Study on Aging (MCSA) cohort.[a]

MCSA (sex and SES group)	National-level ADI[m], BER (95% CI)				State-level ADI, BER (95% CI)				HOUSES[c] Index, BER (95% CI)			
	RF[d]	LR[e]	SVM[f]	NB[g]	RF	LR	SVM	NB	RF	LR	SVM	NB
Male												
High	0.2827[h] (0.23-0.33)	0.3598[h] (0.30-0.41)	0.4419[h] (0.40-0.48)	0.4207[h] (0.37-0.47)	0.2546[h] (0.20-0.31)	0.3368[h] (0.28-0.40)	0.4334[h] (0.39-0.48)	0.4151[h] (0.36-0.47)	0.2688[h] (0.22-0.32)	0.3404[h] (0.28-0.40)	0.4273[h] (0.38-0.47)	0.3907[h] (0.33-0.44)
Low	0.5408[i] (0.17-0.88)	0.5378[i] (0.21-0.88)	0.4531[i] (0.12-0.67)	0.5247[i] (0.20-0.80)	0.3957[i] (0.29-0.50)	0.4482[i] (0.34-0.56)	0.4651[i] (0.38-0.54)	0.4447[i] (0.34-0.54)	0.3767[i] (0.26-0.49)	0.4591[i] (0.33-0.58)	0.4996[i] (0.44-0.54)	0.5708[i] (0.50-0.63)
Female												
High	0.3157[i] (0.26-0.37)	0.3661[h] (0.31-0.42)	0.4790[h] (0.44-0.51)	0.3822[h] (0.33-0.43)	0.3053[h] (0.24-0.37)	0.3709[j] (0.31-0.43)	0.4745[i] (0.44-0.51)	0.3867[h] (0.33-0.45)	0.3269[i] (0.26-0.39)	0.3775[i] (0.31-0.45)	0.4908[i] (0.45-0.53)	0.3964[i] (0.33-0.46)
Low	0.2419[h] (0.10-0.42)	0.4646[i] (0.27-0.67)	0.5147[i] (0.36-0.64)	0.5436[i] (0.33-0.72)	0.3221[i] (0.23-0.42)	0.3715[j] (0.27-0.47)	0.4424[h] (0.36-0.52)	0.4104[i] (0.31-0.51)	0.2750[h] (0.18-0.38)	0.3573[h] (0.25-0.46)	0.4555[h] (0.38-0.52)	0.3921[h] (0.29-0.49)
Male and female												
High	0.3006[h] (0.26-0.34)	0.3625[h] (0.32-0.40)	0.4586[h] (0.43-0.48)	0.4036[h] (0.37-0.44)	0.2784[h] (0.24-0.32)	0.3525[h] (0.31-0.39)	0.4537[h] (0.39-0.51)	0.4012[h] (0.36-0.45)	0.2943[h] (0.25-0.34)	0.3579[h] (0.32-0.40)	0.4558[h] (0.42-0.49)	0.3925[h] (0.35-0.43)
Low	0.3258[i] (0.18-0.49)	0.4844[i] (0.31-0.66)	0.4902[i] (0.35-0.62)	0.5437[i] (0.37-0.69)	0.3627[i] (0.29-0.44)	0.4074[i] (0.34-0.48)	0.4620[i] (0.43-0.49)	0.4292[i] (0.36-0.50)	0.3181[i] (0.24-0.39)	0.3971[i] (0.32-0.48)	0.4721[i] (0.42-0.52)	0.4617[i] (0.39-0.53)

[a]Superscript alphabets indicate a statistically significant difference in BER for the low socioeconomic status (SES) group compared to the high SES group within that stratum, based on Benjamini-Hochberg–corrected $P<.05$.

[b]ADI: Area Deprivation Index.

[c]HOUSES: Housing-Based Socioeconomic Status.

[d]RF: random forest.

[e]LR: logistic regression.

[f]SVM: support vector machine.

[g]NB: Naïve Bayes.

[h]BER significantly lower.

[i]BER significantly higher.

[j]No statistically significant difference.

XSL•FO
RenderX

Table . Balanced error rate (BER) difference for male in high socioeconomic status (SES) as a baseline for Table 3.

MCSA[a] (sex and SES group)	National-level ADI[b], BER				State-level ADI[b], BER				HOUSES[c] Index, BER			
	RF[d]	LR[e]	SVM[f]	NB[g]	RF	LR	SVM	NB	RF	LR	SVM	NB
Male												
High	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base
Low	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]
Female												
High	Higher[h]	Higher[h]	Higher[h]	Lower[i]	Higher[h]	Higher[h]	Higher[h]	Lower[i]	Higher[h]	Higher[h]	Higher[h]	Higher[h]
Low	Lower[i]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Higher[h]	Lower[i]	Higher[h]	Higher[h]	Higher[h]	—[j]

[a]MCSA: Mayo Clinic Study on Aging.

[b]ADI: Area Deprivation Index.

[c]HOUSES: Housing-Based Socioeconomic Status.

[d]RF: random forest.

[e]LR: logistic regression.

[f]SVM: support vector machine.

[g]NB: Naïve Bayes.

[h]Higher: BER significantly higher than baseline.

[i]Lower: BER significantly lower than baseline.

[j]No statistically significant difference from baseline.

XSL•FO
RenderX

Table . Comparison of balanced error rates (BERs) for artificial intelligence–based dementia prediction models in the Rochester Epidemiology Project (REP) cohort.

REP (sex and SES[a] group)	National-level ADI[b], BER (95% CI)				State-level ADI, BER (95% CI)			
	RF[c]	LR[d]	SVM[e]	NB[f]	RF	LR	SVM	NB
Male								
High	0.2550[g] (0.24-0.27)	0.2922[g] (0.27-0.31)	0.3138[g] (0.29-0.33)	0.3500[h] (0.33-0.37)	0.2500[g] (0.23-0.27)	0.2839[g] (0.25-0.32)	0.3415[g] (0.32-0.36)	0.3551[h] (0.33-0.38)
Low	0.3001[h] (0.21-0.39)	0.3113[h] (0.23-0.40)	0.3378[h] (0.25-0.44)	0.3165[g] (0.24-0.40)	0.2756[h] (0.24-0.32)	0.2943[h] (0.27-0.33)	0.3458[h] (0.31-0.38)	0.3302[g] (0.29-0.36)
Female								
High	0.2606[g] (0.24-0.28)	0.2988[g] (0.28-0.32)	0.3266[g] (0.31-0.35)	0.3585[g] (0.34-0.37)	0.2335[g] (0.21-0.26)	0.2815[g] (0.26-0.30)	0.3341[g] (0.32-0.35)	0.3682[h] (0.35-0.39)
Low	0.2648[h] (0.20-0.33)	0.3123[h] (0.25-0.38)	0.3688[h] (0.29-0.42)	0.4067[h] (0.34-0.47)	0.2560[h] (0.23-0.29)	0.2962[h] (0.27-0.33)	0.3584[h] (0.33-0.39)	0.3478[g] (0.32-0.38)
Male and female								
High	0.2581[g] (0.25-0.27)	0.2958[g] (0.28-0.31)	0.3163[g] (0.29-0.33)	0.3547[g] (0.34-0.37)	0.2478[g] (0.23-0.27)	0.2827[g] (0.27-0.30)	0.3376[g] (0.32-0.35)	0.3626[h] (0.35-0.38)
Low	0.2772[h] (0.23-0.33)	0.3124[h] (0.27-0.36)	0.3437[h] (0.29-0.39)	0.3752[h] (0.33-0.42)	0.2637[h] (0.24-0.29)	0.2955[h] (0.28-0.32)	0.3547[h] (0.33-0.38)	0.3411[g] (0.32-0.36)

[a]SES: socioeconomic status.

[b]ADI: Area Deprivation Index.

[c]RF: random forest.

[d]LR: logistic regression.

[e]SVM: support vector machine.

[f]NB: Naïve Bayes.

[g]BER significantly lower.

[h]BER significantly higher.

Table 6 compares BERs within each AI-based dementia prediction model among SES-sex subgroups (using the same approach as Table 4) in the REP cohort; we selected "male in high SES" as the baseline category. From the table, most of both males and females with low SES exhibit higher BER compared to males with high SES. Furthermore, females with high SES, as indicated by the national-level ADI, exhibited higher BER compared to males with high SES. However, females with high SES, as indicated by the state-level ADI, exhibited the lowest BERs across almost all the evaluated models except for the NB model.

XSL•FO
RenderX

Table . Balanced error rate (BER) difference for male high socioeconomic status (SES) as a baseline for Table 5.

REP[a] (sex and SES group)	National-level ADI[b], BER				State-level ADI, BER			
	RF[c]	LR[d]	SVM[e]	NB[f]	RF	LR	SVM	NB
Male								
High	Base	Base	Base	Base	Base	Base	Base	Base
Low	Higher[g]	Higher[g]	Higher[g]	Lower[h]	Higher[g]	Higher[g]	Higher[g]	Lower[h]
Female								
High	Higher[g]	Higher[g]	Higher[g]	Higher[g]	Lower[h]	Lower[h]	Lower[h]	Higher[g]
Low	Higher[g]	Higher[g]	Higher[g]	Higher[g]	Higher[g]	Higher[g]	Higher[g]	Lower[h]

[a]REP: Rochester Epidemiology Project.

[b]ADI: Area Deprivation Index.

[c]RF: random forest.

[d]LR: logistic regression.

[e]SVM: support vector machine.

[f]NB: Naïve Bayes.

[g]Higher: BER significantly higher than baseline.

[h]Lower: BER significantly lower than baseline.

Table 7 presents a comparison of the relative differences in BERs between low and high SES groups, as indicated by both the ADIs and the HOUSES Index, across the 4 AI-based dementia prediction models in the MCSA cohort. The relative difference (value close to 0 is perfectly balanced between low and high SES groups), calculated as (low SES' BER−high SES' BER)/high SES' BER, was first shown for the baseline case with no dataset balancing. Subsequently, the relative differences after balancing the training dataset (separately using national-level ADI, state-level ADI, and the HOUSES Index) are displayed. Within each model and SES measure, if the difference postbalancing was lower than the baseline, the corresponding cell is marked with "i"; otherwise, it is marked with "j". As indicated by "h" in the table, balancing the training dataset using any one of the 3 SES measures was associated with narrower BER differences between high and low SES groups when stratified by that same SES measure. However, in several cases, most notably with the LR and RF models under national-level ADI balancing, the relative difference shifted from positive to negative. This shift indicates an inversion of the disparity, whereby the low SES group achieved a lower BER than the high SES group. Furthermore, when the training dataset was balanced using one SES measure but performance was stratified by another SES measure, BER differences were similarly reduced across nearly all cases. The exceptions occurred when SES was stratified by national-level ADI and the training dataset was balanced by state-level ADI or the HOUSES Index under the RF model or when the training dataset was balanced by national-level ADI but evaluated using the HOUSES Index with the SVM and NB models. For instance, balancing the training data on the HOUSES Index increased the BER disparity measured by the national-level ADI from 8.38% to 30.37% with the RF model. This pattern suggests that fairness interventions are not universally beneficial across different definitions of the protected attribute.

Table 8 presents the same comparison as in Table 7 but for the REP cohort. Similarly, as indicated by footnote "g" in the table, balancing the training dataset using any one of the 3 SES measures consistently reduced the BER differences between high and low SES groups when stratified by that same SES measure. A similar pattern of disparity inversion was observed, where balancing on the national-level ADI for the RF model and the state-level ADI for the SVM model also resulted in negative relative differences.

XSL•FO
RenderX

Table . Comparison of balanced error rate (BER) differences between low and high socioeconomic status (SES) groups under various data balancing strategies in artificial intelligence (AI)–based dementia prediction models (Mayo Clinic Study on Aging [MCSA] cohort).

MCSA (sex and balance on)	National-level ADI[a], (low SES' BER−high SES' BER)/high SES' BER (percentile)				State-level ADI, (low SES' BER−high SES' BER)/high SES' BER (percentile)				HOUSES[b] Index, (low SES' BER−high SES' BER)/high SES' BER (percentile)			
	RF[c]	LR[d]	SVM[e]	NB[f]	RF	LR	SVM	NB	RF	LR	SVM	NB
Male and female												
N/A[g]	8.38%	33.63%	6.89%	34.71%	30.28%	15.57%	1.83%	6.98%	8.09%	10.95%	3.58%	17.63%
National-level ADI	−5.84%[h,i]	−7.27%[h,i]	3.49%[h,i]	5.11%[h,i]	2.53%[i]	2.14%[i]	1.13%[i]	2.94%[i]	4.27%[i]	9.79%[i]	5.90%[j]	18.13%[j]
State-level ADI	−30.37%[j]	−28.64%[i]	−5.31%[i]	−22.96%[i]	3.91%[h,i]	1.11%[h,i]	−0.27%[h,i]	−1.12%[h,i]	−1.95%[i]	−0.20%[i]	1.03%[i]	−9.75%[i]
HOUSES Index	25.13%[j]	−14.89%[i]	−2.33%[i]	1.15%[i]	15.89%[i]	2.45%[i]	1.52%[i]	−0.50%[i]	2.24%[h,i]	3.43%[h,i]	−0.86%[h,i]	3.83%[h,i]

[a]ADI: Area Deprivation Index.

[b]HOUSES: Housing-Based Socioeconomic Status.

[c]RF: random forest.

[d]LR: logistic regression.

[e]SVM: support vector machine.

[f]NB: Naïve Bayes.

[g]Not applicable.

[h]Results corresponding to models trained on datasets balanced using the same SES measure by which performance is stratified.

[i]Relative BER difference lower.

[j]Relative BER difference higher.

Table . Comparison of balanced error rate (BER) differences between low and high socioeconomic status (SES) groups under various data balancing strategies in artificial intelligence–based dementia prediction models (Rochester Epidemiology Project [REP] cohort).

REP (sex and balance on)	National-level ADI[a],(low SES' BER−high SES' BER)/high SES' BER [percentile])				State-level ADI, (low SES' BER−high SES' BER)/high SES' BER [percentile])			
	RF[b]	LR[c]	SVM[d]	NB[e]	RF	LR	SVM	NB
Male and female								
N/A[f]	7.40%	5.61%	8.66%	5.78%	6.42%	4.53%	5.07%	−5.93%
National-level ADI	−1.80%[g,h]	2.37%[g,h]	3.28%[g,h]	2.63%[g,h]	−4.25%[h]	6.75%[i]	3.11%[h]	−2.70%[h]
State-level ADI	−4.14%[h]	9.26%[i]	−0.05%[h]	6.73%[i]	1.62%[g,h]	0.16%[g,h]	−0.52%[g,h]	−3.28%[g,h]

[a]ADI: Area Deprivation Index.

[b]RF: random forest.

[c]LR: logistic regression.

[d]SVM: support vector machine.

[e]NB: Naïve Bayes.

[f]Not applicable.

[g]Results corresponding to models trained on datasets balanced using the same socioeconomic status (SES) measure by which performance is stratified.

[h]Relative BER difference lower.

[i]Relative BER difference higher.

XSL•FO
RenderX

Discussion

Principal Findings

Our study demonstrates socioeconomic bias in AI-driven dementia prediction models. Across both the research cohort (MCSA) and the real-world cohort (REP), the combined male and female population with lower SES consistently exhibited higher BERs compared to their higher SES counterparts, except when using the NB model in the REP cohort. This discrepancy suggests that AI models, when trained on imbalanced SES datasets, may not generalize well across different socioeconomic groups, potentially reinforcing existing health care disparities. The findings highlight important differences in bias detection across SES measures. Both the ADI and the HOUSES Index were effective in identifying disparities, but they capture different aspects of socioeconomic disadvantage. The HOUSES Index, which is based on housing data, identified a more nuanced variation in model performance. In contrast, the ADI (both national-level and state-level rankings), which is derived from census block group–level data, showed broader disparities but may be less sensitive to individual economic conditions.

One important factor contributing to the observed differences in performance between high and low SES groups is the imbalance in their representation within both cohorts. Individuals from higher SES groups significantly outnumber those from lower SES groups. This skewed distribution can have a direct impact on model performance, as AI algorithms trained on predominantly high SES individuals may not adequately capture the health-related patterns of populations with low SES.

A key implication of this imbalance is that AI models become optimized for the majority (high SES) group, leading to lower accuracy and higher BERs for the understudied (low SES) group. Machine learning models inherently learn patterns based on the data they are exposed to, and if the majority of training samples come from high SES individuals, the model is likely to develop a bias toward their health characteristics and health care access patterns. This bias can manifest in the form of increased BER for participants with low SES. Furthermore, disparities in health care access and utilization between SES groups can compound these biases. Traditionally, it is argued that individuals from higher SES backgrounds have more comprehensive and consistent health records, whereas individuals with lower SES may experience gaps in care that lead to incomplete medical histories [23]. Our findings support this view: the low SES group exhibited a higher average number of years with missed visits compared to the high SES group, as reported in the Results section.

Another important observation is the interaction between SES and sex. In the MCSA cohort, high SES male participants generally exhibited the lowest BERs. In the REP cohort, however, the results depended on the specific SES measure used: high SES female individuals tended to have the lowest BERs when stratified by the state-level ADI, whereas this pattern was not observed with the national-level ADI. This discrepancy between cohorts may result from MCSA's active recruitment and retention process, which likely selected for a subgroup of male participants with higher health literacy and more complete records than the general population—a selection bias less present in the passive data collection of the REP cohort. Furthermore, within the REP cohort, the fact that high SES female individuals did not exhibit this advantage under the national-level ADI highlights the differential validity of SES measures. Broader metrics such as the national-level ADI may fail to capture localized resource variations that drive health outcomes for female participants, nuances that are better reflected by more granular measures such as the state-level ADI.

To address these disparities, we applied the SMOTE-NC to balance the SES representation in the training data. This approach narrowed the observed SES-related disparities in BERs, demonstrating that balanced data distributions can improve AI model fairness. However, the effectiveness of this technique varied across SES measures—while balancing based on the HOUSES Index and state-level ADI consistently reduced BER disparities, national-level ADI balancing sometimes led to unintended increases in BER differences for certain models. This may indicate that the national-level ADI lacks the necessary resolution to distinguish SES levels within specific states effectively. Unlike the state-level ADI or the HOUSES Index, the national-level ADI may not adequately represent the local socioeconomic context required for effective data balancing. Furthermore, the limited concordance observed between the HOUSES Index and the ADI measures may explain why oversampling based on one measure resulted in disparities in the others.

While our balancing approach improved fairness in model performance for lower SES groups, it often came at the cost of a slight reduction in overall model performance. In some cases, this intervention led to an inversion of bias, resulting in the models performing better on the low SES group than the high SES group. The trade-off between fairness and predictive accuracy is a known challenge in AI-driven health care apps, and our results highlight this issue. Most balancing strategies led to improved performance for the low SES group while slightly decreasing overall predictive accuracy. However, an exception was observed in the MCSA cohort when balancing based on the HOUSES Index using LR, SVM, and NB, where both the high and low SES group performances improved. This suggests that housing-based SES metrics, such as the HOUSES Index, may be more effective for bias mitigation than area-level measures such as the ADI.

Beyond statistical significance, it is crucial to consider the clinical importance of these performance disparities. For instance, in the MCSA cohort, the NB model using the HOUSES Index yielded a BER of 39.3% for the high SES group compared to 46.2% for the low SES group. While this represents a modest absolute difference of 6.9 percentage points, its impact at a population level is substantial. In a hypothetical clinical screening scenario involving 1000 patients from each SES group, this disparity would translate to an additional 69 patients from low SES backgrounds being misclassified compared to their high SES counterparts. Such a systematic difference in accuracy, when deployed at scale, could further disadvantage already underserved populations in clinical care. Patients from socioeconomically disadvantaged backgrounds would be

XSL•FO
RenderX

disproportionately subject to the consequences of misclassification, including delayed diagnosis and intervention in the case of false negatives, or unnecessary, costly, and anxiety-inducing follow-up procedures in the case of false positives. This quantification highlights how seemingly small statistical biases can manifest as meaningful clinical harms.

Limitations

While this study provides valuable insights, it is important to recognize limitations. First, although our study cohorts (MCSA and REP) represent Midwestern populations with a dominant White demographic, they are not representative of the broader US population [39]. Therefore, the findings regarding SES- and sex-based bias warrant validation in more representative populations. Second, our AI models do not include certain important clinical factors, such as cognitive test scores, functional assessments, or advanced biomarkers, in dementia prediction. Instead, we focused on commonly available variables in real-world clinical settings because the primary goal of this study was to investigate AI bias rather than to maximize predictive accuracy. Third, relying on ICD codes to identify dementia cases in the REP cohort introduces potential ascertainment bias. In real-world settings, socioeconomic barriers to care may prevent low SES individuals from obtaining a formal diagnosis despite the presence of disease. Consequently, these undiagnosed individuals are liable to be misclassified as not having dementia. This specific form of label noise likely inflates the apparent error rates (specifically false negatives) for low SES groups independent of algorithmic performance. Fourth, the requirement for a 5-year observation window and subsequent follow-up may inadvertently exclude individuals with fragmented care. This selection factor suggests that the performance disparities reported here may underestimate the full extent of algorithmic bias. Additionally, while SMOTE-NC consistently reduced numerical BER disparities, the statistical significance of this reduction was not formally tested. These findings should be interpreted as descriptive evidence of fairness improvement. Finally, we did not consider temporal validation of the AI models over time, particularly in the REP cohort (real-world data). Health care systems are dynamic, and evolving diagnostic criteria for dementia and changes in ICD coding practices may lead to concept drift. Consequently, the models' utility for predicting dementia risk in future patient populations requires further investigation. Nonetheless, our main objective was to assess the SES-driven AI model bias rather than to optimize prediction performance, and our findings successfully demonstrate the influence of SES on AI bias in dementia prediction.

Future Work

In this study, BER was selected as the primary metric for evaluating model performance and fairness. This metric is particularly well suited for classification tasks with imbalanced classes, such as dementia prediction, because it gives equal importance to errors made on the understudied and majority classes, unlike standard accuracy. Our use of BER aligns with a fairness goal of achieving parity in overall error rates between demographic subgroups. However, it is important to acknowledge that "fairness" is a multifaceted concept, and no single metric can capture all its dimensions. Alternative fairness criteria could have been used, each with different implications. Future work should extend this analysis by evaluating these models against a broader suite of fairness metrics to provide a more holistic assessment of their performance trade-offs. We should also investigate whether differences in health care utilization patterns contribute to these variations, as female individuals may engage with health care services differently than male participants, potentially influencing AI model predictions.

Our further research should explore advanced fairness-aware algorithms that optimize both fairness and predictive accuracy, ensuring that model improvements do not disproportionately benefit or disadvantage any subgroup. Additionally, our future work should evaluate alternative debiasing strategies beyond oversampling, such as model recalibration and adversarial learning, to address structural biases inherent in the data.

Conclusions

This study underscores the biases in AI models predicting dementia, with performance disparities observed across SES and sex groups. Individuals from lower SES backgrounds consistently experienced less accurate predictions, as reflected in higher BERs. These disparities suggest that AI models trained on real-world health care data may inadvertently reinforce patterns of underservice among certain populations, underscoring the need for fairness-aware modeling approaches.

Comparisons between SES measures revealed that the HOUSES Index and ADI capture different aspects of socioeconomic disadvantage, with the HOUSES Index potentially offering more granularity at the individual level. The observed sex-based disparities in model bias further emphasize the intersectionality of SDOH, highlighting the importance of multidimensional fairness evaluations in AI-driven health care applications.

Ultimately, this research reinforces the ethical imperative to integrate sociodemographic factors into AI model development and evaluation. Addressing these biases is essential for promoting fair AI-driven solutions in dementia risk prediction, where early and accurate diagnosis is critical for patient outcomes.

Funding

This study was supported by National Institutes of Health R01 AG068007, R01 HL171508, and RF1 AG090341. The Mayo Clinic Study of Aging was supported by National Institutes of Health grants (U01 AG006786, P30 AG062677, R37 AG011378, R01 AG041851, and R01 NS097495), the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic, the Mayo Foundation for Medical Education and Research, the Liston Award, the GHR Foundation, and the Schuler Foundation

XSL•FO
RenderX

and used the resources of the Rochester Epidemiology Project (REP) medical records linkage system, which is supported by the National Institute on Aging (AG 058738), by the Mayo Clinic Research Committee, and by fees paid annually by REP users. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy and legal reasons. The Mayo Clinic Study on Aging and Rochester Epidemiology Project studies provide data to qualified researchers upon reasonable request.

Authors' Contributions

XL: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.
MV: Validation, Resources, Data Curation, Writing – review & editing.
JSS: Validation, Resources, Data Curation, Writing – review & editing.
MG: Validation, Writing – review & editing.
RCP: Validation, Writing – review & editing.
MMM: Validation, Writing – review & editing.
CW: Validation, Writing – review & editing.
YJJ: Validation, Writing – review & editing.
SS: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Conflicts of Interest

MV consulted for F. Hoffmann-La Roche Ltd, unrelated to this manuscript; she currently receives research funding from the National Institutes of Health and has equity ownership in Johnson and Johnson, Merck, Medtronic, and Amgen. RCP serves as a consultant for Roche, Inc.; Eisai, Inc.; Genentech, Inc.; Eli Lilly, Inc.; and Nestle, Inc.; served on a Data Safety Monitoring Board for Genentech; receives royalties from Oxford University Press and UpToDate; and receives National Institutes of Health funding.

References

1. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821. [doi: 10.1002/alz.13809] [Medline: 38689398]
2. Sage growth partners. The state of healthcare automation: urgent need, growing awareness and tremendous potential. 2021. URL: https://go.sage-growth.com/state-of-healthcare-automation [accessed 2026-02-04]
3. National trends in hospital and physician adoption of electronic health records. ASTP - Assistant Secretary for Technology Policy. 2025. URL: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records [accessed 2026-01-22]
4. Artificial intelligence-enabled medical devices. US Food and Drug Administration (FDA). 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2026-01-22]
5. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021 Jan;27(1):136-140. [doi: 10.1038/s41591-020-01192-7] [Medline: 33442014]
6. Steinwachs DM, Stratton K, Kwan LY. Accounting for social risk factors in medicare payment. : National Academies Press; 2017 URL: https://www.nationalacademies.org/read/23635/chapter/1 [accessed 2026-01-22]
7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
8. Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. J Natl Med Assoc 2002 Aug;94(8):666-668. [Medline: 12152921]
9. Dzau VJ, McClellan MB, McGinnis JM, et al. Vital directions for health and health care: priorities from a National Academy of Medicine initiative. JAMA 2017 Apr 11;317(14):1461-1470. [doi: 10.1001/jama.2017.1964] [Medline: 28324029]
10. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882. [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
11. Warnecke RB, Oh A, Breen N, et al. Approaching health disparities from a population perspective: the National Institutes of Health Centers for Population Health and Health Disparities. Am J Public Health 2008 Sep;98(9):1608-1615. [doi: 10.2105/AJPH.2006.102525] [Medline: 18633099]
12. Roberts ET, Zaslavsky AM, Barnett ML, Landon BE, Ding L, McWilliams JM. Assessment of the effect of adjustment for patient characteristics on hospital readmission rates: implications for pay for performance. JAMA Intern Med 2018 Nov 1;178(11):1498-1507. [doi: 10.1001/jamainternmed.2018.4481] [Medline: 30242346]

XSL•FO
RenderX

13. Sills MR, Hall M, Colvin JD, et al. Association of social determinants with children's hospitals' preventable readmissions performance. JAMA Pediatr 2016 Apr;170(4):350-358. [doi: 10.1001/jamapediatrics.2015.4440] [Medline: 26881387]
14. Franks P, Fiscella K. Effect of patient socioeconomic status on physician profiles for prevention, disease management, and diagnostic testing costs. Med Care 2002 Aug;40(8):717-724. [doi: 10.1097/00005650-200208000-00011] [Medline: 12187185]
15. Baker DW, Chassin MR. Holding providers accountable for health care outcomes. Ann Intern Med 2017 Sep 19;167(6):418-423. [doi: 10.7326/M17-0691] [Medline: 28806793]
16. Bernheim SM, Ross JS, Krumholz HM, Bradley EH. Influence of patients' socioeconomic status on clinical management decisions: a qualitative study. Ann Fam Med 2008;6(1):53-59. [doi: 10.1370/afm.749] [Medline: 18195315]
17. Bach PB, Pham HH, Schrag D, Tate RC, Hargraves JL. Primary care physicians who treat Blacks and Whites. N Engl J Med 2004 Aug 5;351(6):575-584. [doi: 10.1056/NEJMsa040609] [Medline: 15295050]
18. Jha AK, Zaslavsky AM. Quality reporting that addresses disparities in health care. JAMA 2014 Jul 16;312(3):225-226. [doi: 10.1001/jama.2014.7204] [Medline: 25027134]
19. Snyder-Mackler N, Burger JR, Gaydosh L, et al. Social determinants of health and survival in humans and other animals. Science 2020 May 22;368(6493):eaax9553. [doi: 10.1126/science.aax9553] [Medline: 32439765]
20. Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. Health Aff (Millwood) 2002;21(2):60-76. [doi: 10.1377/hlthaff.21.2.60] [Medline: 11900187]
21. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019 Dec;9(2):010318. [doi: 10.7189/jogh.09.020318] [Medline: 31788229]
22. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023 Jun;2(6):e0000278. [doi: 10.1371/journal.pdig.0000278] [Medline: 37347721]
23. Juhn YJ, Ryu E, Wi CI, et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. J Am Med Inform Assoc 2022 Jun 14;29(7):1142-1151. [doi: 10.1093/jamia/ocac052] [Medline: 35396996]
24. Cadena-Hernandez AG, Trejo-Castro AI, Celaya-Padilla JM, Tamez-Pena J, Martinez-Torteya A. Longitudinal gender-specific differences in the conversion from mild cognitive impairment to Alzheimer's disease. Presented at: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); Mar 4-7, 2018. [doi: 10.1109/BHI.2018.8333404]
25. Gamberger D, Ženko B, Mitelpunkt A, Shachar N, Lavrač N. Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Brain Inform 2016 Sep;3(3):169-179. [doi: 10.1007/s40708-016-0035-5] [Medline: 27525218]
26. Aditya CR, Pande MBS. Devising an interpretable calibrated scale to quantitatively assess the dementia stage of subjects with Alzheimer's disease: a machine learning approach. Informatics in Medicine Unlocked 2017;6:28-35. [doi: 10.1016/j.imu.2016.12.004]
27. Puaschitz NG, Jacobsen FF, Mannseth J, et al. Factors associated with access to assistive technology and telecare in home-dwelling people with dementia: baseline data from the LIVE@Home.Path trial. BMC Med Inform Decis Mak 2021 Sep 15;21(1):264. [doi: 10.1186/s12911-021-01627-2] [Medline: 34525979]
28. Twait EL, Andaur Navarro CL, Gudnason V, Hu YH, Launer LJ, Geerlings MI. Dementia prediction in the general population using clinically accessible variables: a proof-of-concept study using machine learning. The AGES-Reykjavik study. BMC Med Inform Decis Mak 2023 Aug 28;23(1):168. [doi: 10.1186/s12911-023-02244-x] [Medline: 37641038]
29. Liu Z, Garg M, Fu S, et al. Harnessing transfer learning for dementia prediction: leveraging sex-different mild cognitive impairment prognosis. Presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 5-8, 2023. [doi: 10.1109/BIBM58861.2023.10385516]
30. Wu H, Sylolypavan A, Wang M, Wild S. Quantifying health inequalities induced by data and AI models. 2022 Presented at: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence; Jul 23-29, 2022. [doi: 10.24963/ijcai.2022/721]
31. Roberts RO, Geda YE, Knopman DS, et al. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology 2008;30(1):58-69. [doi: 10.1159/000115751] [Medline: 18259084]
32. Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med 2004 Sep;256(3):183-194. [doi: 10.1111/j.1365-2796.2004.01388.x] [Medline: 15324362]
33. Guze SB. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV). AJP 1995 Aug;152(8):1228-1228. [doi: 10.1176/ajp.152.8.1228]
34. Rocca WA, Yawn BP, St Sauver JL, Grossardt BR, Melton LJ 3rd. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. Mayo Clin Proc 2012 Dec;87(12):1202-1213. [doi: 10.1016/j.mayocp.2012.08.012] [Medline: 23199802]
35. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the Neighborhood Atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458. [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
36. Juhn YJ, Beebe TJ, Finnie DM, et al. Development and initial testing of a new socioeconomic status measure based on housing data. J Urban Health 2011 Oct;88(5):933-944. [doi: 10.1007/s11524-011-9572-7] [Medline: 21499815]

XSL•FO
RenderX

37. Cohen JW, Cohen SB, Banthin JS. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. Med Care 2009 Jul;47(7 Suppl 1):S44-S50. [doi: 10.1097/MLR.0b013e3181a23e3a] [Medline: 19536015]
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
39. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ 3rd, Rocca WA. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. Mayo Clin Proc 2012 Feb;87(2):151-160. [doi: 10.1016/j.mayocp.2011.11.009] [Medline: 22305027]

Abbreviations

ADI: Area Deprivation Index
AI: artificial intelligence
BER: balanced error rate
DEM: dementia
HOUSES: Housing-Based Socioeconomic Status
LR: logistic regression
MCI: mild cognitive impairment
MCSA: Mayo Clinic Study on Aging
NB: Naïve Bayes
REP: Rochester Epidemiology Project
RF: Random Forest
SES: socioeconomic status
SMOTE-NC: Synthetic Minority Oversampling Technique for Nominal and Continuous features
SVM: support vector machine

Edited by A Coristine; submitted 09.Jul.2025; peer-reviewed by J Dai, LM Diaz; accepted 07.Jan.2026; published 17.Feb.2026.

Please cite as:
Liu X, Garg M, Vassilaki M, St Sauver J, Petersen RC, Malik MM, Wi CI, Juhn YJ, Sohn S
Assessing the Impact of Sociodemographic Factors on Artificial Intelligence Models in Predicting Dementia: Retrospective Cohort Study
JMIR Med Inform 2026;14:e80405
URL: https://medinform.jmir.org/2026/1/e80405
doi:10.2196/80405

© Xingyi Liu, Muskan Garg, Maria Vassilaki, Jennifer St Sauver, Ronald C Petersen, Momin M Malik, Chung Il Wi, Young J Juhn, Sunghwan Sohn. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Screening for Clinically Significant Nephrolithiasis Based on Simple Health Checkup Clinical and Urine Parameters in General Populations: Multicenter Machine Learning Study

Hao-Wei Chen[1,2,3], MD, PhD; Pei-Siou Wei[4,5], MS; Yu-Chen Chen[1,2,6], MD; Jeng-Yih Wu[7,8,9], MD; Chia-I Lin[8], MD; Yii-Her Chou[1,2], MD, PhD; Yung-Shun Juan[1,2], MD, PhD; Wen-Jeng Wu[1,2], MD, PhD; Chung-Yao Kao[4], PhD; Jung-Ting Lee[10], PhD

[1]Department of Urology, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan
[10]School of Medicine, National Sun Yat-sen University, 70 Lien-Hai Road, Kaohsiung, Taiwan
[2]Graduate Institute of Clinical Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[3]Center for Big Data Research, Kaohsiung Medical University, Kaohsiung, Taiwan
[4]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[5]United Kaohsiung AI Medical Technology Co, Ltd., Kaohsiung, Taiwan
[6]Regenerative Medicine and Cell Therapy Research Center, Kaohsiung Medical University, Kaohsiung, Taiwan
[7]Department of Internal Medicine, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan
[8]Health Management Center, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan
[9]College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

Corresponding Author:
Jung-Ting Lee, PhD
School of Medicine, National Sun Yat-sen University, 70 Lien-Hai Road, Kaohsiung, Taiwan

Abstract

Background: Nephrolithiasis affects approximately 15% of the population and often remains undetected in asymptomatic individuals. Current diagnostic approaches rely on imaging tools, such as ultrasound or computed tomography, which are costly, operator dependent, or involve radiation, making them unsuitable for large-scale screening. A standardized, practical, and low-cost screening strategy for early identification of clinically significant kidney stones is still lacking.

Objective: This study aimed to develop a low-cost, rapid screening model for clinically significant nephrolithiasis using machine learning (ML) and simple clinical parameters.

Methods: We conducted a multihospital retrospective study using data from 3 hospitals in Kaohsiung, Taiwan (2012 - 2021). Adults without renal colic were included. ML models were trained and tested using 10 routine variables: sex, age, BMI, gout, diabetes, estimated glomerular filtration rate, urine pH, red blood cell count, specific gravity, and bacteriuria. Multiple ML algorithms were trained and evaluated, and the best-performing model was selected based on the area under the receiver operating characteristic curve and the area under the precision-recall curve. To assess model interpretability, Shapley value analysis was performed to determine the relative importance and contribution of each variable to the model's predictive performance.

Results: Among 6528 participants, the best-performing model achieved an area under the receiver operating characteristic curve of 0.968 (95% CI 0.956 - 0.980), an area under the precision-recall curve of 0.936 (95% CI 0.918 - 0.953), a sensitivity of 0.873 (95% CI 0.841 - 0.904), and a specificity of 0.947 (95% CI 0.935 - 0.959). Shapley value analysis identified urine red blood cell count, estimated glomerular filtration rate, and urine specific gravity as the 3 most influential predictors.

Conclusions: This ML-based model enables efficient, noninvasive, and large-scale kidney stone screening using routine health data. It can be integrated into health checkups or telemedicine platforms to facilitate early detection and proactive management. Although the model was developed using an Asian population, future validation in diverse cohorts is warranted to confirm its generalizability.

(JMIR Med Inform 2026;14:e80764) doi:10.2196/80764

KEYWORDS

machine learning; nephrolithiasis; health services; software as a medical device; computer-aided diagnosis

XSL•FO
RenderX

Introduction

Approximately 10% of the global population faces a lifetime risk of developing kidney stones [1], and the prevalence of this disease has been steadily rising over the past two decades [2,3]. The most common symptoms include flank pain and hematuria, and kidney and ureteral stones are significant risk factors for chronic kidney disease [4]. A study reported that in the year 2000 alone, the total medical expenditure for treating kidney and ureteral stones in the United States reached US $2.1 billion [5].

Currently, there is no effective tool available for kidney stone screening. Renal ultrasonography is the most commonly used method for detecting kidney stones in clinical settings and emergency departments [5]. While it is a convenient and radiation-free technique, its accuracy is highly dependent on the operator's skill, thus requiring trained medical professionals for both operation and interpretation. As a result, its use in large-scale screening would demand substantial health care manpower [6,7]. In contrast, abdominal computed tomography (CT) is considered the gold standard for diagnosing kidney and ureteral stones [8]. However, due to its high radiation exposure and limited accessibility for routine or large-scale screening, CT is not suitable for asymptomatic populations or frontline outpatient use [9]. Therefore, existing imaging tools are impractical for large-scale screening of kidney stones. Developing a simple, cost-effective, and accessible screening tool would be highly beneficial for the early detection of kidney stones and for timely medical decision-making.

We previously developed an artificial intelligence (AI) algorithm that uses routine health checkup data, including urine routine and blood creatinine, to preliminarily predict the presence of kidney stones larger than 2 mm in overweight individuals (BMI ≥25 kg/m^2) [6]. This tool has shown potential for use in kidney stone screening. In this study, we aimed to build upon our prior success by leveraging machine learning (ML) techniques and routine health data to further develop a clinically applicable AI model that can predict clinically significant kidney stones in all patients—not limited to those with BMI >24 kg/m^2. The models will be validated using patient cohorts from multiple hospitals.

Methods

Description of Participants

In this cohort study, clinical data were retrospectively collected from three hospitals located in Kaohsiung, Taiwan: Kaohsiung Medical University Hospital, Kaohsiung Municipal Ta-Tung Hospital, and Kaohsiung Municipal Siaogang Hospital. The data collection period spanned from January 2012 to March 2021.

Patients were categorized into 2 groups based on the presence or absence of nephrolithiasis measuring ≥2 mm. This classification followed the methodology of our previously published ML-based nephrolithiasis screening model developed for overweight and obese individuals [6]. In accordance with prior urological literature, stone fragments smaller than 2 mm are considered clinically insignificant, as they are unlikely to cause stone-related events or require intervention [7]. Therefore, a cutoff of 2 mm was adopted in this study to define clinically significant nephrolithiasis. The presence of stones of 2 mm or larger was confirmed using abdominal CT or a combination of renal ultrasonography and plain abdominal radiography (KUB: kidney, ureter, and bladder), consistent with our prior study protocol [6]. Patients were excluded if they experienced renal colic at the time of data collection, had undergone kidney transplantation, had indwelling foreign bodies in the upper urinary tract (eg, ureteral stents or percutaneous nephrostomy tubes), were receiving renal replacement therapy (hemodialysis or peritoneal dialysis), were aged younger than 18 years, or lacked complete medical records. For model development, data from January 2012 to December 2019 were used for training and validation. An independent test cohort comprising data collected between January 2020 and March 2021 was used for the clinical evaluation of model performance.

Ethical Considerations

This study was approved by the Institutional Review Board of Kaohsiung Medical University Hospital (KMUHIRB-E(I)-20210331) and conducted in accordance with the principles outlined in the Declaration of Helsinki. The requirement for informed consent was waived due to the retrospective nature of the study and the minimal risk to participants. All data were obtained from existing electronic health records and were fully deidentified before analysis. No personally identifiable information was accessed by the researchers, and data access was restricted to authorized study personnel only. No compensation was provided to participants, as no direct contact or intervention was involved in this retrospective study.

Variables and Data Processing

A total of 10 predictor variables were included for model development, categorized into sociodemographic, health-related, and clinical domains. The categorical response variable indicated the presence (1) or absence (0) of nephrolithiasis ≥2 mm. Sociodemographic variables included age (numeric) and sex (categorical: male=1 and female =0). Health-related variables comprised BMI (numeric with 2 decimals) and a history of diabetes mellitus and gout (both categorical: with=1 and without=0).

Clinical variables included five parameters derived from urine and blood tests:

1. Urine specific gravity (SG; numeric with 3 decimals)
2. Urine pH (numeric with 1 decimal)
3. Bacteriuria (categorical: present =1, absent =0)
4. Urine red blood cell (RBC) count (treated as a numeric variable; details described as follows): In Taiwan, urine RBC counts are commonly reported in terms of a range (eg, 6 - 10 RBCs per high-power field [HPF]), with the lowest range being "0 to 2 RBCs per HPF." In this study, the urine RBC data we collected consist of the following ranges: 0 to 2, 3 to 5, 6 to 10, 11 to 25, 26 to 50, 51 to 99, and 100 or greater. As such, it may be treated as a categorical variable in this study. However, to accommodate various range scales used by different hospitals and make the model

XSL•FO
RenderX

universally applicable, we adopted the mean of the reported ranges (eg, 8 for the range 6 - 10 RBCs per HPF) as the corresponding urine RBC count and treated the variable as numeric (continuous). The exceptions to the rule are the range of "0 to 2" and the ranges whose mean values exceed 100. For the 0 to 2 range, the value 0 is assigned. This is due to an American Urological Association guideline, which defines microhematuria as "3 or more RBCs per HPF on microscopic examination." Therefore, 0 to 2 RBCs per HPF is generally considered as no hematuria. On the other hand, patients with macroscopic hematuria generally have their urine RBC counts reported as "100 or more RBCs per HPF," or a range with a mean value exceeding 100, or something equivalent to that effect. As such, for ranges whose mean values exceed 100 RBCs per HPF, we applied the capped value of 200 to maintain consistency across the dataset [10].

5. Estimated glomerular filtration rate (eGFR), numeric with 1 decimal, calculated using the Isotope Dilution Mass Spectrometry–traceable Modification of Diet in Renal Disease equation: eGFR (mL/min/1.73 m^2)=175×$(Scr)^{-1.154}$×$(age)^{-0.203}$×(0.742 if female) [8]

For descriptive statistics, the Fisher exact test was used for categorical variables (eg, sex, history of diabetes mellitus, and history of gout), and the Mann-Whitney *U* test was applied for numeric variables (eg, age, BMI, eGFR, urine SG, urine RBC count, and urine pH) between patients with and without nephrolithiasis. To verify the similarity between the training and testing datasets, we compared feature distributions by calculating category percentages for categorical variables and means with SDs for continuous variables. All statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (version 24.0; IBM), and *P* values of less than .05 were considered statistically significant.

Model Development, Fitting, and Evaluation

ML methods were used to build the models [9]. The logistic regression, the k-nearest neighbors algorithm, the support vector machine, and the random forest (RF) algorithm were used as the standard methods for comparison, while an artificial neural network (ANN) model was developed for deep learning classification. All standard models were implemented in Python using the scikit-learn library (version 1.1.3; scikit-learn developers) with recommended settings. The respective algorithms for each method were run using the recommended parameters [11,12].

For the ANN, implemented in Python, the 10-fold cross-validation protocol was performed on the training set to produce 10 models, each structured as a 9-layer fully connected neural network. These models were averaged into 1 final model. To address the imbalance between stone-positive and stone-negative cohorts, the synthetic minority over-sampling technique, as well as the approach of assigning different weights to the 2 patient groups, was adopted when training our models. In addition, a pronounced sex imbalance among the patient cohorts was observed. To ensure equitable accuracy for the 2 gender groups, 2 separate cutoff thresholds were identified for male and female patients. The 2 thresholds were then standardized to 0.5 by applying gender-specific rescalers to the averaged outputs. The resulting combined structure served as the final ANN model with a unified threshold of 0.5. Additional information is provided in Figure S1 in Multimedia Appendix 1.

Data between January 2012 and December 2019 were used for model training and validation. The data were randomly divided into a training dataset (75%) and a validation dataset (25%). Model training was performed using the training data, while validation data helped identify overfitting or bias. Data registered after January 1, 2020, formed the testing dataset (ie, the holdout dataset) for independent evaluation of model performance. Performance on the testing data was assessed by various metrics, including the area under the receiver operating characteristic (ROC) curve (AUROC), the area under the precision-recall (PR) curve (AUPRC), and net benefit by the decision curve analysis. The optimal thresholds for standard models were based on the Youden index [13] from their respective training ROC curves. For the ANN, the classification threshold was preset at 0.5. These thresholds were used to calculate sensitivity, specificity, positive predictive value, and negative predictive value on the validation and testing sets. Model calibration analysis was also performed to assess how accurately the output probabilities of the models reflect the true likelihood of outcomes.

To examine model stability and fairness across different patient subgroups, additional analyses were performed by individual variables (eg, sex, diabetes, gout, bacteriuria, urine RBC count, age, BMI, and eGFR), where model performance metrics (area under the curve [AUC] and accuracy) were calculated independently for each subgroup.

Importance of Feature Variables

To evaluate the predictive power of each feature variable, their relative impact on the ANN model output was assessed by applying the Shapley value analysis [14]. Features with notably high impact were selected, and an ANN model was trained using only these variables. The performance of the model was compared with that of the full-variable ANN model to assess whether key features alone provided sufficient predictive power.

Results

Characteristics of Study Participants

In the training dataset, 1243 (33.11%) participants had nephrolithiasis, and 2511 (66.89%) participants were without nephrolithiasis of 2 mm or greater, among a total of 3754 (100.0%) participants. The validation dataset had 414 (33.09%) participants with and 837 (66.91%) participants without nephrolithiasis of 2 mm or greater, among a total of 1251 (100.0%) participants. The testing dataset had 400 (26.26%) participants with and 1123 (73.74%) participants without nephrolithiasis of 2 mm or greater, among a total of 1523 (100.0%) participants. The characteristics of the participants in these 3 sets are described in detail in Tables 1 and 2. Statistical similarity among the 3 sets was observed. It is also observed that the mean values (or the "percentages," in the case of binary variables) of the variables we selected are statistically different between participants with and without nephrolithiasis. The

XSL•FO
RenderX

Fisher exact and Mann-Whitney *U* tests were applied to verify the statistical significance of these differences, and the resulting *P* values are also reported in Tables 1 and 2.

XSL•FO
RenderX

Table . Descriptive analysis of the training, validation, and testing datasets.

	Training set, n (%)				Validation set, n (%)				Testing set, n (%)			
	w/neph[a] (n=1243, 33.1%)	wo/neph[b] (n=2511, 66.9 %)	Total (N=3754)	*P* value	w/neph (n=414, 33.1%)	wo/neph (n=837, 66.9%)	Total (N=1251)	*P* value	w/neph (n=400, 26.3%)	wo/neph (n=1123, 73.7%)	Total (N=1523)	*P* value
Gender				.02				.19				<.001
Male	820 (66.0)	1750 (69.7)	2570 (68.5)		273 (65.9)	584 (69.8)	857 (68.5)		250 (62.5)	433 (38.6)	683 (44.8)	
Female	423 (34.0)	761 (30.3)	1184 (31.5)		141 (34.1)	253 (30.2)	394 (31.5)		150 (37.5)	690 (61.4)	840 (55.2)	
Diabetes				<.001				<.001				<.001
With	220 (17.7)	148 (5.9)	368 (9.8)		86 (20.8)	59 (7.1)	145 (11.6)		94 (23.5)	15 (1.3)	109 (7.2)	
Without	1023 (82.3)	2363 (94.1)	3386 (90.2)		328 (79.2)	778 (92.9)	1106 (88.4)		306 (76.5)	1108 (98.7)	1414 (92.8)	
Gout				<.001				<.001				<.001
With	49 (3.9)	16 (0.6)	65 (1.7)		20 (4.8)	3 (0.4)	23 (1.8)		12 (3.0)	4 (0.4)	16 (1.1)	
Without	1194 (96.1%)	2495 (99.4)	3689 (98.3)		394 (95.2)	834 (99.6)	1228 (98.2)		388 (97.0)	1119 (99.6%)	1507 (98.9)	
BAC[c]				<.001				<.001				<.001
With	191 (15.4)	249 (9.9)	440 (11.7)		47 (11.4)	78 (9.3)	125 (10.0)		49 (12.3)	145 (12.9)	194 (12.7)	
Without	1052 (84.6)	2262 (90.1)	3314 (88.3)		367 (88.6)	759 (90.7)	1126 (90.0)		351 (87.7)	978 (87.1)	1329 (87.3)	
Urine RBC count[d]				<.001				<.001				<.001
0-2	219 (17.6)	2265 (90.2)	2484 (66.2)		80 (19.3)	752 (89.8)	832 (66.5)		84 (21.0)	921 (82.0)	1005 (66.0)	
3-5	128 (10.3)	169 (6.7)	297 (7.9)		36 (8.7)	63 (7.52)	99 (7.9)		41 (10.3)	138 (12.3)	179 (11.7)	
6-10	104 (8.4)	44 (1.8)	148 (3.9)		35 (8.5)	13 (1.6)	48 (3.8)		35 (8.8)	30 (2.7)	65 (4.3)	
11-25	153 (12.3)	23 (0.9)	176 (4.7)		56 (13.5)	6 (0.72)	62 (5.0)		43 (10.7)	19 (1.7)	62 (4.1)	
26-50	141 (11.3)	5 (0.2)	146 (3.9)		38 (9.2)	2 (0.24)	40 (3.2)		33 (8.2)	6 (0.5)	39 (2.5)	
51-99	121 (9.7)	1 (0.04)	122 (3.3)		47 (11.4)	0 (0.0)	47 (3.8)		32 (8.0)	3 (0.3)	35 (2.3)	
≥100	377 (30.4)	4 (0.16)	381 (10.1)		122 (29.4)	1 (0.12)	123 (9.8)		132 (33.0)	6 (0.5)	138 (9.1)	

[a]w/neph: with nephrolithiasis.

[b]wo/neph: without nephrolithiasis.

[c]BAC: bacteriuria.

[d]RBC: red blood cell.

XSL•FO
RenderX

Table . Characteristics about age, BMI, urine pH, estimated glomerular filtration rate (eGFR), and urine red blood cell (RBC) count of patients in the training, validation, and testing datasets.

	Training set, mean (SD) [95% CI]			Validation set, mean (SD) [95% CI]			Testing set, mean (SD) [95% CI]		
	w/neph[a]	wo/neph[b]	*P* value	w/neph	wo/neph	*P* value	w/neph	wo/neph	*P* value
Age	55.92 (13.30) [55.15 - 56.68]	50.92 (10.10) [50.51 - 51.33]	<.001	55.53 (13.47) [54.34 - 56.71]	51.36 (10.07) [50.73 - 51.98]	<.001	57.31 (13.80) [55.95 - 58.66]	45.52 (9.32) [44.97 - 46.07]	<.001
eGFR	74.67 (29.23) [72.98 - 76.35]	86.31 (17.83) [85.58 - 87.03]	<.001	73.38 (30.44) [70.70 - 76.06]	85.95 (17.64) [84.86 - 87.04]	<.001	69.47 (26.74) [66.84 - 72.10]	95.03 (19.18) [93.90 - 96.15]	<.001
BMI	25.77 (3.81) [25.55 - 25.99]	24.64 (3.63) [24.50 - 24.79]	<.001	25.91 (4.02) [25.56 - 26.27]	24.67 (3.59) [24.45 - 24.89]	<.001	26.65 (4.82) [26.18 - 27.13]	23.68 (3.86) [23.45 - 23.90]	<.001
Urine pH	6.03 (0.79) [5.98 - 6.07]	6.03 (0.74) [6.00 - 6.06]	.93	6.07 (0.80) [5.99 - 6.15]	6.01 (0.75) [5.96 - 6.06]	.21	6.11 (0.84) [6.02 - 6.19]	6.02 (0.79) [5.97 - 6.06]	.06
Urine RBC count	75.42 (84.64) [70.55 - 80.30]	1.03 (8.83) [0.67 - 1.39]	<.001	74.93 (84.60) [67.48 - 82.38]	0.83 (6.82) [0.41 - 1.26]	<.001	78.18 (87.85) [69.54 - 86.82]	2.48 (15.48) [1.58 - 3.39]	<.001
Urine SG[c]	1.014 (0.006) [1.0137 - 1.0144]	1.018 (0.007) [1.018 - 1.0185]	<.001	1.014 (0.007) [1.013 - 1.0145]	1.018 (0.007) [1.018 - 1.0185]	<.001	1.015 (0.008) [1.014 - 1.0154]	1.018 (0.006) [1.018 - 1.0187]	<.001

[a]w/neph: with nephrolithiasis.

[b]wo/neph: without nephrolithiasis.

[c]SG: specific gravity.

Model Performance

The ANN model we developed demonstrated strong discriminatory power in identifying participants with kidney or ureteral stones, using only limited information typically collected at an initial clinical visit. It achieved validation and testing AUROCs of 0.970 and 0.968 and validation and testing AUPRCs of 0.960 and 0.936, respectively. Using a cutoff value of 0.5, the model showed high sensitivity, specificity, positive predictive value, and negative predictive value in both validation and testing sets, with overall correction rates of 92.3% and 92.7%, respectively. These performance measures indicate that the model achieved a good balance in identifying both positive and negative cases and is able to render both high precision and high recall.

The performance of models obtained using standard methods appeared to be somewhat inferior to that of the ANN model, with the exception of the RF algorithm, which performed equally well. The logistic regression, k-nearest neighbors, and support vector machine (linear kernel) models had testing AUROCs of 0.933, 0.916, and 0.910, respectively, and testing AUPRCs of 0.822, 0.876, and 0.790, respectively. The RF model achieved AUROC and AUPRC of 0.965 and 0.934, which are almost the same as those of the ANN model. The overall correction rate was 90.5%. Despite the high AUROC and AUPRC scores, the RF model showed unbalanced performance between sensitivity and specificity, a pattern also observed in the other conventional models.

Table 3 summarizes the performance data of these 5 models. Figure 1A-C presents their training, validation, and testing ROC curves, while Figure 2A-C presents the respective PR curves.

XSL•FO
RenderX

Figure 1. The receiver operating characteristic (ROC) curves of the artificial neural network (ANN), random forest, k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression models. (A) Training ROC curves. (B) Validation ROC curves. (C) Testing ROC curves. AUROC: area under the receiver operating characteristic curve.

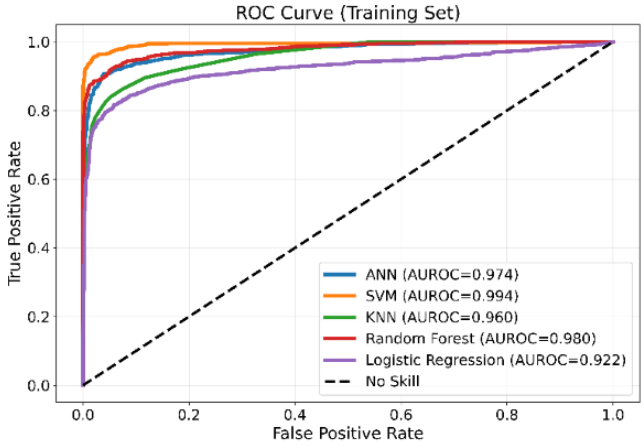


Fig. 1A

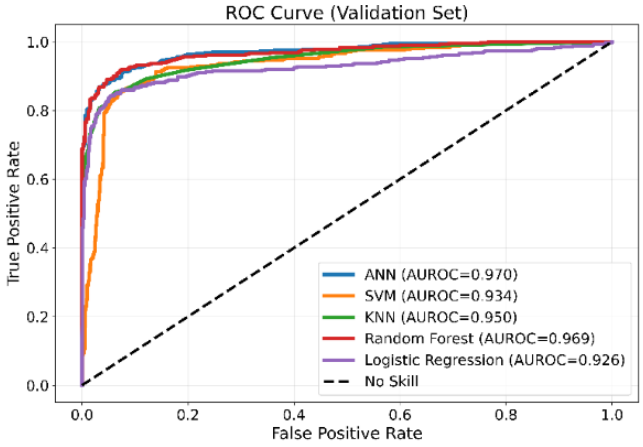


Fig. 1B

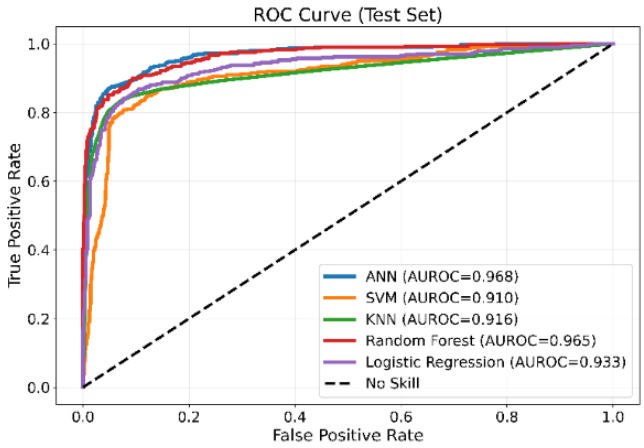


Fig. 1C

XSL•FO
RenderX

Figure 2. Precision-recall (PR) curves of the artificial neural network (ANN), random forest, k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression models. (A) Training PR curves. (B) Validation PR curves. (C) Testing PR curves. AUPRC: area under the precision-recall curve.

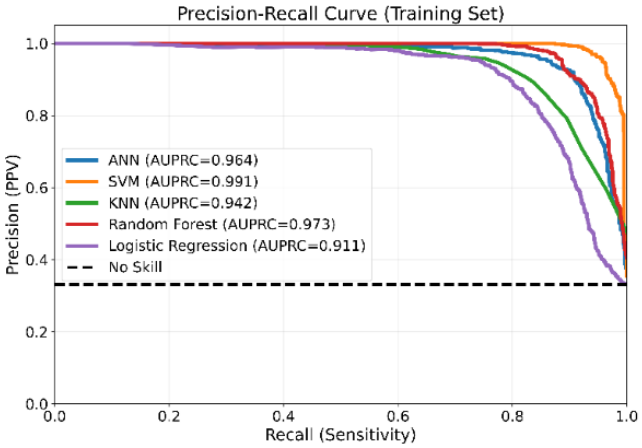


Fig. 2A

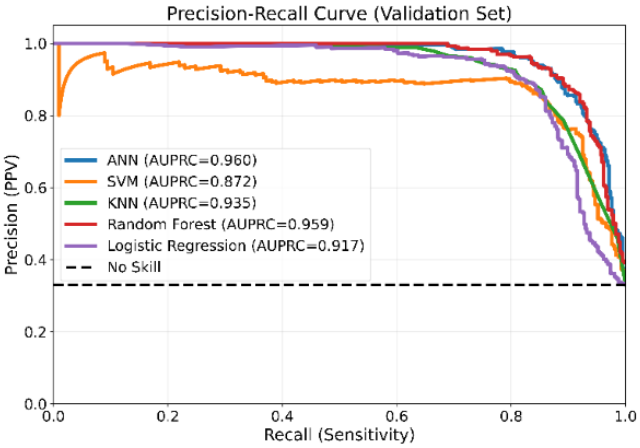


Fig. 2B

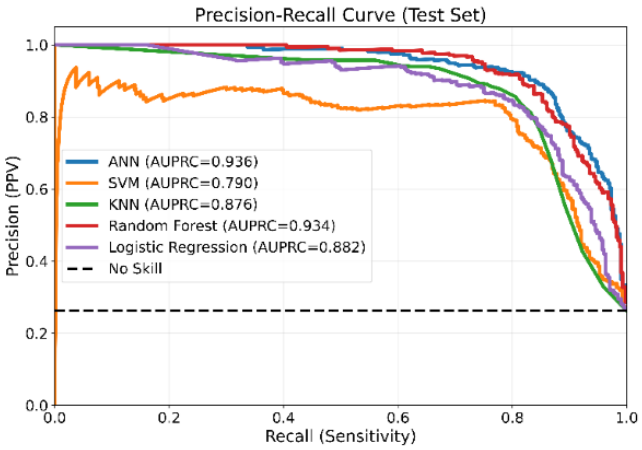


Fig. 2C

XSL•FO
RenderX

Table . Summary of model performances.

Model and dataset	AUROC[a] (95% CI)	AUPRC[b] (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV[c] (95% CI)	NPV[d] (95% CI)
ANN[e]							
Training	0.974 (0.968 - 0.980)	0.964 (0.956 - 0.971)	0.943 (0.935 - 0.951)	0.907 (0.890 - 0.923)	0.961 (0.954 - 0.968)	0.920 (0.903 - 0.937)	0.954 (0.946 - 0.963)
Validation	0.970 (0.958 - 0.982)	0.960 (0.946 - 0.973)	0.923 (0.908 - 0.937)	0.889 (0.860 - 0.918)	0.939 (0.923 - 0.956)	0.878 (0.846 - 0.911)	0.945 (0.928 - 0.962)
Testing	0.968 (0.956 - 0.98)	0.936 (0.918 - 0.953)	0.927 (0.915 - 0.939)	0.873 (0.841 - 0.904)	0.947 (0.935 - 0.959)	0.853 (0.819 - 0.888)	0.954 (0.941 - 0.968)
LR[f]							
Training	0.922 (0.911 - 0.934)	0.911 (0.899 - 0.922)	0.900 (0.891 - 0.910)	0.817 (0.796 - 0.839)	0.942 (0.932 - 0.951)	0.874 (0.855 - 0.893)	0.912 (0.902 - 0.923)
Validation	0.926 (0.906 - 0.945)	0.917 (0.897 - 0.936)	0.908 (0.892 - 0.924)	0.843 (0.808 - 0.878)	0.940 (0.924 - 0.956)	0.875 (0.842 - 0.907)	0.924 (0.906 - 0.942)
Testing	0.933 (0.916 - 0.949)	0.882 (0.859 - 0.904)	0.909 (0.895 - 0.924)	0.800 (0.761 - 0.839)	0.948 (0.935 - 0.961)	0.847 (0.810 - 0.883)	0.930 (0.915 - 0.945)
KNN[g]							
Training	0.960 (0.954 - 0.966)	0.942 (0.932 - 0.951)	0.905 (0.896 - 0.914)	0.861 (0.842 - 0.880)	0.927 (0.917 - 0.937)	0.853 (0.834 - 0.873)	0.931 (0.921 - 0.941)
Validation	0.950 (0.936 - 0.963)	0.935 (0.918 - 0.952)	0.899 (0.882 - 0.915)	0.862 (0.829 - 0.896)	0.916 (0.898 - 0.935)	0.836 (0.801 - 0.871)	0.931 (0.914 - 0.948)
Testing	0.916 (0.896 - 0.936)	0.876 (0.853 - 0.899)	0.892 (0.876 - 0.907)	0.848 (0.812 - 0.883)	0.907 (0.890 - 0.924)	0.765 (0.726 - 0.805)	0.944 (0.930 - 0.957)
SVM[h]							
Training	0.994 (0.991 - 0.996)	0.991 (0.987 - 0.994)	0.972 (0.966 - 0.977)	0.957 (0.946 - 0.969)	0.979 (0.973 - 0.984)	0.957 (0.945 - 0.968)	0.979 (0.973 - 0.985)
Validation	0.934 (0.917 - 0.949)	0.872 (0.848 - 0.895)	0.828 (0.807 - 0.849)	0.930 (0.905 - 0.955)	0.778 (0.750 - 0.806)	0.674 (0.636 - 0.713)	0.957 (0.942 - 0.973)
Testing	0.910 (0.891 - 0.929)	0.790 (0.761 - 0.818)	0.722 (0.700 - 0.745)	0.920 (0.893 - 0.947)	0.652 (0.624 - 0.680)	0.485 (0.449 - 0.520)	0.958 (0.944 - 0.972)
RF[i]							
Training	0.980 (0.975 - 0.984)	0.973 (0.966 - 0.979)	0.939 (0.931 - 0.946)	0.919 (0.904 - 0.934)	0.949 (0.940 - 0.957)	0.899 (0.882 - 0.915)	0.959 (0.952 - 0.967)
Validation	0.969 (0.958 - 0.979)	0.959 (0.945 - 0.972)	0.924 (0.909 - 0.939)	0.918 (0.891 - 0.944)	0.927 (0.910 - 0.945)	0.862 (0.830 - 0.894)	0.958 (0.944 - 0.972)
Testing	0.965 (0.954 - 0.975)	0.934 (0.916 - 0.951)	0.905 (0.890 - 0.920)	0.900 (0.871 - 0.929)	0.907 (0.890 - 0.924)	0.774 (0.736 - 0.812)	0.962 (0.951 - 0.974)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]ANN: artificial neural network.

[f]LR: logistic regression.

[g]KNN: k-nearest neighbor.

[h]SVM: support vector machine.

[i]RF: random forest.

XSL•FO
RenderX

Regarding the practical relevance of models' predicted probabilities, the results of decision curve analysis and model calibration analysis are shown in Figure 3A-C and Figure 4A-C, respectively. Figure 3 indicates that all models rendered higher net benefits than those of the baseline strategies ("treat all" and "treat none") over a large range of threshold probabilities. Among them, the performance of the ANN and RF models appeared to be superior to the others over the validation and testing datasets. Figure 4 indicates that the predicted probabilities of all models were somewhat deviated from the actual events. As such, despite exhibiting good performance for identifying nephrolithiasis, further calibration of the model outputs appears to be required. To further evaluate the robustness and generalizability of the ANN model, subgroup analyses were conducted based on demographic and clinical characteristics.

Figure 3. Decision curve analysis for the artificial neural network (ANN), random forest, k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression models. (A) Net benefits over the training set. (B) Net benefits over the validation set. (C) Net benefits over the testing set.

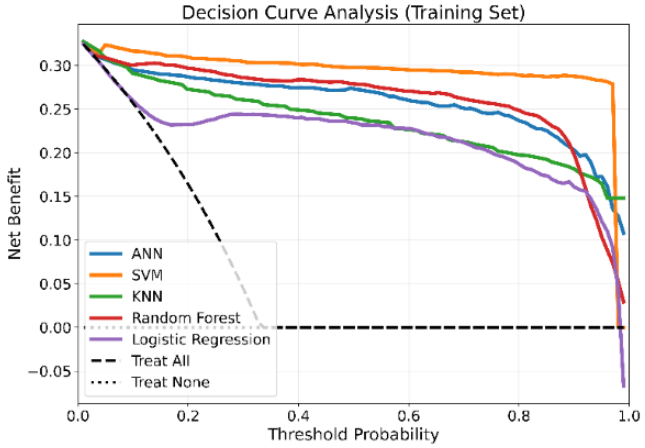


Fig. 3A

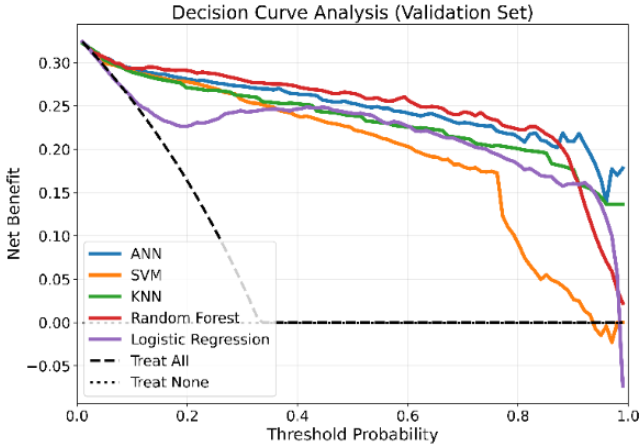


Fig. 3B

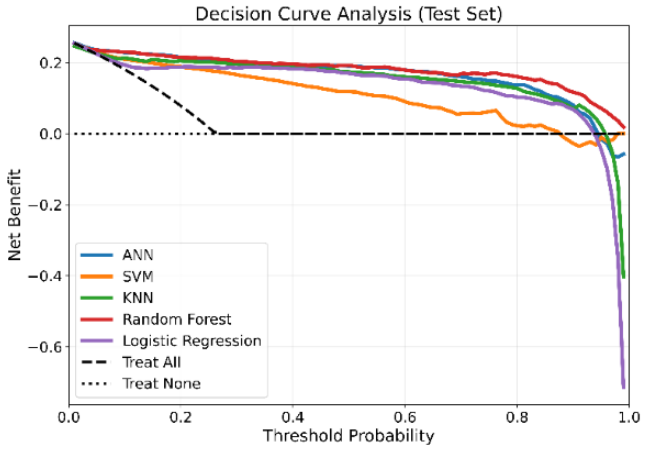


Fig. 3C

XSL•FO
RenderX

Figure 4. Model calibration analysis for the artificial neural network (ANN), random forest, k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression (LR) models. (A) Over the training set. (B) Over the validation set. (C) Over the testing set.

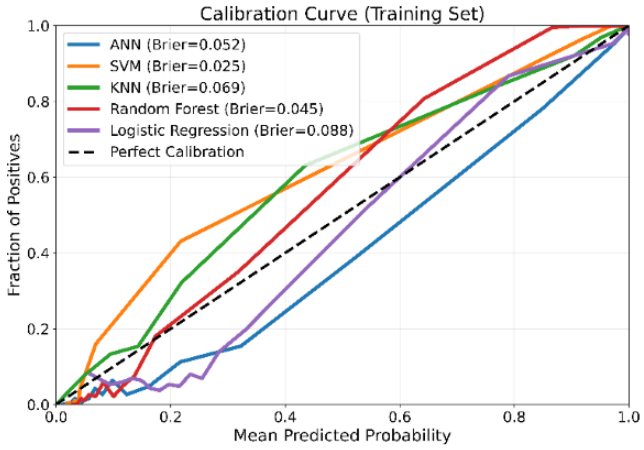


Fig. 4A

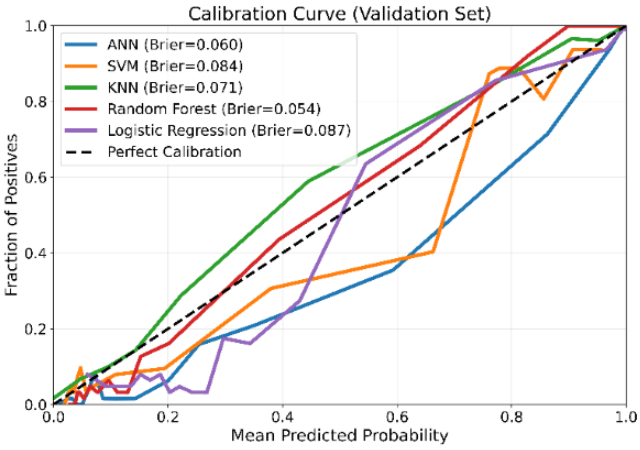


Fig. 4B

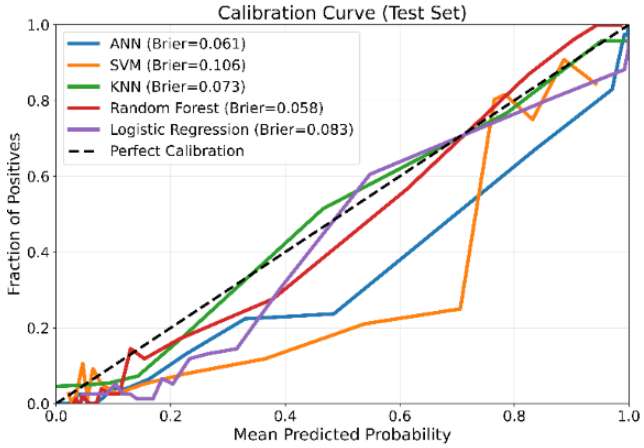


Fig. 4C

In the subgroup analysis, the ANN model achieved an AUC of 0.965 and an accuracy of 0.917 in male patients and an AUC of 0.970 and an accuracy of 0.936 in female patients. Similar consistent performances were observed across other subgroups. In patients with diabetes, the model achieved an AUC of 0.961 and an accuracy of 0.890, whereas in those without diabetes, the AUC and accuracy were 0.963 and 0.930, respectively. In patients with gout, the model reached an AUC of 1.000 and accuracy of 1.000, while in those without gout, the AUC was 0.967, and the accuracy was 0.926. For patients with bacteriuria, the AUC was 0.967, and the accuracy was 0.926, compared with 0.968 and 0.929 in those without bacteriuria. Among urine RBC subgroups, the AUC and accuracy were 0.915 and 0.942 for 0 to 2 RBCs, 0.936 and 0.866 for 3 to 5 RBCs, 0.950 and 0.908 for 6 to 10 RBCs, and 0.936 and 0.916 for 10 or more RBCs. By age category, the AUC and accuracy were 0.972 and 0.949 for patients aged 18 to 40 years, 0.956 and 0.923 for those aged 40 to 65 years, and 0.916 and 0.907 for those older than 65 years. For BMI subgroups, the model achieved 0.966 and 0.934 for BMI less than 25 kg/m^2, 0.961 and 0.918 for BMI of 25 to 30 kg/m^2, and 0.964 and 0.913 for BMI of 30 kg/m^2 or greater. Finally, among patients with eGFR less than 60, the AUC and accuracy were 0.992 and 0.981, compared with 0.953 and 0.921 for those with eGFR greater than 60.

Importance of Feature Variables

The results of the Shapley value analysis for the ANN model are shown in Figure 5A, which indicates that the predictive power of the model was predominantly driven by three features: urine RBC count, eGFR, and urine SG. An ANN model trained using only these 3 predictors (referred to as “ANN-3P” thereafter, as opposed to “ANN-10P” for the original 10-predictor ANN model) achieved validation and testing AUROCs of 0.954 and 0.943, and validation and testing AUPRCs of 0.938 and 0.891, respectively; the performance is competitive with that of the ANN-10P model. Figure 5B-E displays the ROC and PR curves of the 2 models, and a comprehensive performance comparison is summarized in Table 4. The outstanding performance of the ANN-3P model confirms the predictive power rendered by the 3 key feature variables.

XSL•FO
RenderX

Figure 5. Results of the Shapley value analysis and receiver operating characteristic (ROC), precision-recall curves of the artificial neural network (ANN) models. (A) Summary plot of the Shapley values. (B) Validation ROC curves. (C) Validation PR curves. (D) Testing ROC curves. (E) Testing PR curves. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; eGFR: estimated glomerular filtration rate; SG: specific gravity; SHAP: Shapley Additive Explanations; urineRBC: urine red blood cell.

XSL•FO
RenderX

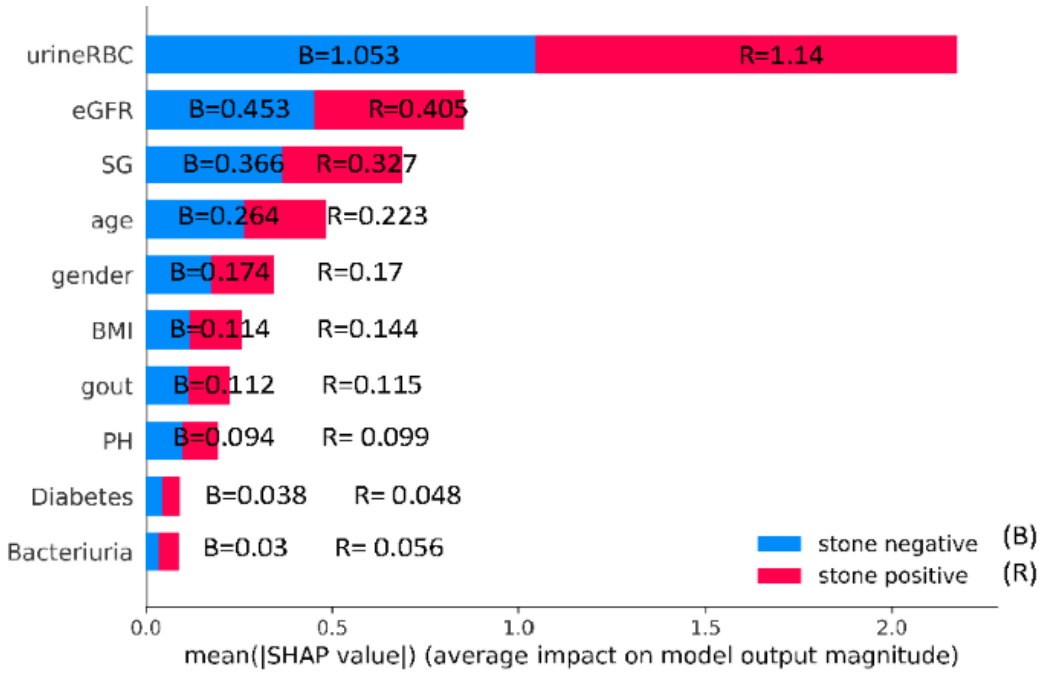


Fig. 5A

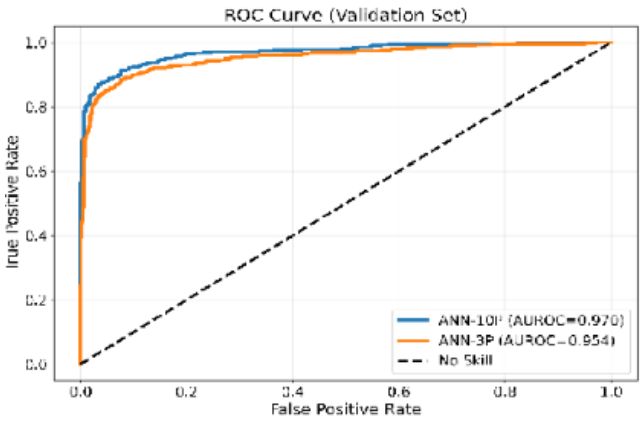


Fig. 5B

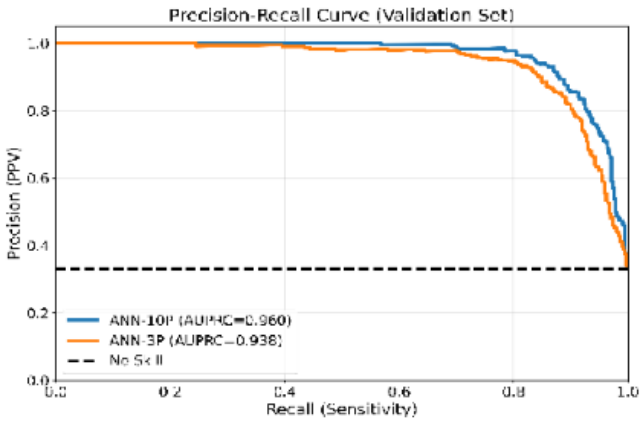


Fig. 5C

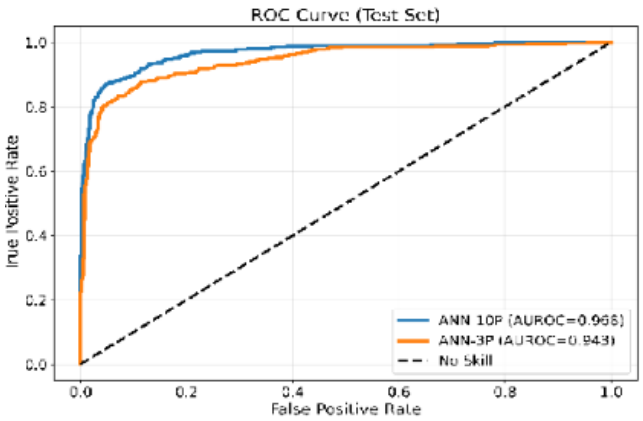


Fig. 5D

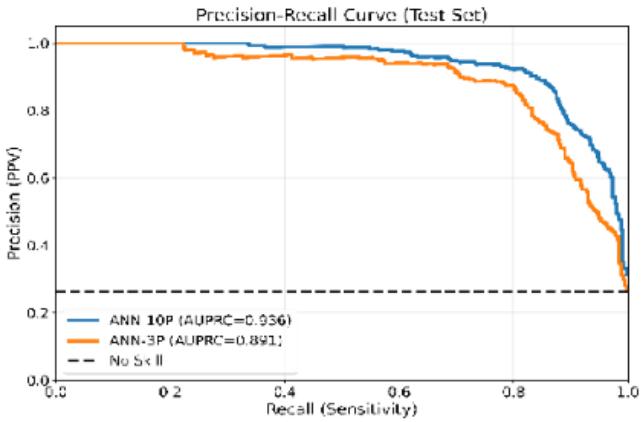


Fig.5E

XSL•FO
RenderX

Table . Performance comparison between artificial neural network (ANN)-3P and ANN-10P models.

Model and dataset	AUROC[a] (95% CI)	AUPRC[b] (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV[c] (95% CI)	NPV[d] (95% CI)
ANN-10P							
Training	0.974 (0.968 - 0.980)	0.964 (0.956 - 0.971)	0.943 (0.935 - 0.951)	0.907 (0.890 - 0.923)	0.961 (0.954 - 0.968)	0.920 (0.903 - 0.937)	0.954 (0.946 - 0.963)
Validation	0.970 (0.958 - 0.982)	0.960 (0.946 - 0.973)	0.923 (0.908 - 0.937)	0.889 (0.860 - 0.918)	0.939 (0.923 - 0.956)	0.878 (0.846 - 0.911)	0.945 (0.928 - 0.962)
Testing	0.968 (0.956 - 0.98)	0.936 (0.918 - 0.953)	0.927 (0.915 - 0.939)	0.873 (0.841 - 0.904)	0.947 (0.935 - 0.959)	0.853 (0.819 - 0.888)	0.954 (0.941 - 0.968)
ANN-3P							
Training	0.953 (0.945 - 0.960)	0.934 (0.924 - 0.943)	0.910 (0.901 - 0.919)	0.866 (0.846 - 0.885)	0.932 (0.922 - 0.942)	0.863 (0.843 - 0.883)	0.934 (0.923 - 0.943)
Validation	0.954 (0.906 - 0.945)	0.938 (0.921 - 0.954)	0.908 (0.892 - 0.924)	0.892 (0.864 - 0.918)	0.916 (0.899 - 0.933)	0.841 (0.809 - 0.872)	0.945 (0.930 - 0.959)
Testing	0.943 (0.927 - 0.956)	0.891 (0.869 - 0.913)	0.883 (0.867 - 0.899)	0.878 (0.845 - 0.909)	0.885 (0.866 - 0.903)	0.731 (0.691 - 0.771)	0.953 (0.940 - 0.965)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

Discussion

Principal Findings

Epidemiological studies indicate that approximately 10% of the global population faces a lifetime risk of developing kidney stones [1], and approximately 80% of those affected experience recurrence after stone treatment [15]. Although kidney stones are not malignant such as cancers, they can cause severe flank pain that often drives patients to seek emergency care [16]. In more serious cases, stones may lead to urinary tract infections, sepsis, and even hospitalization. Furthermore, recent studies have reported a strong association between nephrolithiasis and the development of chronic kidney disease [4]. Despite these risks, most kidney stones are only diagnosed after patients present with symptoms and undergo imaging-based evaluations in clinical settings [17]. Currently, there is no widely adopted, large-scale, low-cost, and simple screening tool available for the early detection of kidney stones [18,19].

In 2024, our team developed an AI-based model using routine health checkup data to screen for kidney stones in overweight and obese individuals. The model achieved excellent performance (AUC=0.96), and its advantages include low cost, rapid computation, minimal computational power requirements, and no need for physician involvement [6]. This enables its application in frontline health screenings to identify high-risk individuals for further evaluation. Building on the success of our previous work, this study extends the application of our AI framework to the general population. Our findings suggest the potential for a scalable, accessible, and noninvasive tool to detect clinically significant kidney stones across diverse patient groups, providing a much-needed advancement in the proactive management of urolithiasis.

In some regions, certain high-end health screening programs incorporate renal ultrasound (US), plain abdominal radiography (KUB), or even CT to detect asymptomatic kidney stones. While US is a fast, user-friendly, and radiation-free imaging modality, its accuracy is highly operator dependent. Studies comparing US with CT have reported widely varying sensitivity (24% - 69%) and specificity (53% - 90%) levels [20-22]. This variability limits its reliability, especially in large-scale screening, as it requires experienced physicians and access to ultrasound equipment, both of which are resource intensive. Plain x-rays or KUB, although useful in identifying some kidney stones, expose patients to radiation (0.7 mSv per scan) [20]. Plain abdominal x-rays (KUB) can identify some radiopaque stones but have limited sensitivity and expose patients to ionizing radiation (~0.7 mSv per scan) [22] . For women of reproductive age, pregnancy status must be considered before imaging. Furthermore, the interpretation of KUB relies on trained radiologists, making it less practical for community-based screening. Currently, abdominal CT is the gold standard for diagnosing kidney and ureteral stones. However, it has notable drawbacks, including significant radiation exposure (10 mSv for standard noncontrast CT and 1 - 3 mSv for low-dose CT) [20,23] and high medical cost [24]. These factors make CT unsuitable for large-scale or repeated screening, particularly in asymptomatic populations. Given the limitations of traditional imaging modalities in mass screening contexts, our AI-based approach offers a compelling alternative. Using only routine clinical parameters from standard health checkups, our model provides a rapid, low-cost, and accessible method for the early detection of kidney stones without the need

XSL•FO
RenderX

for imaging, radiation, or specialized personnel. This has significant implications for population-level screening and preventive nephrology.

To the best of our knowledge, this study is the first to use a large-scale dataset that integrates simple clinical information from both patients with and without kidney stones, applying ML techniques to develop a computer-aided screening tool for detecting stones of various compositions. This approach reflects real-world clinical scenarios and aims to support early detection and decision-making for further diagnostic evaluation. In this study, we expanded upon our previous work by incorporating 10 easily accessible clinical variables to construct an AI-based model capable of identifying individuals at high risk for kidney stones [6]. These variables, readily available during routine health checkups, emergency visits, or outpatient consultations, can be collected and analyzed quickly, with no additional burden on patients or health care resources.

One of the key advantages of this tool lies in its simplicity and cost-effectiveness. Unlike imaging-based diagnostics, which require expensive equipment and specialized interpretation, our AI model only needs a standard internet-connected computer and does not rely on urologists or radiologists for execution. This enhances its feasibility and scalability, particularly in resource-limited settings or rural areas. Conceptually, our AI screening model functions similarly to the fecal occult blood test used for colorectal cancer screening [25]—offering a rapid, noninvasive, and low-cost method to identify high-risk individuals who would then proceed to confirmatory imaging tests such as ultrasound, x-rays, or CT scans.

Furthermore, because our model relies on only 10 standard health checkup parameters, patients do not require any additional testing. This makes the tool highly compatible with existing clinical workflows. Given its accessibility and minimal operational requirements, it could be readily implemented in annual health screenings or even in the follow-up of patients who have undergone treatment for kidney stones [26,27]. By enabling earlier detection of stones—even when they are still small and asymptomatic—the model could facilitate timely intervention, reduce the risk of complications, and simplify surgical management.

Importantly, this approach also holds significant potential for use in telemedicine. By leveraging basic clinical data and our AI software, health care providers could remotely screen for kidney stones in patients living far from medical centers, enabling early identification and reducing the likelihood of emergency presentations. Overall, this screening tool may serve as a practical and scalable solution for improving access to care, minimizing disease burden, and promoting proactive management of urolithiasis across diverse health care settings.

Our AI model is designed as a practical screening tool to identify kidney stones using only routine blood and urine data collected during health checkups. In real-world workflows, it can be seamlessly integrated into existing laboratory information systems, allowing automated risk predictions once test results are available. For individuals with a high predicted probability of nephrolithiasis, clinicians can proceed with confirmatory imaging such as ultrasonography or CT for diagnosis, while those with low risk can safely avoid unnecessary imaging. This stepwise approach facilitates early detection and appropriate resource allocation. Furthermore, it should be noted that the algorithm only uses laboratory data from routine urinalysis and no additional (expensive) tests are required. As such, the extra cost for using our algorithm is minimal, making it highly affordable. Early identification enables timely lifestyle interventions or pharmacologic prevention, potentially reducing emergency visits and treatment costs associated with advanced stone disease.

The variables selected in our model were based on previously validated clinical factors strongly associated with nephrolithiasis. Although several of these factors showed significant differences (P<.001) between cases and controls, which may contribute to potential overestimation of model performance, we minimized this bias by applying stratified sampling and reweighting techniques during model training. Consistently high AUC and accuracy values across all subgroups further support the robustness and generalizability of our model.

As every participant in this study underwent imaging confirmation to verify the presence or absence of stones, conducting multiethnic and cross-national validation would require substantial resources and logistical coordination. Therefore, this work focuses on an Asian cohort as an initial step. Although our dataset was comprehensive—comprising a large number of patients from multiple hospitals—and we further validated the model using an independent testing cohort in a clinical trial setting, one important limitation must be acknowledged. All patients included in this study were of Asian ethnicity. As a result, the predictive performance of our AI model may be limited when applied to populations of different racial or ethnic backgrounds. Differences in genetics, diet, and environmental factors could potentially affect the clinical presentation and risk profiles of kidney stone disease. To reduce racial and environmental bias, variables known to vary with ethnicity or diet—such as urinary calcium, oxalate, citrate, or lifestyle factors—were excluded. Instead, our model incorporated clinically relevant indicators of stone risk, including hematuria, bacteriuria, and serum creatinine. These features are biologically plausible and less population dependent, suggesting that the model may be applicable across diverse groups, although future international validation is warranted to confirm its performance. Future research should focus on collecting and integrating data from diverse populations to further refine and recalibrate the model. This will be essential to enhance the generalizability and global applicability of the AI-based kidney stone screening tool.

Conclusions

In this multihospital study, we developed an ML-based model that accurately identifies clinically significant kidney stones using only 10 routine clinical and urine parameters. This low-cost, noninvasive tool enables large-scale, manpower-free screening and can be readily integrated into health checkups or telemedicine, offering a practical solution for early detection and proactive management of nephrolithiasis.

XSL•FO
RenderX

Funding

This work was supported in part by the National Science and Technology Council of Taiwan (NSTC 112-2221-E-037-006, NSTC 113-2221-E-037-006, NSTC 113-2222-E-110-007, and NSTC 114-2221-E-037-006) and Kaohsiung Medical University Hospital (KMUH113-R006 and KMUH114-4R60). The funders do not have any role in the writing of the manuscript or the decision to submit it for publication.

Data Availability

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

HWC: Conceptualization, methodology, data curation, funding acquisition, validation, writing – original draft, writing – review & editing
JTL: Conceptualization, funding acquisition, formal analysis, validation, visualization, writing – original draft, writing – review & editing
PSW: Data curation, methodology, software, formal analysis, validation, writing – review & editing
YCC: Project administration, resources, investigation, writing – review & editing
JYW: Resources, investigation, writing – review & editing
CIL: Resources, investigation, writing – review & editing
YHC: Resources, investigation, writing – review & editing
YSJ: Resources, investigation, writing – review & editing
WJW: Supervision, resources, investigation, writing – review & editing
CYK: Conceptualization, methodology, formal analysis, validation, writing – original draft, writing – review & editing

Conflicts of Interest

PSW reports serving as CEO and shareholder of United Kaohsiung AI Medical Technology Co. Ltd., which has been licensed the intellectual properties partly developed via his doctoral work at National Sun Yat-sen University. In this study, he was only involved in developing and training the AI model and did not participate in patient data collection, data labeling, or the independent validation of the testing cohort. These procedures were conducted by separate clinical teams to maintain objectivity and minimize potential bias. All related institutional and ICMJE conflict of interest policies were followed, and measures were taken to ensure objectivity, confidentiality, and transparency throughout the research and commercialization processes. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Details regarding the construction of the artificial neural network model.
[DOCX File, 30 KB - medinform_v14i1e80764_app1.docx]

References

1. Raheem OA, Khandwala YS, Sur RL, Ghani KR, Denstedt JD. Burden of urolithiasis: trends in prevalence, treatments, and costs. Eur Urol Focus 2017 Feb;3(1):18-26. [doi: 10.1016/j.euf.2017.04.001] [Medline: 28720363]
2. Roberson D, Sperling C, Shah A, Ziemba J. Economic considerations in the management of nephrolithiasis. Curr Urol Rep 2020 Mar 31;21(5):18. [doi: 10.1007/s11934-020-00971-6] [Medline: 32236700]
3. Abufaraj M, Xu T, Cao C, et al. Prevalence and trends in kidney stone among adults in the USA: analyses of National Health and Nutrition Examination Survey 2007–2018 data. Eur Urol Focus 2021 Nov;7(6):1468-1475. [doi: 10.1016/j.euf.2020.08.011] [Medline: 32900675]
4. Chuang TF, Hung HC, Li SF, Lee MW, Pai JY, Hung CT. Risk of chronic kidney disease in patients with kidney stones-a nationwide cohort study. BMC Nephrol 2020 Jul 22;21(1):292. [doi: 10.1186/s12882-020-01950-2] [Medline: 32698782]
5. Pearle MS, Calhoun EA, Curhan GC, Urologic Diseases of America Project. Urologic diseases in America project: urolithiasis. J Urol 2005 Mar;173(3):848-857. [doi: 10.1097/01.ju.0000152082.14384.d7] [Medline: 15711292]
6. Chen HW, Lee JT, Wei PS, et al. Machine learning models for screening clinically significant nephrolithiasis in overweight and obese populations. World J Urol 2024 Mar 9;42(1):128. [doi: 10.1007/s00345-024-04826-4] [Medline: 38460023]
7. Rippel CA, Nikkel L, Lin YK, et al. Residual fragments following ureteroscopic lithotripsy: incidence and predictors on postoperative computerized tomography. J Urol 2012 Dec;188(6):2246-2251. [doi: 10.1016/j.juro.2012.08.040] [Medline: 23083650]
8. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 5;150(9):604-612. [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]

XSL•FO
RenderX

9. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT press; 2016. URL: http://www.deeplearningbook.org [accessed 2026-01-23]
10. Barocas DA, Boorjian SA, Alvarez RD, et al. Microhematuria: AUA/SUFU guideline. J Urol 2020 Oct;204(4):778-786. [doi: 10.1097/JU.0000000000001297] [Medline: 32698717]
11. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text] [doi: 10.5555/1953048.2078195]
12. Kramer O. Machine Learning for Evolution Strategies: Springer; 2016. [doi: 10.1007/978-3-319-33383-0]
13. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J 2008 Jun;50(3):419-430. [doi: 10.1002/bimj.200710415] [Medline: 18435502]
14. Kamath U, Liu J. Explainable artificial intelligence: an introduction to interpretable machine learning: Springer; 2021. [doi: 10.1007/978-3-030-83356-5]
15. Fisang C, Anding R, Müller SC, Latz S, Laube N. Urolithiasis--an interdisciplinary diagnostic, therapeutic and secondary preventive challenge. Dtsch Arztebl Int 2015 Feb 6;112(6):83-91. [doi: 10.3238/arztebl.2015.0083] [Medline: 25721435]
16. Scales CD Jr, Lin L, Saigal CS, et al. Emergency department revisits for patients with kidney stones in California. Acad Emerg Med 2015 Apr;22(4):468-474. [doi: 10.1111/acem.12632] [Medline: 25779695]
17. Rule AD, Lieske JC, Pais VM Jr. Management of kidney stones in 2020. JAMA 2020 May 19;323(19):1961-1962. [doi: 10.1001/jama.2020.0662] [Medline: 32191284]
18. Ellison JS, Tasian GE. Engaging patients in clinical trials on kidney stones: lessons learned. Eur Urol Focus 2025 Sep;11(5):712-715. [doi: 10.1016/j.euf.2025.03.005] [Medline: 40113514]
19. Chen HW, Chen YC, Lee JT. Re: Jonathan S. Ellison, Gregory E. Tasian. Engaging patients in clinical trials on kidney stones: lessons learned Eur Urol Focus. In press. https://doi.org/10.1016/j.euf.2025.03.005. Eur Urol Focus 2025 May 23:S2405-4569(25)00120-8. [doi: 10.1016/j.euf.2025.04.034] [Medline: 40413156]
20. Fulgham PF, Assimos DG, Pearle MS, Preminger GM. Clinical effectiveness protocols for imaging in the management of ureteral calculous disease: AUA technology assessment. J Urol 2013 Apr;189(4):1203-1213. [doi: 10.1016/j.juro.2012.10.031] [Medline: 23085059]
21. Ulusan S, Koc Z, Tokmak N. Accuracy of sonography for detecting renal stone: comparison with CT. J Clin Ultrasound 2007 Jun;35(5):256-261. [doi: 10.1002/jcu.20347] [Medline: 17373690]
22. Fowler KAB, Locken JA, Duchesne JH, Williamson MR. US for detecting renal calculi with nonenhanced CT as a reference standard. Radiology 2002 Jan;222(1):109-113. [doi: 10.1148/radiol.2221010453] [Medline: 11756713]
23. Hamm M, Knopfle E, Wartenberg S, Wawroschek F, Weckermann D, Harzmann R. Low dose unenhanced helical computerized tomography for the evaluation of acute flank pain. J Urol 2002 Apr;167(4):1687-1691. [Medline: 11912388]
24. Quhal F, Seitz C. Guideline of the guidelines: urolithiasis. Curr Opin Urol 2021 Mar 1;31(2):125-129. [doi: 10.1097/MOU.0000000000000855] [Medline: 33399368]
25. Hewitson P, Glasziou P, Irwig L, Towler B, Watson E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. Cochrane Database Syst Rev 2007 Jan 24;2007(1):CD001216. [doi: 10.1002/14651858.CD001216.pub2] [Medline: 17253456]
26. Tzelves L, Geraghty R, Lombardo R, et al. Duration of follow-up and timing of discharge from imaging follow-up, in adult patients with urolithiasis after surgical or medical intervention: a systematic review and meta-analysis from the European Association of Urology Guideline Panel on urolithiasis. Eur Urol Focus 2023 Jan;9(1):188-198. [doi: 10.1016/j.euf.2022.06.016] [Medline: 35851252]
27. Tzelves L, Juliebø-Jones P, Skolarikos A, Somani BK. Imaging follow-up for patients with kidney stones: European Association of Urology versus American Urological Association guidelines. Eur Urol Focus 2025 Sep;11(5):733-735. [doi: 10.1016/j.euf.2025.03.015] [Medline: 40175244]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
AUC: area under the curve
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CT: computed tomography
eGFR: estimated glomerular filtration rate
HPF: high-power field
KUB: kidney, ureter, and bladder
ML: machine learning
RBC: red blood cell
RF: random forest
ROC: receiver operating characteristic curve

XSL•FO
RenderX

SG: specific gravity
US: ultrasound

Edited by A Benis; submitted 18.Jul.2025; peer-reviewed by M Jaradat, SB Duan; revised version received 27.Oct.2025; accepted 01.Jan.2026; published 19.Feb.2026.

Please cite as:
Chen HW, Wei PS, Chen YC, Wu JY, Lin CI, Chou YH, Juan YS, Wu WJ, Kao CY, Lee JT
Screening for Clinically Significant Nephrolithiasis Based on Simple Health Checkup Clinical and Urine Parameters in General Populations: Multicenter Machine Learning Study
JMIR Med Inform 2026;14:e80764
URL: https://medinform.jmir.org/2026/1/e80764
doi:10.2196/80764

© Hao-Wei Chen, Pei-Siou Wei, Yu-Chen Chen, Jeng-Yih Wu, Chia-I Lin, Yii-Her Chou, Yung-Shun Juan, Wen-Jeng Wu, Chung-Yao Kao, Jung-Ting Lee. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prediction of First and Multiple Antiretroviral Therapy Interruptions in People Living With HIV: Comparative Survival Analysis Using Cox and Explainable Machine Learning Models

Donald Salami[1], PhD; Emily Koech[2], MD; Janet M Turan[3], PhD; Kristen A Stafford[1], PhD; Lilly Muthoni Nyagah[4], PhD; Stephen Ohakanu[1], BEng; Anthony K Ngugi[5], PhD; Manhattan Charurat[1], Prof Dr

[1]Center for International Health, Education and Biosecurity, Institute of Human Virology, School of Medicine, University of Maryland, 725 W Lombard St, Baltimore, MD, United States

[2]Center for International Health, Education, and Biosecurity, Kenya, Nairobi, Kenya

[3]Department of Health Policy and Organization, School of Public Health, The University of Alabama, Birmingham, AL, United States

[4]Department of Strategic Information, Ministry of Health's National AIDS and STI Control Program Kenya, Nairobi, Kenya

[5]Department of Population Health, Aga Khan University Nairobi, Nairobi, Kenya

Corresponding Author:
Donald Salami, PhD
Center for International Health, Education and Biosecurity, Institute of Human Virology, School of Medicine, University of Maryland, 725 W Lombard St, Baltimore, MD, United States

Abstract

Background: The Cox proportional hazards (CPH) model is a common choice for analyzing time-to-treatment interruptions in patients on antiretroviral therapy (ART), valued for its straightforward interpretability and flexibility in handling time-dependent covariates. Machine learning (ML) models have increasingly been adapted for handling temporal data, with added advantages of handling complex, nonlinear relationships and large datasets, and providing clear practical interpretations.

Objective: This study aims to compare the predictive performance of the traditional CPH model and ML models in predicting treatment interruptions among patients on ART, while also providing both global and individual-level explanations to support personalized, data-driven interventions for improving treatment retention.

Methods: Using data from 621,115 patients who started ART between 2017 and 2023, in Kenya, we compared the performance of the CPH with the following ML models—gradient boosting machine, extreme gradient boosting, regularized generalized linear models (Ridge, Lasso, and Elastic-Net), and recursive partitioning—in predicting first and multiple treatment interruptions. Explainable surrogate technique (model-agnostic) was applied to interpret the best performing model's predictions globally, using variable importance and partial dependence profiles, and at individual level, using breakdown additive, Shapley Additive Explanations, and ceteris paribus.

Results: The recursive partitioning model achieved the best performance with a predictive concordance index score of 0.81 for first treatment interruptions and 0.89 for multiple interruptions, outperforming the CPH model, which scored 0.78 and 0.87 for the same scenarios, respectively. Recursive partitioning's performance can be attributed to its ability to model nonlinear relationships and automatically detect complex interactions. The global model-agnostic explanations aligned closely with the interpretations offered by hazard ratios in the CPH model, while offering additional insights into the impact of specific features on the model's predictions. The breakdown additive and Shapley Additive Explanations explainers demonstrated how different variables contribute to the predicted risk at the individual patient level. The ceteris paribus profiles further explored the time-varying model to illustrate how changes in a patient's covariates over time could impact their predicted risk of treatment interruption.

Conclusions: Our results highlight the superior predictive performance of ML models and their ability to provide patient-specific risk predictions and insights that can support targeted interventions to reduce treatment interruptions in ART care.

(JMIR Med Inform 2026;14:e78964) doi:10.2196/78964

KEYWORDS

antiretroviral therapy; explainable machine learning; machine learning; recursive partitioning; treatment interruptions.

XSL•FO
RenderX

Introduction

The Cox proportional hazards (CPH) model has long been a foundational tool for time-to-event analysis, widely used in health care research to examine treatment interruptions, including among patients on antiretroviral therapy (ART). The CPH model is a favorable choice due to its straightforward interpretability; its coefficients can be easily understood as hazard ratios, which quantify the relative effect of each covariate on the likelihood of an event, offering clear insights for clinical decision-making and risk assessment [1,2]. In addition, its flexibility in handling time-dependent covariates allows it to incorporate variables that change over time, providing a more accurate reflection of patient dynamics [3-5]. Our previous work [6] focused on using time-dependent covariates within the CPH model to enhance its application in modeling the complex patterns of multiple treatment interruptions in patients on ART, aiming to capture the nuanced and changing influences on treatment adherence.

Despite its popularity, the CPH model has limitations. A key disadvantage is its handling of high-dimensional data with many covariates, as it can be prone to overfitting and may yield unreliable results due to multicollinearity [4]. This risk is particularly pronounced in time-dependent CPH models, where changing covariates can amplify the potential for erroneous inferences [4]. In addition, as a semiparametric model, it does not fully specify the baseline hazard function, potentially limiting its ability to capture complex, nonlinear relationships [7]. These limitations highlight the need for more flexible modeling approaches, especially when working with complex health care datasets with nonlinear, time-varying, or high-dimensional characteristics.

Machine learning (ML) models are increasingly used in time-to-event analysis due to their ability to capture complex, nonlinear relationships at scale. In HIV and ART research, however, most survival ML applications such as random survival forests and gradient boosting primarily rely on baseline or time-invariant covariates, limiting their ability to reflect the dynamic nature of ART care. Although some methods, including recursive partitioning (RP), have been adapted to incorporate temporal structures and time-varying effects [8-10], support for fully time-varying covariates remains limited in many survival ML frameworks. RP provides a flexible alternative by modeling nonlinear effects and interactions with time-varying covariates, yet its application to large-scale, longitudinal ART cohorts has been limited. In addition, advances in model interpretability have made ML models more explainable, helping to bridge the gap between predictive accuracy and practical applicability. Through methods such as permutation importance, partial dependence plots, and Shapley Additive Explanations (SHAP) values [11-13], ML models now provide insights into variable importance and individual predictions, making them increasingly useful for clinical decision-making and patient-centered care.

The aim of this paper is first to compare the performance of traditional CPH model and ML models in predicting treatment interruptions among patients on ART. This comparison aims to evaluate whether ML models, with their flexibility in handling nonlinear relationships and time-dependent covariates, can provide more accurate and reliable predictions than the CPH model. Second, provide clear, intuitive explanations for the predictions at both the global level to understand overall patterns and key risk factors and the local level for individualized predictions for a patient. By combining predictive accuracy with interpretability, this study aims to support health care providers in identifying patients at high risk for treatment interruption and in implementing personalized interventions to enhance treatment retention.

Methods

Study Population and Data

The analysis used longitudinal, deidentified electronic medical record (EMR) data from patients initiated on ART between January 1, 2017, and November 30, 2023, across 2156 facilities in all 47 counties of Kenya. Eligible patients had at least 2 follow-up visits or drug pickups after ART initiation, were followed for more than 28 days postinitiation, and had returned to care if they had experienced an initial treatment interruption.

Data Preprocessing and Definitions

The preprocessing, cleaning, feature engineering, and covariate definitions are detailed in our previous work [6]. Treatment interruption was defined as no ART drug pickup and no clinical contact for greater than 28 days after the last expected contact [14]. Time-invariant covariates included age, gender, education, marital and employment status, noncommunicable disease, baseline regimen, and ART initiation year, while time-varying covariates were adherence, alcohol intake, World Health Organization stage, viral load, clinical stability assessment, regimen line, multimonth dispensing (MMD), differentiated service delivery (DSD) model, and prevention with positives package. Full list of covariates and description is included in Table S1 in Multimedia Appendix 1. A key preprocessing step included creating an “unknown” category for categorical variables with missing data. Engineered features included the COVID-19 pandemic period—dummy variable to indicate the COVID-19 pandemic; missed appointment—failure to attend a scheduled visit within 1-3 days of the scheduled date; defaulted appointment—failure to attend a scheduled visit within 4-28 days; and the duration of treatment interruption—average duration of any treatment interruption before a patient returned to care.

Analysis and Models

We performed 2 sets of modeling, that is, time-invariant model, for first-ever treatment interruption, and a time-varying model, for multiple interruptions. The data structure differs between the models: in a time-invariant model, covariates were fixed, with 1 row per patient containing the time to interruption and interruption status. In a time-varying model, covariates change over time, with multiple rows per patient representing different time intervals and multiple interruption statuses.

Our base reference model was the traditional CPH model. The time-invariant CPH assessed the effect of fixed covariates on time to first interruption, with an assumption that the hazard ratio remains constant over time, while the time-varying allowed

XSL•FO
RenderX

covariates to change over time, capturing the dynamic effects of these covariates on the hazard at different intervals.

Our choice of ML models was based on algorithms with the ability to fit time-varying covariates and their computational feasibility at scale. We evaluated 3 classes of algorithms: boosting methods for their efficiency and predictive performance, regularized shrinkage models for handling multicollinearity and improving generalization, and RP for its ability to capture nonlinear relationships and interactions while remaining interpretable. Although random survival forests are widely used in survival analysis, their computational demands and limited support for fully time-varying covariates constrained their applicability in this study. The trained models are as described:

- Gradient boosting machine (GBM): an ensemble learning method that builds a series of decision trees, where each new tree corrects the errors of the previous ones. It is powerful for time-invariant survival analysis due to its ability to handle complex, nonlinear relationships and improve predictive accuracy by minimizing loss iteratively.
- Extreme gradient boosting (XGBoost): an optimized version of GBM, known for its speed and performance, especially with large datasets. It is well suited for our analysis due to its efficiency, scalability, and ability to manage missing data, offering robust predictions in survival tasks. The Accelerated Failure Time model was trained for a time-varying dataset.
- Regularized generalized linear models: Extends traditional regression models with regularization techniques—Ridge (L2), Lasso (L1), and Elastic Net (a combination of both)—to prevent overfitting and handle multicollinearity. These models are beneficial in time-invariant survival analysis for selecting relevant covariates while improving model generalization, especially with high-dimensional data.
- Recursive partitioning (LTRCtrees/Rpart): Uses decision trees to model survival data, splitting the dataset based on covariates to form homogeneous subgroups. It is advantageous due to its interpretability and ability to handle nonlinear relationships without making parametric assumptions.

We initially trained each model using default hyperparameters as a baseline, followed by hyperparameter tuning through a 10-fold cross-validation over a grid of parameter combinations. The optimal set of hyperparameters was selected based on the highest Harrell's concordance index (C-index) score, reflecting the study's primary objective of ranking patients by treatment interruption risk. The optimal parameter sets were then used for final model training. All analysis and modeling were performed using R Statistical Software (version 4.4.0; R Foundation for Statistical Computing).

Model Evaluation

We evaluated model performance using an 80:20 train-test split (including the Cox models to ensure comparability). Likewise, a 10-fold cross-validation was performed on the train set, and models were trained on k-1 folds and validated on the remaining fold, repeating this process 10 times, with each fold serving as the validation set once.

Final model evaluation and comparison were performed using the C-index, a widely used metric for assessing the performance of survival models. The C-index measures the model's ability to correctly rank pairs of observations based on survival times, with values ranging from 0.5 (random prediction) to 1.0 (perfect ranking). We calculated the C-index for each model on both the training and test sets to evaluate predictive accuracy. This approach ensured that the models performed well during training and on validation data, providing a reliable measure of generalizability. Furthermore, to assess the sensitivity and uncertainty of the predictive performance measures, we conducted 200 bootstrap resampling iterations to estimate the IQR distribution of the scores for our best performing model.

Model Interpretability

To interpret our model predictions, we used the DALEX package in R [15] and applied 2 levels of model-agnostic explanation, that is, model level global explanations—using permutation variable importance and partial dependence plots, and individual-level explanations—using variable attribution—breakdown and SHAP values and ceteris paribus (CP) analysis.

Global-Level Explanations

The global-level explanations are listed as follows:

1. Variable importance: Using a permutation-based approach, we assessed how removing the effect of selected variables through multiple permutations impacts model performance.
2. Partial dependence profiles: We explored the effects of selected variables by examining how the model's predicted values change as a function of individual variables, based on randomly selected observations.

Individual-Level Explanations

The individual-level explanations are listed as follows:

1. Variable attributions: This explains how a model's prediction for a single observation differs from the average prediction and distributes the difference among explanatory variables. We applied two methods: (1) Breakdown additive attribution, a procedure that decomposes the model's prediction into contributions from each explanatory variable. (2) SHAP, the averaging of variable attribution values over all (or several) possible orderings to show their individual impacts.
2. CP profiles: This explores how the model's prediction changes when the value of a single explanatory variable is altered, while keeping all other variables constant.

Ethical Considerations

This study was reviewed and approved as nonhuman subject research by the institutional review boards of the University of Maryland, Baltimore (HP-00108126), and the Aga Khan University Nairobi (2023/ISERC-94 (v2)). The analysis used secondary EMR data that were fully deidentified prior to receipt by the study team in accordance with Kenya's Data Protection

XSL•FO
RenderX

Act and institutional guidelines; therefore, written informed consent was not required.

Results

Our dataset comprised 621,115 unique patients on ART, with 432,041 first treatment interruption events and 959,170 multiple treatment interruption events. Data were split into an 80:20 ratio for the training and test sets, respectively. The hyperparameters for each model, along with the final parameter combinations used in the final model, are shown in Table 1. The predictive performance of ML models and their comparison with the CPH are shown in Table 2 and Figure 1, summarized by the estimated C-index scores and 95% CIs for both the train and test sets.

XSL•FO
RenderX

Table . Hyperparameters for the machine learning models.

Model (R package) and parameter	Chosen value (time-invariant model)	Chosen value (time-varying model)
Cox proportional hazards (survival)		
N/A[a]	N/A	N/A
Gradient boosting machine (gbm3)		
distribution	coxPH (ties = Efron)	coxph (ties = Efron)
n.trees	100	3000
interaction.depth	3	3
n.minobsinnode	5	20
shrinkage	0.01	0.001
bag.fraction	1	0.5
cv.folds	10	10
id	Unique value of patient ID	Unique value of patient ID
Extreme gradient boosting (xgboost)		
Objective	survival:cox	survival:aft—accelerated failure time
eval_metric	cox-nloglik	aft-nloglik
max_depth	5	6
eta	0.2	0.3
colsample_bytree	0.8	1
min_child_weight	5	1
subsample	1	1
n_folds	10	10
best nrounds	99	N/A
aft_loss_distribution	N/A	normal
aft_loss_distribution_scale	N/A	1
Regularized generalized linear models (glmnet)		
lambda	Length of 100, ranging logarithmically from 10^{10} to 10^{-2}. With minimum lambda = c (Ridge: 0.01, Lasso: 0.023, Elastic Net: 0.01748)	Length of 100, ranging logarithmically from 10^{10} to 10^{-2}. With minimum lambda=c (Ridge: 0.01, Lasso: 0.01, Elastic Net: 0.01)
type.measure	"C"—Harrel's concordance measure	"C"—Harrel's concordance measure
nfolds	10	10
maxit	1.00E+05	1.00E+05
alpha	c (0, 0.5, 1) where 0—Ridge penalty, L2 regularization only[b]. 0.5—Elastic-Net, L1 and L2 regularization[c]. 1—Lasso, L1 regularization only[d].	c (0, 0.5, 1) where 0—Ridge penalty, L2 regularization only[b]. 0.5—Elastic-Net, L1 and L2 regularization[c]. 1—Lasso, L1 regularization only[d].
Recursive partition (LTRCtrees/rpart)		
complexity parameter	0.0001	0.0214
xval (cross validations)	10	10
maxsurrogate	5	0
maxdepth	20	30
minsplit	20	10
maxcompete	4	4

XSL•FO
RenderX

[a]N/A: not applicable.

[b]Ridge—applies L2 regularization which penalizes the sum of the squared coefficients and is effective in cases where multicollinearity is present in the predictor variables.

[c]Elastic-Net regression—applies a combination of L1 and L2 regularization.

[d]Lasso—applies L1 regularization which can shrink some coefficients to zero, effectively performing feature selection.

Figure 1. Models' comparison: C-Index with 95% CIs. (A) C-Index comparison for the time-invariant models. (B) C-Index comparison for the time-varying models.

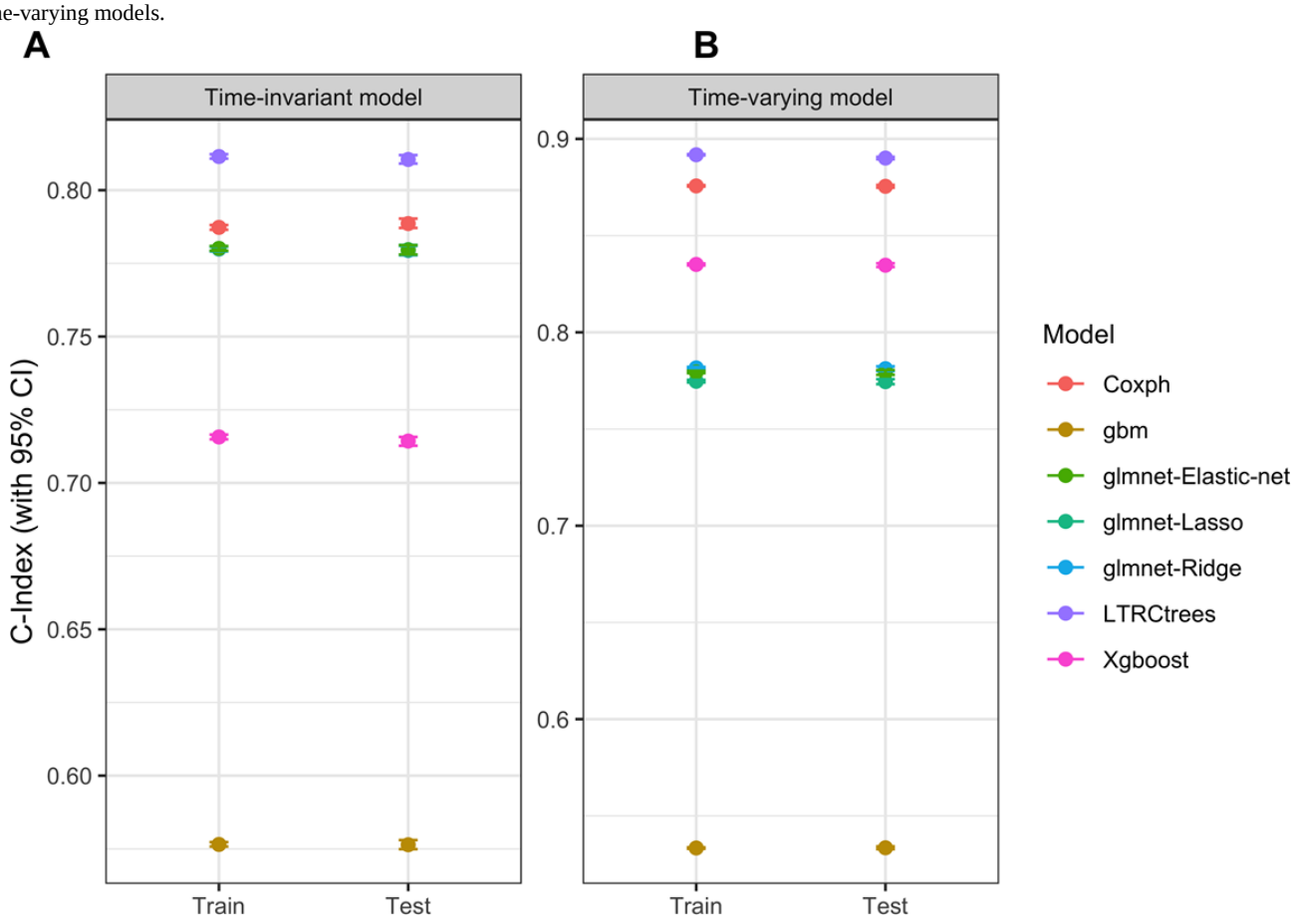


Table . Models' comparison of the time-invariant and time-varying models, using C-index scores for train and test sets.

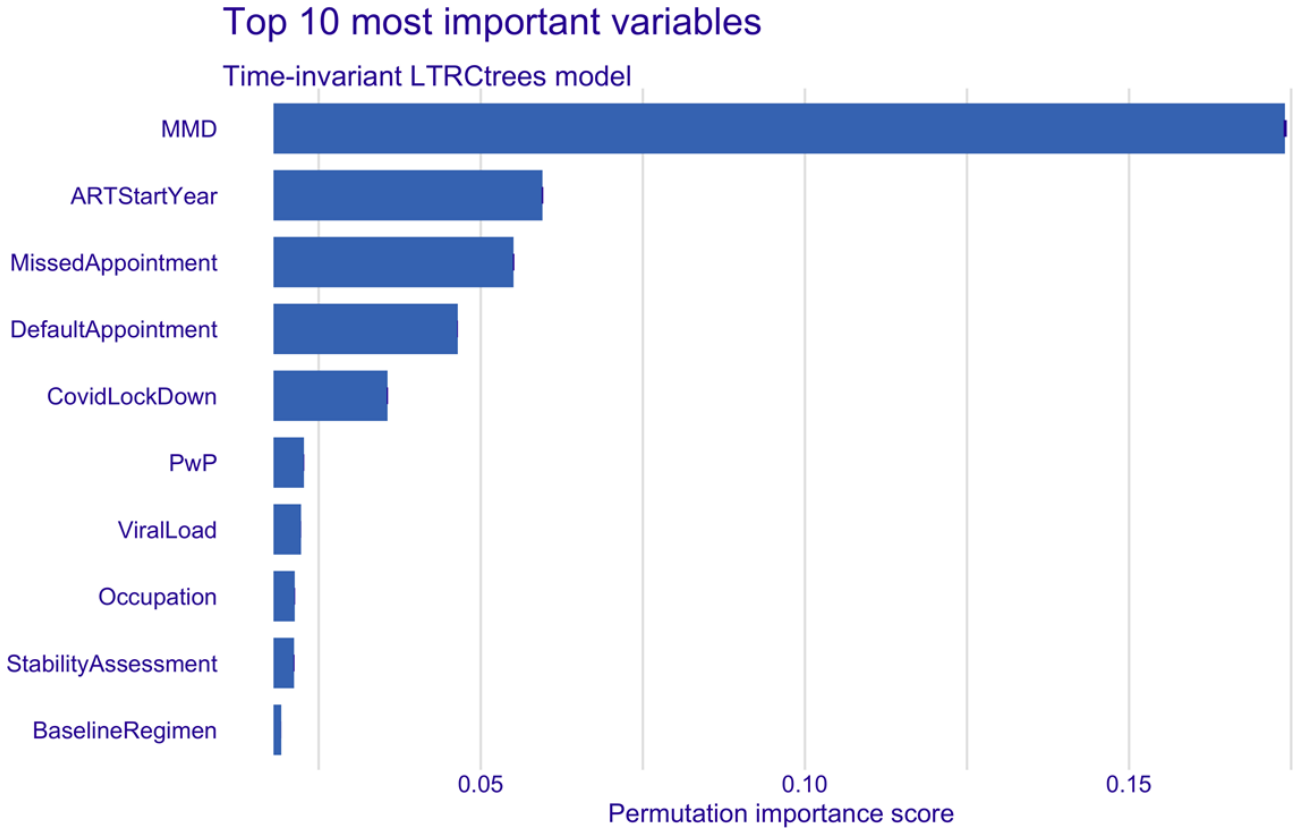
	Time-invariant model		Time-varying model	
Model	C-Index—train (95% CI)	C-Index—test (95% CI)	C-Index—train (95% CI)	C-Index—test (95% CI)
Coxph	0.7873 (0.7865 - 0.7881)	0.7886 (0.7871 - 0.7903)	0.8757 (0.8754 - 0.8761)	0.8755 (0.8748 - 0.8762)
glmnet—Lasso	0.7801 (0.7793 - 0.7809)	0.7797 (0.7780 - 0.7813)	0.7747 (0.7741 - 0.7754)	0.7745 (0.7732 - 0.7758)
glmnet—Ridge	0.7799 (0.7791 - 0.7807)	0.7794 (0.7778 - 0.7810)	0.7816 (0.7810 - 0.7821)	0.7812 (0.7800 - 0.7824)
glmnet—Elastic-net	0.7801 (0.7793 - 0.7809)	0.7796 (0.7780 - 0.7812)	0.7795 (0.7789 - 0.7801)	0.7793 (0.7780 - 0.7805)
gbm	0.5765 (0.5758 - 0.5773)	0.5764 (0.5749 - 0.5780)	0.5334 (0.5330 - 0.5337)	0.5335 (0.5328 - 0.5342)
Xgboost	0.7157 (0.7149 - 0.7165)	0.7143 (0.7127 - 0.7157)	0.8351 (0.8346 - 0.8356)	0.8347 (0.8338 - 0.8357)
LTRCtrees	0.8115 (0.8108 - 0.8123)	0.8105 (0.8091 - 0.8120)	0.8918 (0.8915 - 0.8921)	0.8901 (0.8895 - 0.8907)

In the time-invariant analysis, the glmnet models—Lasso, Ridge, and Elastic-Net—produced comparable results to the CPH model, with train C-index values ranging from 0.7799 to 0.7801 and test C-index values ranging from 0.7794 to 0.7796. XGBoost performed slightly worse than the CPH model, while the GBM model had the lowest performance. The LTRCtrees model outperformed all other ML models and the CPH model with a C-index value of 0.8115 for the train set and 0.8105 for the test set. The LTRCtrees model remained the top performer in the time-varying analysis with a C-index of 0.8918 for the train set and 0.8901 for the test set. XGBoost was the next best performing ML model and slightly comparable with the CPH model. We further evaluated the best-performing model (ie, LTRCtrees), using a 200-bootstrap resamples; the C-index interquartile range remained stable across resamples.

XSL•FO
RenderX

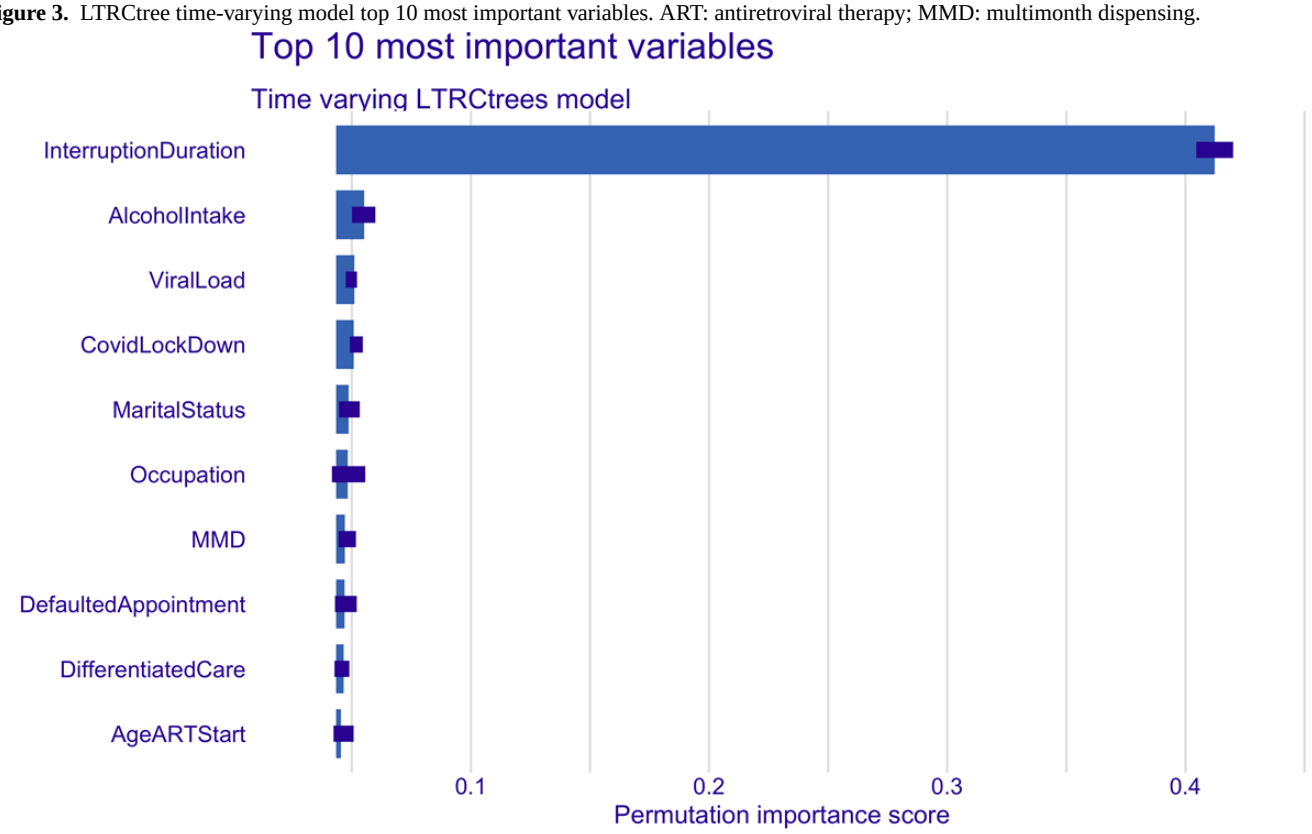
Figures 2 and 3 show the permutation-based variable importance for the top 10 variables in both the time-invariant and time-varying models. MMD emerged as the most important variable in the time-invariant model, while the duration of treatment interruption was most important in the time-varying model. In addition, variables such as viral load, ART start year, stability assessment, baseline regimen, and occupation were consistently ranked among the top 10 most influential variables in both models. The permutation-based variable importance results were similar to the ranking of the absolute values of the coefficients of the CPH models (Table 3). MMD, missed and defaulted appointments, year of ART initiation, occupation, and stability assessment consistently ranked among the top 10 variables in both the time-invariant LTRCtree and CPH models. In the time-varying models, duration of interruption, MMD, year of ART initiation, baseline regimen, and occupation were consistently top influencers in both LTRCtree and CPH models.

Figure 2. LTRCtree time-invariant model top 10 most important variables. ART: antiretroviral therapy; MMD: multimonth dispensing; PwP: prevention with positives package.



XSL•FO
RenderX

Figure 3. LTRCtree time-varying model top 10 most important variables. ART: antiretroviral therapy; MMD: multimonth dispensing.



XSL•FO
RenderX

Table . Hazard ratio of the top 10 variables from the time-invariant and time-varying Cox proportional hazards models.

Covariates	Time invariant—first treatment interruption N=497,208; events=346,118[a]		Time varying—multiple treatment interruption N=7,352,030; events=842,500	
	Hazard ratio (95% CI)[b]	*P* value	Hazard ratio (95% CI)[c]	*P* value
Marital status				
Other	Reference category		Reference category	
Cohabiting	1.02 (0.97-1.07)	.39	1.02 (0.99-1.05)	.26
Divorced	1.00 (0.95-1.04)	.91	0.98 (0.95-1.01)	.21
Married, monogamous	1.00 (0.95-1.05)	.89	0.99 (0.96-1.02)	.45
Married, polygamous	1.07 (1.02-1.12)	.01	0.86 (0.84-0.89)	<.001
Separated	1.10 (1.03-1.18)	<.001	1.18 (1.13-1.23)	<.001
Single	1.06 (1.01-1.11)	.02	1.01 (0.98-1.04)	.45
Widowed	1.07 (1.02-1.13)	<.001	1.00 (0.97-1.03)	.91
Employment status				
Employed	Reference category		Reference category	
Unemployed	1.00 (0.99-1.01)	.72	1.00 (0.99-1.01)	.75
Unknown	1.34 (1.33-1.36)	<.001	1.24 (1.23-1.25)	<.001
Adherence				
Good	Reference category		Reference category	
Fair	1.26 (1.23-1.29)	<.001	0.97 (0.95-0.99)	<.001
Bad	1.81 (1.69-1.94)	<.001	1.09 (1.03-1.15)	<.001
Unknown	1.13 (1.10-1.17)	<.001	0.95 (0.93-0.97)	<.001
Alcohol intake				
Never	Reference category		Reference category	
Monthly or less	0.93 (0.89-0.97)	<.001	0.98 (0.95-1.00)	.08
2-4 times a month	0.95 (0.89-1.02)	.13	1.02 (0.98-1.07)	.26
2-3 times a week	0.99 (0.94-1.05)	.85	1.02 (0.98-1.05)	.36
4 or more times a week	0.96 (0.91-1.01)	.09	1.01 (0.98-1.04)	.59
Unknown	0.99 (0.95-1.03)	.53	1.18 (1.15-1.21)	<.001
Viral load				
<200	Reference category		Reference category	
200 - 999	1.10 (1.09-1.11)	<.001	1.00 (1.00-1.01)	.27
>1000	1.14 (1.12-1.16)	<.001	1.10 (1.09-1.11)	<.001
Unknown	0.82 (0.81-0.83)	<.001	1.11 (1.10-1.11)	<.001
Clinical stability assessment				
Stable	Reference category		Reference category	
Unstable	1.31 (1.30-1.32)	<.001	1.09 (1.08-1.09)	<.001
Baseline ART[d] regimen				
Other	Reference category		Reference category	
2NRTIs[e] + Boosted PI[f].	0.85 (0.83-0.87)	<.001	0.89 (0.88-0.91)	<.001
2NRTIs + INSTI[g].	1.11 (1.08-1.13)	<.001	1.02 (1.01-1.04)	<.001
2NRTIs + NNRTI[h]	1.59 (0.79-3.17)	.19	0.99 (0.71-1.39)	.97
NRTI + INSTI	1.14 (1.04-1.25)	.01	1.30 (1.24-1.36)	<.001

XSL•FO
RenderX

Covariates	Time invariant—first treatment interruption N=497,208; events=346,118[a]		Time varying—multiple treatment interruption N=7,352,030; events=842,500	
	Hazard ratio (95% CI)[b]	*P* value	Hazard ratio (95% CI)[c]	*P* value
Multimonth dispensing (months)				
1	Reference category		Reference category	
2	0.57 (0.57-0.58)	<.001	0.75 (0.75-0.76)	<.001
3 - 5	0.28 (0.28-0.28)	<.001	0.46 (0.46-0.46)	<.001
6	0.15 (0.14-0.15)	<.001	0.27 (0.27-0.28)	<.001
6+	0.04 (0.04-0.05)	<.001	0.22 (0.21-0.22)	<.001
Prevention with positives package				
0	Reference category		Reference category	
1	0.86 (0.85-0.86)	<.001	0.93 (0.92-0.94)	<.001
2	0.61 (0.60-0.62)	<.001	0.85 (0.84-0.86)	<.001
3 or more	0.59 (0.59-0.60)	<.001	0.85 (0.84-0.85)	<.001
ART initiation year				
2017	Reference category		Reference category	
2018	1.21 (1.20-1.22)	<.001	1.20 (1.19-1.21)	<.001
2019	1.50 (1.48-1.52)	<.001	1.48 (1.47-1.49)	<.001
2020	1.68 (1.66-1.71)	<.001	1.83 (1.82-1.85)	<.001
2021	2.02 (1.99-2.05)	<.001	2.22 (2.20-2.25)	<.001
2022	2.08 (2.05-2.12)	<.001	2.64 (2.61-2.68)	<.001
2023	0.63 (0.61-0.65)	<.001	3.51 (3.42, 3.60)	<.001
COVID-19 pandemic period				
Non–COVID-19 period	Reference category		Reference category	
COVID-19 period	1.17 (1.16-1.18)	<.001	0.92 (0.91-0.92)	<.001
Prior missed appointment				
No	Reference category		N/A[i]	N/A
Yes	0.35 (0.34-0.35)	<.001	N/A	N/A
Prior default appointment				
No	Reference category		N/A	N/A
Yes	0.52 (0.51-0.53)	<.001	N/A	N/A
Missed appointments				
0	N/A	N/A	Reference category	
1-3	N/A	N/A	0.93 (0.93-0.94)	<.001
4-6	N/A	N/A	0.81 (0.80-0.82)	<.001
7+	N/A	N/A	0.73 (0.72-0.75)	<.001
Default appointments				
0	N/A	N/A	Reference category	
1-3	N/A	N/A	1.00 (1.00-1.01)	.1
4-6	N/A	N/A	0.97 (0.96-0.98)	<.001
7+	N/A	N/A	0.95 (0.94-0.96)	<.001

XSL•FO
RenderX

Covariates	Time invariant—first treatment interruption N=497,208; events=346,118[a]		Time varying—multiple treatment interruption N=7,352,030; events=842,500	
	Hazard ratio (95% CI)[b]	*P* value	Hazard ratio (95% CI)[c]	*P* value
Duration of treatment interruptions				
0 (No interruptions)	N/A	N/A	Reference category	
<30 days	N/A	N/A	29.94 (29.63-30.26)	<.001
30-197 days	N/A	N/A	19.63 (19.51-19.75)	<.001
180-364 days	N/A	N/A	8.15 (8.08-8.23)	<.001
365 days or more	N/A	N/A	2.85 (2.82-2.88)	<.001

[a]N=number of observations in the train dataset, that is, 80% of the entire dataset; events=number of treatment interruptions.

[b]Concordance=0.79 (se=0).

[c]Concordance=0.88 (se=0).

[d]ART: antiretroviral therapy.

[e]NRTIs: nucleoside analog reverse transcriptase inhibitors.

[f]PI: protease inhibitor.

[g]INSTI: integrase strand transfer inhibitor.

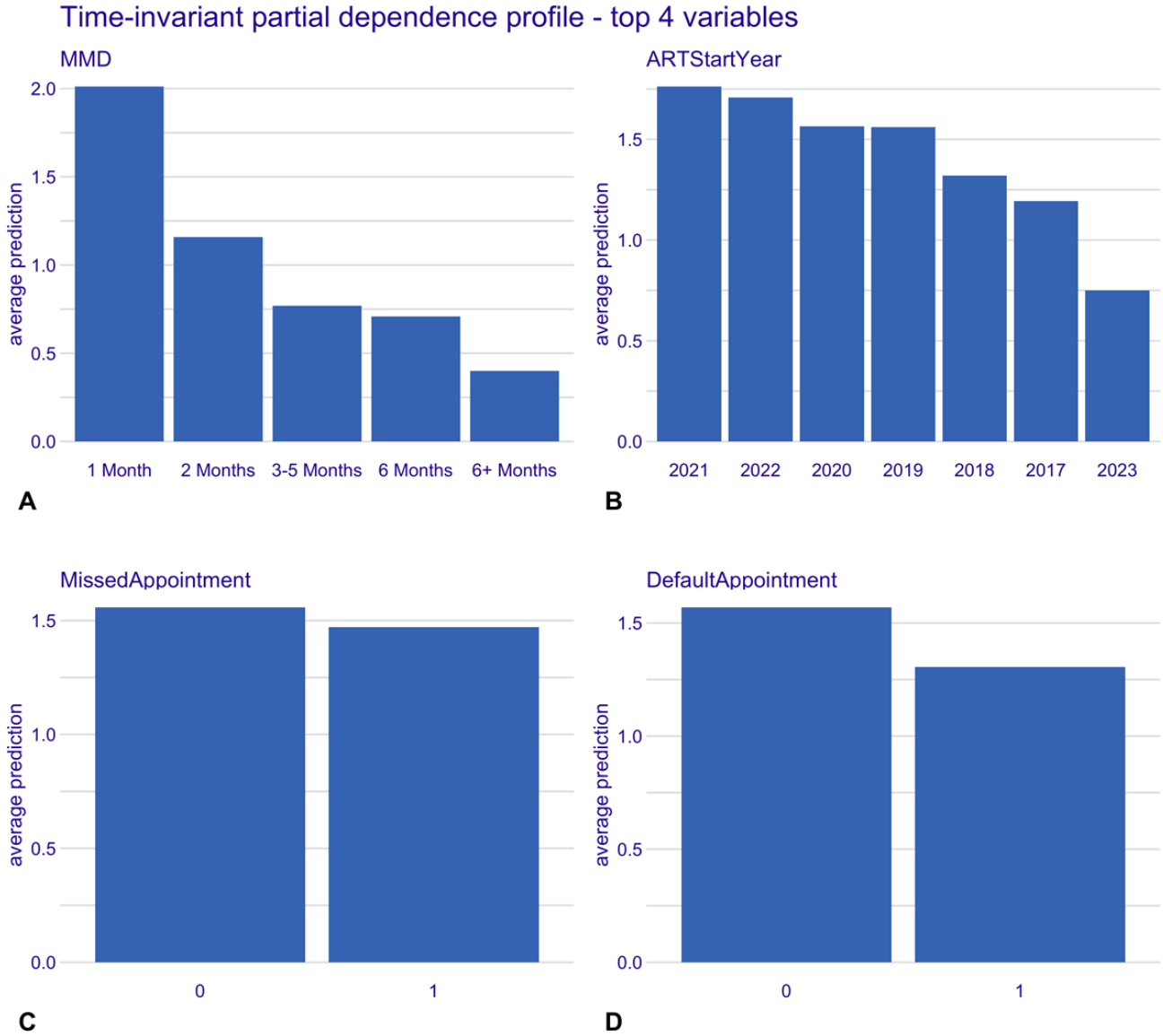
[h]NNRTI: nonnucleoside analog reverse transcriptase inhibitor.

[i]N/A: not applicable.

The full results of the time-invariant and time-varying CPH models are included in Table S2 in Multimedia Appendix 1. The partial dependence plots for the top 4 variables from both models are displayed in Figures 4 and 5. These plots show the average predicted risk for each of these variables. Short dispensing interval (MMD of 1 month), initiating ART in 2021, and having not missed or defaulted on an appointment were associated with higher predicted risks of first treatment interruption. Conversely, longer dispensing intervals (6+ months MMD), initiating ART in 2023, and having missed or defaulted on an appointment had a lower predicted risk of first treatment interruption. Short interruption durations (<30 days), unknown viral load, separated marital status, and unspecified alcohol use were all associated with higher predicted risk of multiple treatment interruption, while having no prior interruptions (ie, interruption duration of zero), lower viral load (<200), and no alcohol consumption showed lower predicted risk of multiple treatment interruption.

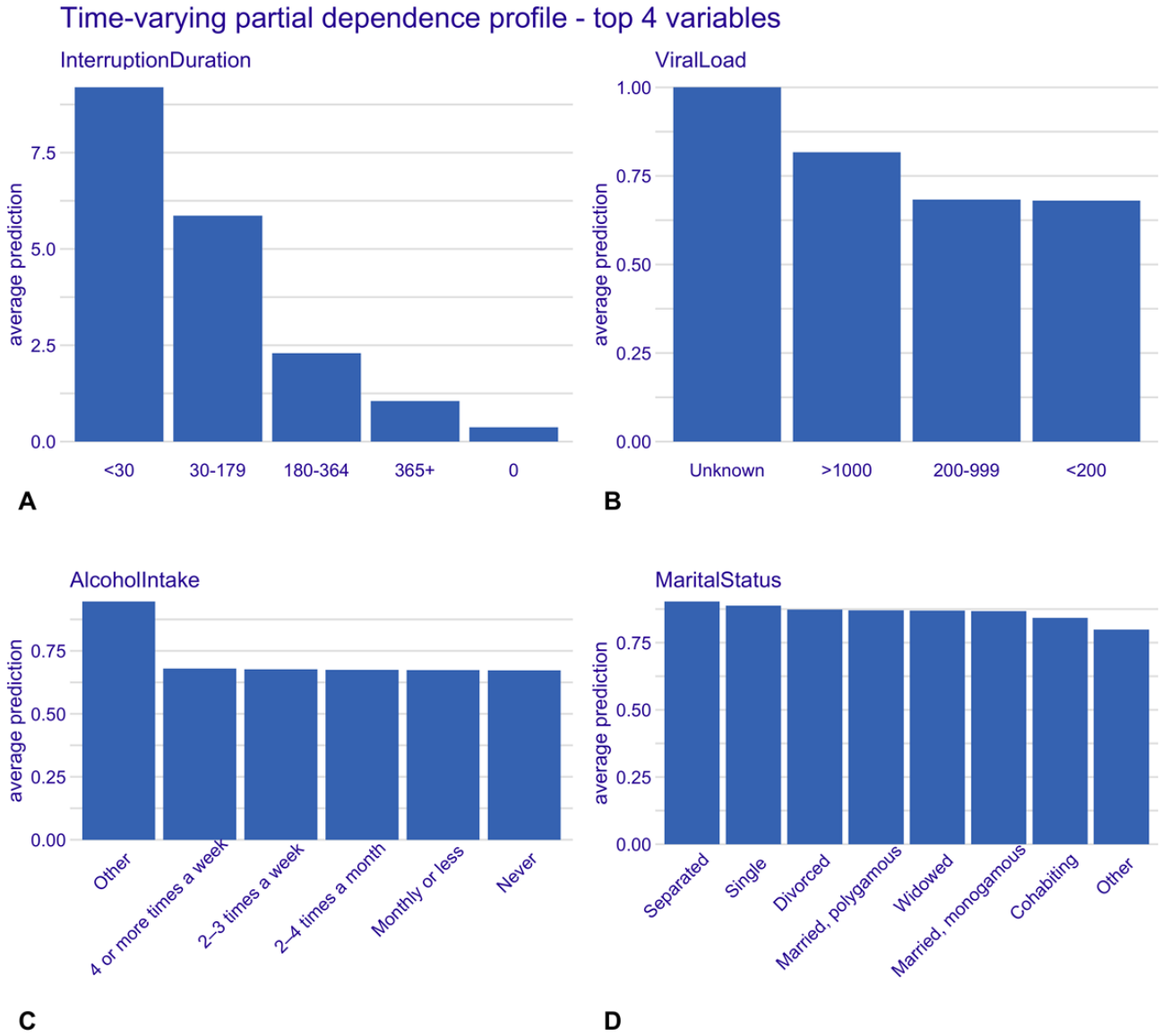
XSL•FO
RenderX

Figure 4. Time-invariant model; partial dependence plots for top 4 variables. Panels show the marginal effect of each variable on the average predicted risk of treatment interruption. (A) Multimonth dispensing duration, (B) ART initiation year, (C) missed appointment, and (D) defaulted appointment. ART: antiretroviral therapy; MMD: multimonth dispensing.



XSL•FO
RenderX

Figure 5. Time-varying model; partial dependence plots for top 4 variables. Panels show the marginal effect of each variable on the average predicted risk of treatment interruption. (A) Treatment interruption duration, (B) viral load count, (C) alcohol use, and (D) marital status.



To examine individual-level predictions for the time-invariant model, we randomly selected 2 observations from the dataset: one representing a patient (labeled "patient A") who experienced a first treatment interruption, and another representing a patient (labeled "patient B") who did not experience an interruption by the end of the follow-up period.

The row labeled "intercept" represents the overall average prediction value (1.516) for the model—baseline prediction (Figure 6A). The subsequent rows show how the mean prediction changes when a specific covariate is fixed. Green bars indicate an increase in the risk of treatment interruption, while red bars show a decrease in the risk. The final row (prediction), shown with a blue bar, displays the overall mean value and change, giving the final predicted risk of treatment interruption for the patient. Patient A had a high predicted risk of treatment interruption, with a risk score of 2.051, which is 0.535 points higher than the model's average prediction (Figure 6). Covariates influencing this prediction are short MMD intervals (1 month), viral load of 200 - 999, did not receive a prevention with positive package, not defaulted on appointment, non–COVID-19 pandemic period, aged 25 - 34 years at the start of ART, and baseline regimen of 2NRTIs + NNRTI. Patient B, on the other hand, had a low predicted risk (0) of interruption, largely driven by the long MMD interval (6+ months), enrolled in the DSD fast-track model, and had prior missed and defaulted appointments contributing to a reduced risk (Figure 7).

XSL•FO
RenderX

Figure 6. Time-invariant model breakdown and SHAP attribution plots for patient A. **(A)** Breakdown plot decomposing the model prediction into cumulative covariate contributions. Intercept—model's overall average prediction value; green bars, with (+) sign indicate increased risk of treatment interruption; red bars, with (−) sign indicate decreased risk of interruption; and blue bar, "prediction" is the individual predicted risk of interruption for the patient. **(B)** SHAP plot summarizing the direction and magnitude of individual covariate effects on the predicted risk of treatment interruption. Note: variable labels shown in the plots represent categorical levels of the corresponding covariates; see Table S1 in Multimedia Appendix 1 for full description. ART: antiretroviral therapy; MMD: multimonth dispensing; NNRTI: nonnucleoside analog reverse transcriptase inhibitor; NRTIs: nucleoside analog reverse transcriptase inhibitors; PwP: prevention with positives package; SHAP: Shapley Additive Explanations.

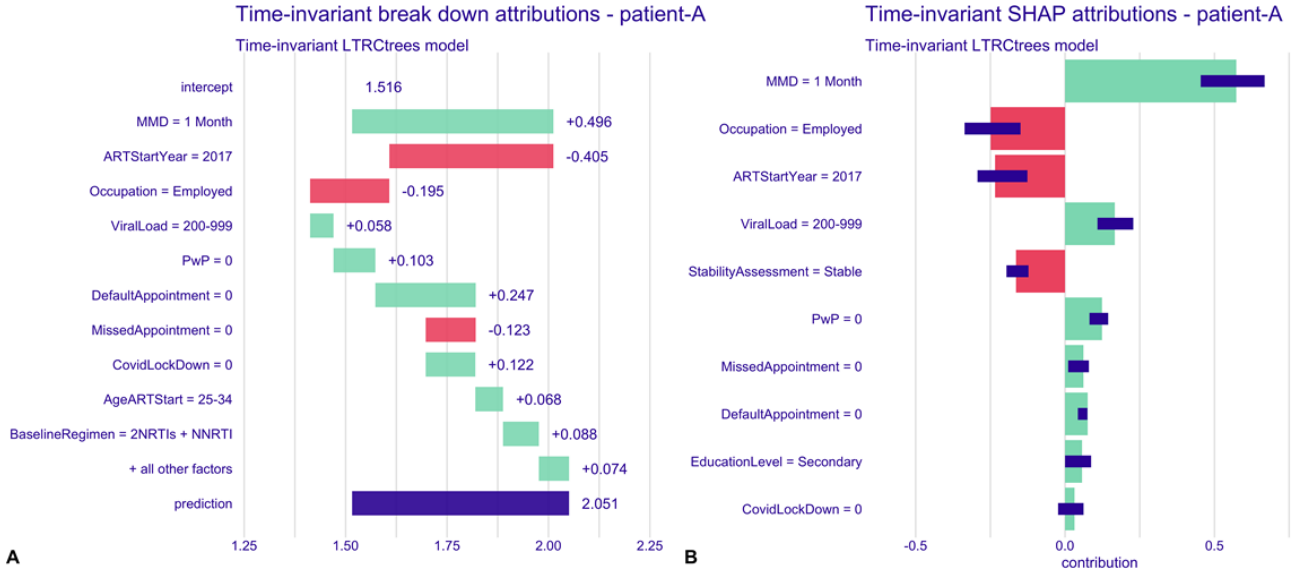
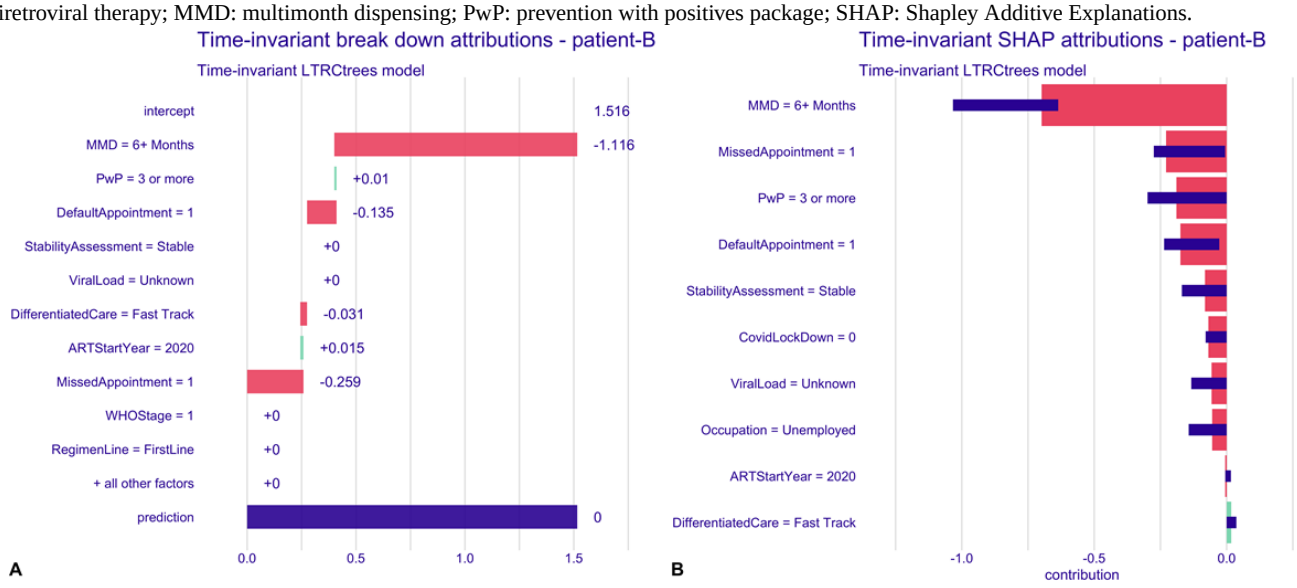


Figure 7. Time-invariant model breakdown and SHAP attribution plots for patient B. **(A)** Breakdown plot decomposing the model prediction into cumulative covariate contributions. Intercept—model's overall average prediction value; green bars, with (+) sign indicate increased risk of treatment interruption; red bars, with (−) sign indicate decreased risk of interruption; and blue bar, "prediction" is the individual predicted risk of interruption for the patient. **(B)** SHAP plot summarizing the direction and magnitude of individual covariate effects on the predicted risk of treatment interruption. ART: antiretroviral therapy; MMD: multimonth dispensing; PwP: prevention with positives package; SHAP: Shapley Additive Explanations.



The right plots of both Figures 6 and 7 show the SHAP plots for the patients, where the SHAP values represent the computed mean attribution across 25 random orderings of the explanatory variables. The red bars indicate an average increase in risk, while the green bars show a decrease in risk. The violet box plots summarize the distribution of the attributions over the 25 orderings. The plots reveal similar influential variables as in the breakdown attributions, with a slight reordering of importance and the inclusion of additional variables—such as education level and stability assessment for patient A, and COVID-19 pandemic period for patient B.

To examine the time-varying model's prediction for a single time-interval instance for a patient, we randomly selected 2 patients: patient C, who had experienced 1 or more interruptions and subsequently returned to care; patient D, who had no history of treatment interruption. Figure 6 shows the breakdown and SHAP attribution plots for a single time interval for patient C. After considering all variable contributions, the final predicted risk of subsequent treatment interruption for this patient was 2.43, which is 1.614 points above the average prediction. The key covariate contributing to this risk prediction is the patient's previous interruption duration (duration of 30 - 179 days). In contrast, being on a 6-month MMD interval and having not

XSL•FO
RenderX

missed or defaulted on any appointments were associated with a lower risk of subsequent interruptions. The SHAP plot reflects similar variable contributions (Figure 8B). The predicted risk of treatment interruption for patient D was 0.181, which is 0.635 lower than the average model prediction (Figure 9A). Key covariates contributing to this reduced risk include no prior treatment interruptions (interruption duration of zero), being on a 3 - to 5-month MMD interval, having a viral load between 200 and 999 cp/mL, the COVID-19 pandemic period, no defaulted appointments, good adherence, and starting ART between the ages of 35 and 44 years. SHAP plots (Figure 9B) further highlighted covariates associated with a lower risk for this patient.

Figure 8. Time-varying model, breakdown, and SHAP attribution plots for patient C. **(A)** Breakdown plot decomposing the model prediction into cumulative covariate contributions. Intercept—model's overall average prediction value; green bars, with (+) sign indicate increased risk of treatment interruption; red bars, with (−) sign indicate decreased risk of interruption; and blue bar, "prediction" is the individual predicted risk of interruption for the patient. **(B)** SHAP plot summarizing the direction and magnitude of individual covariate effects on the predicted risk of treatment interruption. ART: antiretroviral therapy; MMD: multimonth dispensing; PwP: prevention with positives package; SHAP: Shapley Additive Explanations.

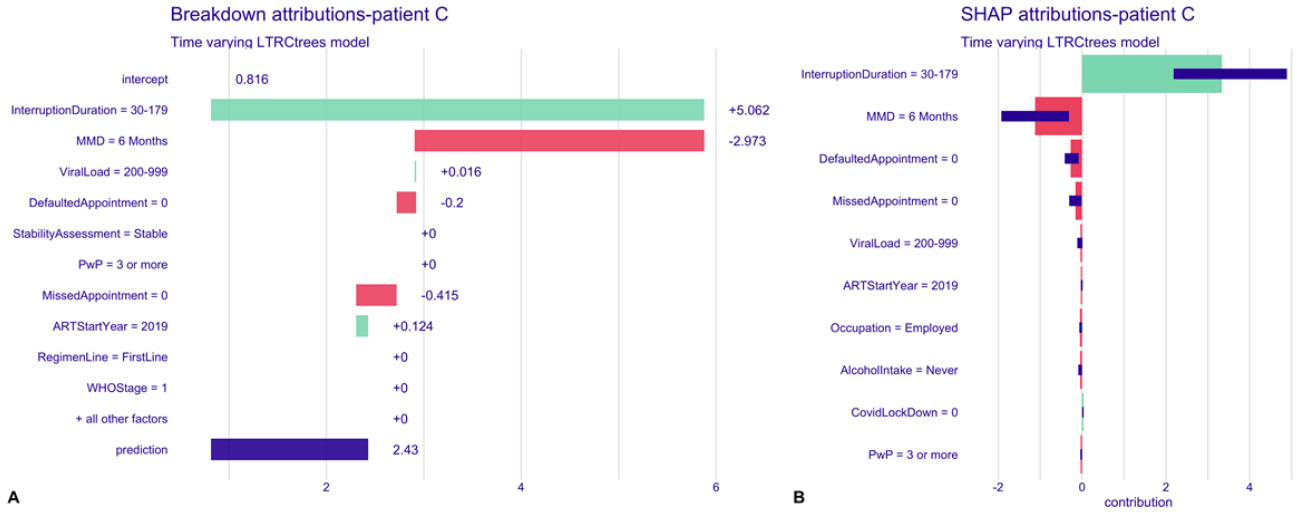
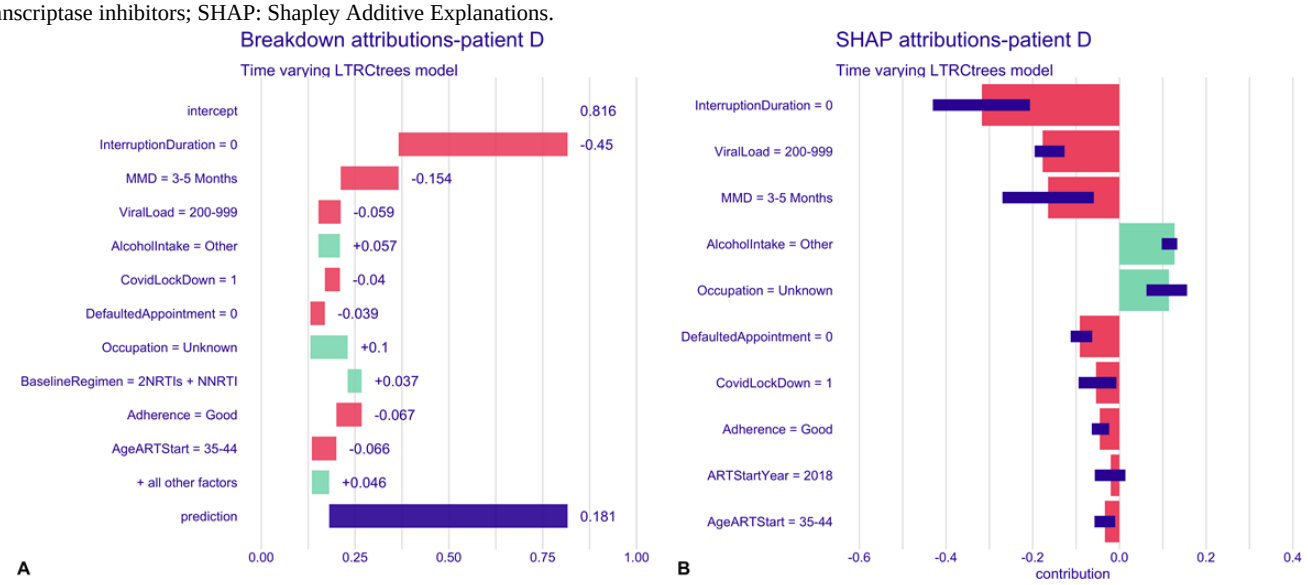


Figure 9. Time-varying model, breakdown, and SHAP attribution plots for patient D. **(A)** Breakdown plot decomposing the model prediction into cumulative covariate contributions. Intercept—model's overall average prediction value; green bars, with (+) sign indicate increased risk of treatment interruption; red bars, with (−) sign indicate decreased risk of interruption; and blue bar, "prediction" is the individual predicted risk of interruption for the patient. **(B)** SHAP plot summarizing the direction and magnitude of individual covariate effects on the predicted risk of treatment interruption. ART: antiretroviral therapy; MMD: multimonth dispensing; NNRTI: nonnucleoside analog reverse transcriptase inhibitor; NRTIs: nucleoside analog reverse transcriptase inhibitors; SHAP: Shapley Additive Explanations.



Using CP profile plots, we further examined how the model's predictions for these patients would shift if key covariates—such as the duration of the previous interruption, viral load, DSD model, MMD, missed appointments, and defaulted appointments—were altered. This allowed us to explore how changes in time-varying covariates impact the risk of multiple treatment interruptions. The blue dot and dotted line in the CP plot (Figures 10 and 11) represent the values of the covariates and the corresponding prediction for the patient. Each green bar in the plot shows the predicted change in value and direction for a specific covariate, with bars pointing to the right of the dotted line indicating an increased risk of interruption and bars pointing to the left indicating a decreased risk.

XSL•FO
RenderX

Figure 10. Ceteris paribus profile plots for patient C. Panels show the marginal effect of changes of each variable on the predicted risk of treatment interruption. (A) Defaulted appointments, (B) differentiated care model, (C) interruption duration, (D) missed appointments, (E) multimonth dispensing duration, and (F) viral load. Blue dot is the value of the covariate; blue dotted line corresponding prediction for the patient; green bars—predicted change in value and direction for a specific covariate; and bars right of the dotted line indicate increased risk, while those to the left indicate decreased risk. ART: antiretroviral therapy; HCW: health care worker; MMD: multimonth dispensing.

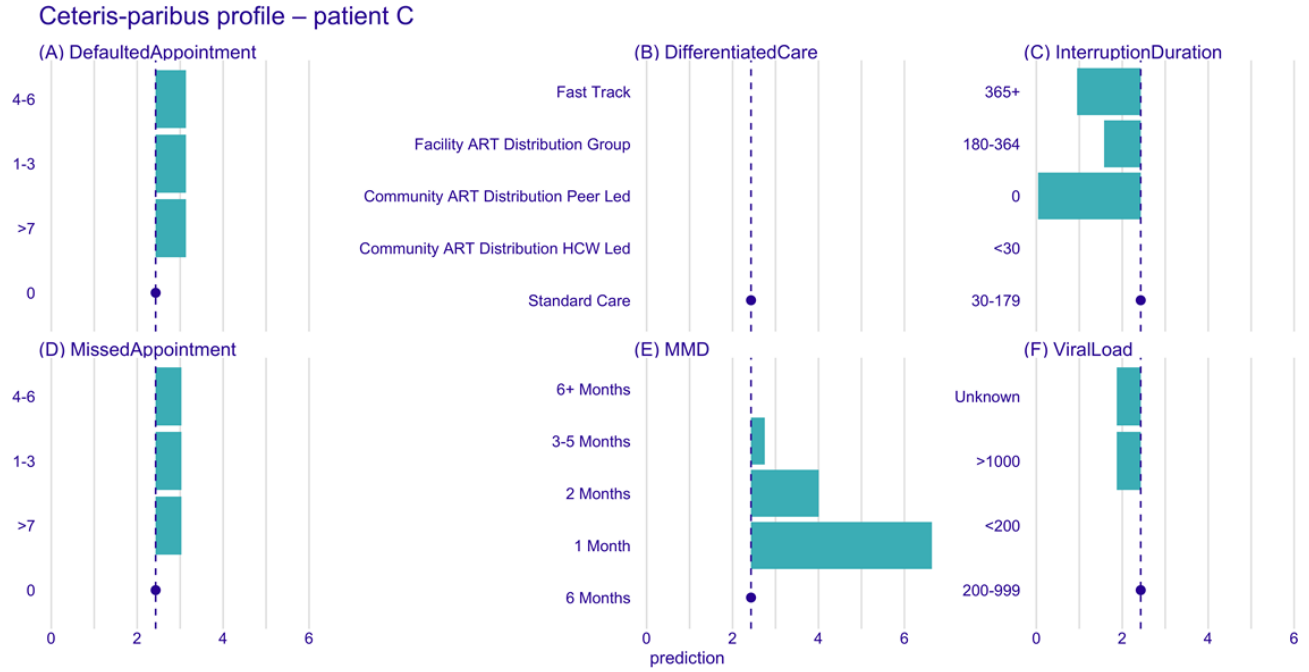
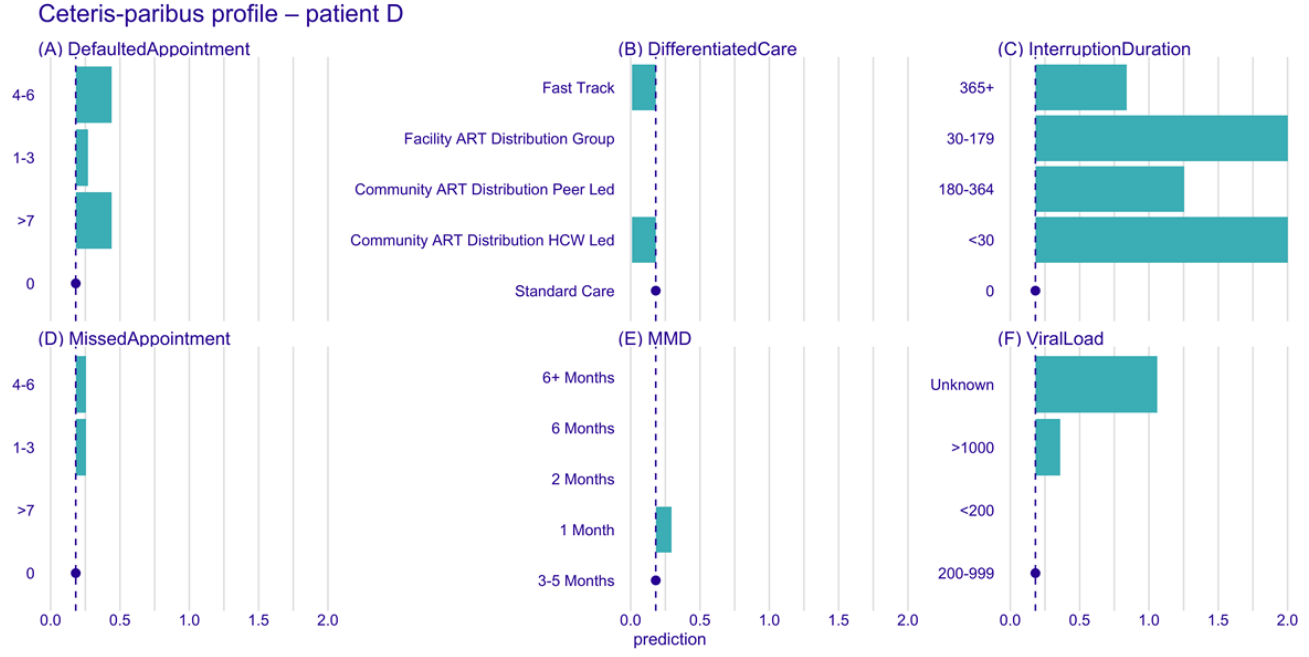


Figure 11. Ceteris paribus profile plots for patient D. Panels show the marginal effect of changes of each variable on the predicted risk of treatment interruption. (A) Defaulted appointments, (B) differentiated care model, (C) interruption duration, (D) missed appointments, (E) multimonth dispensing duration, and (F) viral load. Blue dot is the value of the covariate; blue dotted line corresponding prediction for the patient; green bars—predicted change in value and direction for a specific covariate; and bars right of the dotted line indicate increased risk, while those to the left indicate decreased risk. ART: antiretroviral therapy; HCW: health care worker; MMD: multimonth dispensing.



The CP plot (Figure 10) for patient C indicates that their risk of future treatment interruptions would increase significantly if they were to miss or default on an appointment or be placed on a shorter MMD interval. However, the predicted risk remains unchanged if they were to transition across different DSD models. The future risk of treatment interruption would increase for patient D if they miss or default on an appointment, had an increase in their viral load, were placed on a 1-month MMD interval, or experienced a treatment interruption, irrespective of how long it took to return to care (Figure 11). Transitioning from their current standard care DSD model to fast track or community ART distribution health care worker–led models would slightly reduce the risk of interruption.

XSL•FO
RenderX

Discussion

Principal Findings

We compared the traditional CPH model with 6 different ML models to predict the risk of treatment interruption in 621,115 patients receiving ART in 256 facilities across all 47 counties in Kenya. The analysis was conducted in 2 parts: time-invariant models for predicting time-to-first treatment interruptions and time-varying models for predicting time-to-multiple treatment interruptions. Model evaluation was performed using an 80:20 train-test split, and Harrell's C-index was applied to compare the performance of the models on both the train and test sets. Prediction from the best-performing model was then interpreted at both global and individual levels.

The RP model outperformed all other models in both the time-invariant and time-varying analyses. It achieved C-index values of 0.8115 for the train set and 0.8105 for the test set in the time-invariant model, and 0.8918 for the train set and 0.8901 for the test set in the time-varying model. These results align with findings from previous studies that have compared CPH models with ML models, demonstrating the effectiveness of RP models in survival analysis. Previous studies have reported C-index values ranging from 0.63 to 0.96, primarily from ML models such as GBM, random survival forest, and survival support vector machines in predicting survival for patients with cancer [16-19]. Other studies that focused on predicting treatment interruptions in patients on ART had typically reported metrics such as sensitivity, specificity, and area under the curve, with values ranging between 61.9% and 76.0% [20-24]. Few of these studies have incorporated time-varying covariates [16,25]. Even when compared with prior studies that evaluated ML models against CPH [17], our RP approach achieved higher discrimination, with C-index values of 0.8105 and 0.8901 compared with reported XGBoost performance of approximately 0.73. Notably, our study stands out as one of the few to apply RP within a time-varying survival framework, effectively constructing a survival tree that accommodates time-varying covariates.

The RP method offers flexibility in detecting complex interactions without assuming proportional hazards. They automatically select covariates and split points, allowing more adaptive modeling of survival times [26]. This method is advantageous for identifying nonlinear relationships and interactions, making it particularly useful in our case where covariate effects changed over time. The RP model also had the advantage of simple hyperparameter tuning and faster computational times compared with the other ML models we evaluated. We attempted to train a random survival forest model but were unsuccessful due to the large computational resources and extensive time required to handle our large dataset. Likewise, the widely used randomForestSRC R package does not currently support time-varying covariates directly. While random forests are known for their high predictive performance, their significant computational requirements make them less practical for real-world applications [16].

To ensure that our results were fully comparable with the traditional CPH, we implemented a cohesive framework for explaining the ML predictions with meaningful insights. Using a model-agnostic [15] explainer framework, we gained detailed insights into how specific variables influenced individual-level predictions and overall global outcomes, similar to the interpretations provided by hazard ratios in the CPH model. The explainer framework offers additional advantages by capturing complex, nonlinear relationships that traditional CPH may overlook.

On a global level, the permutation-based variable importance from the RP model was similar to the ranking of the absolute values of the coefficients of the CPH model for both time-invariant and time-varying models. In addition, the effects of the selected covariates we considered in the partial dependence plots were consistent with the direction of hazard ratios in the CPH model. This combination of techniques offers diverse insights into our model predictive performance. They reveal the most significant effects and risks associated with treatment interruption across various covariates without the assumptions of linearity, hence capturing potential nonlinear effects that traditional CPH models might overlook. Overall, these methods provide more comprehensive global explanations, particularly for complex variable interactions.

Further exploration using breakdown attribution and SHAP plots reveals how individual variables influence each patient's predicted risk within specific time intervals. On a global level, the time-invariant model was primarily influenced by MMD. Considering this at the patient level, the breakdown and SHAP plots provide detailed insights into how MMD increases or decreases risk, explaining why patient A was at a higher risk than patient B based on MMD. In addition, these plots highlight the most influential covariates for predicting the risk of a first treatment interruption for each patient, emphasizing individual variability in risk factors. For the time-varying model, we furthered the instance-level exploration using CP profiles. These plots provided insights into how a patient's predicted risk would change if certain variables were adjusted. For example, with patient C—who has no history of treatment interruption—we explored the predicted risk of an interruption if the patient were to miss or default on an appointment, transition to a less intensive DSD model, or receive a different drug-dispensing interval. In contrast, for patient D—a patient with a history of interruptions—the focus was on understanding how the predicted risk of another interruption might change if they transitioned to a less intensive care model, adjusted their dispensing interval, or missed an appointment. The changes highlighted for each patient provide actionable insights to guide transitions between DSD models and targeted interventions to prevent treatment interruptions. Following ART initiation, CP profiles can identify patients whose interruption risk under standard care would decrease with earlier DSD enrollment or longer dispensing intervals. For patients returning to care after an interruption, CP profiles quantify how changes in follow-up intensity or MMD duration influence recurrent interruption risk, supporting tailored rather than routine care. These scenario-based insights align with recent World Health Organization and global guidance [27,28] emphasizing person-centered, targeted care packages for patients reengaging in treatment.

XSL•FO
RenderX

Bringing the various components together, our model could be integrated into an EMR system to predict a patient's risk of treatment interruption in real time. This integration would allow the model to continuously process each patient's historical data, current visit date, and next scheduled visit date, creating a new data point for each interaction. With each visit, the EMR system would automatically update risk predictions, enabling health care providers to proactively identify high-risk patients before interruptions occur. The Kenya EMR currently includes an ML module that generates interruption risk scores based on a limited set of patient-centric covariates. By leveraging our model's broader range of covariates, the current system can be optimized for improved prediction accuracy and comprehensive risk assessment.

Limitations

Our study had some limitations. First is the issue of missing data, as highlighted in our previous work [6]. This was addressed by creating an "unknown" category for affected variables, given that the data were not missing at random. Variables with more than 60% missingness, such as baseline and follow-up CD4 counts, were excluded, which may have introduced bias. Sensitivity analyses comparing complete-case models with alternative "unknown" category specifications were conducted in prior work [6] and showed consistent results. Multiple imputation was not considered, given data were not missing at random. Second, while the explainer we used is model-agnostic, it lacks built-in support for survival-specific metrics such as Brier score and area under the curve. Survex explainer is specifically designed for survival models, but it currently does not support the LTRCtrees package or time-varying survival models. A potential work-around could be treating the model as a classification task, focusing on whether there was an interruption or not within a specified fixed time, although this approach will sacrifice the temporal dimension of survival analysis or future studies can consider implementing other Survex-supported algorithms with the ability to fit time-varying covariates. Finally, generating breakdown and SHAP attribution profiles for individual predictions was computationally intensive, particularly for SHAP, which may limit real-time EMR use. However, subsampling across observations or feature permutations can substantially reduce computational cost while preserving accurate SHAP approximations.

Conclusions

Our study compared the predictive performance of the traditional Cox model with 6 different ML models for predicting the risk of treatment interruptions in patients on ART. Overall, our results demonstrate that ML models outperform the traditional Cox model, with RP, in particular, effectively capturing nonlinear effects and complex interactions without relying on proportional hazards assumptions, making it well suited for time-varying survival analyses. Our model-agnostic explanation aligns well with the Cox model results while also offering additional, patient-specific insights for targeted interventions. The combination of enhanced predictive performance and interpretability demonstrates how ML models can effectively support patient-centered care strategies to reduce the likelihood of treatment interruptions in ART.

Acknowledgments

The authors acknowledge the technical inputs of Davies Kimanga (Centers for Disease Control and Prevention—Kenya) and Jacob Odhiambo (The Palladium Group Kenya) during the early phases of the study design. The authors also acknowledge the following staff of The Palladium Group Kenya, that is, Mary Gikura, Kennedy Muthoka, Ashley Odhiambo, and Evans Munene, for their support with data access and validation. Finally, the authors also thank members of the INSIGHT technical working group from CIHEB Kenya—Sam Wafula, Rebecca Wangusi, Patrick Awuor, Violet Makokha, Dinah Nakhanu Mamai, Sarah Nzioka, and Emmah Momanyi—for their valuable contribution and insights into the Kenya treatment program.

Funding

This work was supported by the National Institutes of Health (NIH) D43 Fogarty Global health research training program and Integrated Networks of Scholars in Global Health Research Training (INSIGHT D43TW012274). The findings and conclusions in this paper are solely those of the authors and do not reflect the official positions of their respective institutions. All authors take full responsibility for the decision to submit this manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of covariates and descriptions.
[DOCX File, 66 KB - medinform_v14i1e78964_app1.docx]

References

1. Cox DR. Regression models and life-tables. J R Stat Soc Series B 1972 Jan 1;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]

XSL•FO
RenderX

2. Kalbfleisch JD, Schaubel DE. Fifty years of the Cox model. Annu Rev Stat Appl 2023 Mar 10;10(1):1-23. [doi: 10.1146/annurev-statistics-033021-014043]
3. Cox DR, Oakes D. Analysis of Survival Data: Chapman and Hall; 1987. URL: https://www.routledge.com/Analysis-of-Survival-Data/Cox-Oakes/p/book/9780412244902?srsltid=AfmBOop9h4oTLfhsnPee_a42cacTol4DIAPZfcQtE7YWoTGaGLqw_ai1 [accessed 2025-02-11]
4. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health 1999;20(1):145-157. [doi: 10.1146/annurev.publhealth.20.1.145] [Medline: 10352854]
5. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med 2018 Apr;6(7):121. [doi: 10.21037/atm.2018.02.12] [Medline: 29955581]
6. Salami D, Koech E, Turan JM, et al. Using big data to evaluate the impact of differentiated care on first and recurrent HIV treatment interruptions in Kenya. SSRN. Preprint posted online on Dec 23, 2025. [doi: 10.2139/ssrn.5954474]
7. Andersen PK. Fifty years with the Cox proportional hazards regression model. J Indian Inst Sci 2022 Oct;102(4):1135-1144. [doi: 10.1007/s41745-021-00283-9]
8. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008 Sep;2(3):841-860. [doi: 10.1214/08-AOAS169]
9. Bender A, Rügamer D, Scheipl F, Bischl B. A general machine learning framework for survival analysis. Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Sep 14-18, 2020. [doi: 10.1007/978-3-030-67664-3_10]
10. Nagpal C, Jeanselme V, Dubrawski A. Deep parametric time-to-event regression with time-varying covariates. Presented at: Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021; Mar 22-24, 2021 URL: https://proceedings.mlr.press/v146/nagpal21a.html [accessed 2026-01-22]
11. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
12. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
13. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
14. Monitoring, evaluation, and reporting indicator reference guide MER 2.0. : PEPFAR US President's Emergency Plan for AIDS Relief; 2024 URL: https://www.state.gov/wp-content/uploads/2025/01/FY25-MER-v2.8-Indicator-Reference-Guide_508-Compliant.pdf [accessed 2024-12-19]
15. Maksymiuk S, Gosiewska A, Biecek P. Landscape of R packages for explainable artificial intelligence. arXiv. Preprint posted online on Sep 24, 2020. [doi: 10.48550/arXiv.2009.13248]
16. Cygu S, Seow H, Dushoff J, Bolker BM. Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. Sci Rep 2023 Jan 25;13(1):1370. [doi: 10.1038/s41598-023-28393-7]
17. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep 2021 Mar 26;11(1):6968. [doi: 10.1038/s41598-021-86327-7]
18. Tran TT, Lee J, Gunathilake M, et al. A comparison of machine learning models and Cox proportional hazards models regarding their ability to predict the risk of gastrointestinal cancer based on metabolic syndrome and its components. Front Oncol 2023;13:1049787. [doi: 10.3389/fonc.2023.1049787] [Medline: 36937438]
19. Xu L, Cai L, Zhu Z, Chen G. Comparison of the Cox regression to machine learning in predicting the survival of anaplastic thyroid carcinoma. BMC Endocr Disord ;23(1):129. [doi: 10.1186/s12902-023-01368-5]
20. Esra RT, Carstens J, Estill J, et al. Historical visit attendance as predictor of treatment interruption in South African HIV patients: extension of a validated machine learning model. PLOS Global Public Health 2023;3:e0002105. [doi: 10.1371/journal.pgph.0002105]
21. Ogbechie MD, Fischer Walker C, Lee MT, et al. Predicting treatment interruption among people living with HIV in Nigeria: machine learning approach. JMIR AI 2023 May 12;2(1):e44432. [doi: 10.2196/44432] [Medline: 38875546]
22. Maskew M, Sharpey-Schafer K, De Voux L, et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Sci Rep 2022 Jul 26;12(1):12715. [doi: 10.1038/s41598-022-16062-0] [Medline: 35882962]
23. Pettit AC, Bian A, Schember CO, et al. Development and validation of a multivariable prediction model for missed HIV health care provider visits in a large US clinical cohort. Open Forum Infect Dis 2021 Jul;8(7):ofab130. [doi: 10.1093/ofid/ofab130] [Medline: 34327249]
24. Ramachandran A, Kumar A, Koenig H, et al. Predictive analytics for retention in care in an urban HIV clinic. Sci Rep 2020 ;10(1):6421. [doi: 10.1038/s41598-020-62729-x]
25. Suresh K, Severn C, Ghosh D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol 2022 Jul 26;22(1):207. [doi: 10.1186/s12874-022-01679-6] [Medline: 35883032]
26. Fu W, Simonoff JS. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. Biostatistics 2017 Apr 1;18(2):352-369. [doi: 10.1093/biostatistics/kxw047] [Medline: 28025180]

XSL•FO
RenderX

27. Supporting re-engagement in HIV treatment services. World Health Organization. 2024. URL: https://www.who.int/publications/i/item/9789240097339 [accessed 2026-01-22]
28. Differentiation at re-engagement in HIV care: a multi-country workshop. International AIDS Society (IAS). 2024. URL: https://www.differentiatedservicedelivery.org/resources/differentiation-at-re-engagement-in-hiv-care-a-multi-country-workshop/ [accessed 2025-11-17]

Abbreviations

ART: antiretroviral therapy
C-index: concordance index
CP: ceteris paribus
CPH: Cox proportional hazards
DSD: differentiated service delivery
EMR: electronic medical record
GBM: gradient boosting machine
ML: machine learning
MMD: multimonth dispensing
RP: recursive partitioning
SHAP: Shapley Additive Explanations
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 16.Jun.2025; peer-reviewed by RS A, Y Shi; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 04.Feb.2026.

Please cite as:
Salami D, Koech E, Turan JM, Stafford KA, Nyagah LM, Ohakanu S, Ngugi AK, Charurat M
Prediction of First and Multiple Antiretroviral Therapy Interruptions in People Living With HIV: Comparative Survival Analysis Using Cox and Explainable Machine Learning Models
JMIR Med Inform 2026;14:e78964
URL: https://medinform.jmir.org/2026/1/e78964
doi:10.2196/78964

© Donald Salami, Emily Koech, Janet M Turan, Kristen A Stafford, Lilly Muthoni Nyagah, Stephen Ohakanu, Anthony K Ngugi, Manhattan Charurat. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prospective Diagnostic Accuracy and Technical Feasibility of Artificial Intelligence-Assisted Rib Fracture Detection on Chest Radiographs: Observational Study

Shu-Tien Huang[1,2,3], MSc, MD; Liong-Rung Liu[1,2,3], MSc, MD; Ming-Feng Tsai[2,3,4], MD; Ming-Yuan Huang[1,3*], MSc, MD; Hung-Wen Chiu[2,5*], PhD

[1]Department of Emergency Medicine, Mackay Memorial Hospital, Taipei, Taiwan

[2]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 9F, Education & Research Building, Shuang-Ho Campus, No. 301, Yuantong Rd., Zhonghe Dist., New Taipei City, Taiwan

[3]College of Medicine, Mackay Medical University, New Taipei City, Taiwan

[4]Division of Plastic Surgery, Mackay Memorial Hospital, Taipei, Taiwan

[5]Clinical Big Data Research, Taipei Medical University Hospital, Taipei City, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hung-Wen Chiu, PhD
Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 9F, Education & Research Building, Shuang-Ho Campus, No. 301, Yuantong Rd., Zhonghe Dist., New Taipei City, Taiwan

Abstract

Background: Rib fractures are present in 10% - 15% of thoracic trauma cases but are often missed on chest radiographs, delaying diagnosis and treatment. Artificial intelligence (AI) may improve detection and triage in emergency settings.

Objective: This study aims to evaluate diagnostic accuracy, processing speed, and technical feasibility of an artificial intelligence–assisted rib fracture detection system using prospectively collected data within a real-world, high-volume emergency department workflow.

Methods: We conducted an observational feasibility study with prospective data collection of a faster region-based convolutional neural network–based AI model deployed in the emergency department to analyze 23,251 real-world chest radiographs (22,946 anteroposterior; 305 oblique) from April 1 to July 2, 2023. This study was approved by the Institutional Review Board of MacKay Memorial Hospital (IRB No. 20MMHIS483e). AI operated passively, without influencing clinical decision-making. The reference standard was the final report issued by board-certified radiologists. A subset of discordant cases underwent post hoc computed tomography review for exploratory analysis.

Results: AI achieved 74.5% sensitivity (95% CI 0.708-0.780), 93.3% specificity (95% CI 0.930-0.937), 24.2% positive predictive value, and 99.2% negative predictive value. Median inference time was 10.6 seconds versus 3.3 hours for radiologist reports (paired Wilcoxon signed-rank test W=112 987.5, P<.001). The analysis revealed peak imaging demand between 08:00 and 16:00 and Thursday-Saturday evenings. A 14-day graphics processing unit outage underscored the importance of infrastructure resilience.

Conclusions: The AI system demonstrated strong technical feasibility for real-time rib fracture detection in a high-volume emergency department setting, with rapid inference and stable performance during prospective deployment. Although the system showed high negative predictive value, the observed false-positive and false-negative rates indicate that it should be considered a supportive screening tool rather than a stand-alone diagnostic solution or a replacement for clinical judgment. These findings support further clinician-in-the-loop studies to evaluate clinical feasibility, workflow integration, and impact on diagnostic decision-making. However, interpretation is limited by reliance on radiology reports as the reference standard and the system's passive, non-interventional deployment.

(JMIR Med Inform 2026;14:e77965) doi:10.2196/77965

KEYWORDS

artificial intelligence; chest radiograph; clinical decision support; digital health; emergency radiology; feasibility study; faster R-CNN; real-world data; rib fracture

XSL•FO
RenderX

Introduction

Digital health technologies, particularly artificial intelligence (AI), are increasingly used to address diagnostic delays in high-acuity clinical settings. In emergency departments, timely identification of injuries is essential, yet radiographic interpretation remains constrained by heavy workloads and the inherent complexity of imaging—especially for subtle findings such as rib fractures.

Rib fractures are a frequent consequence of thoracic trauma, occurring in 10% - 15% of trauma patients and often indicating more serious underlying injuries [1-3]. When missed, they may lead to inadequate pain management, delayed respiratory support, pneumonia, or even preventable intensive care unit admissions. Beyond clinical harm, undetected fractures also carry medicolegal implications and increase health care costs.

Despite their significance, rib fractures are notoriously difficult to detect on chest radiographs (CXRs)—the first-line imaging modality in most emergency departments—due to overlapping anatomical structures and subtle fracture lines. Reported sensitivities for radiologist detection can be as low as 15%, with up to half of fractures potentially missed in high-volume settings [4,5]. Although computed tomography (CT) and ultrasound can improve accuracy, they are resource-intensive and not always feasible for frontline triage [6-8]. These limitations highlight an urgent need for AI-driven tools that can assist clinicians by rapidly identifying suspected rib fractures in routine CXRs, enabling more effective prioritization and timely intervention.

Recent advances in AI, particularly deep learning, have demonstrated strong potential in automating image analysis tasks across medical domains, including dermatology, ophthalmology, and pulmonary imaging [9-11]. Deep learning models, especially convolutional neural networks, can automatically extract complex image features and have shown superior performance compared to traditional machine learning methods in various image classification tasks [11-13]. Transfer learning further enables the adaptation of pretrained convolutional neural networks—originally developed for natural images—for medical image classification tasks, including bone fracture detection [14,15].

Although prior studies have applied deep learning to rib fracture detection with promising results, most were retrospective, limited in scale, and did not assess feasibility in operational emergency department workflows [7,16,17]. These proof-of-concept efforts did not address the practical barriers to integrating AI into emergency radiology workflows, such as inference latency, system interoperability, or artifact handling.

To address this gap, we conducted an observational feasibility study with prospective data collection, evaluating an AI model for rib-fracture detection on CXRs. The system was passively deployed in parallel with routine emergency department imaging workflows using real-world data, without influencing clinical decisions. This design allowed the assessment of diagnostic performance, processing speed, and operational characteristics within standard clinical workflows.

Methods

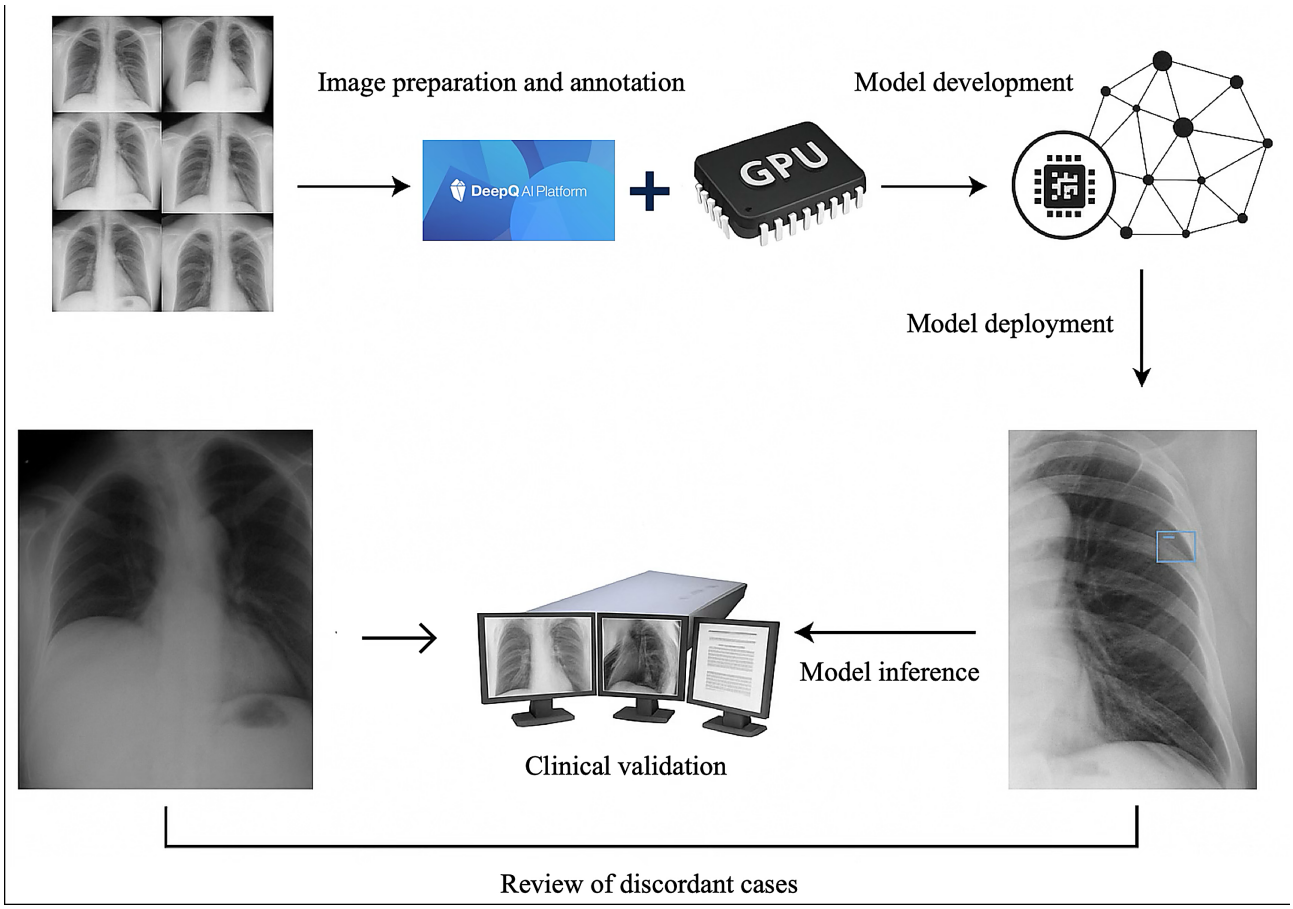
Study Design

The observational feasibility study protocol was reviewed and approved by the Institutional Review Board of MacKay Memorial Hospital (IRB No. 20MMHIS483e) prior to the initiation of data collection. MacKay Memorial Hospital is a tertiary referral and level 1 trauma center in northern Taiwan. The AI system functioned passively in real time without influencing clinical decisions or patient management. As the system functioned in a noninterventional, observational manner, prospective trial registration was not required.

From April 1 to July 2, 2023, all chest and rib radiographs acquired in the emergency department were automatically processed by the AI system in near-real-time. Both standard CXRs and rib-only views acquired during the study period were automatically analyzed, as both modalities are routinely used for suspected thoracic trauma. During the study period, a temporary 14-day graphics processing unit (GPU) hardware outage occurred, during which radiographs were not processed in real time; these examinations were excluded from turnaround time analysis but retained for diagnostic accuracy, as formal radiology reports were available. No additional exclusion criteria were applied beyond the 14-day system outage; all eligible emergency radiographs during the study period were included in the analysis. The system operated passively alongside routine clinical workflows, without influencing clinical decisions. AI-identified suspected rib fractures were highlighted using bounding boxes on a backend interface, which was accessible only for research evaluation and remained hidden from the clinical care team (Figure 1).

XSL•FO
RenderX

Figure 1. Study workflow of a prospective observational feasibility study evaluating an artificial intelligence–assisted rib fracture detection system using chest radiographs in patients admitted to the emergency department at a high-volume tertiary medical center (April 1-July 2, 2023). GPU: graphics processing unit.



Ethical Considerations

This study was approved by the Institutional Review Board of MacKay Memorial Hospital (IRB No. 20MMHIS483e) and conducted in accordance with the Declaration of Helsinki. Informed consent was waived because the study involved secondary analysis of routinely acquired, deidentified clinical imaging data, and the AI system operated passively without influencing patient management.

All data were deidentified prior to analysis and processed on secure institutional servers with access limited to authorized research personnel. No compensation was provided to participants. All images included in the manuscript were fully anonymized, and no identifiable patient information is disclosed.

Consistent with the approved observational study design, all AI outputs—including discordant cases—were withheld from treating clinicians and did not influence patient management.

AI Model Development

CXRs were retrospectively collected from the hospital picture archiving and communication system (PACS) for model training. All images were deidentified and preprocessed using histogram equalization and image inversion to improve fracture conspicuity. Fracture locations were annotated using bounding boxes by a board-certified emergency physician with 18 years of clinical experience via the DeepQ AI platform [18].

A deep learning model was developed using PyTorch (v1.13) with GPU acceleration. The architecture was based on faster region-based convolutional neural network (R-CNN) [19], incorporating a ResNet-50 backbone for feature extraction, a region proposal network for candidate region generation, and a classification head for fracture detection.

The dataset comprised 2079 CXRs (1065 fracture-positive and 1014 normal) collected between 2010 and 2020. Images were randomly divided into training (80%) and validation (20%) sets at the image level, as each radiograph represented an independent study. When multiple images were obtained from the same patient encounter, each radiograph was treated as an independent sample. Data augmentation—including random rotation, flipping, brightness, and contrast adjustment—was applied to improve generalization. To address the inherent class imbalance given the low fracture prevalence, class-weighted loss and oversampling of fracture-positive images were employed.

Model Validation

Model performance was assessed on a hold-out test set of 262 CXRs containing 724 expert-annotated rib fractures. Evaluation metrics were reported at both the case and object levels.

At the case level, the unit of analysis was the radiographic study. A study was considered positive if at least 1 rib fracture was detected, regardless of the number of fractures present. The model correctly identified fractures in 230 of 257

XSL•FO
RenderX

fracture-positive studies, achieving a sensitivity of 89.5%. With only 8 false-positive cases, precision reached 96.6%, yielding an overall F_1-score of 0.93.

At the object level, performance reflected per-lesion detection accuracy. The model correctly localized 680 of 724 annotated rib fractures (recall=94.0%) and generated 55 false-positive boxes. The mean average precision at an intersection-over-union threshold of ≥0.5 was 0.65, indicating robust lesion-level localization (Table 1).

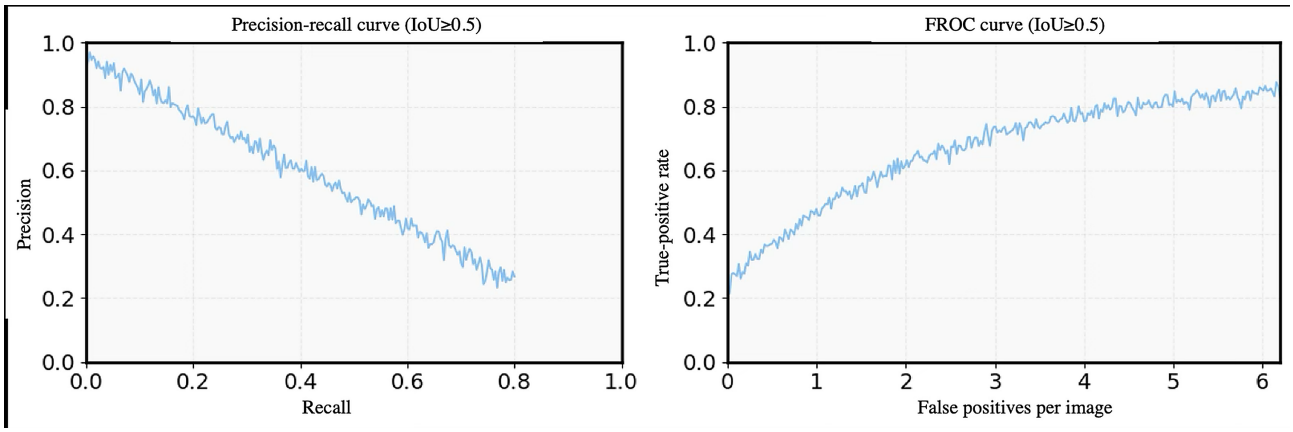
Table . Dataset composition and performance of the artificial intelligence (AI) model using retrospective emergency department chest radiographs, including case-level detection and object-level localization (intersection-over-union [IoU]≥0.50).

Category and metric	Value
Dataset	
Total images	262
Ground-truth boxes	724
Case-level detection (%)	
Sensitivity (recall)	89.6
Precision	96.6
F_1-score	0.93
Object-level localization	
Recall	94.0%
mAP[a] (IoU≥0.5)	0.65

[a]mAP: mean average precision.

Curve-based analyses further characterized the model's detection behavior (Figure 2). The precision-recall curve (Figure 2A) maintained precision ≥0.90 until recall fell below 0.55 (precision-recall area under the curve=0.65), demonstrating high reliability across a broad sensitivity range. The free-response receiver operating characteristic (FROC) curve (Figure 2B) showed true-positive rates of 0.77 at 1 false positive per image and 0.88 at three, representing practical trade-offs between sensitivity and alert frequency in potential clinical deployment.

Figure 2. The performance of the artificial intelligence (AI)–assisted rib fracture detection model was evaluated in the retrospective model development and validation dataset using emergency department chest radiographs. (A) Precision-recall curve demonstrating case-level detection performance at an intersection-over-union (IoU) threshold of ≥0.50 (area under the curve=0.65). (B) Free-response receiver operating characteristic (FROC) curve showing lesion-level sensitivity as a function of false positives per image, with localization performance assessed at the same IoU threshold (mean average precision=0.65; recall=94%).



Prospective Evaluation of AI Model in Emergency Department Workflow

The trained model was prospectively evaluated in parallel with clinical workflow, performing automated inference on incoming emergency radiographs. During the automated inference process, all incoming radiographs were standardized and resized to a fixed resolution of 512×512 pixels. The prospective evaluation cohort (April-July 2023) was temporally and operationally independent from the retrospective training and validation dataset (2010-2020), and no patient overlap existed between the 2 cohorts. During the prospective evaluation phase, all chest and rib radiographs from emergency department encounters were automatically processed by the AI system without disrupting clinical workflows. The bounding box outputs were logged for research analysis but were not disclosed to radiologists or used for patient management.

XSL•FO
RenderX

Performance Assessment Using NLP-Derived Labels

To evaluate AI performance in the real-world setting, output was compared to formal radiology reports issued by board-certified radiologists. A rule-based natural language processing (NLP) pipeline was developed to extract structured rib fracture labels (positive, negative, or ambiguous) from free-text reports. The algorithm combined keyword detection (eg, “rib fracture,” “fx”) and negation handling (eg, “no evidence of,” “no definite fracture”).

To validate NLP accuracy, a random sample of 200 radiology reports was manually reviewed by 2 emergency physicians blinded to both NLP and AI results. The NLP classification achieved 96.5% agreement (193/200) with manual review, with a Cohen κ of 0.91, indicating excellent concordance. Most discrepancies were due to ambiguous language or complex negation structures (Table 2).

Table . Confusion matrix comparing natural language processing (NLP)–extracted radiology report labels with manual expert review in a randomly selected subset of 200 emergency department chest radiographs, used to assess labeling accuracy in the retrospective dataset.

Manual review	Fracture present	Fracture absent	Ambiguous	Total (NLP prediction)
NLP: fracture present	95	2	1	98
NLP: fracture absent	3	90	2	95
NLP: ambiguous	1	2	4	7
Total (manual review)	99	94	7	200

Formal radiology reports served as the primary reference standard for AI evaluation. This approach may have underestimated AI sensitivity because CT confirmation was not performed systematically. In select discordant cases—where the AI flagged fractures not documented in the reports—subsequent CT scans confirmed some of these findings. These discrepancies were retrospectively reviewed to investigate potential underreporting by radiologists. While informative, these exploratory adjudications were not used as a universal reference standard due to inconsistent CT availability. Nonetheless, radiology reports remained the definitive benchmark for all performance metrics. Additionally, the selected misclassified cases were examined to identify recurring patterns of diagnostic oversight among frontline physicians.

Targeted Adjudication of Discordant Cases

To further explore potential underreporting within the report-based reference standard, we performed a focused review of discordant cases in which the AI system flagged suspected rib fractures not documented in the corresponding formal radiology reports. Because not all discordant cases had confirmatory CT imaging, an illustrative subset of 11 cases was selected based on the availability of same-encounter chest CT and clinical relevance for qualitative adjudication. Each case was reviewed to determine whether the AI-predicted fractures corresponded to true fractures confirmed on CT. These adjudications were exploratory and intended to contextualize the potential clinical value of AI detection beyond the report-based benchmarking.

Data Analysis

Case-level performance was assessed at the radiographic study (accession) level by comparing AI outputs to NLP-derived labels from radiology reports. A study was considered positive if at least 1 image within the same examination was flagged as having a rib fracture by the AI system; otherwise, the study was classified as negative. Key metrics included sensitivity, specificity, accuracy, positive predictive value, negative predictive value (NPV), and F_1-score. Ninety-five percent CIs were calculated using nonparametric bootstrap resampling (1000 iterations). The results were summarized in confusion matrices and diagnostic performance plots.

Statistical Analysis

All statistical analyses were performed using Python 3.11 (pandas v2.2, scikit-learn v1.4) and R 4.3.2. Continuous variables were reported as mean (SD) or median with IQR. Categorical variables were summarized as counts and percentages. A 2-tailed $P<.001$ was considered statistically significant.

This study was reported in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guideline, with the completed checklist provided as Checklist 1.

Results

Study Cohort

From April 1 to July 2, 2023, all chest and dedicated rib radiographs acquired in the emergency department were automatically processed by the AI system in a parallel workflow, yielding 23,251 imaging studies from 20,908 unique patient visits. Population demographics are summarized in Table 3. The mean age was 55.9 years (SD 22.3; range 0 - 106), with 10,770 (51.5%) male and 10,138 (48.5%) female patients. A radiologist review identified 589 rib-fracture cases (prevalence 2.8%).

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients in the emergency department included in a prospective observational study of artificial intelligence (AI)–assisted rib fracture detection (April 1-July 2, 2023).

Characteristic	Value
Total cases	20,908
Age (y), mean (SD; range)	55.9 (22.3; 0 - 106)
Sex, n (%)	
Male	10,770 (51.5)
Female	10,138 (48.5)
Radiologist-confirmed rib fractures, n (%)	589 (2.8)

AI Model Performance

AI model outputs were compared on a per-case basis against structured rib-fracture labels derived from board-certified radiology reports. At the selected operating point—corresponding to approximately one false-positive per image on the FROC curve—the system achieved a sensitivity of 0.745 (95% CI 0.708 - 0.780) and specificity of 0.933 (95% CI 0.930 - 0.937). Positive predictive value was 0.242 (95% CI 0.223 - 0.262), and negative predictive value was 0.992 (95% CI 0.991 - 0.994). The overall F_1-score was 0.365 (95% CI 0.340 - 0.390) with an accuracy of 0.928 (Table 4).

Table . Case-based diagnostic performance of an artificial intelligence–assisted rib fracture detection system in a prospective observational emergency department study[a].

Metric	Estimate (95% CI)
Sensitivity	0.745 (0.708 - 0.780)
Specificity	0.933 (0.930 - 0.937)
PPV[b]	0.242 (0.223 - 0.262)
NPV[c]	0.992 (0.991 - 0.994)
F_1-score	0.365 (0.340 - 0.390)
Accuracy	0.928 (N/A[d])

[a]Performance metrics are reported with 95% CIs using final radiologist reports as the reference standard.

[b]PPV: positive predictive value.

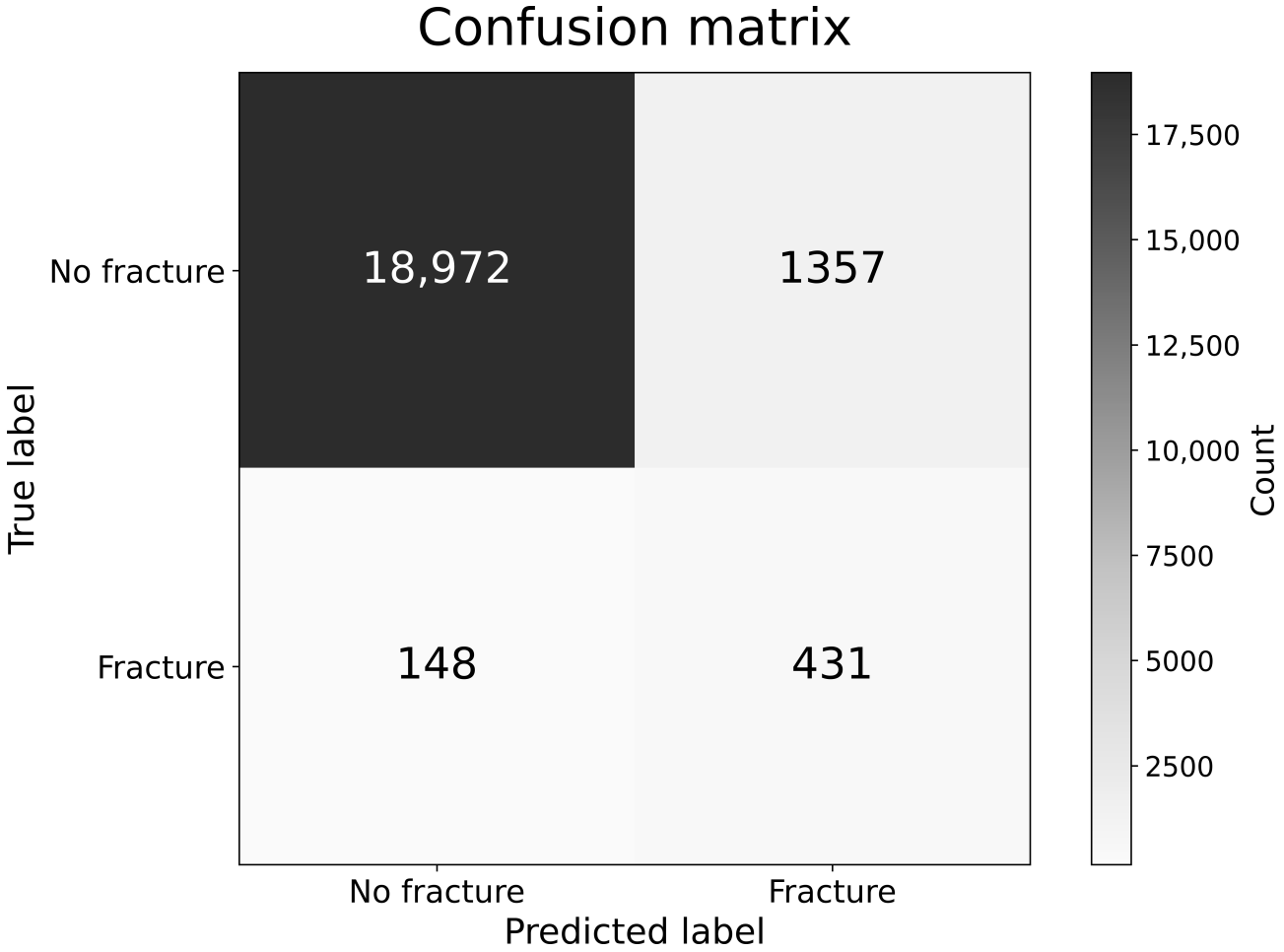
[c]NPV: negative predictive value.

[d]N/A: not available.

As shown in Figure 3, the AI system correctly identified 431 (74.5%) fracture-positive cases while producing 1357 (6.1%) false positives and 148 (0.7%) false negatives across 23,251 studies. This distribution demonstrates the model's high true-negative count (n=18,972, 93.3%) and its strong negative predictive value during deployment in the emergency department. No temporal drift or learning curve effects were observed, as the deployed model remained fixed throughout the study period.

XSL•FO
RenderX

Figure 3. Case-level confusion matrix of the artificial intelligence (AI)–assisted rib fracture detection system during prospective emergency department deployment, using final radiologist reports as the reference standard. Darker blue indicates a higher number of cases (count), as shown in the color bar.

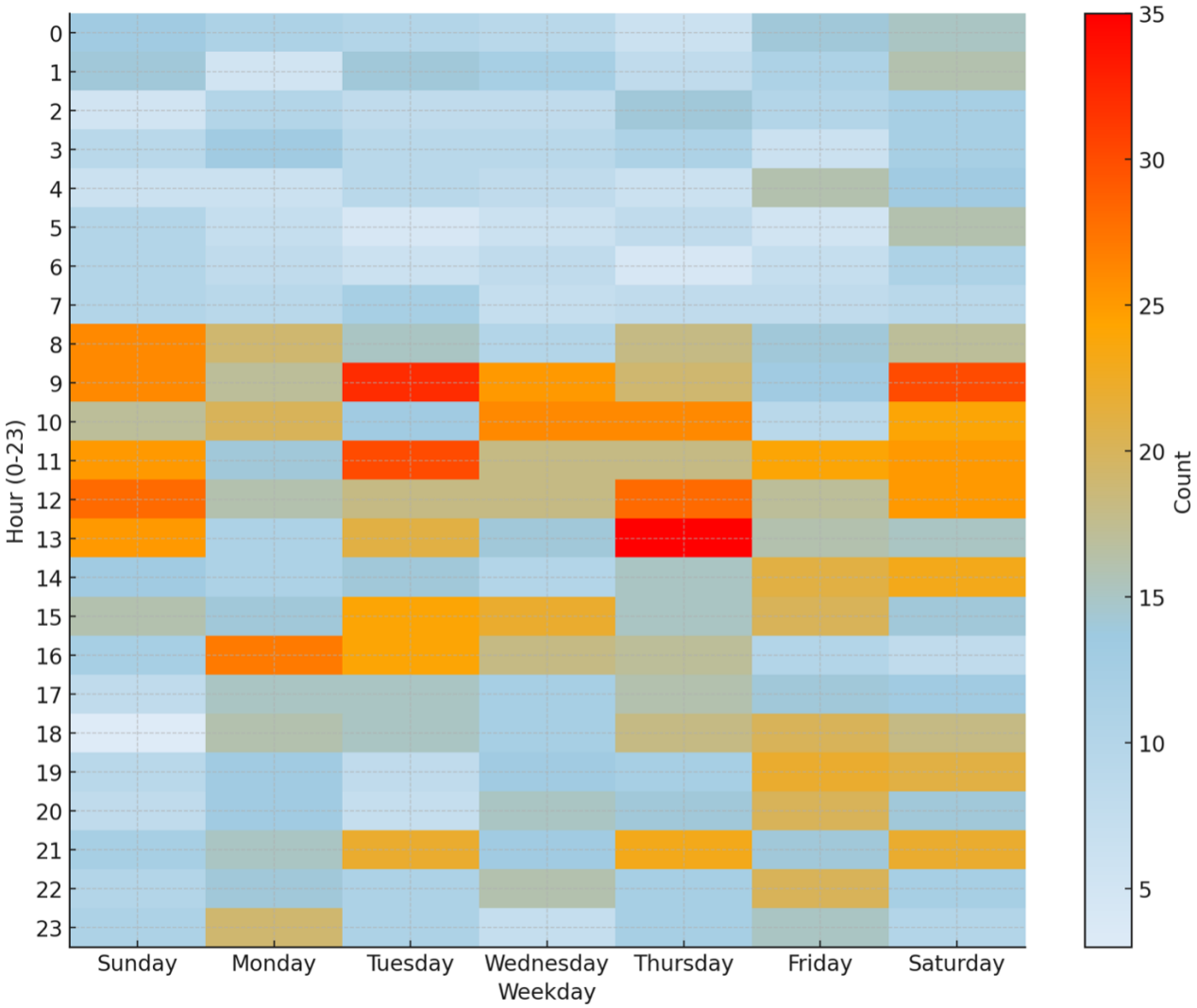


Imaging Workload Patterns

Analysis of imaging demand revealed predictable diurnal and weekly patterns, with peak volumes between 08:00 and 16:00 daily and secondary surges on Thursday to Saturday evenings. Demand was the lowest between 00:00 and 07:00 across all days of the week (Figure 4).

XSL•FO
RenderX

Figure 4. Heatmap illustrating the temporal distribution of chest radiograph imaging workload by hour of day and weekday during prospective emergency department deployment. Color intensity represents the number of chest radiographs acquired per hour, highlighting peak imaging periods across weekdays and weekends.



Inference Turnaround Time

A total of 19,641 paired cases were included to compare AI inference and radiologist report turnaround times. As shown in Table 5, the AI system achieved a median processing time of 10.6 seconds per image (IQR 9.0 - 14.0; range 3 - 35 s), compared with a median of 3.3 hours (IQR 1.31 - 4.80; range 0.08 - 72 h) for radiologist reports. This represents a more than 1000-fold reduction in turnaround time. Figure 5 illustrates this disparity using boxplots on a logarithmic scale. A paired Wilcoxon signed-rank test confirmed that AI inference was significantly faster than radiologist reporting (W=112,987.5; P<.001). This median reporting time reflects the full clinical workflow, including overnight and backlog delays typical of high-volume emergency radiology settings.

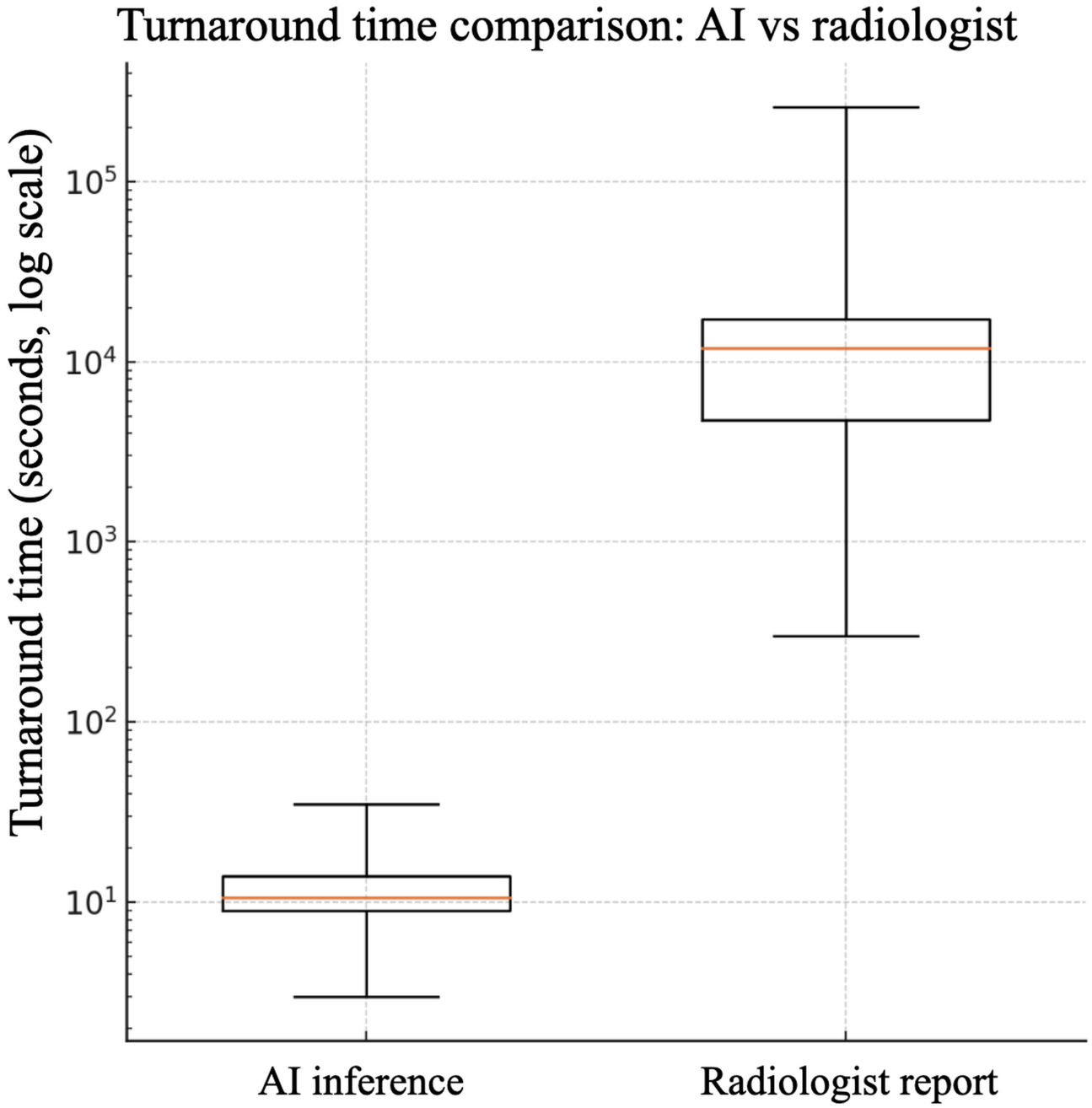
Table . Turnaround times for artificial intelligence (AI) inference versus radiologist reporting during prospective emergency department deployment (n=19,641)[a].

Metric	AI inference time (s)	Radiologist report time (s)
Mean	10.9	10,877
Standard deviation	3.0	4008
Median (50%)	10.6	11,880
IQR (25% - 75%)	9.0 - 14.0	4728 - 17,280
Min	3.0	300
Max	35.0	259,200

[a]Times are summarized using descriptive statistics, including median and interquartile range.

XSL•FO
RenderX

Figure 5. Comparison of turnaround times between artificial intelligence (AI) inference and radiologist reporting during prospective emergency department deployment (April 1-July 2, 2023). Boxplots on a logarithmic scale illustrate differences in processing time distributions between real-time AI inference and routine clinical reporting.



This processing speed highlights the potential of AI-assisted triage systems to complement radiology workflows by rapidly identifying cases for prioritized review, especially in high-volume emergency settings.

System Reliability

During the prospective evaluation of the AI system operating in a parallel clinical workflow, the AI platform experienced 1 service interruption—from 27 April to 10 May 2023—caused by GPU overload that halted all inference operations. A total of 3610 studies acquired during this 14-day outage were excluded from the turnaround time analysis but retained in the diagnostic accuracy evaluation (their formal radiology reports remained available). Following hardware replacement, full functionality was restored on 11 May, and no further outages occurred over the remainder of the study period.

Illustrative Review of Discordant Cases

To further evaluate the AI system's potential diagnostic value beyond report-based benchmarking, a targeted adjudication was conducted on 11 representative discordant cases in which the AI system flagged suspected rib fractures not described in the corresponding radiology reports (Table 6). Among these, 7 cases (cases 1 - 7) were subsequently confirmed as true fractures on CT ("AI-CT concordant"), indicating that several AI-labeled false positives in the report-based analysis represented true fractures missed in the reference standard. The remaining 4 cases (cases 8 - 11) were confirmed negative on CT, primarily attributable to nonfracture anatomical structures or imaging

XSL•FO
RenderX

artifacts. Including these CT-confirmed fractures as true positives would modestly increase the model's effective sensitivity and positive predictive value, highlighting the underestimation inherent in report-based benchmarking.

Table . Targeted post hoc computed tomography (CT) adjudication of representative discordant cases in which artificial intelligence (AI) flagged suspected rib fractures not described in the corresponding radiology reports during prospective emergency department deployment.

Case	AI output	Radiologist report	Emergency physician	Outcome	Note (AI significance)
1	Flagged right rib fracture	No fracture	Noted with POCUS[a]	CT-confirmed	Triage value, prompting clinicians to perform US[b]
2	Flagged right rib fracture	No fracture	Missed	CT-confirmed	AI-CT concordance
3	Flagged left fifth rib fracture	No fracture	Missed	CT-confirmed	AI-CT concordance
4	Flagged rib fracture post chest tube	No fracture	Missed	CT-confirmed	Chest tube artifact did not impair detection
5	Flagged lower-rib fracture	No fracture	Missed	CT-confirmed	AI-CT concordance
6	Flagged fracture near hardware	No fracture	Missed	CT-confirmed	Detected fracture adjacent to surgical hardware
7	Flagged fracture under scapular shadow	No fracture	Missed	CT-confirmed	AI-CT concordance
8	Flagged fracture at scapula border	No fracture	No fracture	False positive	Scapular margin misidentified
9	Flagged rib fracture at bra clasp	No fracture	No fracture	False positive	Bra hardware artifact
10	Flagged fracture at chest tube marker	No fracture	No fracture	False positive	Chest tube marker misinterpreted
11	Flagged fracture (image noise)	No fracture	No fracture	False positive	Image noise

[a]POCUS: point-of-care ultrasound.

[b]US: ultrasound.

In one representative case (Case 3), the AI system correctly identified a subtle nondisplaced fracture of the left fifth rib that was not documented in the radiology report but later verified on 3D CT reconstruction (Figure 6). In contrast, Figure 7 illustrates the main sources of false positives, including scapular margin misinterpretation, chest-tube hardware artifacts, and motion-induced noise.

XSL•FO
RenderX

Figure 6. Representative true-positive rib fracture detected by artificial intelligence (AI) and confirmed by computed tomography (CT) during prospective emergency department deployment. (A) Chest radiograph showing an AI-flagged fracture of the left fifth rib that was not described in the initial radiology report. (B) Corresponding CT image confirming the fracture at the same anatomical location.

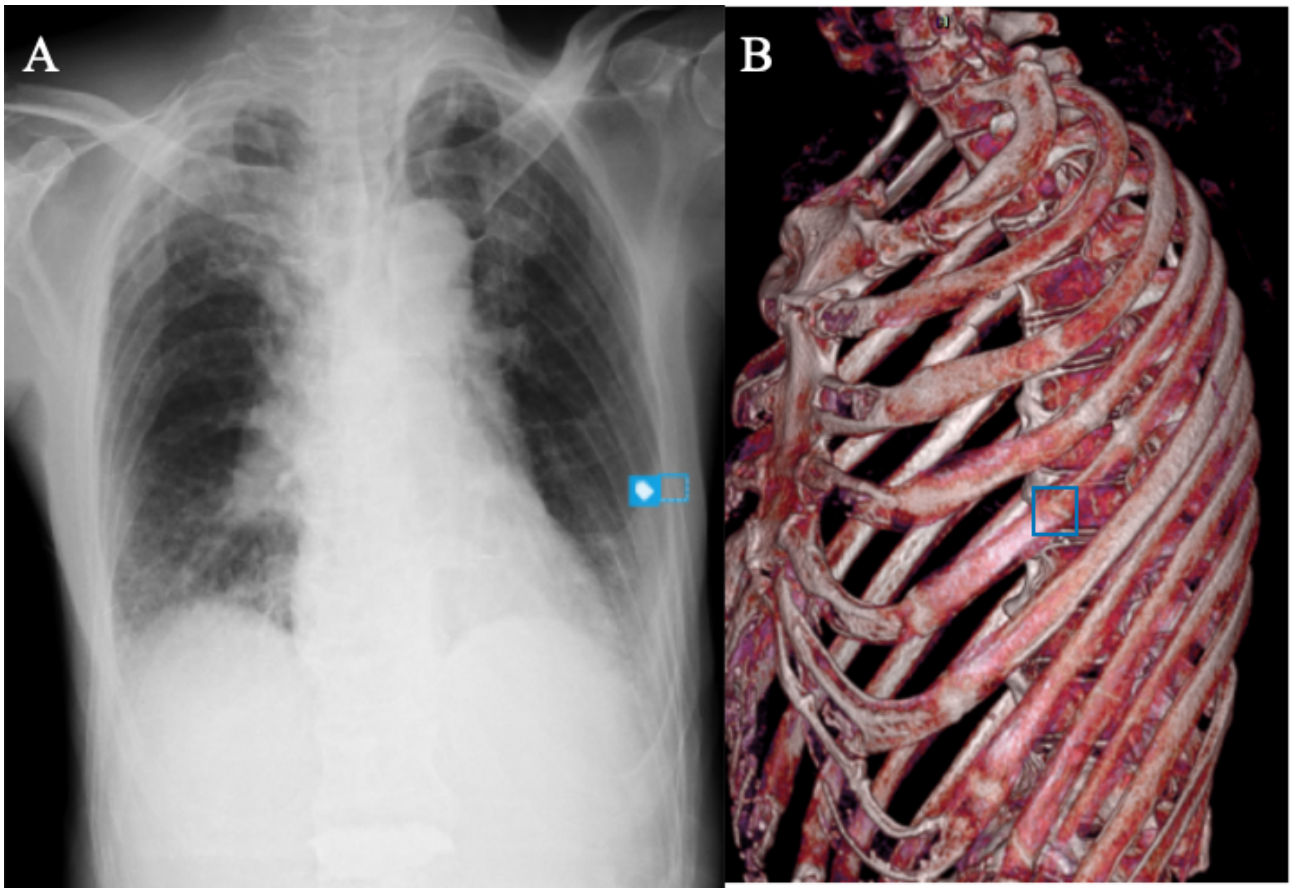
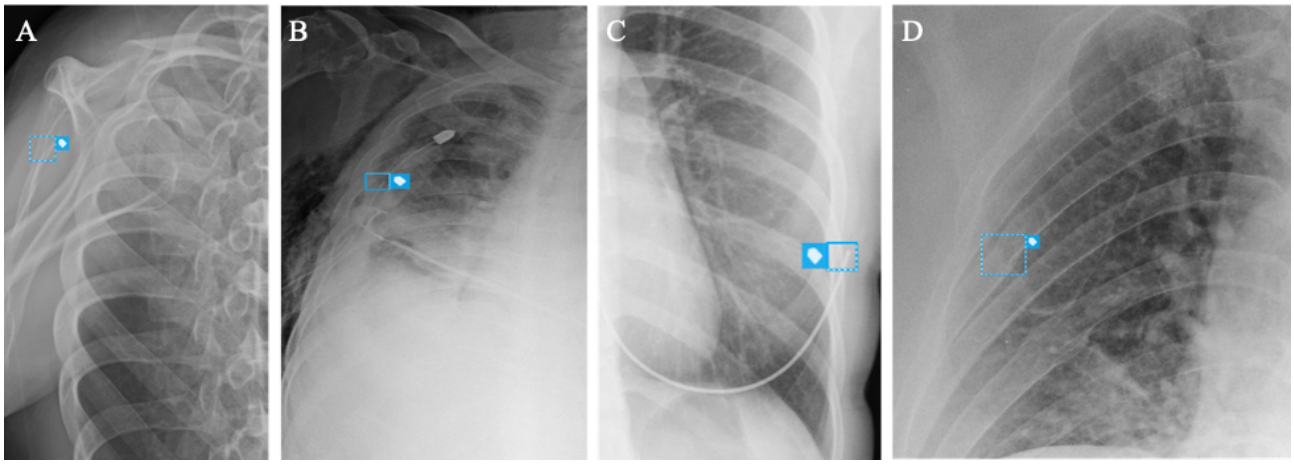


Figure 7. Representative false-positive detections generated by an artificial intelligence (AI)–assisted rib fracture detection system during prospective emergency department deployment. The examples illustrate common sources of false-positive signals on chest radiographs. (A) Scapular margin overlap misinterpreted as a rib fracture. (B) Chest tube marker misidentified as a rib discontinuity. (C) Bra hardware producing a linear opacity mimicking a fracture. (D) Image noise and low-contrast regions leading to spurious detection.



This targeted CT adjudication underscores the potential of AI-assisted screening to augment clinical vigilance by identifying subtle or overlooked fractures, while also emphasizing the need to improve artifact robustness and optimize false-positive suppression for practical clinical integration.

Common Pitfalls in Frontline Rib Fracture Detection

Our review of discordant cases identified 3 principal drivers of missed rib fractures by emergency physicians: first, non-thoracic presenting complaints (eg, catheter malfunction or abdominal pain) led interpreters to focus on unrelated findings and overlook subtle rib breaks; second, the absence of classic chest pain—patients describing only mild discomfort or a vague “pop”—lowered clinical vigilance for nondisplaced fractures; and third, competing urgent injuries (facial, limb, or soft-tissue

XSL•FO
RenderX

trauma) diverted attention from the chest, resulting in underappreciated fractures.

Discussion

Principal Findings

In this observational feasibility study using real-world emergency department imaging data, we demonstrated that a faster R-CNN–based AI system can operate in parallel with routine clinical workflows to provide near-instantaneous rib fracture triage without influencing patient care. During a 3-month evaluation period, the model automatically processed 23,251 CXRs with a median inference time of 10.6 seconds per image, achieving a turnaround time reduction exceeding 3 orders of magnitude compared with formal radiologist reporting, while maintaining 74.5% sensitivity and 93.3% specificity. These findings position AI as a potential automated screening aid capable of rapidly identifying low-risk examinations and generating signals that could inform future prioritization strategies. The observed discrepancy between the 10.6-second AI inference time and the 3.3-hour radiologist turnaround time reflects a critical clinical bottleneck in busy emergency departments. Although these metrics represent different stages—technical processing versus final clinical documentation—the delay in official reporting highlights the diagnostic gap AI aims to address. In this context, near-instantaneous AI alerts may support case prioritization before formal reporting, although no clinician-facing alerts were implemented in this study.

However, the current findings primarily demonstrate technical feasibility rather than full clinical feasibility, as the AI system operated passively without direct clinician interaction. Future clinician-in-the-loop evaluations will be necessary to assess workflow integration, usability, and impact on diagnostic behavior or patient outcomes.

Although the system's positive predictive value was relatively low (24.2%), this trade-off aligns with its design as a triage support tool rather than a stand-alone diagnostic system. In high-volume emergency departments, the ability to rapidly identify examinations with a low likelihood of fracture is crucial. The model's high NPV of 99.2% allows clinicians to focus on a smaller, higher-risk subset of cases, thereby improving efficiency and reducing cognitive load. Given the observed sensitivity of 74.5%, false-negative cases remain possible, and the system should not be used as a stand-alone rule-out mechanism or as a substitute for clinical judgment.

Comparison to Prior Work

Similar findings have been reported by Yao et al [20], who demonstrated that deep learning systems with high NPV on chest CT can reduce radiologists' workload by effectively identifying nonfracture cases. A recent systematic review by van den Broek et al [21] further emphasized the triage value of AI in fracture detection, underscoring its potential across multiple imaging modalities.

Our AI system also demonstrated robust performance across circadian and weekly imaging surges, with peak volumes observed between 08:00 and 16:00 and during Thursday to Saturday evenings. However, a 14-day GPU hardware outage during the study period highlighted a real-world challenge of maintaining the AI system's reliability in clinical environments. This incident underscores the need for infrastructure redundancy, real-time monitoring, and failover protocols—key considerations for sustainable AI deployment. These practical aspects of AI deployment remain underreported in most published studies [22].

Compared with prior approaches such as the PACS-AI platform [23], our system offered full automation, operating continuously in real time and without the need for manual image selection. This better reflects the demands of frontline emergency radiology. Herpe et al [24] demonstrated improved diagnostic accuracy with PACS-integrated AI for limb fractures; however, their study did not evaluate scalability or autonomous triage capability under high-throughput conditions. In contrast, our study incorporated prospective data collection within a real-world emergency department workflow, allowing the assessment of AI performance, reliability, and operational feasibility under authentic clinical conditions.

Focused adjudication of discordant cases revealed that the AI system correctly identified rib fractures that were missed by both radiologists and emergency physicians in 7 cases, all subsequently confirmed on follow-up CT ("AI-CT concordant"). These findings highlight the potential of AI to strengthen diagnostic vigilance in complex clinical scenarios. While prior studies—such as Zhou et al [25]—have shown that AI can detect rib fractures overlooked in initial CT interpretations, with confirmation on follow-up imaging, these investigations have largely centered on CT-based workflows. Although CT is highly sensitive, its routine use is limited by concerns over radiation, cost, and logistics. In contrast, our CXR-focused approach targets the most widely used imaging modality in acute care, offering a more scalable and practical solution for real-world emergency triage. Notably, Brady et al [26] have emphasized that diagnostic errors and discrepancies are not uncommon in radiology, with daily error rates estimated at 3% - 5%, reinforcing the importance of AI as a complementary tool to enhance diagnostic accuracy.

Four false-positive cases revealed predictable pitfalls, including the misinterpretation of scapular margins, chest-tube hardware, and motion artifacts. Similar findings were reported by Sun et al [27], who noted frequent false positives in an AI model for rib fracture detection on CXR, often due to anatomical overlap and imaging artifacts. These results support the need for artifact-aware retraining and preprocessing optimization to reduce false alerts and improve clinical integration.

This targeted CT adjudication further highlights the complementary role of AI in identifying subtle or overlooked fractures and underscores the inherent limitation of using report-based labels as the reference standard in real-world studies.

Future Directions

In addition, integrating AI-generated alerts into emergency radiology workflows will require careful calibration of alert thresholds to minimize false positives and prevent alert fatigue

XSL•FO
RenderX

among clinicians. Human-centered design, interface refinement, and iterative feedback from end users will be critical to achieving effective and sustainable adoption.

While most prior prospective studies have emphasized diagnostic performance or radiologist feedback, our findings extend beyond these metrics to include diagnostic efficacy, operational resilience, and system failure contingencies. These real-world insights support the feasibility and clinical value of embedding AI into routine emergency department workflows. Recent work has highlighted the importance of not only measuring accuracy but also assessing robustness across patient and workflow variability [28]. Furthermore, the need for deployment frameworks that address hardware resilience, continuous quality monitoring, and interpretability safeguards is increasingly recognized as essential for sustainable AI adoption in high-acuity settings [29].

Recent reports and position statements have highlighted a persistent gap between the promising diagnostic performance of AI systems and their limited demonstrated clinical benefit. Robust, prospective, and randomized clinical studies remain urgently needed to justify large-scale implementation [30,31]. Even high-performing AI models (area under the curve≈0.85) have failed to surpass standard clinical practice in improving patient outcomes [32,33]. These findings reinforce ongoing concerns that most AI or machine learning devices, despite regulatory authorization, are primarily validated using retrospective data and therefore remain susceptible to selection bias, distributional shift, and overestimation of generalizability [34,35].

Clinical decision-making in emergency care is inherently multimodal: physicians integrate imaging findings with the mechanism of injury, examination, and vital signs to guide judgment. In contrast, this AI system analyzes images in isolation and is designed not to replace but to support clinicians as a rapid screening aid—enhancing vigilance in high-volume, high-pressure environments where missed fractures may occur. Incorporating multimodal clinical data in future models could further improve diagnostic relevance and workflow integration.

Although this was an observational feasibility study, it represents one of the largest evaluations of an AI-assisted rib fracture detection system in real-world emergency radiology. The findings demonstrate that such a system can provide meaningful diagnostic support, maintain consistent performance at scale, and potentially enhance patient safety. Future implementation should therefore shift from technical to clinical feasibility, focusing on clinician-in-the-loop impact studies, PACS-integrated trials, and workflow efficiency assessments. Although user perception was not formally assessed, informal feedback from emergency physicians indicated strong interest in AI-supported flagging—particularly for subtle fractures and during periods of high patient volume.

Future research should prioritize prospective, multicenter studies to validate generalizability and quantify AI impact on workflow, resource utilization, and patient outcomes. Model improvements—including artifact-aware retraining, expanded fracture coverage in challenging scenarios such as subtle or anatomically obscured fractures, and continuous learning—will be critical to enhance diagnostic precision. Finally, building infrastructure resilience and integrating effective alert management into radiologist workflows are essential for sustainable clinical adoption.

Limitations

First, this single-center observational study may limit generalizability to other institutions with different imaging protocols, patient populations, or workflow environments. Second, during retrospective model development, training and validation were performed at the image level rather than the patient level. This may have resulted in optimistic internal validation estimates due to potential within-patient similarity, which likely contributes to the observed performance gap between retrospective validation and prospective real-world deployment. Additionally, stratified performance analyses by age group and sex were not performed due to the low prevalence of rib fractures in certain subgroups, particularly pediatric patients. Similarly, the small fraction of oblique views (approximately 1.3%) prevented a dedicated analysis by imaging view, as the limited sample size would yield unstable estimates for these specific cohorts.

The prospective evaluation period also did not include winter months. Seasonal variation in trauma mechanisms or imaging artifacts may influence fracture detectability in certain settings; therefore, caution is warranted when generalizing these findings across different seasonal contexts.

Additionally, the use of a 512×512 resolution for model inference represents a technical trade-off; while it facilitates rapid processing, the associated downsampling may limit the detection of very subtle cortical disruptions.

Third, using radiology reports as the reference standard—while pragmatic—may underestimate the AI system's true performance, as subtle or occult fractures can be underreported in clinical practice. A focused CT review of representative discordant cases further supported this concern, revealing instances where AI-predicted fractures were subsequently confirmed as true fractures on CT. This approach likely yielded conservative performance estimates, since NLP-derived labels may not capture subtle fractures identified by AI or CT.

Regarding the study design, although post hoc adjudication is generally more appropriate for hypothesis generation than for definitive performance reassessment, modifying the reference standard after study completion may introduce bias. Accordingly, in this study, performance evaluation was anchored to the contemporaneous clinical reference standard used in routine practice. More comprehensive adjudication strategies—such as consensus radiologist review of AI-positive, report-negative cases—may provide additional insights when implemented within a separately designed study.

Finally, because AI predictions were not disclosed to clinicians, we did not assess downstream clinical outcomes, including changes in diagnostic behavior, time to intervention, or patient management. As for the data pipeline, although we used NLP to extract rib fracture labels from radiology reports, which may introduce misclassification in ambiguous cases, the pipeline demonstrated high agreement with manual review (κ=0.91).

XSL•FO
RenderX

Given the observed 3.5% discrepancy rate, any residual label noise propagating through the large-scale dataset may introduce modest uncertainty into performance estimates. However, the substantial sample size (n=23,251) is expected to attenuate the impact of such noise, supporting the stability of the resulting confidence intervals for large-scale clinical benchmarking.

Conclusions

In this observational feasibility study, we evaluated a faster R-CNN–based AI system deployed in parallel with clinical workflows to automatically detect rib fractures on CXRs using real-world emergency department data. Although AI outputs were not visible to clinicians, the system processed over 23,000 studies with high throughput, achieving 74.5% sensitivity and 93.3% specificity and delivering results within seconds—over 1000 times faster than formal radiologist reports.

These findings demonstrate strong technical feasibility of real-time AI-assisted rib fracture detection in emergency radiology. While clinical decisions remained unaffected during this observational phase, future studies should validate clinical feasibility through clinician-in-the-loop evaluation, PACS integration, and workflow optimization to address potential alert fatigue and false-positive management.

Acknowledgments

We thank the emergency department, radiology department, and information technology teams for their support during system deployment and data extraction. Generative artificial intelligence tools were used solely for language editing and grammar checking. All scientific content, analyses, interpretations, and conclusions were developed by the authors. The artificial intelligence system evaluated in this study was deployed on the DeepQ AI platform as a technical deployment environment. The authors have no financial, employment, or equity relationship with DeepQ Technology or HTC, and the platform was used solely for noncommercial research purposes.

Funding

This research was financially supported by National Science and Technology Council, Taiwan, under Grant No. NSTC113-2221-E-038 - 006

Data Availability

The imaging data analyzed in this study are not publicly available due to institutional regulations and patient privacy protections. Deidentified data may be made available from the corresponding author upon reasonable request and subject to institutional review board approval.

Authors' Contributions

Conceptualization: H-WC, S-TH
Methodology: H-WC, S-TH
Data curation: M-YH, S-TH
Formal analysis: M-YH, S-TH
Investigation: M-YH, S-TH
Writing – original draft: M-YH, S-TH
Writing – review & editing: H-WC, L-RL, M-FT, M-YH, S-TH
Supervision: CH-W

Conflicts of Interest

None declared.

Checklist 1
CLAIM (Checklist for Artificial Intelligence in Medical Imaging).
[PDF File, 3 KB - medinform_v14i1e77965_app1.pdf]

References

1. Tignanelli CJ, Rix A, Napolitano LM, Hemmila MR, Ma S, Kummerfeld E. Association between adherence to evidence-based practices for treatment of patients with traumatic rib fractures and mortality rates among US trauma centers. JAMA Netw Open 2020 Mar 2;3(3):e201316. [doi: 10.1001/jamanetworkopen.2020.1316] [Medline: 32215632]
2. Colling KP, Goettl T, Harry ML. Outcomes after rib fractures: more complex than a single number. J Trauma Inj 2022 Dec;35(4):268-276. [doi: 10.20408/jti.2021.0096] [Medline: 39380934]
3. Sharma OP, Oswanski MF, Jolly S, Lauer SK, Dressel R, Stombaugh HA. Perils of rib fractures. Am Surg 2008 Apr;74(4):310-314. [doi: 10.1177/000313480807400406] [Medline: 18453294]

XSL•FO
RenderX

4. Griffith JF, Rainer TH, Ching AS, Law KL, Cocks RA, Metreweli C. Sonography compared with radiography in revealing acute rib fracture. AJR Am J Roentgenol 1999 Dec;173(6):1603-1609. [doi: 10.2214/ajr.173.6.10584808] [Medline: 10584808]
5. Tomas X, Facenda C, Vaz N, et al. Thoracic wall trauma-misdiagnosed lesions on radiographs and usefulness of ultrasound, multidetector computed tomography and magnetic resonance imaging. Quant Imaging Med Surg 2017 Aug;7(4):384-397. [doi: 10.21037/qims.2017.08.02] [Medline: 28932697]
6. Chapman BC, Overbey DM, Tesfalidet F, et al. Clinical utility of chest computed tomography in patients with rib fractures CT chest and rib fractures. Arch Trauma Res 2016 Dec;5(4):e37070. [doi: 10.5812/atr.37070] [Medline: 28144607]
7. Huang ST, Liu LR, Chiu HW, Huang MY, Tsai MF. Deep convolutional neural network for rib fracture recognition on chest radiographs. Front Med 2023;10. [doi: 10.3389/fmed.2023.1178798]
8. Pishbin E, Ahmadi K, Foogardi M, Salehi M, Seilanian Toosi F, Rahimi-Movaghar V. Comparison of ultrasonography and radiography in diagnosis of rib fractures. Chin J Traumatol 2017 Aug;20(4):226-228. [doi: 10.1016/j.cjtee.2016.04.010] [Medline: 28687342]
9. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
10. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv. Preprint posted online on Nov 14, 2017. [doi: 10.48550/arXiv.1711.05225]
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
12. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
13. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019 May;29(2):102-127. [doi: 10.1016/j.zemedi.2018.11.002] [Medline: 30553609]
14. Kora P, Ooi CP, Faust O, et al. Transfer learning techniques for medical image analysis: a review. Biocybern Biomed Eng 2022 Jan;42(1):79-107. [doi: 10.1016/j.bbe.2021.11.004]
15. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017 Dec;88(6):581-586. [doi: 10.1080/17453674.2017.1344459] [Medline: 28681679]
16. Wu J, Liu N, Li X, et al. Convolutional neural network for detecting rib fractures on chest radiographs: a feasibility study. BMC Med Imaging 2023 Jan 30;23(1):18. [doi: 10.1186/s12880-023-00975-x] [Medline: 36717773]
17. Lee K, Lee S, Kwak JS, Park H, Oh H, Koh JC. Development and validation of an artificial intelligence model for detecting rib fractures on chest radiographs. J Clin Med 2024 Jun 30;13(13):3850. [doi: 10.3390/jcm13133850] [Medline: 38999416]
18. Chang EY. DeepQ: advancing healthcare through artificial intelligence and virtual reality. Presented at: MM '17: Proceedings of the 25th ACM international conference on Multimedia New York; Oct 23-27, 2017. [doi: 10.1145/3123266.3130875]
19. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv. Preprint posted online on Jun 4, 2015. [doi: 10.48550/arXiv.1506.01497]
20. Yao L, Guan X, Song X, et al. Rib fracture detection system based on deep learning. Sci Rep 2021 Dec 6;11(1):23513. [doi: 10.1038/s41598-021-03002-7] [Medline: 34873241]
21. van den Broek MCL, Buijs JH, Schmitz LFM, Wijffels MME. Diagnostic performance of artificial intelligence in rib fracture detection: systematic review and meta-analysis. Surgeries 2024;5(1):24-36. [doi: 10.3390/surgeries5010005]
22. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
23. Theriault-Lauzier P, Cobin D, Tastet O, et al. A responsible framework for applying artificial intelligence on medical images and signals at the point of care: the PACS-AI platform. Can J Cardiol 2024 Oct;40(10):1828-1840. [doi: 10.1016/j.cjca.2024.05.025] [Medline: 38885787]
24. Herpe G, Nelken H, Vendeuvre T, et al. Effectiveness of an artificial intelligence software for limb radiographic fracture recognition in an emergency department. J Clin Med 2024 Sep 20;13(18):5575. [doi: 10.3390/jcm13185575] [Medline: 39337062]
25. Zhou Q, Qin P, Luo J, et al. Evaluating AI rib fracture detections using follow-up CT scans. Am J Emerg Med 2023 Oct;72:34-38. [doi: 10.1016/j.ajem.2023.07.018] [Medline: 37478635]
26. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017 Feb;8(1):171-182. [doi: 10.1007/s13244-016-0534-1] [Medline: 27928712]
27. Sun H, Wang X, Li Z, et al. Automated rib fracture detection on chest X-ray using contrastive learning. J Digit Imaging 2023 Oct;36(5):2138-2147. [doi: 10.1007/s10278-023-00868-z] [Medline: 37407842]
28. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS ONE 2018;13(7):e0201016. [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
29. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health 2024;12:1342937. [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]

XSL•FO
RenderX

30. Khera R, Butte AJ, Berkwits M, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023 Sep 5;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
31. Armoundas AA, Narayan SM, Arnett DK, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. Circulation 2024 Apr 2;149(14):e1028-e1050. [doi: 10.1161/CIR.0000000000001201] [Medline: 38415358]
32. Mazor T, Farhat KS, Trukhanov P, et al. Clinical trial notifications triggered by artificial intelligence-detected cancer progression: a randomized trial. JAMA Netw Open 2025 Apr 1;8(4):e252013. [doi: 10.1001/jamanetworkopen.2025.2013] [Medline: 40257799]
33. Mandair D, Elia MV, Hong JC. Considerations in translating AI to improve care. JAMA Netw Open 2025 Apr 1;8(4):e252023. [doi: 10.1001/jamanetworkopen.2025.2023] [Medline: 40257804]
34. Habib AR, Gross CP. FDA regulations of AI-driven clinical decision support devices fall short. JAMA Intern Med 2023 Dec 1;183(12):1401-1402. [doi: 10.1001/jamainternmed.2023.5006] [Medline: 37812411]
35. Chouffani El Fassi S, Abdullah A, Fang Y, et al. Not all AI health tools with regulatory authorization are clinically validated. Nat Med 2024 Oct;30(10):2718-2720. [doi: 10.1038/s41591-024-03203-3] [Medline: 39187696]

Abbreviations

AI: artificial intelligence
AP: anteroposterior
CLAIM: Checklist for Artificial Intelligence in Medical Imaging
CT: computed tomography
CXR: chest radiograph
FROC: free-response receiver operating characteristic
GPU: graphics processing unit
NLP: natural language processing
NPV: negative predictive value
PACS: picture archiving and communication system
R-CNN: region-based convolutional neural network

Edited by A Coristine, A Benis; submitted 22.May.2025; peer-reviewed by D Yoo, E Bai, J Zhang; revised version received 30.Dec.2025; accepted 01.Jan.2026; published 29.Jan.2026.

Please cite as:
Huang ST, Liu LR, Tsai MF, Huang MY, Chiu HW
Prospective Diagnostic Accuracy and Technical Feasibility of Artificial Intelligence-Assisted Rib Fracture Detection on Chest Radiographs: Observational Study
JMIR Med Inform 2026;14:e77965
URL: https://medinform.jmir.org/2026/1/e77965
doi:10.2196/77965

© Shu-Tien Huang, Liong-Rung Liu, Ming-Feng Tsai, Ming-Yuan Huang, Hung-Wen Chiu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning Prediction of Pharmacogenetic Testing Uptake Among Opioid-Prescribed Patients Using Electronic Health Records: Retrospective Cohort Study

Mohammad Yaseliani[1], MSc; Je-Won Hong[2], PharmD; Jiang Bian[3,4], MS, PhD; Larisa Cavallari[2,5], PharmD; Julio D Duarte[2,5], PharmD, PhD; Danielle Nelson[6], MPH, MD; Wei-Hsuan Lo-Ciganic[7,8,9], MSPharm, MS, PhD; Khoa Anh Nguyen[2*], PharmD; Md Mahmudul Hasan[1,10*], PhD

[1]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

[10]Department of Information Systems and Operations Management, Warrington College of Business, University of Florida, Gainesville, FL, United States

[2]Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, United States

[3]Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN, United States

[4]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States

[5]Center for Pharmacogenomics and Precision Medicine, University of Florida, Gainesville, FL, United States

[6]College of Medicine, University of Florida, Gainesville, FL, United States

[7]Division of General Internal Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[8]Center for Pharmaceutical Policy and Prescribing, University of Pittsburgh, Pittsburgh, PA, United States

[9]Geriatric Research Education and Clinical Center, North Florida/South Georgia Veterans Health System, Gainesville, FL, United States

[*]these authors contributed equally

Corresponding Author:
Md Mahmudul Hasan, PhD
Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

Abstract

Background: Opioids are a widely prescribed class of medication for pain management. However, they have variable efficacy and adverse effects among patients, due to the complex interplay between biological and clinical factors. Pharmacogenetic testing can be used to match patients' genetic profiles to individualize opioid therapy, improving pain relief and reducing the risk of adverse effects. Despite its potential, the pharmacogenetic testing uptake (use of pharmacogenetic testing) remains low due to a range of barriers at the patient, health care provider, infrastructure, and financial levels. Since testing typically involves a shared decision between the provider and patient, predicting the likelihood of a patient undergoing pharmacogenetic testing and understanding the factors influencing that decision can help optimize resource use and improve outcomes in pain management.

Objective: This study aimed to develop machine learning (ML) models, identifying patients' likelihood of pharmacogenetic uptake based on their demographics, clinical variables, medication use, and social determinants of health.

Methods: We used electronic health record data from a single center health care system to identify patients prescribed opioids. We extracted patients' demographics, clinical variables, medication use, and social determinants of health, and developed and validated ML models, including a neural network, logistic regression, random forest, extreme gradient boosting (XGB), naïve Bayes, and support vector machines for pharmacogenetic testing uptake prediction based on procedure codes. We performed 5-fold cross-validation and created an ensemble probability-based classifier using the best-performing ML models for pharmacogenetic testing uptake prediction. Various performance metrics, uptake stratification analysis, and feature importance analysis were used to evaluate the performance of the models.

Results: The ensemble model using XGB and support vector machine–radial basis function classifiers had the highest *C*-statistics at 79.61%, followed by XGB (78.94%), and neural network (78.05%). While XGB was the best-performing model, the ensemble model achieved a high accuracy (32,699/48,528, 67.38%), recall (537/702, 76.50%), specificity (32,162/47,826, 67.25%), and negative predictive value (32,162/32,327, 99.49%). The uptake stratification analysis using the ensemble model indicated that it can effectively distinguish across uptake probability deciles, where those in the higher strata are more likely to undergo pharmacogenetic testing in the real world (320/4853, 6.59% in the highest decile compared to 6/4853, 0.12% in the lowest).

XSL•FO
RenderX

Furthermore, Shapley Additive Explanations value analysis using the XGB model indicated age, hypertension, and household income as the most influential factors for pharmacogenetic testing uptake prediction.

Conclusions: The proposed ensemble model demonstrated a high performance in pharmacogenetic testing uptake prediction among patients using opioids for pain. This model can be used as a decision support tool, assisting clinicians in identifying patients' likelihood of pharmacogenetic testing uptake and guiding appropriate decision-making.

(JMIR Med Inform 2026;14:e81048) doi:10.2196/81048

KEYWORDS

ensemble model; ML; machine learning; opioids; opioid prescription; pain management; pharmacogenetic testing

Introduction

Opioid prescribing presents a complex therapeutic challenge due to variable efficacy and the risk of adverse effects, often resulting in a trial and error approach for prescription. These variations in opioid response often arise from interindividual differences in pharmacokinetics and pharmacodynamics, influenced in part by genetic variations [1]. Over recent decades, pharmacogenetic testing has emerged as a promising strategy to tailor opioid therapy to an individual genetic profile, with the goal of enhancing pain relief, minimizing adverse effects, and improving overall patient outcomes. Several opioids, including codeine, tramadol, hydrocodone, and oxycodone, are metabolized to varying extents by cytochrome P450 family 2 subfamily D member 6 (CYP2D6), which is an enzyme involved in the metabolism of many drugs [2]. A pragmatic trial demonstrated improved composite pain intensity outcomes with CYP2D6-guided pain management, highlighting the potential benefits of personalized therapy based on pharmacogenetic testing [2]. Based on the current evidence, the Clinical Pharmacogenetics Implementation Consortium (CPIC) provides guidelines on using CYP2D6 genotype results for prescribing tramadol, hydrocodone, and codeine [2]2 yet their real-world impact depends on whether patients offered pharmacogenetic testing actually complete it. In opioid prescribing contexts, uptake hinges on perceived use and timing, out-of-pocket costs, health literacy, trust, social context, and clinician offering practices, as well as workflow constraints that differ across acute, perioperative, and chronic pain settings.

Electronic health records (EHRs) contain rich information relevant to testing acceptance: prior and current opioid exposures, pain diagnoses, comorbidity burden, care setting, prescriber specialty, monitoring and follow-up patterns, previous laboratory or genetic testing, and engagement with the health system. Machine learning (ML) such as neural network (NN), random forest (RF), and logistic regression (LR) can synthesize these heterogeneous and high-dimensional features [3] to produce calibrated, patient-level probability estimates of pharmacogenetic testing uptake at the moment testing is offered. Such predictions could help health systems, including patient care teams and insurers, tailor communication (eg, language, framing, and channel), select appropriate clinical checkpoints (such as before initiating certain opioids or at dose escalation), and allocate resources where they are most likely to improve test uptake without adding unnecessary burden. These predictions may be especially valuable for patients who are expected to benefit most from testing (though this was not the focus of this study), enabling patient care teams to dedicate additional resources to support uptake in this population.

Prior work has developed ML predictive models to support uptake for vaccines [4], cancer screening [5], and genetic counseling [6]; however, pharmacogenetics-specific models for opioid-related testing are scarce, and few studies evaluate calibration, clinical use, fairness, and workflow integration together. In this study, we aimed to predict the likelihood that patients receiving opioid prescriptions will undergo pharmacogenetic testing after an offer in routine care. We hypothesized that ML models could accurately predict the likelihood of pharmacogenetic testing uptake among opioid-prescribed patients by identifying patterns in demographic, clinical, and health care use factors associated with testing behavior under current practice conditions. Our objectives were to achieve strong discrimination and predictive performance of the ML models, and reliable model calibration through patients' stratification in subgroups based on the predicted probability of test uptake. We emphasized model explainability to support clinician understanding and sought to identify factors influencing the likelihood of pharmacogenetic testing uptake in opioid therapy. Ultimately, we aimed to complement, rather than replace, clinical judgment, potentially facilitating better use of pharmacogenetic testing for precision pain management. By focusing on actionable predictions at the point of offer, this paper may help programs deliver the right message through the right channel at the right time: improving test uptake, informing safer opioid prescribing, and advancing equitable delivery of pharmacogenetic testing.

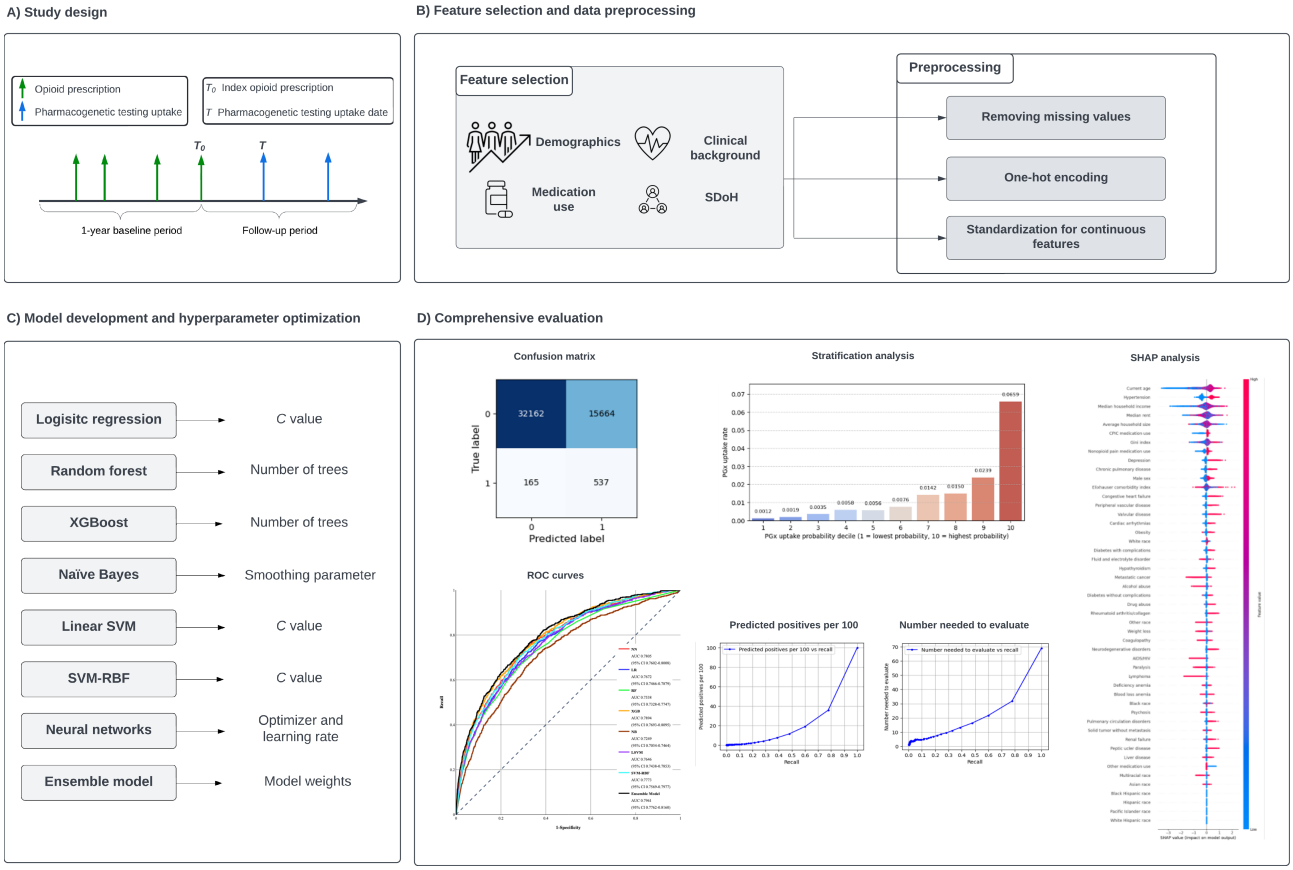
Methods

Overview

The overall ML pipeline for pharmacogenetic testing uptake prediction is shown in Figure 1. The pipeline involves 4 steps, including study design, data preprocessing and feature selection, model development, and comprehensive evaluation. Each step involves several tasks that are critical to building and evaluating the final predictive model. The details of each step are described in this section.

XSL•FO
RenderX

Figure 1. Overall machine learning pipeline for pharmacogenetic testing uptake prediction. CPIC: Clinical Pharmacogenetics Implementation Consortium; ROC: receiver operating characteristic; SDoH: social determinants of health; SHAP: Shapley Additive Explanations; SVM: support vector machine; SVM-RBF: support vector machine–radial basis function; XGBoost: extreme gradient boosting.



Ethical Considerations

We used real-world deidentified EHR data available from the University of Florida Health Integrated Data Repository (IDR) to develop the models. These data included information on patients' demographics, clinical variables, and medication use from 2011 to 2023. The study was conducted in compliance with ethical standards and was approved by the University of Florida Institutional Review Board under the reference number IRB202301927. This study is a secondary analysis of existing deidentified EHR data, and the approved IRB covers secondary analyses without additional consent. Therefore, no informed consent was obtained.

Study Design

Patients' age range was 1 to 89 years, and we did not exclude any patients based on age. Patients were included if they were aged between 18 and 89 years and had an opioid prescription order for noncancer treatment from 2011 to 2023. Patients were stratified into an intervention group (patients with a minimum of one pharmacogenetic testing order) and a control group (those with no history of pharmacogenetic testing order). We used current procedural terminology codes to determine patients' pharmacogenetic testing. To create the cohort for ML model development, we defined the index dates separately for intervention and control groups. The index date for the intervention group was defined as the most recent date of an opioid prescription prior to pharmacogenetic testing. The index date for the control group was defined as the earliest date of opioid prescription. All baseline covariates were collected in the 1-year period prior to the index date. The outcome was a binary target variable, indicating whether the patient was in the intervention or control group based on their pharmacogenetic testing record. If a patient was in the intervention group, their outcome value was encoded as 1 and otherwise 0.

Feature Selection and Data Preprocessing

We selected four different categories of features for our study sample: (1) demographics, (2) clinical history, (3) medication use, and (4) social determinants of health (SDoH). Demographics included age, sex, and race, where age was continuous, sex was binary, and race was multicategory which was one-hot-encoded to get separate binary inputs for each racial group and prevent the ordinality assumption by the model. We used the *International Classification of Diseases, Ninth Revision* (*ICD-9*) and *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes to determine the clinical history variables, including a diagnosis of AIDS/HIV, alcohol abuse, blood loss anemia, cardiac arrhythmias, chronic pulmonary disease, coagulopathy, systolic (congestive) heart failure, deficiency anemia, depression, diabetes with complications, diabetes without complications, drug abuse, fluid and electrolyte disorder, hypertension, hypothyroidism, liver disease, lymphoma, metastatic cancer, neurodegenerative disorders, obesity, paralysis, peptic ulcer disease, peripheral vascular disease, psychosis, pulmonary circulation disorders, renal failure, rheumatoid arthritis or collagen, solid tumor without metastasis, valvular disease, and weight loss. Furthermore, we

XSL•FO
RenderX

included the Elixhauser comorbidity index [7] to calculate the overall health care burden for each patient. In addition, the SDoH information included the average household size, Gini index as a measure of income inequality, median household income, and median gross rent, all linked to the University of Florida Health EHR based on year and ZIP code using the Agency for Healthcare Research and Quality SDoH database [8]. We removed all the missing data, so all patients had complete information for model development.

Model Development and Hyperparameter Optimization

We adopted a comprehensive strategy for model development using LR, extreme gradient boosting (XGB) [9], RF [10], support vector machine (SVM) [11] with linear and radial basis function (RBF) kernels (ie, linear support vector machine [LSVM] and SVM-RBF), and NN to capture predictive performance across a wide range of models for pharmacogenetic testing uptake prediction. LR was first developed due to its ability to capture linear relationships in the data and providing baseline performance. Similarly, the naïve Bayes (NB) model was developed due to its simplicity and effectiveness in high-dimensional spaces. However, considering LR's and NB's limitations in capturing complex nonlinear relationships, we trained tree-based models, including XGB and RF, given their ability to capture nonlinear interactions that may be critical to accurately predicting the pharmacogenetic testing uptake. LSVM and SVM-RBF were developed to capture high-dimensional complex interactions among input features. Finally, the NN was trained to explore more complex interactions among features that may be missed by traditional models.

To train the models, we randomly partitioned the data with a stratified 80/20 train/test split. To mitigate the risk of overfitting and get the best performance out of ML models, we conducted 5-fold cross-validation. For LR, we tuned the hyperparameter *C* with values of 0.01, 0.1, 1, and 10. To tune NB, we used smoothing parameters ranging from 10^{-9} to 10^{-5} with 10x increments. Both XGB and RF models were tuned with 100, 200, 300, 400, and 500 trees to explore the effect of different numbers of trees on performance. LSVM and SVM-RBF were optimized with *C* values equal to 0.01, 0.1, 1, and 10. Finally, NNs were trained with all possible combinations of various optimizers and learning rates. The optimizers included stochastic gradient descent [12], Adam [13], and Nadam [14], and learning rates ranged from 10^{-5} to 10^{-1} with 10x increases. The best hyperparameters were selected based on the highest area under the receiver operating characteristic curve (AUC) or *C*-statistics achieved on the training set during the cross-validation process.

We trained all the models using the best-performing hyperparameters and evaluated their performance on the test set. Notably, we used a balanced class weighting approach, assigning proportionally higher weights to the minority class, to decrease the risk of overfitting when training each of the models. As an additional analysis, we used the synthetic minority over-sampling technique (SMOTE) [15] with a 1:1 class ratio between minority and majority samples, creating a balanced dataset for model training and evaluation. To improve the performance and enhance generalizability of the models, we created a weighted probability-based ensemble classifier. We used a grid search optimization procedure, searching for the best combination of model weights, conditioned on having a total sum of weights equal to 1. The ensemble model achieving the highest AUC or *C*-statistics was selected as the final best model for pharmacogenetic testing uptake prediction.

Comprehensive Evaluation

We followed TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence) guidelines [16] to evaluate the performance of ML models. AUC or *C*-statistics demonstrates the discriminative ability of the model, considering the trade-off between sensitivity and specificity at different threshold values. DeLong test [17,18] is used to identify whether there is a statistically significant difference between AUC values. Accuracy provides the overall performance of the model by calculating the proportion of true negative and true positive cases compared to all the samples in the data. However, this measure may be misleading for rare events and not suitable for highly imbalanced datasets. To address this, we calculated specificity, which measures the model's ability to correctly identify true negative cases and recall or sensitivity, indicating the model's performance for identifying true positive cases. Since model performance may not be optimal, we used the Youden index [19-21] to obtain the best performing classification threshold for each model, where the highest value of "specificity+sensitivity-1" is achieved. To further evaluate model performance for real-world applications, we calculated the number needed to evaluate and predictive positives per 100 patients, showing how many predictions must be made to identify 1 actual positive case and the number of pharmacogenetic testing uptake predictions per 100 individuals, respectively. Moreover, we did stratification analysis, where pharmacogenetic testing uptake probabilities were categorized into 10 deciles in ascending order and the percentage of actual pharmacogenetic testing uptake in the data was identified in each decile to evaluate the ability of the model in identifying more uptakes in the higher strata. In addition, we used the Shapley Additive Explanations (SHAP) [22] analysis to report feature importance and identify the features that are most influential to pharmacogenetic testing uptake.

Results

Patient Characteristics

We used data from the University of Florida Health IDR with patients from a wide range of demographic and clinical backgrounds. The data included a total of 455,773 patients, 7645 (1.68%) with a recorded pharmacogenetic testing and 448,128 (98.32%) without. After removing 160,914 further patients who had no opioid prescription, there were 294,859 patients using opioids in the cohort. We further excluded patients with no race information and SDoH variables (after matching with the Agency for Healthcare Research and Quality database). Overall, the final cohort included 242,640 patients, where 3510 (1.45%) had a recorded pharmacogenetic testing uptake and 239,130 (98.55%) did not. The cohort had an average age of 58 (SD 17.47) years, a gender distribution comprising 41.49%

XSL•FO
RenderX

(100,670/242,640) as male participants and 58.51% (141,970/242,640) as female participants, and a racial distribution of 64.94% (157,567/242,640) White individuals, 28.77% (69,810/242,640) Black individuals, and 6.29% (15,263/242,640) belonging to other racial groups (ie, Hispanic, White Hispanic, Black Hispanic, Asian, Pacific Islander, American Indian, Multiracial, and other). In addition, the cohort had an average household size of 2.58 (SD 0.28), a Gini index of 0.45 (SD 0.05), and an average median income of US $46,324. Table 1 presents the key sociodemographic characteristics of patients in the cohort. Moreover, the distribution of patient characteristics in the training and test sets is provided in Table 2.

Table . Sociodemographic summary of patients.

Characteristics	Pharmacogenetic testing	Nonpharmacogenetic testing	*P* value[a]
Number of individuals, n (%)	3510 (1.45)	239,130 (98.55)	—[b]
Sex, n (%)			<.001
Male	1627 (46.35)	99,043 (41.42)	
Female	1883 (53.65)	140,087 (58.58)	
Race, n (%)			<.001
American Indian	12 (0.34)	346 (0.14)	
Asian	30 (0.85)	2345 (0.98)	
Black	1010 (28.77)	68,800 (28.77)	
Black Hispanic	0 (0.00)	1 (0.00)	
Hispanic	1 (0.03)	168 (0.07)	
Multiracial	13 (0.37)	1165 (0.49)	
Pacific Islander	0 (0.00)	109 (0.05)	
White	2346 (66.84)	155,221 (64.91)	
White Hispanic	0 (0.00)	45 (0.02)	
Other	98 (2.79)	10,930 (4.57)	
Age (y), mean (SD)	66.04 (12.99)	58.24 (17.51)	<.001
Average household size, mean (SD)	2.60 (0.27)	2.58 (0.28)	<.001
Gini index[c], mean (SD)	0.454 (0.051)	0.451 (0.052)	<.001
Median household income (US $), mean (SD)	46,093.44 (15,015.22)	46,328.29 (15,538)	.374
Median gross rent[d], mean (SD)	898.12 (213.45)	905.81 (223.60)	.034

[a]*P* values were calculated using a chi-square test for categorical variables and independent 2-tailed *t* test with equality of variance analysis for continuous variables.

[b]Not applicable.

[c]The Gini index ranges from 0 to 1, where 0 represents perfect equality and 1 represents perfect inequality.

[d]The median gross rent is per month in US $ and is based on rented cash-paid housing units.

XSL•FO
RenderX

Table . Distribution of patient characteristics in the training and test sets.

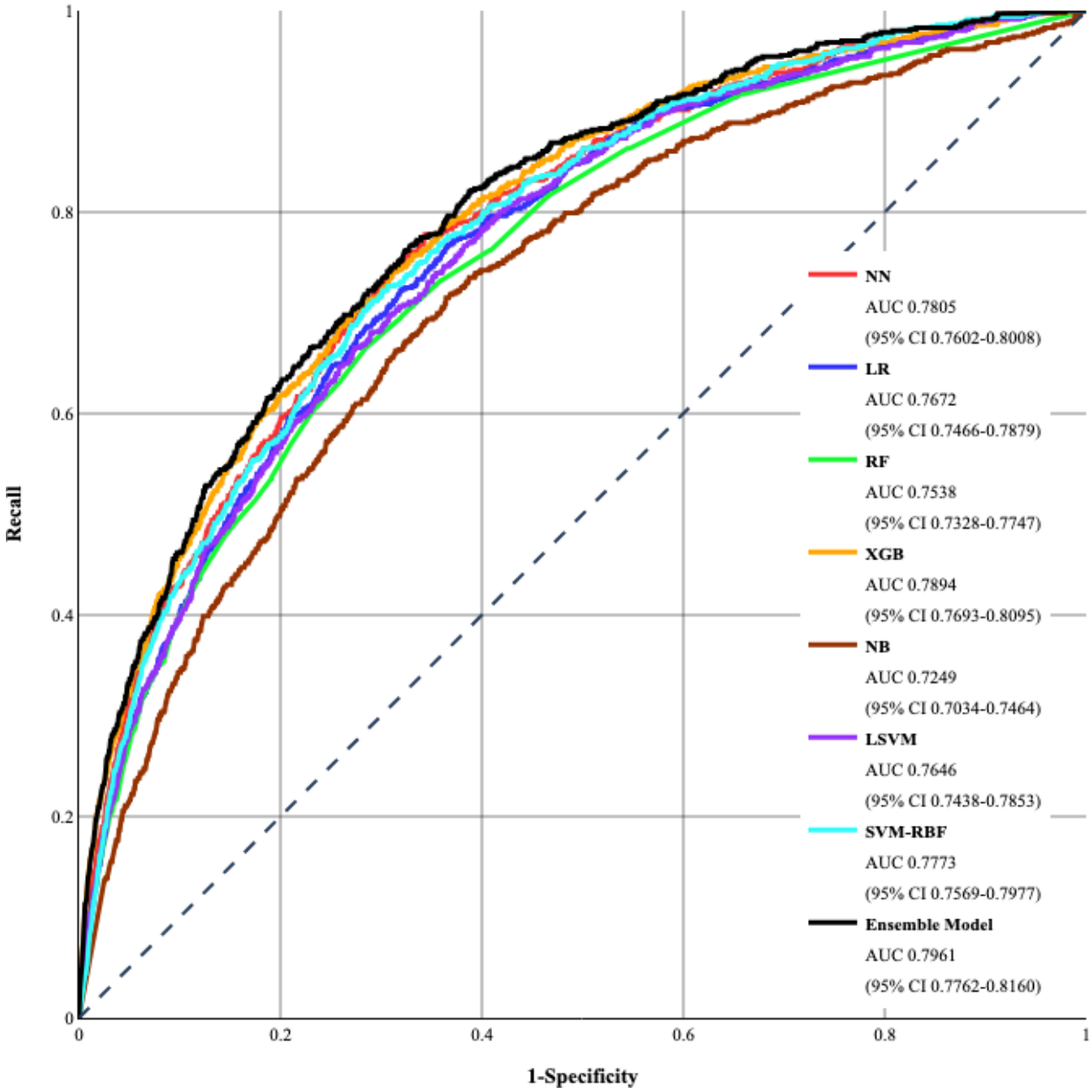
Characteristics	Training (n=194,112)	Testing (n=48,528)
Sex, n (%)		
Male	80,519 (41.48)	20,151 (41.52)
Female	113,593 (58.52)	28,377 (58.48)
Race, n (%)		
American Indian	273 (0.14)	85 (0.18)
Asian	1877 (0.97)	498 (1.03)
Black	55,993 (28.85)	13,817 (28.47)
Black Hispanic	1 (0.00)	0 (0.00)
Hispanic	128 (0.07)	41 (0.08)
Multiracial	919 (0.47)	259 (0.53)
Pacific Islander	89 (0.05)	20 (0.04)
White	125,938 (64.88)	31,629 (65.18)
White Hispanic	33 (0.02)	12 (0.02)
Other	8861 (4.56)	2167 (4.47)
Age, mean (SD)	58.36 (17.48)	58.32 (17.45)
Average household size, mean (SD)	2.58 (0.28)	2.58 (0.28)
Gini index, mean (SD)	0.45 (0.05)	0.45 (0.05)
Median household income (US $), mean (SD)	46,327.05 (15,529.07)	46,316.28 (15,536.70)
Median rent (US $), mean (SD)	905.64 (223.38)	905.92 (223.78)

Model Performance

The ROC curves of all models and their corresponding AUC or *C*-statistics with 95% CIs are provided in Figure 2. The ensemble model achieved the highest *C*-statistics of 79.61% using 0.7 and 0.3 weights for XGB and SVM-RBF, respectively. In contrast, NB achieved the lowest *C*-statistics at 72.49%. Other models had comparable *C*-statistics, equal to 75.38% (RF), 76.72% (LR), 76.46% (LSVM), 77.73% (SVM-RBF), 78.05% (NN), and 78.94% (XGB). Additionally, the DeLong test results indicated that there was a statistically significant difference ($P<.05$) between the *C*-statistics of the ensemble model and all other classifiers, showing its higher performance is less likely to be due to chance.

XSL•FO
RenderX

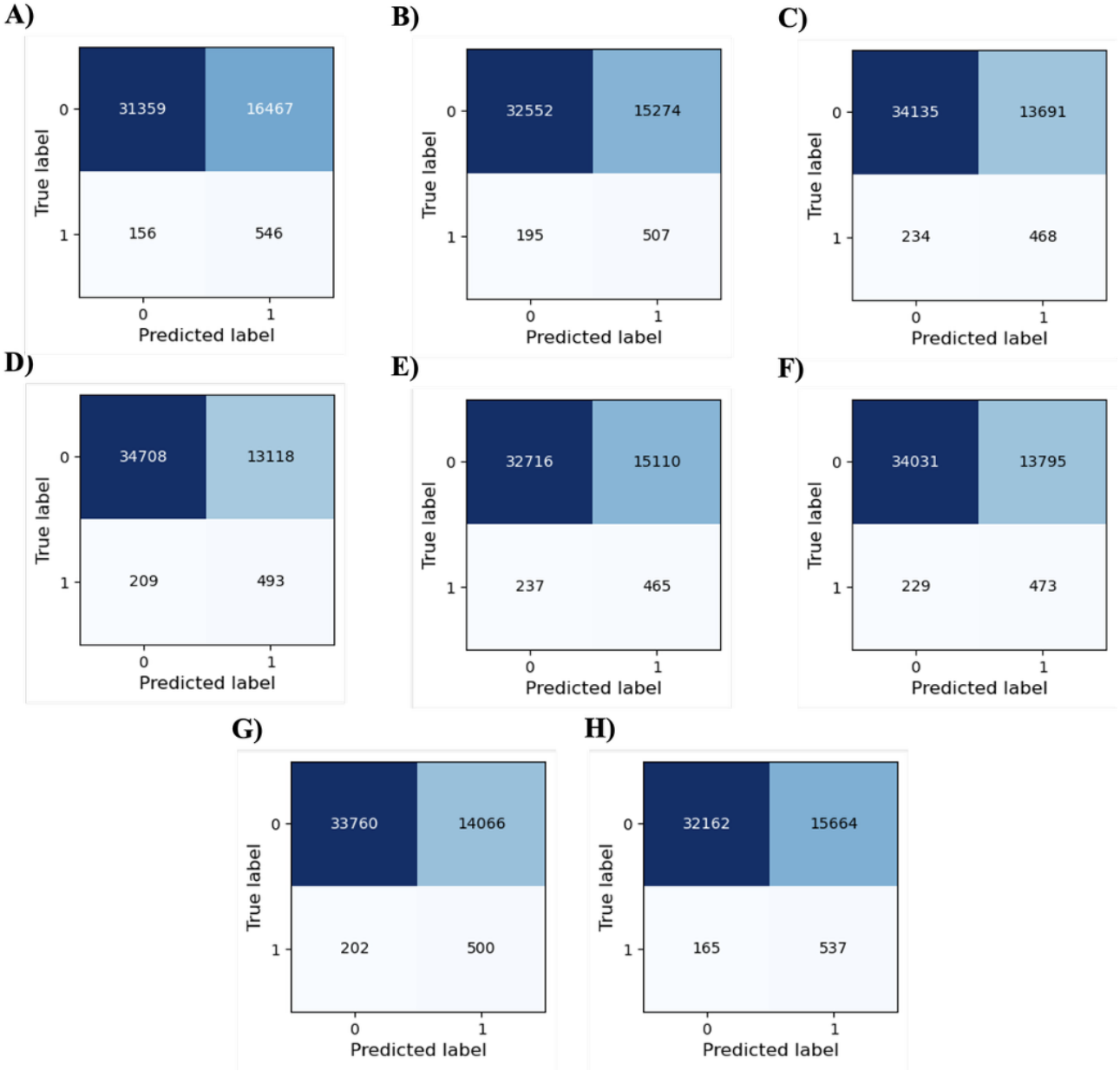
Figure 2. Receiver operating characteristic curves of developed models. AUC: area under the receiver operating characteristic curve, LR: logistic regression, LSVM: linear support vector machine, NB: naïve Bayes, NN: neural network, RF: random forest, SVM-RBF: support vector machine–radial basis function, XGB: extreme gradient boosting.



Since model performance was not optimal, we used the Youden index [19-21] to obtain the best performing classification threshold for each model. Figure 3 portrays the confusion matrices for all developed models based on their Youden index threshold. Table 3 provides a summary of performance for all developed models. Accuracy representing the proportion of correctly classified cases was highest for XGB (35,201/48,528, 72.54%), followed by RF (34,603/48,528, 71.31%), LSVM (34,504/48,528, 71.10%), and SVM-RBF (34,260/48,528, 70.60%). Other classifiers had 65% to 70% accuracy with the ensemble model having an accuracy of 67.38% (32,699/48,528).

XSL•FO
RenderX

Figure 3. Confusion matrix of models based on Youden index: (A) neural network; (B) logistic regression; (C) random forest; (D) gradient boosting; (E) naïve Bayes; (F) linear support vector machine; (G) support vector machines-radial basis function; and (H) Ensemble model.



XSL•FO
RenderX

Table . Performance metrics of developed models based on Youden index.

Model	Accuracy	Recall	Specificity	NPV[a]	PPV[b]	AUC[c]
NN[d]	65.75	77.78	65.57	99.51	3.21	78.05
LR[e]	68.12	72.22	68.06	99.40	3.21	76.72
RF[f]	71.31	66.67	71.37	99.32	3.31	75.38
XGB[g]	72.54	70.23	72.57	99.40	3.62	78.94
NB[h]	68.38	66.24	68.41	99.28	2.99	72.49
LSVM[i]	71.10	67.38	71.16	99.33	3.32	76.46
SVM-RBF[j]	70.60	71.23	70.59	99.41	3.43	77.73
Ensemble	67.38	76.50	67.25	99.49	3.31	79.61

[a]NPV: negative predictive value.

[b]PPV: positive predictive value.

[c]AUC: area under the receiver operating characteristic curve.

[d]NN: neural network.

[e]LR: logistic regression.

[f]RF: random forest.

[g]XGB: extreme gradient boosting.

[h]NB: naïve Bayes.

[i]LSVM: linear support vector machine.

[j]SVM-RBF: support vector machine–radial basis function.

To assess the model performance in identifying true positive cases (ie, those with a recorded pharmacogenetic testing uptake) and true negative cases (ie, those without a recorded pharmacogenetic testing uptake), recall and specificity were calculated. Recall values ranged from 66.24% (465/702) to 77.78% (546/702) with the NN model achieving the highest recall and the ensemble model achieving 76.50% (537/702). Specificity values ranged from 65.57% (31,359/47,826) to 72.57% (34,708/47,826), with the XGB model achieving the highest specificity. To further assess models' performance in distinguishing between true positives and true negatives, negative predictive value (NPV) and positive predictive value (PPV) were calculated. NPV values ranged from 99.28% (32,716/32,953) to 99.51% (31,359/31,515) with the NN model having the highest NPV, while PPV values ranged from 2.99% (456/15,575) to 3.62% (493/13,611) with XGB achieving the highest value. To evaluate model fairness, we used the ensemble model to calculate the performance metrics across different racial groups (Table 4). All the metrics, except NPV, exhibited notable variations across different racial groups with AUC ranging from 74.31% to 83.29%, and accuracy from 65.11% (20,592/31,629) to 79.15% (205/259).

Table . Performance metrics of developed models across racial groups[a].

Race group	Accuracy	Recall	Specificity	NPV[b]	PPV[c]	AUC[d]
White	65.11	75.27	64.95	99.44	3.11	77.73
Black	70.34	81.59	70.18	99.61	3.88	83.29
Multiracial	79.15	75.00	79.22	99.51	5.36	74.31
Asian	74.90	66.67	75.00	99.46	3.15	82.69
Other	78.27	60.87	78.45	99.47	2.94	82.32

[a]The performance metrics were reported using the testing data for the racial groups with at least 4 samples in the positive class (ie, pharmacogenetic testing uptake)

[b]NPV: negative predictive value.

[c]PPV: positive predictive value.

[d]AUC: area under the receiver operating characteristic curve.

The SMOTE results, including ROC curves and performance metrics measured at the Youden index are provided in Multimedia Appendix 1. The results did not show any improvement, and the overall performance significantly reduced. The best-performing model was the ensemble model with 73.63% AUC, which is lower than the AUC without SMOTE (ie, 79.61%). All other models demonstrated a notable decrease

XSL•FO
RenderX

in AUC and other performance metrics, especially accuracy, specificity, and NPV.

Stratification analysis results are presented in Figure 4. The analysis was performed by dividing the predicted probabilities of pharmacogenetic testing uptake into deciles and calculating observed uptake rates within each to assess performance across the probability groups. The lowest decile was decile 1 with the lowest probability values, while decile 10 represented the highest probability values. Within each decile, observed uptake rates were calculated to plot against the probability deciles. The observed event rates demonstrated an increasing trend across the probability deciles (except for the 5th). The event rate in the highest probability decile was 6.59% (320/4853) while 0.12% (6/4853) in the lowest decile.

Figure 4. Stratification analysis for the probability of pharmacogenetic testing uptake.

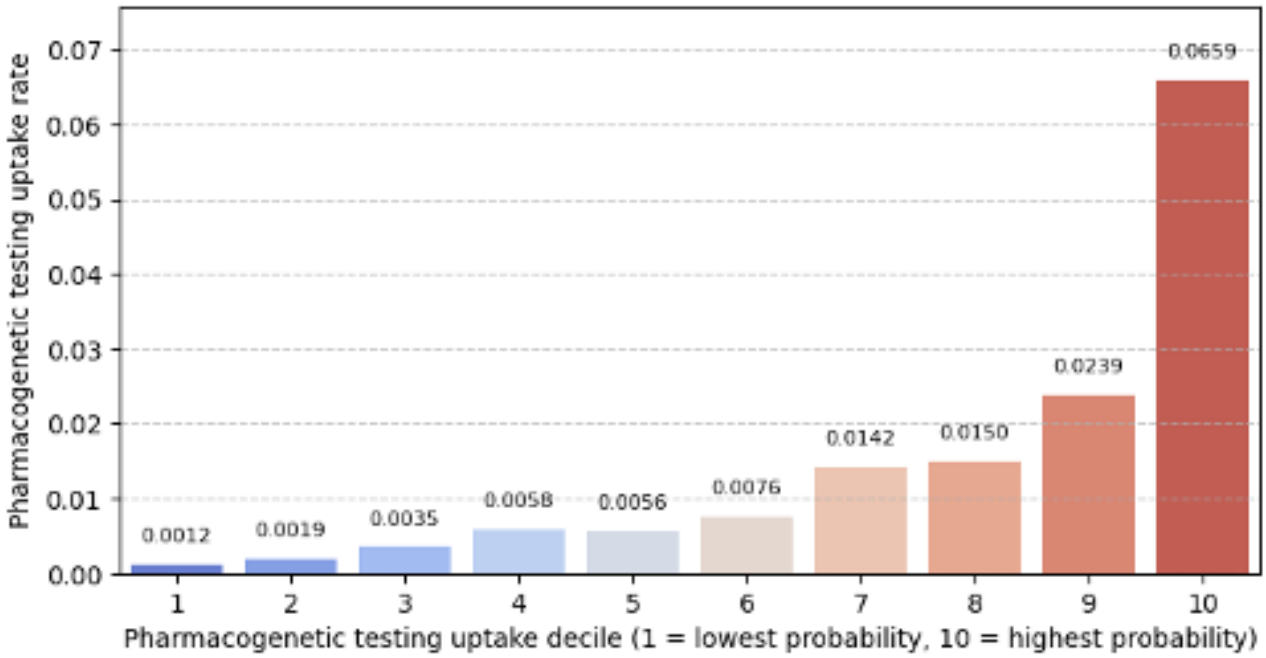
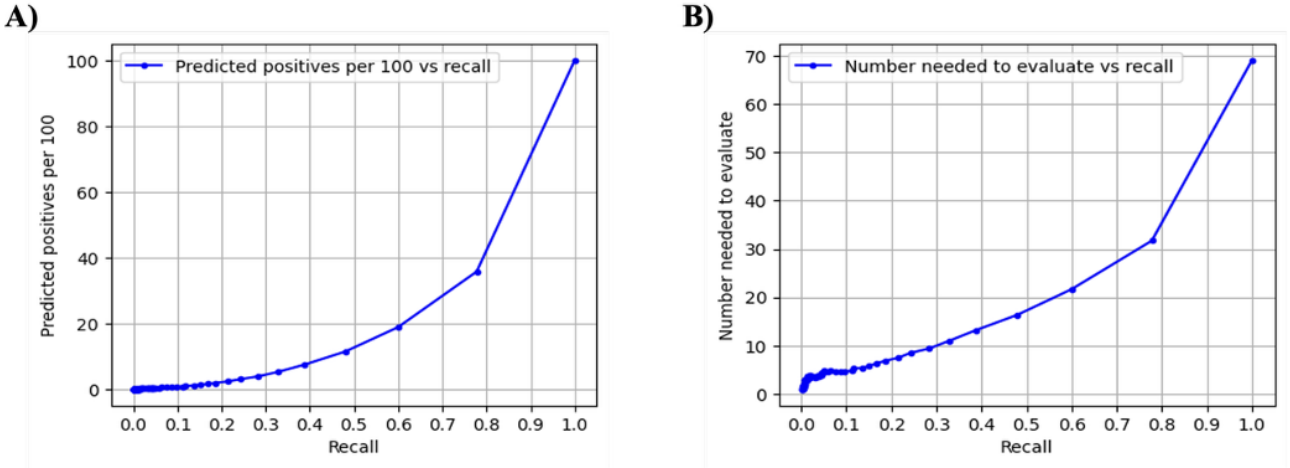


Figure 5 demonstrates the plots of recall values against predictive positives per 100 and the number needed to evaluate. Predictive positives per 100 represent the number of individuals per 100 who are predicted to undergo pharmacogenetic testing across different recall values. The number needed to evaluate indicates the number of individuals that need to be assessed by the model to find 1 true positive (ie, recorded pharmacogenetic testing uptake).

Figure 5. Recall versus (A) predicted positives per 100 and (B) number needed to evaluate.



SHAP analysis for feature importance using the XGB model has been shown in Figure 6. Age had the highest impact on pharmacogenetic testing uptake, where older patients are more likely to undergo pharmacogenetic testing. Hypertension and median household income were the next most important features, both associated with a higher likelihood of pharmacogenetic testing uptake. Other SDoH factors were among the top 7 important factors for pharmacogenetic testing uptake. In contrast, racial group variables (ie, Hispanic, Pacific Islander, and White Hispanic) had the lowest influence on the outcome.

XSL•FO
RenderX

Figure 6. Shapley Additive Explanations (SHAP) analysis for feature importance (the x-axis shows the impact on prediction and the color represents the feature value, ie, red for high and blue for low). CPIC: Clinical Pharmacogenetics Implementation Consortium.

XSL•FO
RenderX

Discussion

Principal Findings

ML has widely been used in pharmacogenetics, where investigations focused on using genetic information to predict drug response [23,24]. However, this is the first study to develop ML models for predicting probability of the use of pharmacogenetic testing using clinical and nonclinical factors such as demographics, comorbidities, medication history, and SDoH. It highlighted the potential of using ML methods to improve the use of pharmacogenetic testing, especially in pain management with opioid medications. The results of this study demonstrated that the ensemble model achieved the highest AUC compared to other models for pharmacogenetic testing uptake prediction. Although the XGB model, when evaluated at its Youden index, performed better than other models in terms of accuracy, recall, specificity, and NPV, we selected the ensemble model as the final classifier. This was because the ensemble model harnesses the strengths of both high-performing XGB and SVM-RBF models for prediction, ultimately improving the performance for pharmacogenetic testing uptake prediction. Overall, the ensemble model proves to be a viable tool for accurately predicting pharmacogenetic testing uptake, supporting more informed, data-driven decision-making.

Several methodological considerations must be addressed when interpreting these promising results. The primary goal of our study was to predict pharmacogenetic testing uptake, which is a clinically favorable outcome. In this scenario, NPV represents the ability of the model to correctly predict no pharmacogenetic testing uptake out of all no-uptake predictions. While our study showed a very high NPV (32,162/32,327; 99.49%), a low NPV could lead to negative consequences, including resource misallocation, alert fatigue, and missed clinical opportunities for those who need pharmacogenetic testing uptake for more effective treatment. In contrast, PPV represents the accuracy of the model in correctly predicting pharmacogenetic testing uptake out of all positive predictions. Thus, a lower PPV leads to lower trustworthiness of positive predictions and inadequate resource allocation to patients who are wrongly predicted to undergo pharmacogenetic testing. Since our data was highly imbalanced with a low percentage of patients having a recorded pharmacogenetic testing uptake, we used a balanced class weighting approach for final model development, assigning proportionally higher weights to the minority class (ie, pharmacogenetic testing uptake). However, the model achieved a much lower PPV compared to NPV. To address this issue, threshold moving was used as a strategy to balance clinical context and consequences of false positives compared to false negatives. Since lowering and raising the threshold can decrease or increase PPV and NPV, the optimal threshold should be selected based on the trade-off between minimizing the consequences of misallocated interventions (ie, higher NPV) and maximizing clinical opportunities for patient care (ie, higher NPV). Overall, the high NPV in our study based on the Youden index could be advantageous in decision-making, as it allows for the allocation of more resources and encouragement plans for patients predicted not to undergo pharmacogenetic testing.

We conducted a comprehensive feature importance analysis to enhance model interpretability and clinical acceptance. The SHAP value analysis using the XGB model highlighted the contribution of each feature to individual predictions, enhancing the model's clinical credibility and trustworthiness. The most influential features were age, hypertension, and household income, suggesting that these features require more attention for increasing pharmacogenetic testing uptake among patients. As mentioned, pharmacogenetic testing uptake is not necessarily due to opioid prescribing and could be due to prescribing of other medications and participation in a research project, potentially affecting the study results. Thus, although the results demonstrate the model's transparency and clinical use, more research is needed before integrating the model into a real-world decision support system. While the SHAP diagram ranks features based on their overall importance, it is crucial to examine critical factors for each patient on a case-by-case basis when making decisions.

Beyond individual prediction, we evaluated the model's ability to stratify patients into meaningful pharmacogenetic testing uptake categories. The uptake stratification analysis showed that the ensemble model effectively differentiated patients across pharmacogenetic testing uptake deciles, with higher uptake probabilities corresponding to higher observed uptake rates. Such an analysis can assist decision-makers in categorizing patients into different deciles based on their likelihood of pharmacogenetic testing uptake and prioritize those in lower deciles (less likely for the pharmacogenetic testing uptake) for optimal resource allocation and improved access to testing. It is important to recognize that stratification analysis often relies on predefined thresholds (ie, 10% in this study) to categorize probabilities into deciles, which may not be optimal for real-world decision-making.

The implementation of the pharmacogenetic testing uptake prediction model addresses the barriers to pharmacogenetic testing in opioid therapy. Our models accurately predict the likelihood of pharmacogenetic testing uptake but do not identify patients who clinically require or should receive testing (this was not the objective of this study). The models capture associations between demographic, clinical, and social determinants with testing behavior under current practice patterns, not clinical appropriateness. This distinction is important as the predictions reflect existing health care access and use patterns rather than evidence-based treatment recommendations. The predictive patterns we observed, favoring older patients with multiple comorbidities and greater health care access, largely reflect existing health care use behaviors and socioeconomic factors rather than clinical need or potential for improved outcomes. Consequently, this model should not be interpreted as a clinical decision support tool for determining which patients should receive pharmacogenetic testing. Rather, the primary applications of this model lie in health care operations and policy rather than direct clinical care. By providing the probability of pharmacogenetic testing uptake at the individual patient level, health systems can identify patients less likely to receive guideline-recommended testing and target barrier-reduction efforts. Insurers can use uptake predictions to forecast costs, allocate resources to underserved populations,

XSL•FO
RenderX

and design coverage policies that promote equitable access to precision medicine. Such predictions could be particularly useful for patients predicted to benefit from testing (not the focus of this study), allowing care teams to prioritize resources to promote uptake in this population. By quantifying uptake likelihood, the model supports data-driven decisions about resource allocation and coverage policies to move beyond trial-and-error opioid prescribing toward more systematic, evidence-based pain management.

Our study has several limitations. First, it is notable that although the study used PPV as a performance metric, it is highly dependent on outcome prevalence and may not generalize across different patient populations or clinical settings, reducing its applicability in decision-making. Second, this study used various clinical and demographic variables to develop more-informed ML models. However, the underrepresentation of certain racial groups, such as Hispanic and Pacific Islander, could potentially affect models' outcomes. This raises concerns about the bias and fairness issues, as ML models may exhibit lower performance for these groups compared to others. Third, while the study incorporated SDoH information as potential predictors of pharmacogenetic testing uptake, this information was based on ZIP code-level data, which may lack the necessary granularity required for precise prediction. Detailed individual-level data on variables such as income, education, and household size could enhance the model's reliability and improve the accurate identification of each feature's contribution to the outcome. Fourth, the model's performance may not directly translate to external populations with varying demographic and clinical backgrounds. Specifically, the model was trained on data from a single EHR system, including minority racial groups, and the results may not be applicable to other racial groups. Therefore, caution is warranted when interpreting the model's outcomes for these populations. Finally, some of the patients in the data who had a recorded pharmacogenetic testing uptake were part of a clinical trial where the cost of testing was covered. This would potentially influence the ability of the model to identify the true impact of patients' economic and income status on undergoing pharmacogenetic testing. To demonstrate the model's broader applicability, future work should focus on external validation using independent datasets from different health systems and identification of patients who are more likely to benefit from testing. This would help evaluate the model's robustness and clinical use for real-world decision-making.

Conclusions

Pharmacogenetic testing is a viable tool for matching patients' genetic profiles to suitable opioids for pain treatment. This study proposed ML models for pharmacogenetic testing uptake prediction using data from an EHR system. Results demonstrated that the ensemble ML model combining XGB and SVM-RBF classifiers achieved the highest AUC at 79.61%, making it a reliable prediction model for pharmacogenetic testing uptake prediction. Additionally, the uptake stratification and feature importance analysis using SHAP values further indicated the model's use for real-world applications. Following further validation using an external dataset, this model can be integrated into a data-driven decision support system, enabling health systems and insurers in resource planning and health equity assessment to optimize pain management and improve patient outcomes.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Synthetic minority over-sampling technique results.
[DOCX File, 174 KB - medinform_v14i1e81048_app1.docx]

References

1. Dale O, Moksnes K, Kaasa S. European palliative care research collaborative pain guidelines: opioid switching to improve analgesia or reduce side effects. a systematic review. Palliat Med 2011 Jul;25(5):494-503. [doi: 10.1177/0269216310384902] [Medline: 21708856]
2. Crews KR, Monte AA, Huddart R, et al. Clinical pharmacogenetics implementation consortium guideline for CYP2D6, OPRM1, and COMT genotypes and select opioid therapy. Clin Pharmacol Ther 2021 Oct;110(4):888-896. [doi: 10.1002/cpt.2149] [Medline: 33387367]
3. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930. [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
4. Vike NL, Bari S, Stefanopoulos L, et al. Predicting COVID-19 vaccination uptake using a small and interpretable set of judgment and demographic variables: cross-sectional cognitive science study. JMIR Public Health Surveill 2024 Mar 18;10:e47979. [doi: 10.2196/47979] [Medline: 38315620]
5. Zhang B, Shi H, Wang H. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. J Multidiscip Healthc 2023;16:1779-1791. [doi: 10.2147/JMDH.S410301] [Medline: 37398894]
6. Nguyen MH, Applegate CD, Murray B, et al. Assessing genetic counseling efficiency with natural language processing. J Am Med Inform Assoc 2025 Nov 10:ocaf190. [doi: 10.1093/jamia/ocaf190] [Medline: 41211696]

XSL•FO
RenderX

7. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care 2009 Jun;47(6):626-633. [doi: 10.1097/MLR.0b013e31819432e5] [Medline: 19433995]
8. Social determinants of health database. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download [accessed 2025-04-04]
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016. [doi: 10.1145/2939672.2939785]
10. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
11. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 2020 Sep;408:189-215. [doi: 10.1016/j.neucom.2019.10.118]
12. Ruder S. An overview of gradient descent optimization algorithms. arXiv. Preprint posted online on Sep 15, 2016. [doi: 10.48550/arXiv.1609.04747]
13. Kingma DP, Ba LJ. Adam: a method for stochastic optimization. Presented at: International Conference on Learning Representations; May 7-9, 2015 URL: https://dare.uva.nl/search?identifier=a20791d3-1aff-464a-8544-268383c33a75 [accessed 2025-12-18]
14. Dozat T. Incorporating Nesterov momentum into Adam. Presented at: International Conference on Learning Representations (ICLR 2016); May 2-4, 2016 URL: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ [accessed 2026-01-07]
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. jair 2002;16:321-357. [doi: 10.1613/jair.953]
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
18. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 2014;21(11):1389-1393. [doi: 10.1109/LSP.2014.2337313]
19. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
20. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. Biom J 2005 Aug;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
21. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden index and the optimal threshold for markers with mass at zero. Stat Med 2008 Jan 30;27(2):297-315. [doi: 10.1002/sim.2993] [Medline: 17624866]
22. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017. [doi: 10.5555/3295222.3295230]
23. Firoozbakht F, Yousefi B, Schwikowski B. An overview of machine learning methods for monotherapy drug response prediction. Brief Bioinform 2022 Jan 17;23(1):bbab408. [doi: 10.1093/bib/bbab408] [Medline: 34619752]
24. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020;4(1):19. [doi: 10.1038/s41698-020-0122-1] [Medline: 32566759]

Abbreviations

AUC: area under the receiver operating characteristic curve
CPIC: Clinical Pharmacogenetics Implementation Consortium
CYP2D6: cytochrome P450 family 2 subfamily D member 6
EHR: electronic health record
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
***ICD-9*:** *International Classification of Disease, Ninth Revision*
IDR: Integrated Data Repository
LR: logistic regression
LSVM: linear support vector machine
ML: machine learning
NB: naïve Bayes
NN: neural network
NPV: negative predictive value
PPV: positive predictive value
RF: random forest

XSL•FO
RenderX

SDoH: social determinants of health
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority over-sampling technique
SVM: support vector machine
SVM-RBF: support vector machine–radial basis function
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence
XGB: extreme gradient boosting

Edited by J Klann; submitted 21.Jul.2025; peer-reviewed by A Berihun, R Mpofu; revised version received 13.Nov.2025; accepted 26.Nov.2025; published 21.Jan.2026.

Please cite as:
Yaseliani M, Hong JW, Bian J, Cavallari L, Duarte JD, Nelson D, Lo-Ciganic WH, Nguyen KA, Hasan MM
Machine Learning Prediction of Pharmacogenetic Testing Uptake Among Opioid-Prescribed Patients Using Electronic Health Records: Retrospective Cohort Study
JMIR Med Inform 2026;14:e81048
URL: https://medinform.jmir.org/2026/1/e81048
doi:10.2196/81048

© Mohammad Yaseliani, Je-Won Hong, Jiang Bian, Larisa Cavallari, Julio Duarte, Danielle Nelson, Wei-Hsuan Lo-Ciganic, Khoa Anh Nguyen, Md Mahmudul Hasan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of Venous Thromboembolism Risk Prediction Models Based on Whole Blood Gene Expression Profiling Using 20 Machine Learning Algorithms: Comprehensive Analysis Study

Yedong Huang[1*], MD, PhD; Xiaoyun Chen[2*], MD, PhD; Guannan Bai[3], MD; Yajun Zhao[4], MD; Dapeng Kuang[5], MD; Lin Zhang[6,7], Prof Dr; Wei Lu[8], Prof Dr

[1]Department of Radiation Oncology, Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou, Fujian, China

[2]The School of Clinical Medicine, Fujian Medical University; Department of Respiratory and Critical Care Medicine, Fujian Provincial Geriatric Hospital, No. 147 Beihuan Middle Road, Fuzhou, China

[3]Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China

[4]Department of Health Management Centre, Zhongshan Hospital, Fudan University, Shanghai, China

[5]Department of Emergency and Critical Care, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[6]Department of Epidemiology and Preventive Medicine, The School of Public Health and Preventive Medicine, Monash University, Victoria, Australia

[7]Suzhou Industrial Park Monash Research Institute of Science and Technology, Monash University, Suzhou, China

[8]Department of Cardiovascular Surgery, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou City, No.100 of Minjiang Road, Wenzhou, China

[*]these authors contributed equally

Corresponding Author:
Wei Lu, Prof Dr
Department of Cardiovascular Surgery, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou City, No.100 of Minjiang Road, Wenzhou, China

Abstract

Background: There is a lack of venous thromboembolism (VTE) risk prediction models based on gene expression information.

Objective: This study aimed to construct a VTE prediction model based on whole blood gene expression profiling, by performing a comprehensive analysis of 20 machine learning (ML) algorithms.

Methods: Two transcriptome datasets containing patients with VTE and healthy controls were obtained by searching the Gene Expression Omnibus database and used as the training and validation sets, respectively. Feature selection for model construction was performed on the training set using the least absolute shrinkage and selection operator and random forest, followed by the selection of the intersection of the chosen features. Subsequently, recursive feature elimination was applied to further refine the selected features. The selected features underwent model construction using 20 ML algorithms. The performance of the models was evaluated using various methods such as receiver operating characteristic and confusion matrix. The validation set was used for external model validation.

Results: The final results demonstrated that all algorithm models, except for k-nearest neighbor, exhibited good performance in VTE prediction. External validation data indicated that 9 algorithm models had an area under the curve greater than 0.75. The confusion matrix analysis revealed that the algorithm models maintained high specificity in the external validation cohort.

Conclusions: This study used 20 ML algorithms to construct VTE prediction models based on whole blood gene expression information, with 9 of these models demonstrating good diagnostic performance in external validation cohorts. The above models, when used in conjunction with D-dimer, may provide more valuable references for VTE diagnosis.

(JMIR Med Inform 2026;14:e75565) doi:10.2196/75565

KEYWORDS

venous thromboembolism; gene expression profiling; machine learning; biomarkers; predictive models; ROC; receiver operating characteristic curve; transcriptome

Introduction

Venous thromboembolism (VTE) is a prevalent thrombotic disorder in clinical practice, consisting primarily of deep venous thrombosis and pulmonary embolism. VTE remains a major global health burden, affecting approximately 10 million people annually and ranking among the most common vascular diseases worldwide [1]. The pathogenesis of VTE is considered to be

XSL•FO
RenderX

multifaceted and intricate, encompassing a multitude of factors, including but not limited to congenital genetic factors, malignancies, pregnancy in women, surgical trauma, and the use of oral contraceptives [2,3]. Despite the existence of relatively comprehensive prevention and assessment systems, the high mortality rate associated with VTE remains a significant concern among clinical practitioners. According to a study by Heit et al [4], the short-term (30-d) survival rate for deep vein thrombosis is approximately 94.5%, whereas the short-term survival rate for pulmonary embolism is only 66.8%.

In clinical practice, the diagnostic modalities most frequently used for VTE include D-dimer measurement, ultrasonography, and venography [5]. Among these methods, venography is considered the gold standard for diagnosing VTE. However, due to its invasive nature and the potential risk of contrast-induced renal impairment and allergic reactions, its clinical application is limited [6]. The limitations of ultrasonography include its subjectivity and difficulty in detecting deep venous thrombosis [7]. D-dimer testing is known to lack specificity, as elevated D-dimer levels may be caused by conditions such as infection, autoimmune diseases, pregnancy, and childbirth, leading to a higher likelihood of false positive results [8]. Therefore, it is crucial to construct a VTE risk prediction model with high performance.

With the advancement of science and technology, artificial intelligence (AI) has become a widely studied and highly regarded field in recent years, particularly in the medical domain. As one of the methods for achieving AI, machine learning (ML) algorithms have tremendous potential in medical research [9-11]. The aim of this study was to construct a VTE risk prediction model based on whole blood gene expression information, using a comprehensive analysis of 20 ML algorithms and external validation, with the goal of providing a reference for clinical decision-making and VTE prevention. These models are not only designed to differentiate between existing VTE cases and controls, but also intended to detect high-risk patients in early or asymptomatic stages, thereby serving as predictive tools that may assist in clinical decision-making before definitive diagnostic imaging is performed.

Methods

Data and Source

The gene expression data of patients with VTE and healthy controls were retrieved and downloaded from the Gene Expression Omnibus (GEO) database and used as the training and validation groups, respectively, in this study. The GSE19151 dataset was used as the training set, while the GSE48000 dataset served as the validation set. Both datasets were generated using the Affymetrix Human Genome U133 Plus 2.0 microarray platform (GPL570). Table 1 shows the basic information of the 2 datasets, GSE19151 and GSE48000. The GSE19151 dataset comprises gene expression data from 70 patients with VTE and 63 healthy controls [12]; the GSE48000 dataset contains gene expression data from 107 patients with VTE and 25 healthy controls [13]. Data analysis was performed using the *Scikit-learn* (sklearn) module in Python (version 3.6; Python Software Foundation).

Table . Details of the Gene Expression Omnibus (GEO)[a] datasets used in this study.

Dataset	GEO ID	Sample size (n)	PubMed ID	Number of genes	Country
Training set	GSE19151	133	21737128	13,515	USA
Validation set	GSE48000	132	25684211	31,412	USA

[a]GEO: Gene Expression Omnibus.

Feature Selection

Least absolute shrinkage and selection operator (LASSO) regression analysis method was developed by Robert Tibshirani [14]. The principle of this method is to construct a penalty function to achieve the goal of shrinking regression coefficients and refining the model. While the principle of feature selection using random forest (RF) is mainly based on evaluating the contribution of each feature in the RF on each tree. In this study, we used the mean decrease in Gini impurity to measure feature importance in the RF model. The above 2 methods were used to select modeling features in the training set, and the intersection analysis was subsequently performed to reveal the features selected by both methods. This intersection strategy was designed to combine the strengths of both linear (LASSO) and nonlinear (RF) approaches and to reduce potential selection bias by focusing on consistently selected features. Although this intersection approach may sacrifice sensitivity, it enhances stability and reduces overfitting risk in high-dimensional settings. All feature selection procedures were performed exclusively on the training dataset to prevent information leakage from the validation data. The results were then visualized by a Venn plot.

Recursive Feature Elimination for Model Simplification

The main idea of the recursive feature elimination (RFE) method is to repeatedly construct models and remove the most significant features, ranking the features according to the order in which they are eliminated. To achieve a more streamlined model without compromising performance, the features selected by the aforementioned method were further screened and simplified using the RFE method in our study.

Construction of a VTE Risk Prediction Model Using 20 ML Algorithms

After completing the feature selection process, this study constructed the model using commonly used ML algorithms in the field, which included the following: adaptive boosting, artificial neural network, bagging, Bayesian ridge, decision tree, elastic net, extra tree (extremely randomized trees), gradient boosting, k-nearest neighbors (KNN) algorithm, LASSO, linear

XSL•FO
RenderX

LASSO, linear regression, logistic regression, naïve Bayes, RF, ridge regression, ridge cross-validation (ridge regression with cross-validation), stochastic gradient descent, support vector machine, and voting (voting classifier). The algorithm code is taken from the sklearn website [15]. All models were constructed using 5-fold cross-validation.

External Validation of the Model

In this study, GSE48000 was used as the external validation cohort to evaluate the generalization ability of the 20 ML models, ensuring the reproducibility of the model. The performance of the external validation was evaluated by the receiver operating characteristic (ROC) curve. All feature selection and model training were exclusively performed on the training dataset. The final trained models were then directly applied to the validation dataset without any retraining or adjustment, thereby ensuring a true external validation.

Decision Curve Analysis

To evaluate the potential clinical use of the constructed models, we performed decision curve analysis (DCA) across both internal and external evaluation settings. The net benefit for each model was calculated using the standard formula:

NetBenefit=TPN-FPN*PT1-PT

where TP and FP denote the numbers of true and false positives at a given threshold probability, and *n* is the total number of samples. DCA was conducted over a range of clinically relevant threshold probabilities from 0.05 to 0.5. The decision curves were generated using Python with appropriate packages compatible with sklearn.

Ethical Considerations

This study involved secondary analysis of publicly available, deidentified transcriptomic datasets obtained from the GEO database. As per institutional and GEO data usage policies, such analysis does not require ethical approval or informed consent. For more information, please refer to the GEO data policy [16].

Results

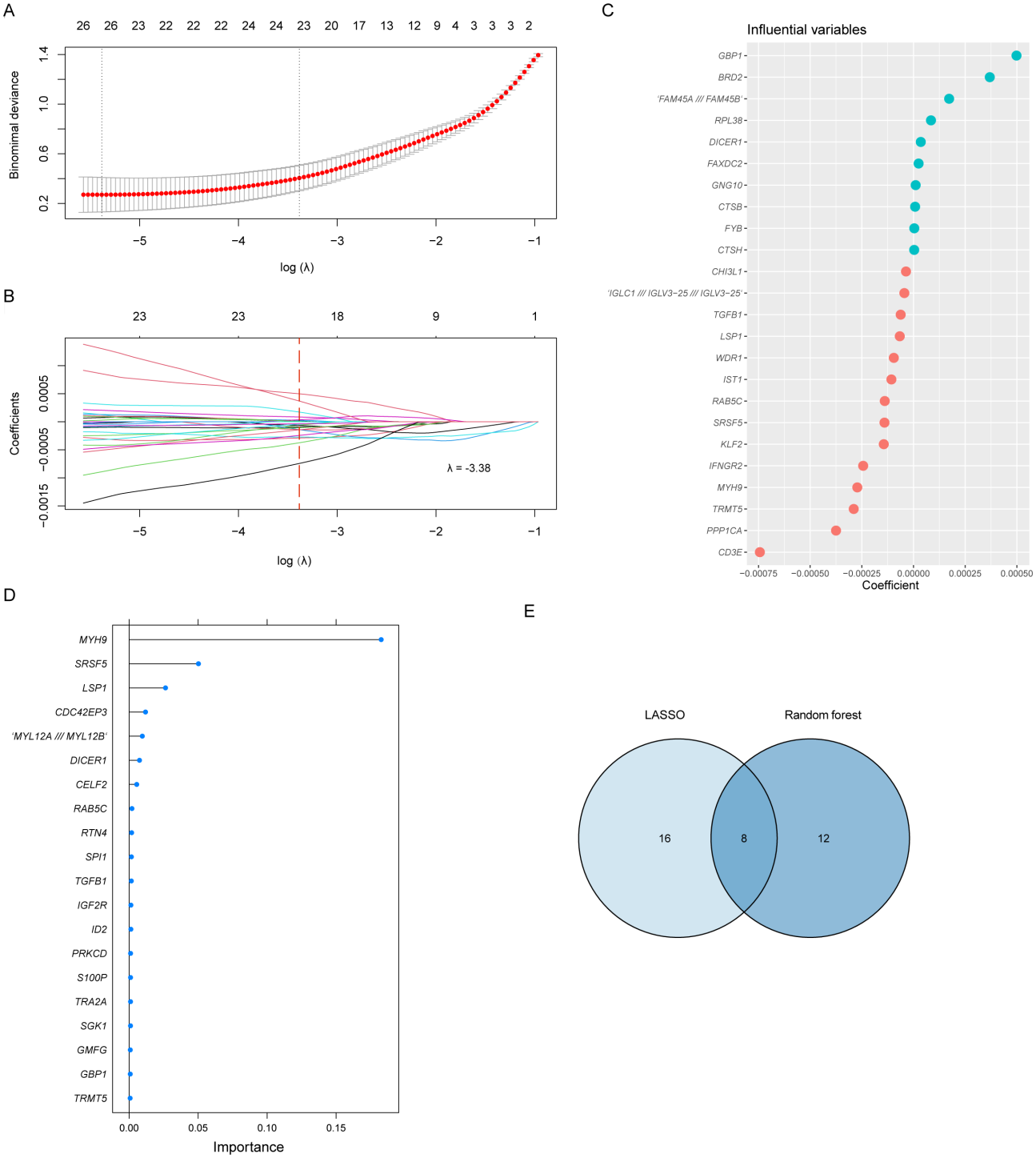
The Features Selected by LASSO Regression and RF

To determine the optimal regularization strength in LASSO regression, we used 5-fold cross-validation to identify the λ value corresponding to the “one standard error rule” (λ_1 se), which balances model simplicity and predictive performance. The regularization path was explored using the LASSO cross-validation function in sklearn, which automatically tests a logarithmic range of λ values. We selected 24 features corresponding to λ=−3.38.

The process of feature selection using LASSO regression is illustrated in Figure 1A and B and Figure 1C illustrates the 24 features selected by LASSO regression and their corresponding coefficients. The screening results of the RF are shown in Figure 1D, where the lollipop plot displays the 20 features selected by RF, with the length of the horizontal axis representing the corresponding importance of each gene. The Venn diagram in Figure 1E shows the intersection of the features selected by LASSO regression and RF methods. The 8 features selected by the intersection analysis are as follows: *TRMT5*, *TGFB1*, *SRSF5*, *RAB5C*, *MYH9*, *LSP1*, *GBP1*, and *DICER1*.

XSL•FO
RenderX

Figure 1. The process of selecting modeling features using least absolute shrinkage and selection operator (LASSO) and random forest methods.

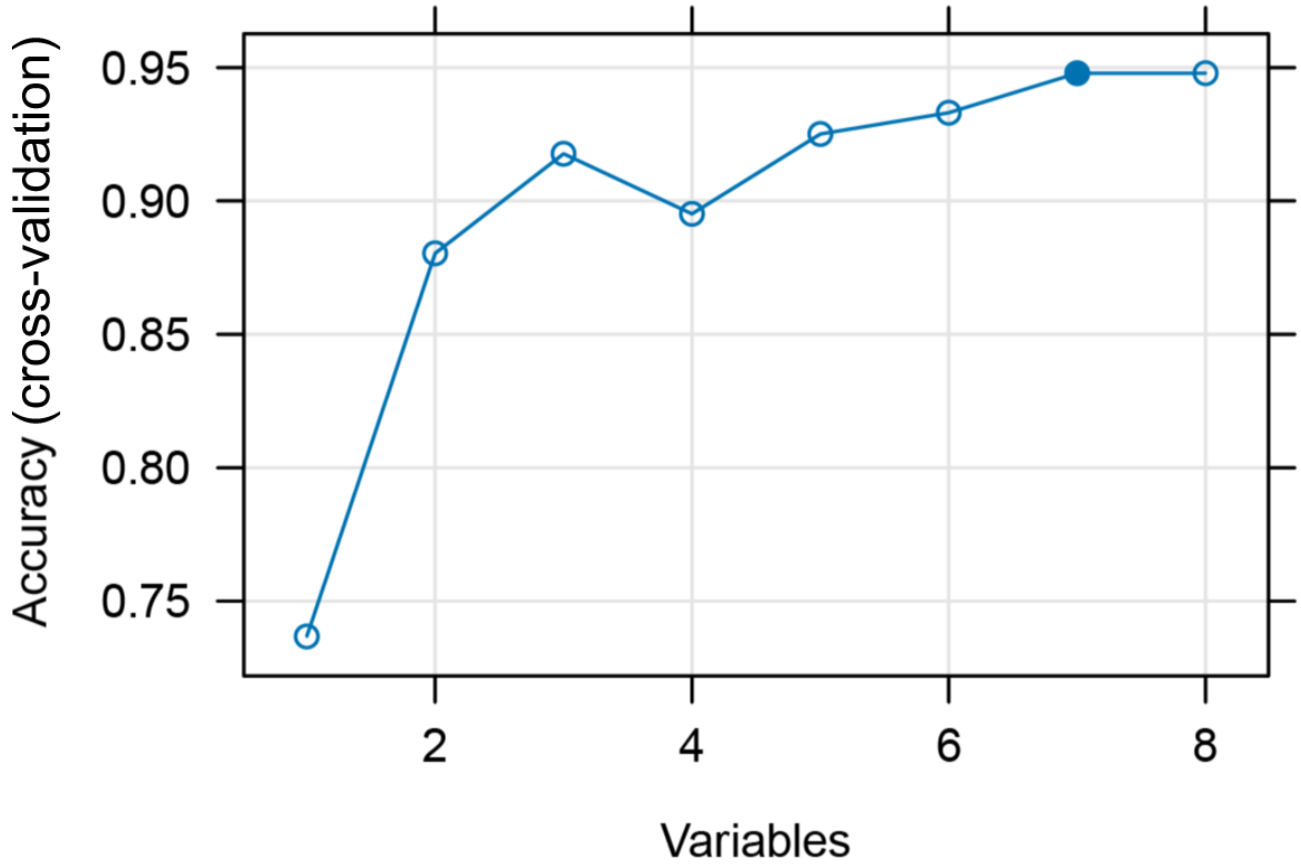


Recursive Feature Elimination to Simplify the Model

The above 8 features were further simplified using the RFE algorithm, which reduces costs and improves the generalizability of the model while removing redundant features. Figure 2 displays the results of the RFE calculations using a line graph. The results indicate that the accuracy of the model in cross-validation no longer significantly changes when the number of features is reduced to 7. The features further refined by the RFE algorithm include 7 genes: *SRSF5*, *MYH9*, *LSP1*, *RAB5C*, *TGFB1*, *DICER1*, and *GBP1*.

XSL•FO
RenderX

Figure 2. The process of simplifying modeling features using recursive feature elimination.



Construction of VTE Risk Prediction Models Using 20 ML Algorithms

Based on the 7 features selected, our study used 20 commonly used ML algorithms to construct the models and used 5-fold cross-validation to improve the accuracy of model evaluation and reduce random error. Figure 3A displays the ROC curves and corresponding area under the curve (AUC) values of the 20 ML models using different colored lines. Figure 3B depicts a radar chart showcasing the accuracy, F_1-score, precision, and recall for 20 ML algorithms. The accuracy reflects the proportion of correctly predicted samples; precision reflects the proportion of actual VTE cases among samples classified as VTE by the model; recall represents the proportion of VTE cases correctly classified by the model among all VTE samples; the F_1-score is a value used to comprehensively evaluate both recall and precision. As illustrated in Figure 3, most models exhibit an AUC value greater than 0.9, and the radar chart based on 4 parameters evaluation indicates that, except for the KNN model, all other ML models can make relatively accurate predictions for VTE.

XSL•FO
RenderX

Figure 3. Performance comparison of 20 machine learning models using receiver operating characteristic (ROC) curves and radar plots. (**A**) ROC curves and area under the curve values based on 20 machine learning models. (**B**) Radar plots comparing accuracy, F_1-score, precision, and recall across models. AdaBoost: adaptive boosting; ANN: artificial neural network; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; SGD: stochastic gradient descent; SVM: support vector machine.

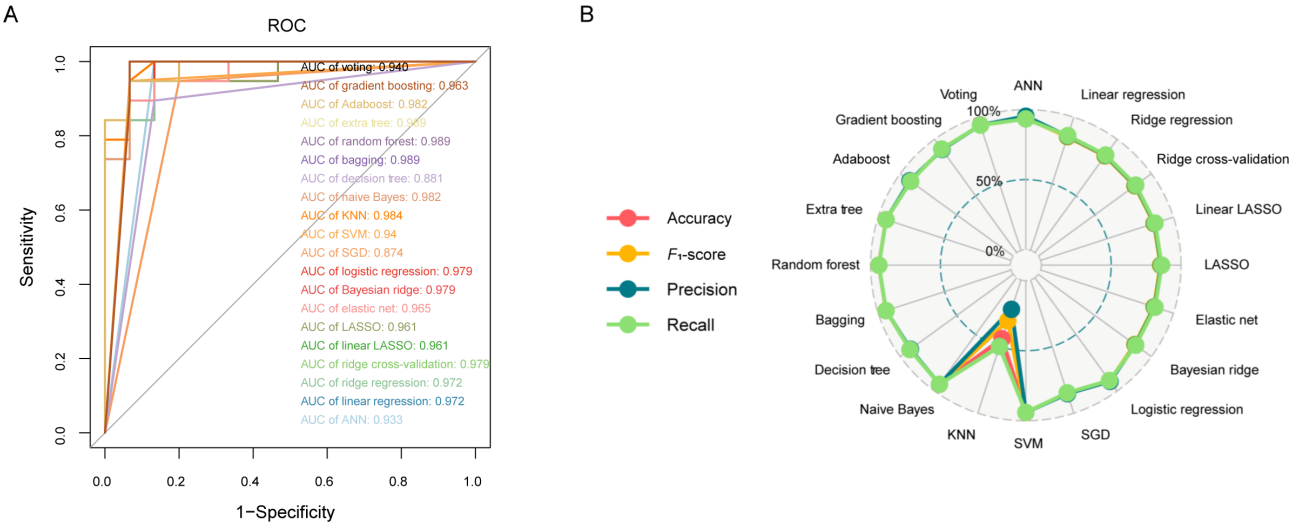
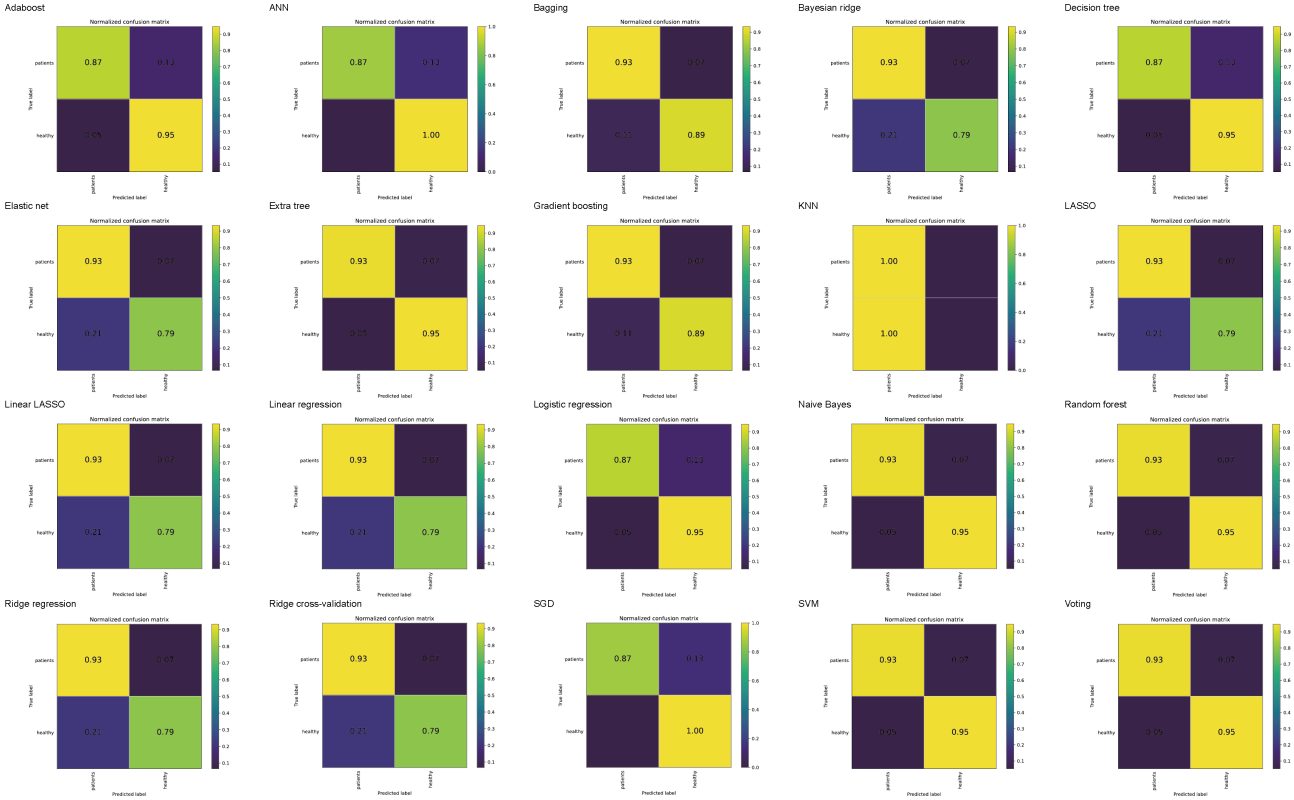


Figure 4 presents the confusion matrices for 20 ML algorithms, among which the sensitivity of all models except for the KNN fluctuated between 87% and 93%, and the specificity fluctuated between 79% and 100%. The findings above indicate that, with the exception of the KNN model, the remaining 19 ML models possess favorable predictive efficacy for VTE. Multimedia Appendix 1 presents the DCA for 20 ML algorithms.

Figure 4. Confusion matrices for 20 machine learning algorithms. The top-left cell indicates the true positive, top-right indicates the false negative, bottom-left indicates the false positive, and bottom-right indicates the true negative. AdaBoost: adaptive boosting; ANN: artificial neural network; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; SGD: stochastic gradient descent; SVM: support vector machine.



External Validation of the Models

To ensure the reproducibility and generalizability of the model, this study performed external validation of the 20 ML models using the GSE48000 dataset. Figure 5 illustrates the ROC curves and the corresponding AUC based on the external validation dataset. The results revealed a significant decrease in the AUC of all models during external validation, but the AUC of the following 9 models remained above 0.75: extra tree, RF, Bayesian ridge, elastic net, LASSO, linear LASSO, ridge cross-validation, ridge regression, and linear regression. This result indicates that the predictive performance of the models

XSL•FO
RenderX

corresponding to the 9 ML algorithms is relatively stable, and they exhibit strong generalizability.

Figure 5. The external validation receiver operating characteristic (ROC) curve plot for the 20 machine learning models. AdaBoost: adaptive boosting; ANN: artificial neural network; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; SGD: stochastic gradient descent; SVM: support vector machine.

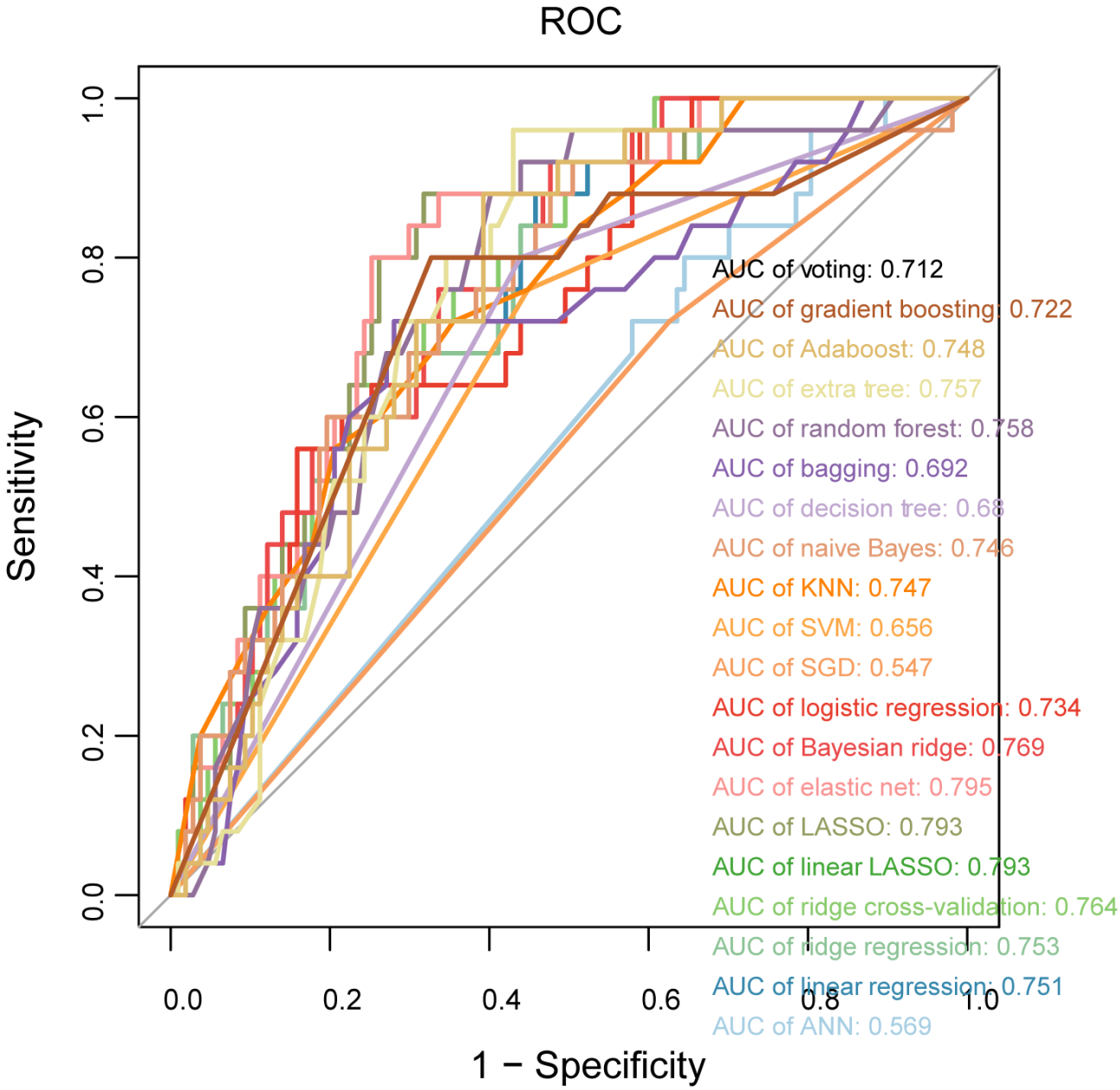
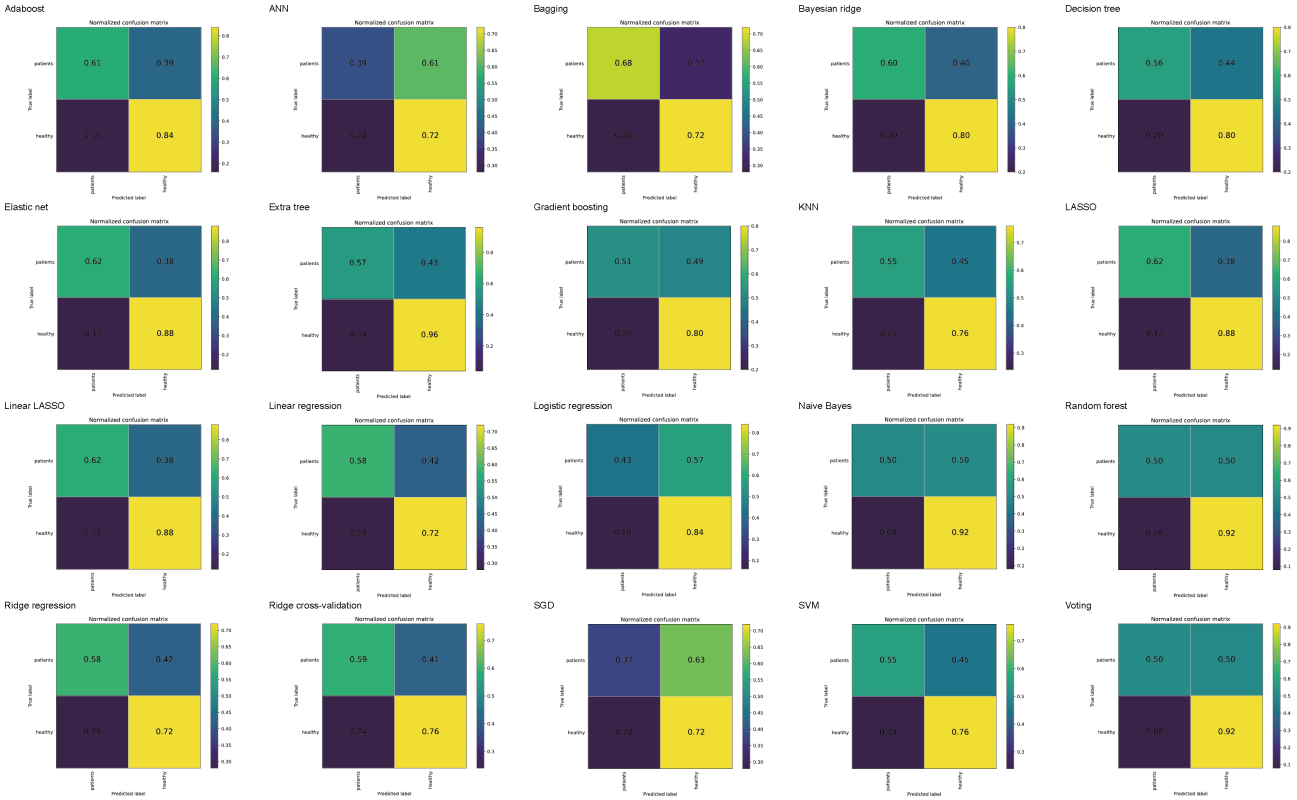


Figure 6 displays the confusion matrices of 20 ML algorithms during external validation. The sensitivity of all models decreased significantly, ranging from 37% to 62%, whereas the specificity remained relatively stable, fluctuating between 72% and 96%. These findings suggest that the models still possess high specificity in the external validation dataset. Multimedia Appendix 2 displays the DCA for 20 ML algorithms in external validation.

XSL•FO
RenderX

Figure 6. Confusion matrices of 20 machine learning algorithms during external validation. The top-left cell indicates the true positive, top-right indicates the false negative, bottom-left indicates the false positive, and bottom-right indicates the true negative. AdaBoost: adaptive boosting; ANN: artificial neural network; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; SGD: stochastic gradient descent; SVM: support vector machine.



Discussion

Principal Findings

As a rapidly advancing field in computer science, AI has the ability to process vast and intricate medical data for the diagnosis and prediction of potential clinical outcomes [17]. ML, as a component of AI, can extract valuable information and patterns from vast amounts of data through learning and analysis, thereby achieving AI [18]. In the medical field, ML is predominantly used for medical image analysis, disease prediction and risk assessment, health care management, decision support, and drug development, among other applications [19-21]. As the first study to construct a VTE risk prediction model based on whole blood gene expression information, the present research aims to empower the diagnosis of VTE through the use of ML algorithms. It is important to note that the current model was trained and validated only on healthy controls versus patients with VTE and does not assess its discriminative ability in populations with other inflammatory or thrombotic conditions. On the other hand, the external validation results indicate that the models developed in this study are characterized by low sensitivity and high specificity. According to the "No free lunch" theory proposed by Wolpert et al [22], a perfect model does not exist. In other words, if a certain ML model outperforms other algorithms in a particular evaluation metric, it must inevitably sacrifice some performance in other metrics as a trade-off. As a result of this constraint, it is not feasible to determine a flawless ML model as the conclusive outcome. Rather, a more fitting model is chosen by taking into account a comprehensive evaluation of diverse metrics.

At present, the models used for predicting the risk of VTE occurrence are predominantly built upon clinical features. Darzi et al [23] conducted a meta-analysis that included 17 studies, which indicated that VTE may be associated with the following factors: advanced age, elevated C-reactive protein levels, increased D-dimer and fibrinogen levels, tachycardia, thrombocytosis, leukocytosis, fever, leg swelling, malignancy, immobility, infection, and so on. In 2021, Pandor et al [24] conducted a systematic review of 24 VTE risk assessment models to evaluate their predictive performance. The results showed that the *C*-statistic was generally less than 0.7 for all models, with only a few models having favorable *C*-statistics (>0.8). Among these models, sensitivity fluctuated between 12% and 100%, while specificity varied between 7.2% and 100%. Based on the aforementioned findings, the study suggests that the existing VTE risk prediction models have generally weak predictive ability and high heterogeneity, and thus, does not recommend the use of any specific prediction model. Our study is an exciting and innovative attempt to discover modeling features based on gene expression analysis. In the internal validation, most of the models developed in this study exhibited good performance (AUC >0.90), while in the external validation, the data showed that 9 models had an AUC greater than 0.75. Considering clinical applicability, we note that models such as LASSO, ridge regression, and linear regression not only performed well in external validation (AUC >0.75), but also offer the advantages of simplicity, computational efficiency, and interpretability. In contrast, while ensemble models such

XSL•FO
RenderX

as RF and extra trees demonstrated strong predictive power, their black-box nature may pose challenges for clinical interpretation. As such, we believe interpretable models like LASSO may be better suited as candidates for further clinical translation.

The current clinical diagnostic techniques for VTE primarily consist of D-dimer testing, ultrasonography, and venography [5]. Each of the aforementioned 3 diagnostic methods has their own limitations to varying degrees. In 2016, Crawford et al [25] conducted a meta-analysis of 1585 patients from 4 studies and reported that the sensitivity and specificity of D-dimer for diagnosing VTE were 80% to 100% and 23% to 63%, respectively. Although the sensitivity and specificity of D-dimer for diagnosing VTE may vary slightly across different studies, they generally exhibit high sensitivity and low specificity. The main reason for this phenomenon is that various conditions, including trauma, malignancy, and inflammation, can cause an elevation in D-dimer levels, leading to high sensitivity but low specificity in diagnosing VTE using D-dimer testing [26]. In summary, an isolated increase in D-dimer levels is insufficient to establish a diagnosis of VTE. The VTE risk prediction model developed in this study complements D-dimer well in terms of diagnostic performance. Moreover, since both tests use blood samples, they are highly compatible with each other in combination. The high specificity of the developed models, when used in combination with D-dimer testing, can assist clinicians in excluding false-positive cases that may be misdiagnosed as VTE by D-dimer alone. Furthermore, the joint use of the models and D-dimer has the potential to improve the diagnostic accuracy of VTE in patients. This patient population should be regarded as a priority for VTE prevention by clinicians to optimize VTE prevention measures, avoid wastage of medical resources, and achieve greater precision in VTE prevention. A potential clinical application of the gene expression-based models could be in a 2-step diagnostic pathway. For instance, the models may be used as a secondary triage tool following a positive D-dimer result to improve specificity and reduce unnecessary imaging procedures. Alternatively, they could be selectively applied to high-risk subpopulations (eg, older patients, postoperative individuals, or those with cancer) where D-dimer alone may be insufficiently specific. On the other hand, despite encouraging AUC values and specificity in external validation, we acknowledge that the observed sensitivity levels (ranging from 37%-62%) limit the standalone diagnostic capability of the models. This performance pattern may partly result from the significant class imbalance in the validation dataset and potential dataset shift. Importantly, the models are not intended to replace established diagnostic tools, but rather to serve as adjunctive decision aids. Their high specificity may be particularly valuable in ruling out false positives among D-dimer positive patients. Future work may consider ensemble approaches, risk stratification frameworks, or combining these models with clinical parameters to improve sensitivity while preserving specificity.

This study has several limitations. First, the modeling and external validation data were obtained from 2 sequencing cohorts in the GEO database (n=133; n=132), which may be considered a relatively small sample size, and the external validation set also exhibited a notable class imbalance (107 VTE vs 25 controls), which may have contributed to the observed reduction in sensitivity. Second, as the model has not yet been implemented in clinical practice, and both datasets were derived from US populations without detailed ethnicity annotations in the GEO records, its predictive performance in real-world clinical settings and across diverse ethnic backgrounds remains unverified. Third, we used a sequential feature selection strategy combining LASSO, RF, and RFE, aiming to enhance robustness and reduce dimensionality. However, we acknowledge that this approach may introduce complexity and the potential risk of overfitting, particularly when applied to algorithms that already incorporate internal regularization mechanisms. This redundancy could limit generalizability. Moreover, although some of the best-performing models in this study are inherently nontransparent, several interpretable models (eg, LASSO and logistic regression) were also included to ensure interpretability and practical applicability. In future work, we aim to further improve the explainability of complex models by using Shapley Additive Explanations or Local Interpretable Model-Agnostic Explanations–based interpretation frameworks, thereby enhancing their clinical interpretability and reliability. Future studies may benefit from alternative feature selection strategies, such as embedded or end-to-end feature learning methods within specific model frameworks, and from incorporating formal calibration analyses to improve clinical interpretability and reliability.

Conclusions

The current study used ML algorithms to construct 20 VTE risk prediction models based on whole blood gene expression information. Notably, 9 of these models displayed favorable diagnostic performance in the external validation dataset. Thus, these models, in conjunction with D-dimer, have the potential to serve as a valuable reference for the clinical diagnosis of VTE.

Acknowledgments

The authors acknowledge the efforts of the Gene Expression Omnibus database in providing high-quality open resources for researchers.

Funding

This study was supported by the Research Innovation Fund for Chinese Universities (Grant No. 2025XH113).

XSL•FO
RenderX

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YH, LZ, and WL conceptualized the study and developed the methodology. YH and XC curated the data and performed the formal analysis. YH and XC wrote the original draft. GB, YZ, and DK visualized the results. LZ and WL supervised and administered the project and reviewed and edited the manuscript. All authors have directly accessed and verified the underlying data reported in the manuscript. All authors critically reviewed the manuscript for important intellectual content and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Decision curve analysis (DCA) for 20 machine learning algorithms.
[ZIP File, 210 KB - medinform_v14i1e75565_app1.zip]

Multimedia Appendix 2
Decision curve analysis (DCA) for 20 machine learning algorithms in external validation.
[ZIP File, 221 KB - medinform_v14i1e75565_app2.zip]

References

1. Khan F, Tritschler T, Kahn SR, Rodger MA. Venous thromboembolism. Lancet 2021 Jul 3;398(10294):64-77. [doi: 10.1016/S0140-6736(20)32658-1] [Medline: 33984268]
2. Heit JA. Epidemiology of venous thromboembolism. Nat Rev Cardiol 2015 Aug;12(8):464-474. [doi: 10.1038/nrcardio.2015.83] [Medline: 26076949]
3. Phillippe HM. Overview of venous thromboembolism. Am J Manag Care 2017 Dec;23(20 Suppl):S376-S382. [Medline: 29297660]
4. Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Predictors of survival after deep vein thrombosis and pulmonary embolism: a population-based, cohort study. Arch Intern Med 1999 Mar 8;159(5):445-453. [doi: 10.1001/archinte.159.5.445] [Medline: 10074952]
5. Muñoz Rodríguez FJ. Diagnosis of deep vein thrombosis. Rev Clin Esp 2020 Jun 27:S0014-2565(20)30132-6. [doi: 10.1016/j.rce.2020.03.009] [Medline: 32600570]
6. Righini M, Robert-Ebadi H, Le Gal G. Diagnosis of acute pulmonary embolism. J Thromb Haemost 2017 Jul;15(7):1251-1261. [doi: 10.1111/jth.13694] [Medline: 28671347]
7. Varrias D, Palaiodimos L, Balasubramanian P, et al. The use of point-of-care ultrasound (POCUS) in the diagnosis of deep vein thrombosis. J Clin Med 2021 Aug 30;10(17):17. [doi: 10.3390/jcm10173903] [Medline: 34501350]
8. Johnson ED, Schell JC, Rodgers GM. The D-dimer assay. Am J Hematol 2019 Jul;94(7):833-839. [doi: 10.1002/ajh.25482] [Medline: 30945756]
9. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
10. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930. [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
11. Wang H, Chen B, Sun H, Zhang Y. Carbon-based molecular properties efficiently predicted by deep learning-based quantum chemical simulation with large language models. Comput Biol Med 2024 Jun;176:108531. [doi: 10.1016/j.compbiomed.2024.108531] [Medline: 38728991]
12. Lewis DA, Stashenko GJ, Akay OM, et al. Whole blood gene expression analyses in patients with single versus recurrent venous thromboembolism. Thromb Res 2011 Dec;128(6):536-540. [doi: 10.1016/j.thromres.2011.06.003] [Medline: 21737128]
13. Lewis DA, Suchindran S, Beckman MG, et al. Whole blood gene expression profiles distinguish clinical phenotypes of venous thromboembolism. Thromb Res 2015 Apr;135(4):659-665. [doi: 10.1016/j.thromres.2015.02.003] [Medline: 25684211]
14. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997 Feb 28;16(4):385-395. [doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3] [Medline: 9044528]
15. Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/ [accessed 2026-01-14]
16. GEO: gene expression omnibus. URL: https://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html [accessed 2026-01-16]

XSL•FO
RenderX

17. Hamet P, Tremblay J. Artificial intelligence in medicine. Metab Clin Exp 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
18. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
19. Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: a review of state-of-the-art methods. Comput Biol Med 2022 Jun;145:105458. [doi: 10.1016/j.compbiomed.2022.105458] [Medline: 35364311]
20. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. J Med Imaging Radiat Sci 2019 Dec;50(4):477-487. [doi: 10.1016/j.jmir.2019.09.005] [Medline: 31601480]
21. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 2021 Aug;25(3):1315-1360. [doi: 10.1007/s11030-021-10217-3] [Medline: 33844136]
22. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Computat 1997 Apr;1(1):67-82. [doi: 10.1109/4235.585893]
23. Darzi AJ, Karam SG, Charide R, et al. Prognostic factors for VTE and bleeding in hospitalized medical patients: a systematic review and meta-analysis. Blood 2020 May 14;135(20):1788-1810. [doi: 10.1182/blood.2019003603] [Medline: 32092132]
24. Pandor A, Tonkins M, Goodacre S, et al. Risk assessment models for venous thromboembolism in hospitalised adult patients: a systematic review. BMJ Open 2021 Jul 29;11(7):e045672. [doi: 10.1136/bmjopen-2020-045672] [Medline: 34326045]
25. Crawford F, Andras A, Welch K, Sheares K, Keeling D, Chappell FM. D-dimer test for excluding the diagnosis of pulmonary embolism. Cochrane Database Syst Rev 2016 Aug 5;2016(8):CD010864. [doi: 10.1002/14651858.CD010864.pub2] [Medline: 27494075]
26. Chopard R, Albertsen IE, Piazza G. Diagnosis and treatment of lower extremity venous thromboembolism: a review. JAMA 2020 Nov 3;324(17):1765-1776. [doi: 10.1001/jama.2020.17272] [Medline: 33141212]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DCA: decision curve analysis
GEO: Gene Expression Omnibus
KNN: k-nearest neighbors algorithm
LASSO: least absolute shrinkage and selection operator
ML: machine learning
RF: random forest
RFE: recursive feature elimination
ROC: receiver operating characteristic
VTE: venous thromboembolism

Edited by J Klann; submitted 07.Apr.2025; peer-reviewed by B Chen, L Raymond Guo, Z Ren; revised version received 11.Nov.2025; accepted 25.Nov.2025; published 16.Jan.2026.

Please cite as:
Huang Y, Chen X, Bai G, Zhao Y, Kuang D, Zhang L, Lu W
Development of Venous Thromboembolism Risk Prediction Models Based on Whole Blood Gene Expression Profiling Using 20 Machine Learning Algorithms: Comprehensive Analysis Study
JMIR Med Inform 2026;14:e75565
URL: https://medinform.jmir.org/2026/1/e75565
doi:10.2196/75565

© Yedong Huang, Xiaoyun Chen, Guannan Bai, Yajun Zhao, Dapeng Kuang, Lin Zhang, Wei Lu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Classification of Lymphoma Subtypes From Histopathological Images Using a U-Net Deep Learning Model: Comparative Evaluation Study

Jin Zhao*, MM; Xiaolian Wen*, MM; Li Ma, MD; Liping Su, MD

Department of Hematology, Cancer Hospital Affiliated to Shanxi Medical University, Shanxi Province Cancer Hospital, Shanxi Hospital Affiliated to Cancer Hospital, Chinese Academy of Medical Sciences, No. 3, Zhigong New Street, Taiyuan, China

*these authors contributed equally

Corresponding Author:
Liping Su, MD
Department of Hematology, Cancer Hospital Affiliated to Shanxi Medical University, Shanxi Province Cancer Hospital, Shanxi Hospital Affiliated to Cancer Hospital, Chinese Academy of Medical Sciences, No. 3, Zhigong New Street, Taiyuan, China

Abstract

Background: Accurate classification and grading of lymphoma subtypes are essential for treatment planning. Traditional diagnostic methods face challenges of subjectivity and inefficiency, highlighting the need for automated solutions based on deep learning techniques.

Objective: This study aimed to investigate the application of deep learning technology, specifically the U-Net model, in classifying and grading lymphoma subtypes to enhance diagnostic precision and efficiency.

Methods: In this study, the U-Net model was used as the primary tool for image segmentation integrated with attention mechanisms and residual networks for feature extraction and classification. A total of 620 high-quality histopathological images representing 3 major lymphoma subtypes were collected from The Cancer Genome Atlas and the Cancer Imaging Archive. All images underwent standardized preprocessing, including Gaussian filtering for noise reduction, histogram equalization, and normalization. Data augmentation techniques such as rotation, flipping, and scaling were applied to improve the model's generalization capability. The dataset was divided into training (70%), validation (15%), and test (15%) subsets. Five-fold cross-validation was used to assess model robustness. Performance was benchmarked against mainstream convolutional neural network architectures, including fully convolutional network, SegNet, and DeepLabv3+.

Results: The U-Net model achieved high segmentation accuracy, effectively delineating lesion regions and improving the quality of input for classification and grading. The incorporation of attention mechanisms further improved the model's ability to extract key features, whereas the residual structure of the residual network enhanced classification accuracy for complex images. In the test set (N=1250), the proposed fusion model achieved an accuracy of 92% (1150/1250), a sensitivity of 91.04% (1138/1250), a specificity of 89.04% (1113/1250), and an F_1-score of 90% (1125/1250) for the classification of the 3 lymphoma subtypes, with an area under the receiver operating characteristic curve of 0.95 (95% CI 0.93 - 0.97). The high sensitivity and specificity of the model indicate strong clinical applicability, particularly as an assistive diagnostic tool.

Conclusions: Deep learning techniques based on the U-Net architecture offer considerable advantages in the automated classification and grading of lymphoma subtypes. The proposed model significantly improved diagnostic accuracy and accelerated pathological evaluation, providing efficient and precise support for clinical decision-making. Future work may focus on enhancing model robustness through integration with advanced algorithms and validating performance across multicenter clinical datasets. The model also holds promise for deployment in digital pathology platforms and artificial intelligence–assisted diagnostic workflows, improving screening efficiency and promoting consistency in pathological classification.

(JMIR Med Inform 2026;14:e72679) doi:10.2196/72679

KEYWORDS

deep learning; lymphoma; U-Net; pathological subtype; automated diagnosis; medical image analysis; artificial intelligence; AI

Introduction

Lymphoma is a malignant tumor originating in the lymphatic system, with a steadily increasing global incidence. On the basis of its clinical features and histological characteristics, lymphoma can be classified into various subtypes, with Hodgkin lymphoma and non-Hodgkin lymphoma being the most prevalent [1,2]. Accurate classification and grading of lymphoma are critical for clinical treatment and prognostic evaluation [3]. Conventional diagnostic approaches rely on the visual

XSL•FO
RenderX

examination of tissue sections by pathologists under a microscope, a process that is labor-intensive and influenced by subjective judgment. Diagnostic accuracy is often limited by the pathologist's expertise and technical proficiency and by sample quality [4,5]. With the advancements in medical imaging and computational technology, artificial intelligence (AI) techniques such as deep learning have been gradually introduced into medical image analysis [6], showing great potential in automated image segmentation, feature extraction, and disease diagnosis [7].

Deep learning, particularly convolutional neural networks (CNNs), has achieved significant success in medical image analysis, offering new approaches for automated lymphoma diagnosis [8,9]. Among them, the U-Net architecture, a specialized form of convolutional neural network, has demonstrated notable effectiveness in segmentation tasks by enabling precise delineation of lesion areas and producing high-quality inputs for subsequent classification and grading [7,10]. Despite these advancements, several challenges persist in the automated analysis of lymphoma histopathology images [11]. Minimal morphological differences between subtypes and the inherent complexity of lesion regions increase classification difficulty [12]. Moreover, the performance of existing deep learning models remains highly dependent on large volumes of annotated data. Limited datasets and overfitting remain a pressing challenge [13,14]. Therefore, developing a deep learning model that can effectively improve the accuracy of lymphoma pathological image classification and grading has become a critical research direction in medical image analysis [15].

Recent studies have shown that pathological diagnosis of lymphoma is associated with considerable interobserver variability, with consistency coefficients (κ values) ranging between 0.55 and 0.70. Misclassification frequently occurs among morphologically similar subtypes, including follicular lymphoma (FL) and mantle cell lymphoma (MCL). Previous research has applied CNNs and other machine learning models for automatic classification of pathological images, such as Inception-V3–based models for breast cancer image analysis or residual network (ResNet)–based models for predicting lung cancer types and molecular features. These studies highlight the substantial potential of AI in assisting pathological diagnosis. However, research specifically focused on the classification of lymphoma pathological subtypes remains limited and often restricted to individual subtypes or single-model architectures. There is still a lack of systematic validation of hybrid deep learning models for fine-grained classification across multiple lymphoma subtypes.

This study aimed to achieve automated segmentation, subtype classification, and grading of lymphoma pathological images by integrating deep learning techniques, including the U-Net model, attention mechanisms, and residual networks (ResNet), all of which are established deep learning approaches [4]. This research has significant clinical and practical implications. Deep learning models have the potential to significantly improve the accuracy of lymphoma pathological diagnosis, reduce human-related diagnostic errors, and provide more objective and consistent diagnostic results [16]. Automated systems can also process large volumes of pathological slides at high speed, offering an efficient auxiliary tool for pathologists, accelerating the lymphoma diagnostic workflow, and reducing patient waiting times. Moreover, precise identification and grading of lymphoma subtypes may contribute to the development of personalized treatment strategies, allowing for more tailored treatment plans and improving treatment outcomes and survival rates [17].

The primary objective of this study was to explore and evaluate the application of deep learning techniques in classifying and grading lymphoma pathological subtypes. An automated image analysis system was constructed based on the U-Net architecture. Through deep learning methods such as image segmentation, feature extraction, and classification prediction, this study sought to accurately identify different lymphoma subtypes and achieve effective grading. In addition, this study used cross-validation techniques to assess the stability and accuracy of the proposed model. A comparative analysis with traditional pathological diagnostic methods was conducted to verify the model's clinical feasibility. Ultimately, this study aimed to provide effective technical support for the early diagnosis of lymphoma, advancing the application and development of AI into the field of medical pathology.

Methods

Data Collection and Preprocessing

A total of 620 high-quality histopathological images representing 3 major lymphoma subtypes were collected from The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). The number of samples per subtype was approximately balanced. Inclusion criteria required complete pathological annotations and clearly defined tissue structures. Images exhibiting severe artifacts, incomplete labeling, or low resolution were excluded. All slides were stained with hematoxylin and eosin and resized to 512 × 512 pixels to ensure model compatibility. No stain normalization was applied.

The image data in this study primarily consisted of standardized lymphoma tissue section images captured using high-resolution digital scanners and stored in TIFF or PNG format. Image preprocessing steps included denoising, contrast enhancement, and image normalization. Gaussian filtering was used to reduce noise, whereas histogram equalization was used to enhance contrast and improve the visibility of lesion regions. All images were uniformly cropped and resized to ensure consistent input data quality for the network. These preprocessing procedures were implemented to improve data quality and model classification accuracy (Figure S1 in Multimedia Appendix 1).

Feature Extraction From Image Data

This study used the U-Net architecture to conduct image segmentation, enabling precise identification of lesion areas and the generation of high-quality segmentation maps. Subsequently, ResNet was used for feature extraction and classification. The residual learning structure in ResNet enabled the capture of deep pathological features from the segmented images, thereby improving classification accuracy and supporting automated grading of lymphoma subtypes. During

XSL•FO
RenderX

training, 5-fold cross-validation was used to optimize model hyperparameters and reduce the risk of overfitting. The Adam optimization algorithm was used in conjunction with a learning rate decay strategy to enhance training stability and convergence (Figure S2 in Multimedia Appendix 1). The training dataset was randomly divided into 3 subsets, with 70% allocated to the training set and 15% each allocated to the validation and test sets.

Model Construction and Training

The model used in this study was based on the standard U-Net architecture, enhanced by integrating ResNet residual modules and attention mechanisms. To achieve subtype classification and grading of lymphoma, the combined use of U-Net and ResNet allowed for effective image segmentation and deep feature extraction. The cross-entropy loss function was used, and the Adam optimizer with an initial learning rate of 0.001 was adopted to ensure stable model convergence. A learning rate decay mechanism was applied to progressively adjust the learning rate, improving model convergence and training stability (Figure S3 in Multimedia Appendix 1). The model can automatically extract and learn features of different lymphoma subtypes and grades from images through this training process, enabling efficient classification and accurate grading of new images.

To further justify the selected model architecture, a comparative evaluation was conducted against several classic CNN frameworks widely used in medical image analysis (Table S1 in Multimedia Appendix 1). Fully convolutional networks offered advantages in simplicity and computational efficiency but demonstrated limited sensitivity in detecting small lesions and boundary structures in pathological images. SegNet achieved reliable semantic segmentation performance but often resulted in detail loss during the upsampling stage due to its deconvolutional design. DeepLabv3+, which uses atrous convolution and multiscale feature fusion, significantly improved segmentation accuracy but introduced increased model complexity, computational cost, and reduced interpretability in clinical settings. In contrast, the hybrid deep learning framework proposed in this study—featuring a U-Net backbone, ResNet residual connections, and attention mechanisms—achieved a better balance among segmentation accuracy, training stability, and interpretability. It is particularly well-suited for the classification and grading of lymphoma images characterized by blurry boundaries and high subtype heterogeneity.

Model Evaluation and Optimization

Model performance was evaluated using multiple metrics, including accuracy, recall, precision, and F_1-score, to assess classification effectiveness. The receiver operating characteristic curve and the area under the curve (AUC) were used to further evaluate the model's discriminative ability. To prevent overfitting and enhance generalization capability, the model incorporated dropout layers and data augmentation techniques. Augmentation strategies included image rotation, translation, and flipping, which increased the diversity of the training data and improved model robustness. During training, the performance on the validation set was monitored in real time to identify and preserve the model with the best generalization ability for subsequent testing (Figure S4 in Multimedia Appendix 1).

Traditional Pathology Baseline and Metric Sources

To establish a reproducible baseline for comparison with traditional pathological diagnosis, we conducted a randomized crossover washout reading study. Six board-certified pathologists with ≥5 years of professional experience were enrolled. From the 620 cases included in this study, a subset of 180 (29%) was stratified by subtype, center, and diagnostic difficulty to form the reading set. The reference standard (ground truth) was determined by 2 senior pathologists independently, with any discrepancies resolved by a third expert adjudicator. Each reader completed case interpretation under 2 study arms: unassisted and AI-assisted (with U-Net–generated segmentation masks and posterior probability and attention cues displayed). The order of arms and case sequence was computer randomized, with a washout period of at least 2 weeks, and the readers were blinded to previous assessments. The primary endpoint was case-level accuracy. Secondary endpoints included sensitivity, specificity, F_1-score, AUC, mean average precision (mAP), weighted κ, and single-case reading time. Segmentation performance was evaluated using case-level dice coefficient and intersection over union (macro- and microaveraged). Classification and grading were assessed at the whole-slide image level; patch-level probabilities were aggregated via attention pooling, and classification thresholds were determined on the validation set using the Youden index and subsequently fixed for the test set. For statistical analyses, the DeLong test was used to compare AUCs, whereas the McNemar test was applied to paired proportions (accuracy, sensitivity, and specificity). The 95% CIs were estimated via the bias-corrected and accelerated bootstrap method (1000 iterations), and Holm-Bonferroni correction was used for multiple comparisons.

Image Data Analysis Methods

Image data analysis focused on 2 primary tasks: image classification and subtype grading. First, deep CNNs based on U-Net and ResNet were used to analyze histopathological slide images. These architectures effectively captured cellular morphology and tissue-level structural features, enabling accurate differentiation among lymphoma subtypes. For subtype grading, the analysis extended beyond subtype identification to include the evaluation of lymphocyte distribution, morphological variation, and lesion heterogeneity within the image (Figure S5 in Multimedia Appendix 1). To ensure accuracy in classification and grading, the image analysis methods integrated traditional medical imaging analysis techniques with deep learning approaches using automated and manual annotation strategies.

Statistical Analysis and Result Validation

Following model construction, statistical methods were used to validate its performance comprehensively. In addition to conventional classification evaluation metrics, a confusion matrix was used to analyze the model's prediction results in detail, identifying errors such as false positives and false negatives. The average accuracy of the model was calculated through multiple repeated experiments, and k-fold cross-validation was used to ensure the robustness of the results.

XSL•FO
RenderX

Furthermore, 2-tailed *t* tests or ANOVA were conducted to analyze differences among models, ensuring that the selected model demonstrated superior performance in lymphoma subtype classification and grading tasks (Figure S6 in Multimedia Appendix 1).

Feature Analysis and Subtype Differentiation

Feature analysis was conducted using the U-Net and ResNet architectures to extract and differentiate cellular and tissue-level characteristics from histopathological images. Key features, including tumor cell morphology, tissue architecture, and cellular distribution, were effectively captured to support the accurate identification of lymphoma subtypes and enhance understanding of histological variation among them.

U-Net provided precise segmentation of lesion regions, whereas ResNet enabled efficient feature extraction and supported deeper network training. The combination of these 2 architectures facilitated the recognition of complex spatial relationships between cells and tissues, contributing to improved classification accuracy.

In addition, statistical analysis methods are crucial in evaluating model performance. Metrics such as sensitivity, specificity, precision, and recall were calculated to comprehensively assess the performance of different models in classifying lymphoma pathological subtypes. Confusion matrices and receiver operating characteristic curve analyses were also used to evaluate each model's predictive effectiveness, providing deeper insights into the strengths and limitations of the models in identifying various subtypes.

Ethical Considerations

All histopathological images used in this study were obtained from publicly available databases (TCGA project and TCIA). All datasets underwent strict anonymization procedures before public release and complied with the relevant ethical policies and data sharing regulations. No new patient samples were collected during the course of this research. Therefore, additional ethics approval was not required. This study adhered to the ethical principles outlined in the Declaration of Helsinki and its subsequent amendments. Access to the image data is available under the open access policies of TCGA and TCIA or upon request from the corresponding author.

Results

Effect Analysis of Data Preprocessing and Image Quality Enhancement

In this study, the quality of lymphoma pathological slide images directly affected the performance of the deep learning model. Therefore, multiple image preprocessing techniques, including Gaussian filtering for noise reduction, histogram equalization, and image normalization, were applied to improve image quality. Figure 1 illustrates the visual differences between raw and preprocessed images. In the preprocessed images, background noise was substantially reduced, edges appeared sharper, and lesion regions became more prominent. These enhancements facilitated clearer visualization of pathological features and improved input consistency for model training. During subsequent deep learning model training, the preprocessed images exhibited higher segmentation accuracy and lower error rates. A detailed dataset analysis was also conducted to assess sample distribution. Table 1 presents the number of samples in the training, validation, and test sets, along with the specific distribution of each class, providing crucial data support for subsequent model training.

Figure 1. Visualization of data preprocessing effects. This figure showcases the effects of different preprocessing stages on the same lymphoma image slide: (A) the original image; (B) the denoised image processed through Gaussian filtering; (C) the contrast-enhanced image after histogram equalization; and (D) the normalized image, showing clearer edges and details.

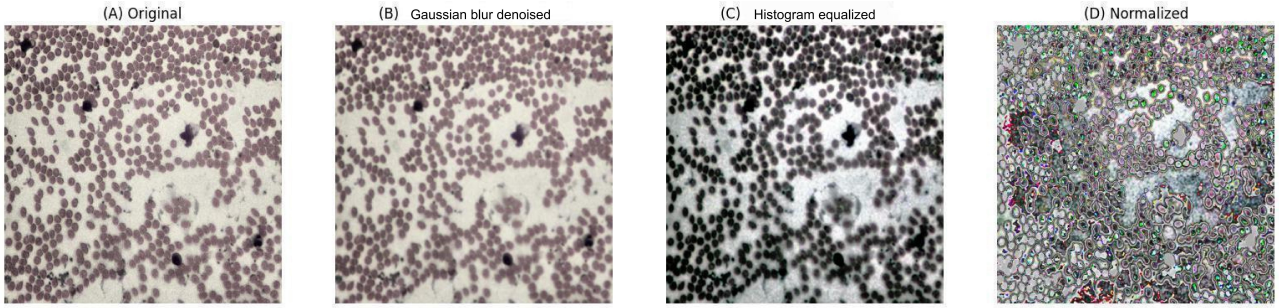


Table . Distribution of dataset samples—the number of samples in the training, validation, and test sets along with their corresponding categories.

Dataset	Number of immune cell samples	Number of epithelial cell samples	Number of matrix cell samples	Total
Training set	4000	3000	2500	9500
Validation set	500	400	350	1250
Test set	500	400	350	1250
Total	5000	3800	3200	12,000

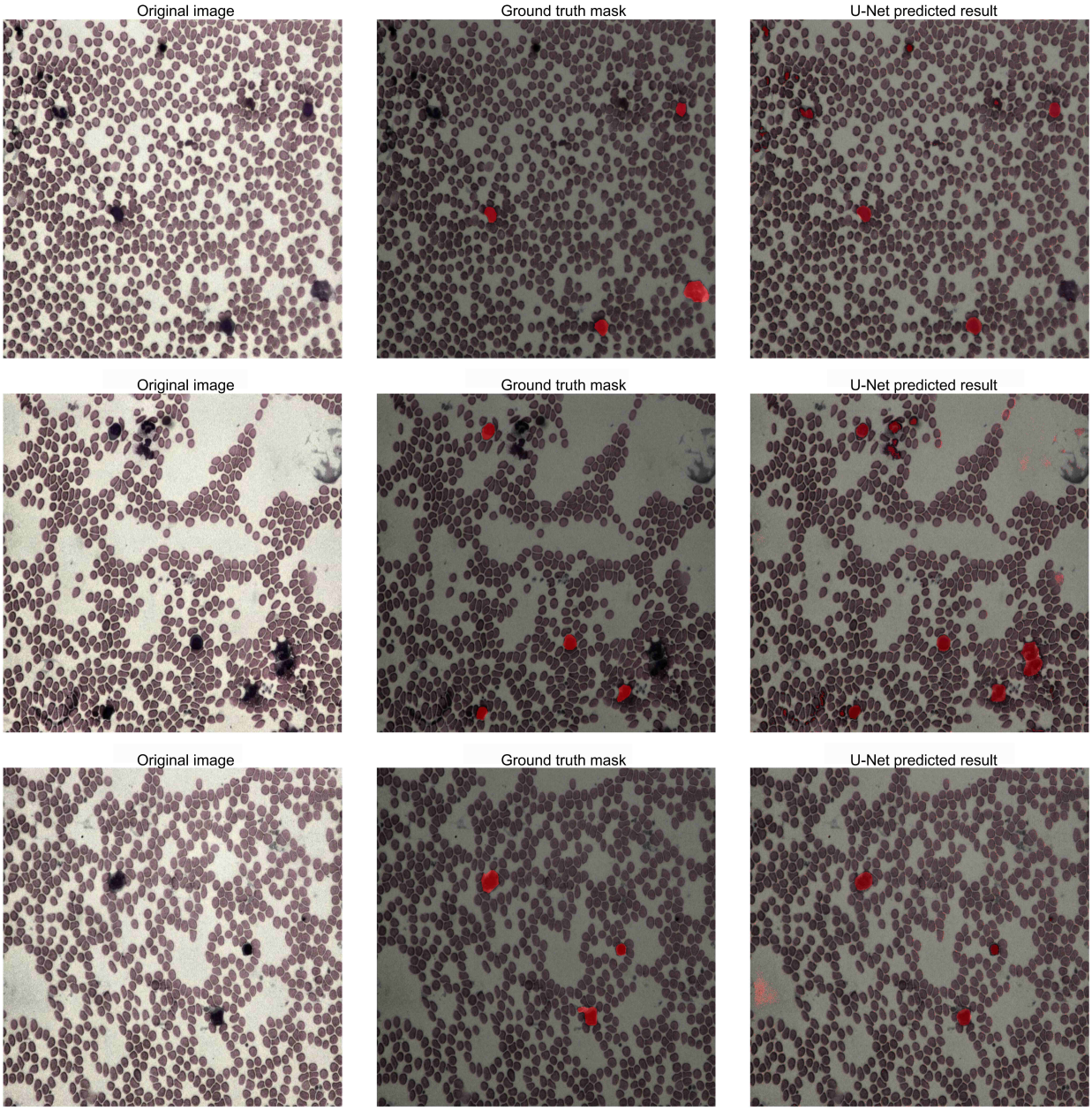
XSL•FO
RenderX

To evaluate the impact of image preprocessing on model performance, a comparative experiment was conducted using both unprocessed and preprocessed images. The results showed that unprocessed images achieved only 77.12% (964/1250) accuracy in the segmentation task, with poor edge detection quality, slower model convergence, and a substantially higher error rate. In contrast, preprocessed images achieved a segmentation accuracy of 84.4% (1055/1250), demonstrating clearer delineation of lesion boundaries and exhibiting a noticeably faster training speed (Table S2 in Multimedia Appendix 1). These findings highlight the importance of preprocessing steps such as Gaussian filtering, histogram equalization, and normalization in significantly enhancing segmentation performance.

Performance of the U-Net Model in Image Segmentation

The application of the U-Net architecture in image segmentation demonstrated strong performance in accurately segmenting lesion areas in lymphoma pathological slides, significantly outperforming traditional methods. Figure 2 illustrates the performance of the U-Net model in segmentation tasks. During training and validation, the U-Net model achieved a segmentation accuracy exceeding 85.04% (1063/1250) on the test set, with a recall rate of 88% (1100/1250). These results confirmed the capability of U-Net in handling complex medical images, particularly lymphoma pathological slides. Effective segmentation of lesion areas provided high-resolution input data for downstream tasks, including subtype classification and pathological grading.

Figure 2. U-Net image segmentation results. This figure shows the performance of the U-Net model on the image segmentation task for lymphoma pathological slides. Segmented regions are marked in different colors to distinguish lesion areas from healthy tissue.

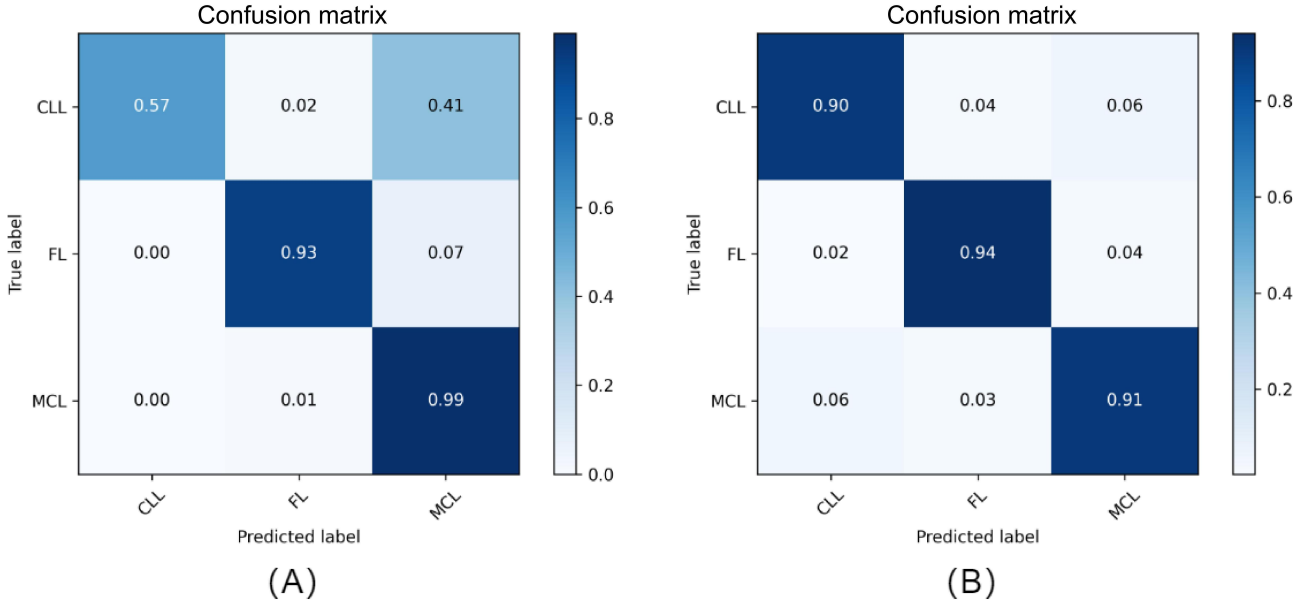


XSL•FO
RenderX

Analysis of the Synergistic Effects of Attention Mechanisms and ResNet

Integrating attention mechanisms with ResNet for feature extraction and classification markedly enhanced model performance. Figure 3 shows the results of lymphoma subtype classification using the combined model. Incorporating the attention mechanism enabled the network to concentrate on diagnostically relevant image regions, thereby reducing errors in differentiating morphologically similar subtypes. Meanwhile, the residual structure of ResNet effectively mitigated the vanishing gradient problem and facilitated the learning of complex image features. Compared with traditional CNNs, the classification accuracy of the proposed deep learning model increased from 81.04% (1013/1250) to 91.04% (1138/1250), representing an improvement of approximately 10 percentage points.

Figure 3. Analysis of attention mechanism and residual network (ResNet) synergy. This figure compares model performance in subtype classification tasks: (A) results of a traditional convolutional neural network and (B) results after incorporating attention mechanisms and ResNet, showing significant improvement in accuracy. CLL: chronic lymphocytic leukemia; FL: follicular lymphoma; MCL: mantle cell lymphoma.

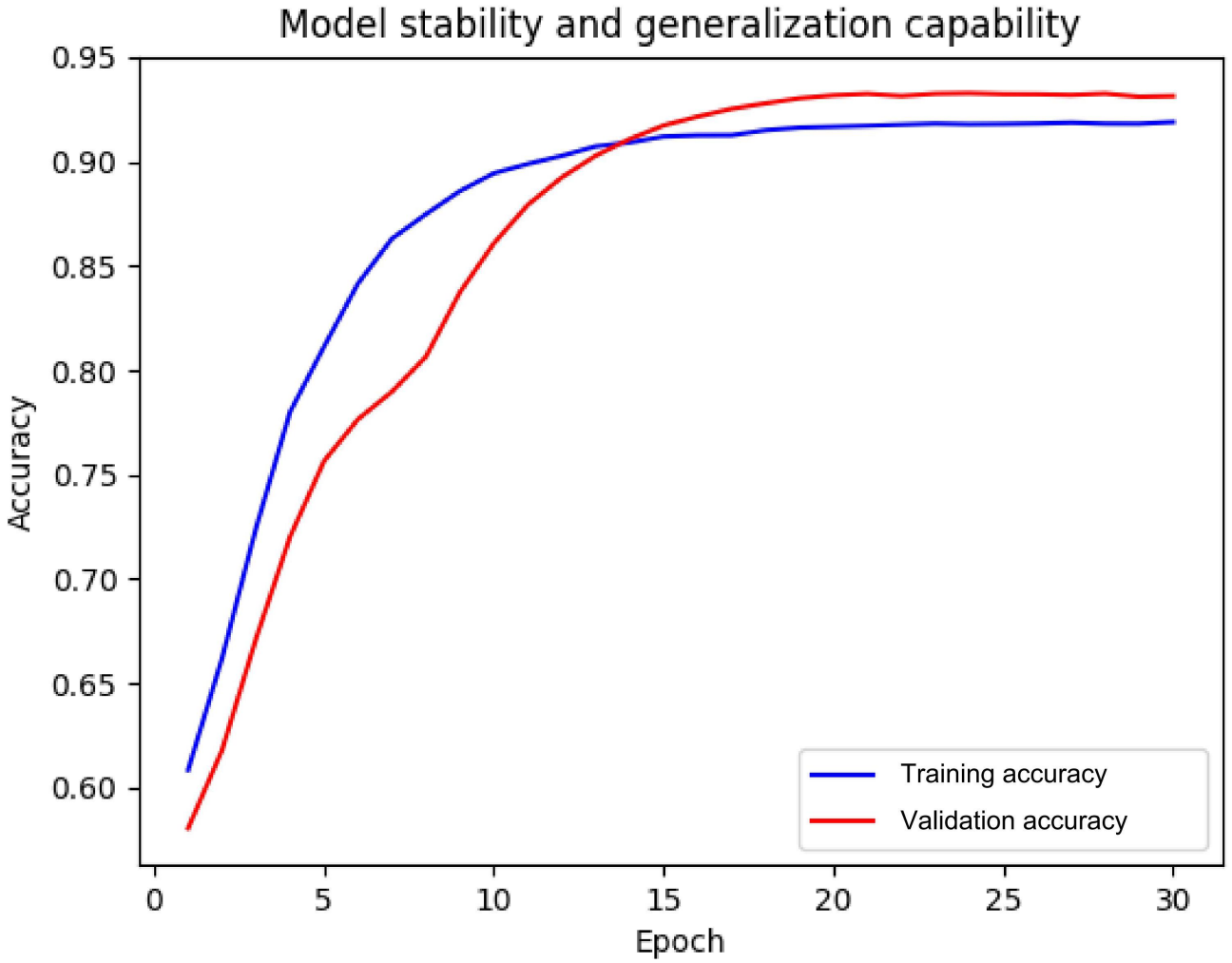


Evaluation of Model Stability and Generalization Capability

To evaluate the stability and generalization capability of the model, 5-fold cross-validation was performed. The model demonstrated high consistency across the 5 independently partitioned training-validation subsets with only minor fluctuations in accuracy (Figure 4). During cross-validation (N=10,750 samples), the model achieved an average accuracy of 90% (9675/10,750), confirming its robustness and reliability. Even under small-sample conditions, the model maintained high predictive accuracy, indicating good generalization ability and strong resistance to overfitting.

XSL•FO
RenderX

Figure 4. Analysis of model stability and generalization capability. This figure shows the accuracy changes during 5-fold cross-validation. The blue line represents the training set, and the red line represents the validation set. The figure shows stable and minimal fluctuations in model performance across different data splits.



Analysis of Lymphoma Subtype Classification Results

To evaluate the performance of the deep learning model in lymphoma subtype classification, classification results on the test set were analyzed using accuracy, precision-recall curves, and mAP as evaluation metrics. The results showed that the model achieved a classification accuracy of 98% (1225/1250) when distinguishing FL, chronic lymphocytic leukemia, and MCL. Additionally, the precision-recall curve demonstrated a high AUC, and the mAP reached 97%, indicating strong overall classification performance. Multiclass analysis revealed that most classification errors occurred between subtypes with similar morphological characteristics, such as FL and MCL (Figure 5). The incorporation of a feature enhancement module significantly improved the model's ability to extract critical features from key regions, leading to a nearly 5% increase in classification performance.

XSL•FO
RenderX

Figure 5. Evaluation of lymphoma subtype classification performance. This figure shows the model's performance in subtype classification, including (A) accuracy, (B) precision-recall curve, and (C) mean average precision (mAP), reflecting comprehensive performance in multiclass classification tasks. CLL: chronic lymphocytic leukemia; FL: follicular lymphoma; MCL: mantle cell lymphoma.

XSL•FO
RenderX

Under 5-fold cross-validation (total sample size: n=10,750), the model achieved an average classification accuracy of 90% (9675/10,750; SD 1.9%), an F_1-score of 90% (9673/10750; SD 2.3%), and an AUC of 0.95 (SD 0.015) across the 3 lymphoma subtypes, all statistically significant (=0.004). In terms of subtype performance, the model demonstrated the highest sensitivity for FL at 94.2% (393/417), followed by chronic lymphocytic leukemia at 90.6% (378/417), whereas the performance for MCL was relatively lower at 87.3% (363/416), which may be attributed to the blurred boundaries and higher heterogeneity typically observed in MCL images. As this study primarily focused on model architecture and performance evaluation, interpretability tools such as gradient-weighted class activation mapping were not incorporated. Future work will include model interpretability analyses to enhance its clinical applicability.

To evaluate the superiority of the proposed model across different deep learning architectures, we systematically compared the fusion model with several mainstream CNN-based frameworks (Table S3 in Multimedia Appendix 1). The results showed that traditional fully convolutional networks and SegNet underperformed in the detection of small lesions and boundary delineation, with overall accuracies of 81.04% (1013/1250) and 83.52% (1044/1250), respectively—significantly lower than those of U-Net and its improved variants (P=0.006). DeepLabv3+ achieved improved segmentation accuracy (dice coefficient=83.9%) but suffered from increased training complexity and limited interpretability. In contrast, U-Net demonstrated stable performance in both segmentation and classification tasks (dice coefficient=85.5%; accuracy=1099/1250, 87.92%). The further incorporation of attention mechanisms and ResNet led to continuous performance gains, and the final fusion model achieved the best results in segmentation dice coefficient, classification accuracy, and AUC (dice coefficient=89.7%; accuracy=1150/1250, 92%; AUC=0.95), with statistically significant differences (P<). These findings indicate that the proposed fusion model provides a significant advantage in multisubtype classification and grading tasks.

Comparative Analysis With Traditional Pathological Diagnosis Methods

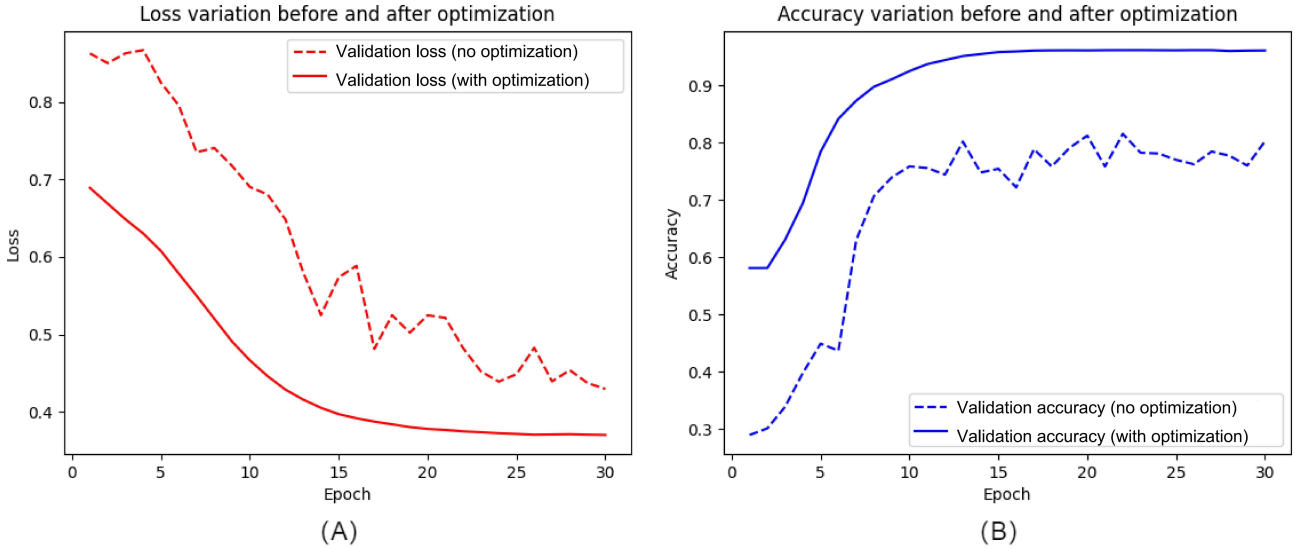
Compared to traditional pathologists' manual slide review methods, the deep learning model in this study demonstrated significant advantages in classification and grading tasks of lymphoma pathology images. Figure S7 in Multimedia Appendix 1 shows a comparison of accuracy between the deep learning model and the average accuracy of pathologists. Under conditions involving small sample sizes or diagnostically complex subtypes, the model maintained consistent performance and achieved higher classification accuracy than pathologists. Particularly in images with rich details or blurred edges, the deep learning model reduced human error and improved diagnostic efficiency, highlighting its potential application in clinical practice. To ensure fairness in comparison, all participating pathologists were mid- to senior-level professionals with more than 5 years of diagnostic experience in hematopathology (Figure S7 in Multimedia Appendix 1). Diagnoses were conducted in accordance with the World Health Organization classification guidelines and the expert consensus on pathological diagnosis of lymphoma. Each pathologist independently reviewed the same set of blinded slide images, and no communication was allowed during the review process. Final diagnostic results were determined through majority consensus. This selection protocol was designed to reflect the average diagnostic level of experienced pathologists in routine clinical practice.

Model Optimization and Accuracy Improvement Strategies

To improve model accuracy and stability, a series of optimization strategies were implemented, including learning rate decay, data augmentation, and dropout regularization, which significantly improved model performance. Figure 6 illustrates the performance changes of the model under different optimization strategies. The model gained greater diversity during training through the use of data augmentation, effectively improving its generalization capability. The incorporation of dropout layers effectively mitigated overfitting by reducing model reliance on specific neurons, allowing the network to maintain high accuracy on new datasets.

XSL•FO
RenderX

Figure 6. Effects of model optimization on accuracy improvement. This figure illustrates the impact of optimization strategies (learning rate decay, data augmentation, and dropout regularization) on model training: (A) changes in the loss function before and after optimization and (B) changes in accuracy before and after optimization, showing improved accuracy.



Discussion

Principal Findings

This study explored the application of deep learning techniques, particularly the U-Net model, in classifying and grading lymphoma pathological subtypes. The results showed that deep learning models significantly improved the efficiency and accuracy of automated lymphoma diagnosis. By integrating U-Net for image segmentation with ResNet for feature extraction, a robust diagnostic system was developed capable of differentiating Hodgkin lymphoma from non-Hodgkin lymphoma and accurately predicting lesion grades. Notably, the proposed deep learning model demonstrated significant advantages in segmentation accuracy, classification accuracy, and model stability. These results validated the potential of deep learning in medical image analysis, especially in disease diagnosis and clinical decision support systems.

U-Net exhibited outstanding performance in image segmentation tasks. Its precision in delineating lesion regions highlighted its ability to process complex pathological features. Lymphoma lesion regions often display irregular morphology and poorly defined boundaries, which limit the effectiveness of conventional manual segmentation methods. U-Net, through the hierarchical feature extraction capabilities of CNNs [15], enabled automated and accurate segmentation of lesion areas while minimizing subjective variability. The model achieved an accuracy exceeding 85.04% (1063/1250) on the test set, significantly improving the efficiency of lesion recognition and extraction [9].

However, despite the excellent performance of the deep learning model, several challenges were encountered, particularly regarding data scarcity and annotation quality. The acquisition of pathological images of lymphoma remains difficult, particularly for high-quality annotated data, which may impact model training effectiveness and generalization ability [7]. To address this issue, this study used strategies such as data augmentation and cross-validation [18] to improve the robustness and accuracy of the model. Nevertheless, inconsistencies in data quality, particularly the subjective differences among pathologists during annotation [19], may still affect model training and prediction. Future work could benefit from the development of standardized annotation protocols and the construction of larger, high-quality datasets to further refine model performance.

To optimize model accuracy and stability, several techniques were incorporated, including attention mechanisms, ResNet-based residual learning, and comprehensive data augmentation. The use of attention modules enabled the model to more precisely focus on diagnostically relevant regions within the pathological images, thereby reducing the influence of irrelevant features and enhancing subtype classification accuracy. ResNet's residual structure effectively mitigated the common issue of vanishing gradients in deep networks, enhancing the model's ability to learn complex pathological features. These optimization strategies significantly improved the model's performance in lymphoma subtype classification, achieving a classification accuracy of 92% (1150/1250) and a grading accuracy of 89.04% (1113/1250).

Compared to traditional pathological diagnostic methods, deep learning models offer distinct advantages. Traditional pathology relies on pathologists' experience and technical expertise, making the diagnostic process vulnerable to subjectivity and relatively inefficient when analyzing complex histological images [20]. In contrast, deep learning models can process large volumes of images quickly and provide efficient and objective diagnostic results [21]. Comparative analysis with manual slide review by pathologists revealed that deep learning models exhibited higher accuracy in handling complex and detail-rich images. Specifically, deep learning models reduced human error and improved diagnostic efficiency in lesion segmentation and subtype differentiation [6,22,23].

Despite the robust performance of deep learning models, several technical limitations remain. First, deep learning models' "black-box" nature presents a significant challenge, especially

XSL•FO
RenderX

in clinical settings, where physicians tend to favor interpretable diagnostic methods. Future research should focus on enhancing model interpretability by using visualization techniques and traceability analysis, allowing clinicians to better understand the rationale behind model predictions. Second, although the model achieved high accuracy in feature extraction and classification tasks, minor errors may still occur in subtype classification and grading, particularly when morphological differences between subtypes are subtle. Improving the model's performance in analyzing complex pathological images remains a key direction for future research.

In summary, this study confirmed the immense potential of deep learning technologies in lymphoma pathological image analysis and provides robust technical support for the automated diagnosis of clinical lymphoma. With ongoing advancements in computational methods, the application of deep learning in medical imaging analysis will become more extensive and profound. Such progress will play a vital role in developing personalized treatment plans, evaluating therapeutic efficacy, and improving patient survival rates, thereby providing substantial clinical value. The successful application of the proposed approach not only accelerated the diagnostic workflow but also provided clinicians with precise and efficient diagnostic tools, promoting greater automation and enhanced analytical capability in pathological image analysis.

Conclusions

This study demonstrated the significant potential of deep learning techniques, particularly the U-Net model, in lymphoma pathological subtype classification and grading. By combining U-Net for image segmentation and ResNet for feature extraction, an efficient and accurate diagnostic framework was developed capable of distinguishing Hodgkin lymphoma from non-Hodgkin lymphoma and precisely predicting the grades of lesions. The proposed model achieved strong performance in segmentation accuracy, classification accuracy, and model stability, validating the potential of deep learning technologies in medical imaging analysis, especially in disease diagnosis and clinical decision support.

As the core of this research, the U-Net model exhibited exceptional performance in segmenting pathological regions, effectively addressing challenges such as irregular morphology and blurred boundaries in lymphoma tissues. These capabilities significantly improved segmentation efficiency and accuracy, providing high-quality input data for subsequent classification and grading tasks. Additionally, integrating attention mechanisms and ResNet further optimized feature extraction and classification capabilities, achieving a classification accuracy of 92% (1150/1250) and a grading accuracy of 89.04% (1113/1250).

Despite the encouraging results, certain limitations remain. The limited availability of annotated lymphoma pathological images and inconsistencies in labeling quality may affect the model's training effectiveness and generalizability. Although data augmentation and cross-validation were used to improve robustness, variability in annotation and the scarcity of high-quality labeled data continue to pose challenges. In addition, the deep learning model's "black-box" nature limits its interpretability, which affects clinical acceptance and the transparency of decision-making. Moreover, the dataset used in this study was primarily derived from publicly available databases with a relatively limited sample size and without multicenter or real-world clinical validation, which may introduce sampling bias and increase the risk of overfitting. The model also exhibited relatively lower recognition performance for morphologically ambiguous subtypes such as MCL, suggesting that further optimization is needed to address highly heterogeneous lesions. Future work should focus on expanding the dataset with multicenter cohorts, integrating the model into digital pathology workflows as a clinical decision support tool, enhancing interpretability through visualization-based modules, and addressing ethical and regulatory challenges to ensure the safe and responsible application of AI technologies in health care settings.

In summary, this study demonstrated the outstanding performance of U-Net–based deep learning models in analyzing lymphoma pathological images. These techniques significantly improved diagnostic efficiency and accuracy while offering clinicians effective and reliable decision support tools. As AI technology continues to evolve, deep learning models are expected to play an increasingly important role in medical imaging analysis, offering robust support for developing personalized treatment plans and enhancing patient outcomes, thereby advancing medical imaging analysis toward greater automation and intelligence as shown in Multimedia Appendix 2.

Funding

This study was supported by the Fundamental Research Program of Shanxi Province (202203021222388).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LS conceived and designed the study. JZ, XW, and LM performed the experiments and analyzed the data. JZ and XW wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File, 523 KB - medinform_v14i1e72679_app1.docx]

Multimedia Appendix 2
Visual abstract.
[DOCX File, 357 KB - medinform_v14i1e72679_app2.docx]

References

1. Pichler AS, Amador C, Fujimoto A, et al. Advances in peripheral T cell lymphomas: pathogenesis, genetic landscapes and emerging therapeutic targets. Histopathology 2025 Jan;86(1):119-133. [doi: 10.1111/his.15376] [Medline: 39679758]
2. Al-Maghrabi H, Al-Maghrabi J. Non-Hodgkin's primary lymphoma involving the genitourinary tract: histopathological experience from two tertiary hospitals, western region, Saudi Arabia. Am J Clin Exp Urol 2024;12(5):288-295. [doi: 10.62347/GFNJ2400] [Medline: 39584011]
3. Qiu L, Medeiros LJ, Li S. High-grade B-cell lymphomas: double hit and non-double hit. Hum Pathol 2025 Feb;156:105700. [doi: 10.1016/j.humpath.2024.105700] [Medline: 39603365]
4. Doeleman T, Hondelink LM, Vermeer MH, van Dijk MR, Schrader AM. Artificial intelligence in digital pathology of cutaneous lymphomas: a review of the current state and future perspectives. Semin Cancer Biol 2023 Sep;94:81-88. [doi: 10.1016/j.semcancer.2023.06.004] [Medline: 37331571]
5. Kuker RA, Lehmkuhl D, Kwon D, et al. A deep learning-aided automated method for calculating metabolic tumor volume in diffuse large B-cell lymphoma. Cancers (Basel) 2022 Oct 25;14(21):5221. [doi: 10.3390/cancers14215221] [Medline: 36358642]
6. Haghofer A, Fuchs-Baumgartinger A, Lipnik K, et al. Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing. Sci Rep 2023 Nov 9;13(1):19436. [doi: 10.1038/s41598-023-46607-w] [Medline: 37945699]
7. Naji H, Sancere L, Simon A, et al. HoLy-Net: segmentation of histological images of diffuse large B-cell lymphoma. Comput Biol Med 2024 Mar;170:107978. [doi: 10.1016/j.compbiomed.2024.107978] [Medline: 38237235]
8. Yang YF, Zhao E, Shi Y, Zhang H, Yang YY. Multicenter investigation of preoperative distinction between primary central nervous system lymphomas and glioblastomas through interpretable artificial intelligence models. Neuroradiology 2024 Nov;66(11):1893-1906. [doi: 10.1007/s00234-024-03451-7] [Medline: 39225815]
9. Shi T, Jiang H, Wang M, Diao Z, Zhang G, Yao YD. Metabolic anomaly appearance aware U-net for automatic lymphoma segmentation in whole-body PET/CT scans. IEEE J Biomed Health Inform 2023 May;27(5):2465-2476. [doi: 10.1109/JBHI.2023.3248099] [Medline: 37027631]
10. Weisman AJ, Kieler MW, Perlman S, et al. Comparison of 11 automated PET segmentation methods in lymphoma. Phys Med Biol 2020 Nov 27;65(23):235019. [doi: 10.1088/1361-6560/abb6bd] [Medline: 32906088]
11. Isavand P, Aghamiri SS, Amin R. Applications of multimodal artificial intelligence in non-Hodgkin lymphoma B cells. Biomedicines 2024 Aug 5;12(8):1753. [doi: 10.3390/biomedicines12081753] [Medline: 39200217]
12. Steinbuss G, Kriegsmann M, Zgorzelski C, et al. Deep learning for the classification of non-Hodgkin lymphoma on histopathological images. Cancers (Basel) 2021 May 17;13(10):2419. [doi: 10.3390/cancers13102419] [Medline: 34067726]
13. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 2021 Sep 27;13(1):152. [doi: 10.1186/s13073-021-00968-x] [Medline: 34579788]
14. Schmidt-Barbo P, Kalweit G, Naouar M, et al. Detection of disease-specific signatures in B cell repertoires of lymphomas using machine learning. PLoS Comput Biol 2024 Jul;20(7):e1011570. [doi: 10.1371/journal.pcbi.1011570] [Medline: 38954728]
15. Jiang C, Chen K, Teng Y, et al. Deep learning-based tumour segmentation and total metabolic tumour volume prediction in the prognosis of diffuse large B-cell lymphoma patients in 3D FDG-PET images. Eur Radiol 2022 Jul;32(7):4801-4812. [doi: 10.1007/s00330-022-08573-1] [Medline: 35166895]
16. Jiang H, Diao Z, Shi T, et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Comput Biol Med 2023 May;157:106726. [doi: 10.1016/j.compbiomed.2023.106726] [Medline: 36924732]
17. Hamdi M, Senan EM, Jadhav ME, Olayah F, Awaji B, Alalayah KM. Hybrid models based on fusion features of a CNN and handcrafted features for accurate histopathological image analysis for diagnosing malignant lymphomas. Diagnostics (Basel) 2023 Jul 4;13(13):2258. [doi: 10.3390/diagnostics13132258] [Medline: 37443652]
18. Srisuwananukorn A, Salama ME, Pearson AT. Deep learning applications in visual data for benign and malignant hematologic conditions: a systematic review and visual glossary. haematol 2023;108(8):1993-2010. [doi: 10.3324/haematol.2021.280209]
19. Huang Z, Guo Y, Zhang N, et al. Multi-scale feature similarity-based weakly supervised lymphoma segmentation in PET/CT images. Comput Biol Med 2022 Dec;151:106230. [doi: 10.1016/j.compbiomed.2022.106230] [Medline: 36306574]

XSL•FO
RenderX

20. Schoenpflug LA, Chatzipli A, Sirinukunwattana K, et al. Tumour purity assessment with deep learning in colorectal cancer and impact on molecular analysis. J Pathol 2025 Feb;265(2):184-197. [doi: 10.1002/path.6376] [Medline: 39710952]
21. Huang H, Yan Z, Li B, et al. LungPath: artificial intelligence-driven histologic pattern recognition for improved diagnosis of early-stage invasive lung adenocarcinoma. Transl Lung Cancer Res 2024 Aug 31;13(8):1816-1827. [doi: 10.21037/tlcr-24-258] [Medline: 39263012]
22. Ning X, Liu R, Wang N, et al. Development of a deep learning-based model to diagnose mixed-type gastric cancer accurately. Int J Biochem Cell Biol 2023 Sep;162:106452. [doi: 10.1016/j.biocel.2023.106452]
23. Khalid M, Deivasigamani S, V S, Rajendran S. An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images. Sci Rep 2024 Aug 17;14(1):19109. [doi: 10.1038/s41598-024-70117-y] [Medline: 39154091]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
FL: follicular lymphoma
mAP: mean average precision
MCL: mantle cell lymphoma
ResNet: residual network
TCGA: The Cancer Genome Atlas
TCIA: The Cancer Imaging Archive

Edited by A Benis; submitted 15.Feb.2025; peer-reviewed by O Pitsun, S Vasavi; revised version received 11.Oct.2025; accepted 13.Oct.2025; published 06.Jan.2026.

Please cite as:
Zhao J, Wen X, Ma L, Su L
Automated Classification of Lymphoma Subtypes From Histopathological Images Using a U-Net Deep Learning Model: Comparative Evaluation Study
JMIR Med Inform 2026;14:e72679
URL: https://medinform.jmir.org/2026/1/e72679
doi:10.2196/72679

© Jin Zhao, Xiaolian Wen, Li Ma, Liping Su. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Deep Learning for Dynamic Prognostic Prediction in Minimally Invasive Surgery for Intracerebral Hemorrhage: Model Development and Validation Study

Jingxuan Wang[1*], MD; Jian Shi[2*], PhD; Qing Ye[3], MD; Danyang Chen[1], MD; Yuhao Sun[1], MD; Chao Pan[1], MD; Yingxin Tang[1], MD; Ping Zhang[1], MD; Zhouping Tang[1,4], MD

[1]Department of Neurology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan City, China

[2]State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan City, China

[3]Big Data and Artificial Intelligence Office, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan City, China

[4]Key Laboratory for Diagnosis and Treatment of Severe Zoonotic Infectious Diseases, Huazhong University of Science and Technology, Wuhan city, China

[*]these authors contributed equally

Corresponding Author:
Zhouping Tang, MD
Department of Neurology
Tongji Hospital, Tongji Medical College
Huazhong University of Science and Technology
No. 1095, Jiefang Avenue
Wuhan City, 430030
China
Phone: 86 83663337
Email: ddjtzp@163.com

Abstract

Background: The pathological and physiological state of patients with intracerebral hemorrhage (ICH) after minimally invasive surgery (MIS) is a dynamic evolution, and the traditional models cannot dynamically predict prognosis. Clinical data at multiple time points often show the characteristics of different categories, different numbers, and missing data. The existing models lack methods to deal with imbalanced data.

Objective: This study aims to develop and validate a dynamic prognostic model using multi–time point data from patients with ICH undergoing MIS to predict survival and functional outcomes.

Methods: In this study, 287 patients who underwent MIS for ICH were retrospectively collected on the day of surgery, days 1, 3, 7, and 14 after surgery, and the day of drainage tube removal. Their general information, vital signs, laboratory test findings, neurological function scores, head hematoma volume, and MIS-related indicators were collected. In addition, this study proposes a multistep attention model, namely the MultiStep Transformer. The model can simultaneously output 3 types of prediction probabilities for 30-day survival probability, 180-day survival probability, and 180-day favorable functional outcome (modified Rankin Scale [mRS] 0-3) probability. Five-fold cross-validation was used to evaluate the performance of the model and compare it with mainstream models and traditional scores. The main evaluation indexes included accuracy, precision, recall, and F_1-score. The predictive performance of the model was evaluated using receiver operating characteristic (ROC) curves; its calibration was assessed via calibration curves; and its clinical utility was examined using decision curve analysis (DCA). Attributable value analysis was conducted to assess the key predictive features.

Results: The 30-day survival rate, 180-day survival rate, and 180-day favorable functional outcome rate among 287 patients were 92.3%, 88.8%, and 52.3%, respectively. In terms of predictive efficacy for survival and functional outcomes, the MultiStep Transformer model showed a remarkable superiority over traditional scoring systems and other deep learning models. For these three outcomes, the model achieved areas under the receiver operating characteristic curves (AUROCs) of 0.87 (95% CI 0.82-0.92), 0.85 (95% CI 0.77-0.93), and 0.75 (95% CI 0.72-0.78), with corresponding Brier scores of 0.1041, 0.1115, and 0.231. DCA confirmed that the model provided a definite clinical net benefit when threshold probabilities ranged within 0.06-0.26, 0.04-0.5, and 0.21-0.71.

XSL•FO
RenderX

Conclusions: The MultiStep Transformer model proposed in this study can effectively use imbalanced data to construct a model. It possesses good dynamic prediction ability for short-term and long-term survival and functional outcome of patients with ICH undergoing MIS, providing a novel tool for individualized assessment of prognosis among patients with ICH undergoing MIS.

(JMIR Med Inform 2026;14:e86327) doi:10.2196/86327

KEYWORDS

intracerebral hemorrhage; minimally invasive surgery; dynamic features; prognosis prediction; deep learning

Introduction

Intracerebral hemorrhage (ICH) accounts for 10%-15% of all stroke cases [1,2]. It is the leading cause of death and disability worldwide [3], characterized by [4] high morbidity, disability, and mortality rates. Conservative medical treatment of ICH exhibits limited efficacy [5]. Surgery for the removal of hematoma may improve the functional outcome of patients [6]. However, Surgical Trial in Intracerebral Haemorrhage (STICH) [7] and Surgical Trial in Intracerebral Haemorrhage II (STICH II) [8] confirmed the failure of craniotomy in improving functional outcomes among patients with ICH.

Recently, minimally invasive surgery (MIS), characterized by limited trauma, rapid recovery, and other advantages, has received increasing attention. Minimally Invasive Surgery Plus Alteplase for Intracerebral Hemorrhage Evacuation Phase 3 (MISTIE Ⅲ) [9] demonstrated the safety of minimally invasive hematoma evacuation combined with alteplase thrombolysis. The Early Minimally Invasive Removal of Intracerebral Hemorrhage (ENRICH) [10] trial reported that MIS can improve the functional outcomes of patients with lobar ICH at 180 days. Although the ENRICH trial provided encouraging evidence regarding the efficacy of MIS, not all patients with ICH can benefit from this approach. The uncertainty surrounding the therapeutic efficacy of MIS poses challenges for assessing the prognosis of patients.

Moreover, the condition of patients with ICH is a dynamic process. Although some tools have been used to predict death or functional outcome in patients with ICH [11-17], these predictions are mostly based on data at admission or a single time point, which cannot reflect the dynamic changes of the disease and also limit the accuracy of prognostic assessment. Electronic health records (EHRs) containing abundant data reflect the pathological and physiological state of dynamic changes in patients with ICH undergoing MIS. However, data extracted from EHR, without systematic organization, exhibits characteristics such as variable quality, high dimensionality, heterogeneity, temporality, and incompleteness [18], posing significant challenges for predictive modeling. Therefore, this study aimed to leverage multitime point, imbalanced data from patients undergoing MIS for ICH to develop a dynamic prognostic model and validate its performance in predicting survival and functional outcomes.

Methods

Study Design and Population

Patients with ICH who underwent MIS for ICH in the Department of Neurology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology were enrolled in this study. Clinical and imaging data were collected retrospectively on the day of surgery, 1, 3, 7, and 14 days after surgery, and on the day of drainage tube removal. A deep learning algorithm was used to construct a dynamic prognostic model and validate the effect of the model.

In total, 372 patients undergoing MIS for ICH between May 2012 and May 2024 were selected. Among them, 11 cases of ICH caused by cerebrovascular malformation and coagulopathy, 25 cases with a history of neurological deficit and renal failure, and 49 patients with missing outcome data were excluded, and data from 287 patients were analyzed. A flowchart for the patient selection process is shown in Figure 1.

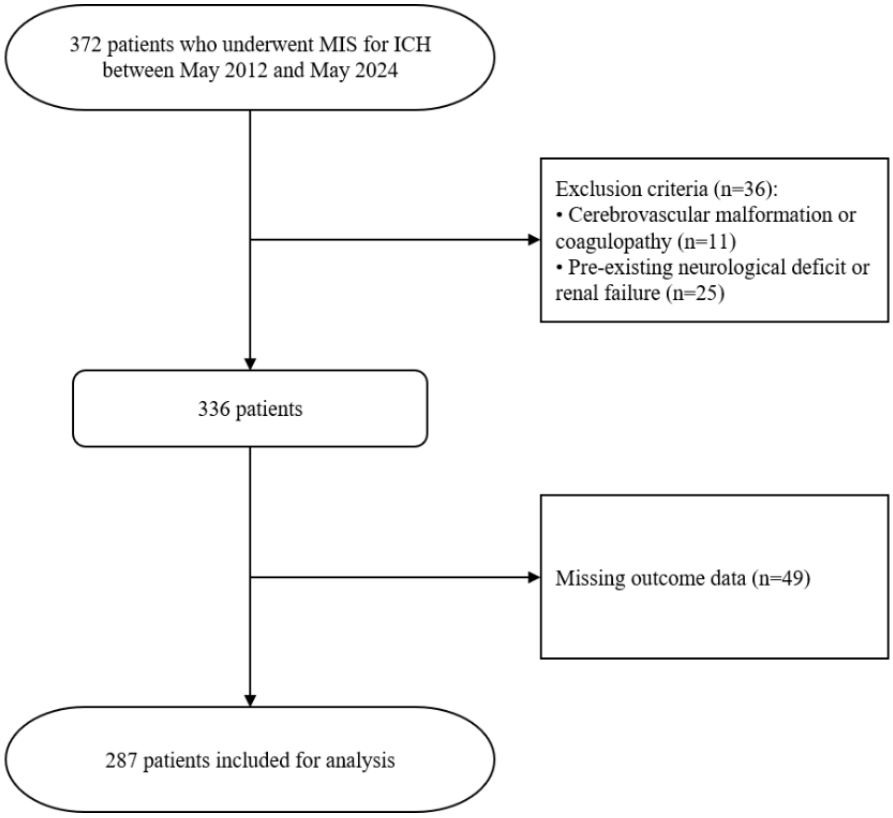
The inclusion criteria were as follows: (1) patients with spontaneous acute ICH in basal ganglia or lobar regions based on cranial computed tomography (CT), who underwent stereotactic intracranial hematoma puncture and drainage surgery after admission; (2) ICH hematoma volume 20-100 mL; (3) age 18-80 years old; (4) surgery within 1 week of onset; and (5) survival for at least 7 days after admission.

The exclusion criteria were as follows: (1) coagulation disorders, such as thrombocytopenia and significant abnormal coagulation indexes caused by hepatitis; (2) traumatic ICH; (3) intracranial infection and other diseases leading to a significant increase in intracranial pressure; (4) previous severe heart, liver, kidney, or lung diseases or functional failure; (5) previous history of neurological dysfunction; (6) ICH was caused by structural abnormalities of the cerebral blood vessels (eg, arteriovenous malformations, intracranial aneurysms, vasculitis, and moyamoya disease); and (7) ICH was caused by brain tumor or cerebral infarction and was treated with thrombolytic therapy.

The model was developed and validated in accordance with the TRIPOD-AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence) checklist.

XSL•FO
RenderX

Figure 1. Patient selection flowchart. ICH: intracerebral hemorrhage; MIS: minimally invasive surgery.



Data Sources and Collection

We electronically extracted clinical data from the EHR of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. Data were collected on general information, vital signs, laboratory test findings, neurological function scores, and MIS-related metrics, such as the time from onset to surgery and the duration of drainage tube placement, along with established neurological scales: the ICH score, Glasgow Coma Scale (GCS), National Institutes of Health Stroke Scale (NIHSS), and the functional outcome in patients with primary intracerebral hemorrhage (FUNC) score. Rapid segmentation and volume calculation of hematomas on cranial CT images were conducted using 3D Slicer.

Follow-up and Clinical Outcomes

Follow-up was conducted via telephone call. The primary outcome was all-cause mortality at 30 days. Secondary outcomes included all-cause mortality at 180 days and functional outcome at 180 days. The outcomes were assessed using the modified Rankin Scale (mRS), which was transformed into a binary variable, favorable functional outcome (mRS 0-3) and unfavorable functional outcome (mRS 4-6) [19,20]. The assessment was conducted by 2 medical doctors (JW and YS), with discrepancies being resolved through arbitration by a senior physician (PZ).

Data Processing and Statistical Analysis

Missing values in this study were predominantly attributed to the absence of imaging features at certain time points, whereas clinical features (vital signs, laboratory studies, etc) were fully documented. Instead of imputing the missing imaging features, the model automatically ignored the corresponding time points via an attention mechanism and completed representation learning and outcome prediction solely based on available temporal imaging features and comprehensive clinical features. To ensure a consistent input structure across patients and time points, we adopted a simple zero-imputation strategy during preprocessing. For patients with missing measurements at specific time points (eg, when certain laboratory tests or clinical assessments were not performed on a given day), the corresponding feature values were set to zero. All features were standardized before model training so that zero-imputed values lie within a well-defined range and are distinguishable from typical observed values.

For continuous variables, intergroup comparisons were conducted using the *t* test. The chi-square test was used for categorical variables. Data are presented as median (IQR), mean (SD), and range. A *P* value of <.05 was considered statistically significant.

Model Development and Training

Modeling Strategy

The changes of clinical indicators after MIS for ICH reflect the dynamic changes of individual status. Reasonable modeling of time series data dependence and cross-time point information interaction is of high importance for survival and prognosis prediction. We collected a data series of patients' multidimensional clinical measures: , where x^{1d}, x^{3d}, x^{7d}, x^{14d} denote the characteristics at 1, 3, 7, and 14 days after surgery, respectively. x^{ft} denotes fixed features (eg, gender and age) that do not change over time. x^{op}denotes the characteristics of the day of surgery. In particular, since not all individuals had

XSL•FO
RenderX

the same drainage tube removal time, the corresponding characteristics were introduced separately.

Our objective was to learn a mapping function $f: X \rightarrow \gamma$, where γ {0,1}represents the survival and prognosis labels, specifically defined as 30-day survival, 180-day survival, and 180-day favorable functional outcome. Crucially, we aimed to establish a mapping function capable of predicting both future data and the final labels from postoperative data of any given day. For instance, when input with first-day postoperative data x^{1d}, the model was meant to successfully predict subsequent data x^{3d}, x^{7d}, x^{14d} and the final ternary label γ. The problem was a multistage, multi–time point, and multivariate time series classification task.

Since the clinical indicators at each time point in X were not entirely consistent, it was not feasible to establish a unified model for cyclic prediction at each time point. Available machine learning or deep learning models exhibit limited capability in handling variable dimensional feature data [21].

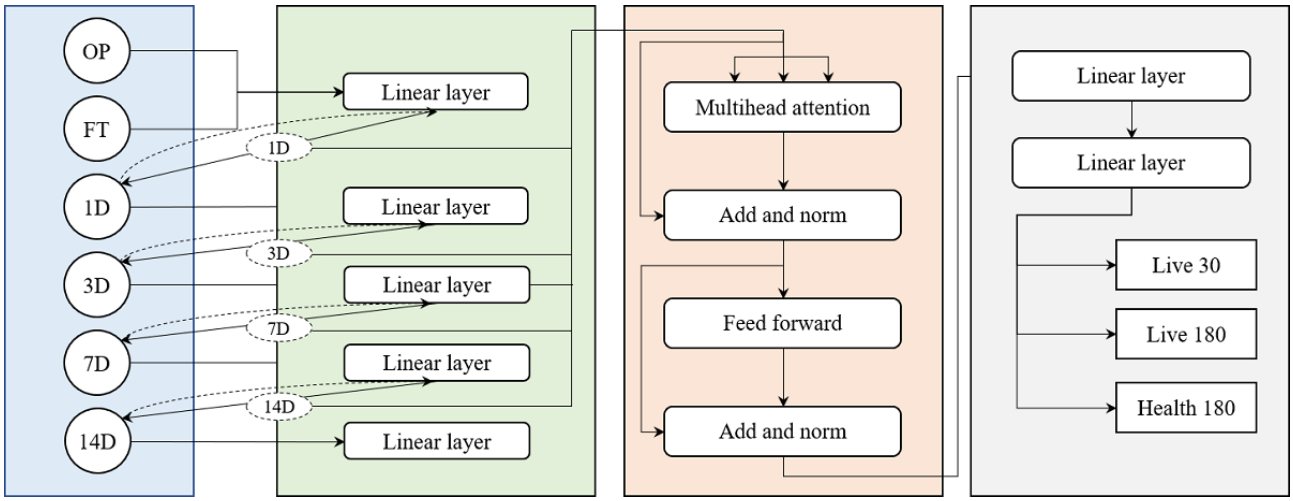
To address this, we proposed a MultiStep Transformer model designed to capture nonlinearly evolving clinical features over time and manage the cross-influences existing between different time points.

MultiStep Transformer Model Architecture

In the MultiStep Transformer model, we introduced a multistage feature prediction module and a time-sharing embedding mechanism. It cooperated with the transformer encoder to realize the reinforcement modeling of multi–time point sequences. Specifically, we constructed a multilayer feature prediction module to predict and train each prognostic data separately. Finally, the Transformer encoder was used to predict 30-day survival probability, 180-day survival probability, and 180-day favorable prognosis probability. The overall process of the model is shown in Figure 2.

The MultiStep Transformer model was able to receive data at any time point as input and output for 3 types of prediction probabilities. The model framework is as follows:

Figure 2. Model diagram.



Multistage Feature Cycle Prediction Module

The core idea of the multistage feature cycle prediction module was to use the information of current and previous time points to recursively infer the feature expression of missing or future time points, which was formulated as

Here, $LR_t(\cdot)$represents a linear mapping layer specific to the t-th time point, responsible for establishing recursive dependencies among the time-series features. This design effectively mitigated the challenges of incomplete clinical data and the isolation of temporal features, thereby enhancing the ability of the model to capture temporal evolutionary trends.

Time-Sharing Feature Embedding

In view of the large difference of input data dimensions and strong heterogeneity of feature distribution at different time points, an independent linear mapping layer was designed:

The high-dimensional projection of the input at each time point was conducted to eliminate the difference in input dimensions and ensure the consistency of the sequence dimensions of the input of the transformer encoder. In addition, the independent mapping parameters W_t and d_t captured the unique statistical characteristics of time points and improved the adaptability of the model to the information of different stages.

Transformer Encoder Structure and Time-Series Dependence Modeling

Through the self-attention mechanism, the transformer encoder flexibly modeled the interactive dependencies between all positions in a sequence. It overcame the limitations of traditional recurrent neural networks and possessed the capabilities of parallel computation and long-range dependency capture. In our framework, the same temporal feature sequence was used simultaneously as the query, key, and value in the self-attention operation. This design allows feature extraction to be performed directly on the observed sequence without forcing all patients to have the same number of time points and therefore naturally supports variable-length temporal inputs.

XSL•FO
RenderX

Input sequence: concatenate the embedding vectors at all time points to form an input sequence:

Here, t= 6 represents the number of input feature layers, and B denotes the batch size for each training or inference step. The experimental framework supports arbitrary combinations of time points as input, for example using only Day-1 data or Days 1, 3, 7, and 14, depending on data availability and clinical requirements.

Encoding process: the multilayer stacked transformer encoder consisted of multihead self-attention and a feed-forward network. A single layer was formulated as

Where the same sequence serves as Q, K, and V in the self-attention module. The model performs multistep, progressive computation; features from earlier time points are encoded first, and then information from subsequent time points is gradually integrated. After multilayer coding, the model comprehensively considered the influence between different time points and its feature correlation, and the output sequence was expressed as:

$$Z = \text{TransformerEncoder}(H) \quad (6)$$

The encoding vector Z^d at the final time point of the sequence was taken as the representation of the comprehensive temporal dynamic state of patients.

Survival Prediction Head

The final classification layer was mapped using single-layer linearity:

Here, σ denotes the sigmoid activation function for generating binary classification probability outputs. The output data format was (30 live, 180 live, and 180 heal), representing the probability of 30-day survival, 180-day survival, and 180-day favorable functional outcome, respectively.

Training and Prediction Processes

For model training, data were input uniformly, and the model was trained step by step. Specifically, features from all time points $X = \{x^{ft}, x^{op}, x^{1d}, x^{3d}, x^{7d}, x^{14d}\}$were input into the model. The model first processed the features $[x^{ft}, x^{op}]$ through the linear layer LR_{1d} to compute , and calculate the loss using x^{1d}, followed by backpropagation to optimize LR_{1d}. Subsequently, stepwise prediction and optimization were conducted sequentially based on the mapping layers in the multistage feature prediction module. All features, including known inputs and predicted outputs, were combined into a complete feature sequence . Finally, the transformer encoded all temporal features to obtain the patient's dynamic state representation Z, which was then classified using a linear layer.

During inference, given the first i temporal features X_i (eg, $X_i = \{x^{ft}, x^{op}, x^{1d}, x^{3d}\}$), the model applied masking to the existing feature time points based on the number of feature layers and directly computed subsequent features , finally predicting survival and prognosis (30 live, 180 live, and 180 heal).

Model hyperparameters were selected empirically based on preliminary experiments on the training folds. The network was optimized using the Adam optimizer with a fixed learning rate of 0.001 and trained for 200 epochs per fold. We used the cross-entropy loss for outcome prediction, combined with the feature prediction loss for intermediate temporal features in the multistage prediction module. The batch size, number of transformer layers, hidden dimensions, and dropout rate were tuned within a reasonable range to balance model performance and overfitting risk.

Model Evaluation

The predictive performances of different models and scores were compared using accuracy, precision, recall, and F_1-score. The predictive performance of the model was evaluated by the area under the receiver operating characteristic curve (AUROC). Calibration curves were plotted to assess the consistency between the predicted probabilities and the observed outcomes of the model. In addition, decision curve analysis (DCA) was performed to evaluate the clinical utility of the model.

Model Interpretation

To gain insight into the decision-making process of the model, we used attribution values, a technique used to quantify how much each input feature contributes to the output of the model. Attribution values revealed the importance of clinical characteristics in predicting the outcome and helped us determine which characteristics had the greatest impact on the model's decision-making process.

Ethical Considerations

This study was approved by the Medical Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (TJ-IRB202502129). Given the retrospective nature of the study and the absence of any direct interaction with participants, informed consent was exempted by the ethics committee, and no compensation of any form was provided. This study adheres to the ethical principles outlined in the Declaration of Helsinki.

Results

Baseline Characteristics of the Study Population

The mean age of the 287 patients was 53.2 (10.6) years, and 67.9% of the patients were male. Their 30-day survival rate was 92.3%, their 180-day survival rate was 87.8%, and the rate of favorable functional outcome at 180 days was 52.3%. The presence of intraventricular hemorrhage, GCS, NIHSS, ICH score, and FUNC score showed significant differences across the 3 groups of patients (Tables S1-S3 in Multimedia Appendices 1-3).

XSL•FO
RenderX

Evaluation of the Model Performance

Due to variations in clinical assessment items across different time points, the feature categories at each time point were not entirely consistent. In the collected dataset, there were 48 features on the day of surgery, 47 features on postoperative days 1 and 3, a total of 46 features on postoperative day 7, a total of 10 features on postoperative day 14, and 15 features on the day of drainage tube removal (feature labels are in Multimedia Appendix 4). Based on feature dimensions of the patients, we constructed a MultiStep Transformer model with 4 feature prediction modules.

Given the limited sample size (287 patients) and single-center design, we adopted a 5-fold cross-validation scheme for internal validation. The entire cohort was randomly partitioned into 5 approximately equal subsets. In each iteration, one subset was held out as the validation (test) set, and the remaining 4 subsets were used for model training. This process was repeated 5 times so that each subset served as the validation set exactly once. Model performance metrics were averaged across the 5 folds and reported together with their SDs, providing an estimate of variability and robustness. The outcome labels were not evenly distributed across classes. To mitigate the impact of class imbalance, we ensured that the proportion of outcome categories in each fold of the cross-validation was approximately consistent with that of the full cohort by using stratified splitting according to the target labels. In addition, class weights inversely proportional to class frequencies in the training data were applied in the loss function to reduce bias toward the majority class. This strategy helped the model to learn more balanced decision boundaries across different outcome categories.

We calculated the confusion matrices of the results to obtain accuracy, precision, recall, and F_1-score. Regarding the prediction results, actual labels were defined as follows: true positive (TP), positive labels correctly predicted; false positive (FP), negative labels incorrectly predicted; true negative (TN), negative labels correctly predicted; false negative (FN), positive labels incorrectly predicted. The evaluation metrics were calculated as follows:

For comparison, we selected the commonly used deep learning models, including temporal convolutional network (TCN), gated recurrent unit (GRU), and transformer. Table 1 presents the comparison of prediction performance between the MultiStep Transformer model and each deep learning model.

Table 1. Comparison of model performance.

Model and scene	Accuracy (%), mean (SD)	Precision (%), mean (SD)	Recall (%), mean (SD)	F_1-score (%), mean (SD)
GRU[a]				
live_30	77.34 (28.89)	100 (0)	75.66 (30.88)	81.17 (28.30)
live_180	66.5 (27.68)	98.1 (0.99)	63.39 (31.91)	70.24 (33.74)
heal_180	73.70 (7.20)	75.61 (6.44)	71.92 (10.52)	73.49 (8.05)
TCN[b]				
live_30	85.55 (19.8)	100 (0)	84.49 (21.23)	89.83 (15.30)
live_180	74.04 (19)	97.92 (2.55)	71.57 (21.23)	80.96 (15.30)
heal_180	74.68 (5.86)	77.09 (2.51)	71.94 (12.72)	73.90 (8.03)
Transformer				
live_30	88.83 (12.53)	100 (0)	88 (13.40)	93.01 (8.50)
live_180	76.33 (15.52)	98.18 (2.28)	74.16 (16.59)	83.54 (12.12)
heal_180	78.95 (5.91)	81.75 (3.65)	76.43 (13.4)	78.28 (7.81)
MultiStepTransformer (Ours)				
live_30	92.43 (1.68)	96.47 (2.15)	95.37 (1.83)	95.89 (0.92)
live_180	88.80 (3.24)	93.99 (1.68)	93.22 (3.90)	93.55 (1.20)
heal_180	80.13 (3.68)	81.16 (2.13)	82.02 (6.05)	81.42 (2.82)

[a]GRU: gated recurrent unit.

[b]TCN: temporal convolutional network.

The proposed method achieved the best performance in terms of accuracy, recall, and F_1-score for the 3 categories of outcomes. The accuracy of the MultiStep Transformer model was 92.43 (1.68) for 30-day survival, 88.80 (3.24) for 180-day survival, and 80.13 (3.68) for 180-day favorable functional outcome, respectively.

Furthermore, we separated the 4 scores (GCS, NIHSS, ICH score, and FUNC score) from the data to predict the three categories of outcomes. The FUNC score is associated with the volume of hematoma, but there were some missing data, and it could not be combined with the other scores for classification. We compared the single-score classification and the combined classification of scores (ICH score, GCS, and NIHSS),

XSL•FO
RenderX

respectively. We also used support vector machine (for classification). Using the same 5-fold cross validation, their confusion matrices were calculated to obtain the corresponding indicators (Table 2).

As can be seen, the NIHSS predicts the best results of the traditional scores for 30-day survival. The FUNC score was superior to the other 2 scores in the accurate classification of 180-day survival and 180-day favorable functional outcome. It significantly improved the prediction of 180-day favorable functional outcome, indicating that the FUNC score was better at predicting the long-term prognosis of patients. The combination of the ICH score, GCS, and NIHSS features significantly improved 30-day survival prediction, but the FUNC score still exhibited some advantages in long-term prediction.

Figure 3 presents a comparison of classification accuracy between the MultiStep Transformer model, traditional scoring methods, and other deep learning models. The results indicated that compared to either traditional scoring methods or other deep learning models, our proposed method offers superior performance in both short-term and long-term predictions of the 3 outcome categories. Furthermore, the MultiStep Transformer model yielded the smallest SD (eg, mean 92.43, SD 1.68 for accuracy in 30-day survival prediction), reflecting its consistent performance across different data subsets, enhanced stability, and strong generalization capability.

Figures 4A, D, and G present the receiver operating characteristic (ROC) curves of the model for different outcomes, namely 30-day survival, 180-day survival, and 180-day favorable functional outcome. The model achieved AUROCs of 0.87 (95% CI 0.82-0.92), 0.85 (95% CI 0.77-0.93), and 0.75 (95% CI 0.72-0.78) for these 3 outcomes, respectively.

Furthermore, the calibration curves demonstrated good calibration performance of the model (Figures 4B, E, and H), with corresponding Brier scores of 0.1041, 0.1115, and 0.231, respectively. DCA indicated that the model yielded a net benefit when the threshold probabilities ranged within 0.06-0.26, 0.04-0.5, and 0.21-0.71, respectively (Figures 4C, F, and I). The above findings confirm that this model exhibits favorable generalizability.

XSL•FO
RenderX

Table 2. Comparison of traditional scores.

Score and scene	Accuracy (%), mean (SD)	Precision (%), mean (SD)	Recall (%), mean (SD)	F_1-score (%), mean (SD)
ICH[a] score				
live_30	73.36 (16.52)	56.56 (4.66)	66.00 (11.4)	54.22 (10.68)
live_180	71.07 (15.30)	55.96 (4.90)	58.51 (7.03)	53.88 (9.33)
heal_180	60.52 (5.78)	66.05 (6.37)	59.67 (3.89)	55.65 (5.53)
GCS[b]				
live_30	78.61 (3.63)	55.27 (5.30)	63.4 (12.93)	55.47 (6.98)
live_180	76.96 (2.85)	57.64 (7.18)	61.71 (8.53)	57.92 (8.23)
heal_180	57.92 (3.74)	58.73 (4.62)	56.89 (4.59)	54 (5.61)
NIHSS[c]				
live_30	83.22 (1.60)	58.03 (4.79)	67.72 (11.77)	59.60 (6)
live_180	78.31 (2.72)	55.88 (7.82)	61.67 (11.79)	56.94 (9.08)
heal_180	56.92 (2.45)	57.43 (4.79)	55.31 (3.29)	52.36 (4.51)
FUNC[d] score				
live_30	81.57 (16.08)	70 (24.81)	66.58 (27.95)	67.46 (26.83)
live_180	79.56 (18.99)	68.75 (25.62)	65.75 (28.37)	65.99 (27.79)
heal_180	71.56 (23.01)	74.40 (20.49)	70.83 (19.43)	68.10 (23.25)
Combination of the ICH score, GCS, and NIHSS				
live_30	85.86 (1.61)	60.67 (6.36)	70.98 (13.28)	62.93 (8.07)
live_180	79.29 (3.12)	58.39 (5.30)	63.17 (10.61)	59.02 (6.64)
heal_180	57.58 (3.75)	57.37 (6.33)	56.00 (5.09)	52.91 (6.86)
MultiStep Transformer (ours)				
live_30	92.43 (1.68)	96.47 (2.15)	95.37 (1.83)	95.89 (0.92)
live_180	88.80 (3.24)	93.99 (1.68)	93.22 (3.90)	93.55 (1.20)
heal_180	80.13 (3.68)	81.16 (2.13)	82.02 (6.05)	81.42 (2.82)

[a]ICH: intracerebral hemorrhage.

[b]GCS: Glasgow Coma Scale.

[c]NIHSS: National Institutes of Health Stroke Scale.

[d]FUNC: functional outcome in patients with primary intracerebral hemorrhage.

XSL•FO
RenderX

Figure 3. Comparison of classification accuracy across models and traditional scoring methods. FUNC: functional outcome in patients with primary intracerebral hemorrhage; GCS: Glasgow Coma Scale; GRU: gated recurrent unit; ICH: intracerebral hemorrhage; NIHSS: National Institutes of Health Stroke Scale; TCN: temporal convolutional network.

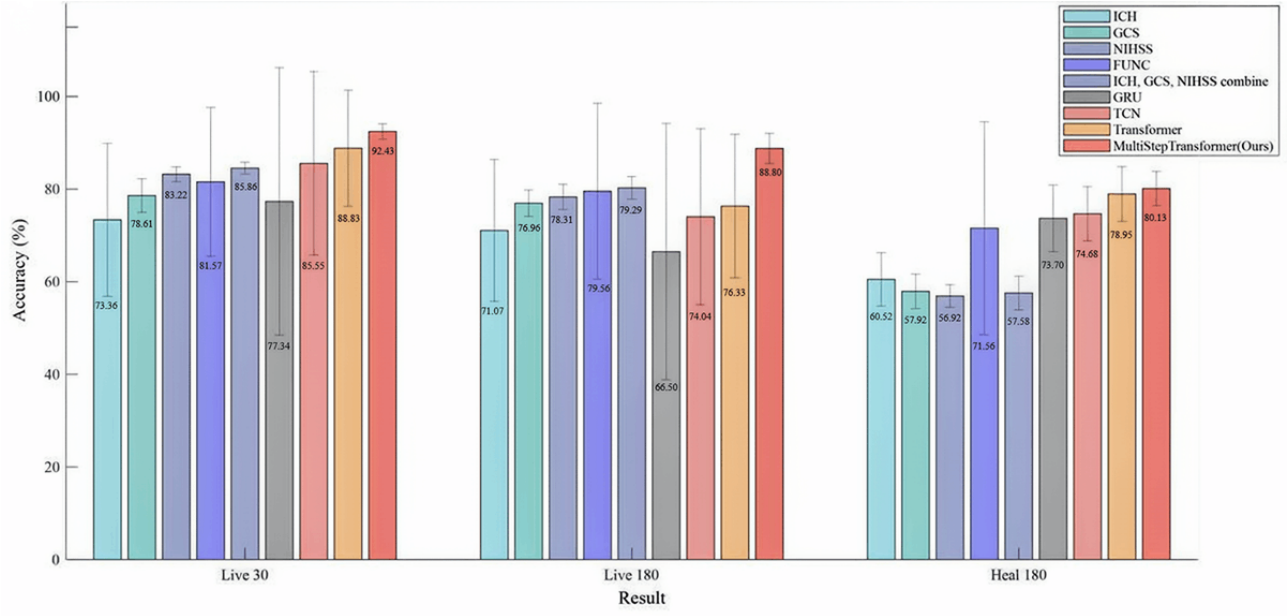
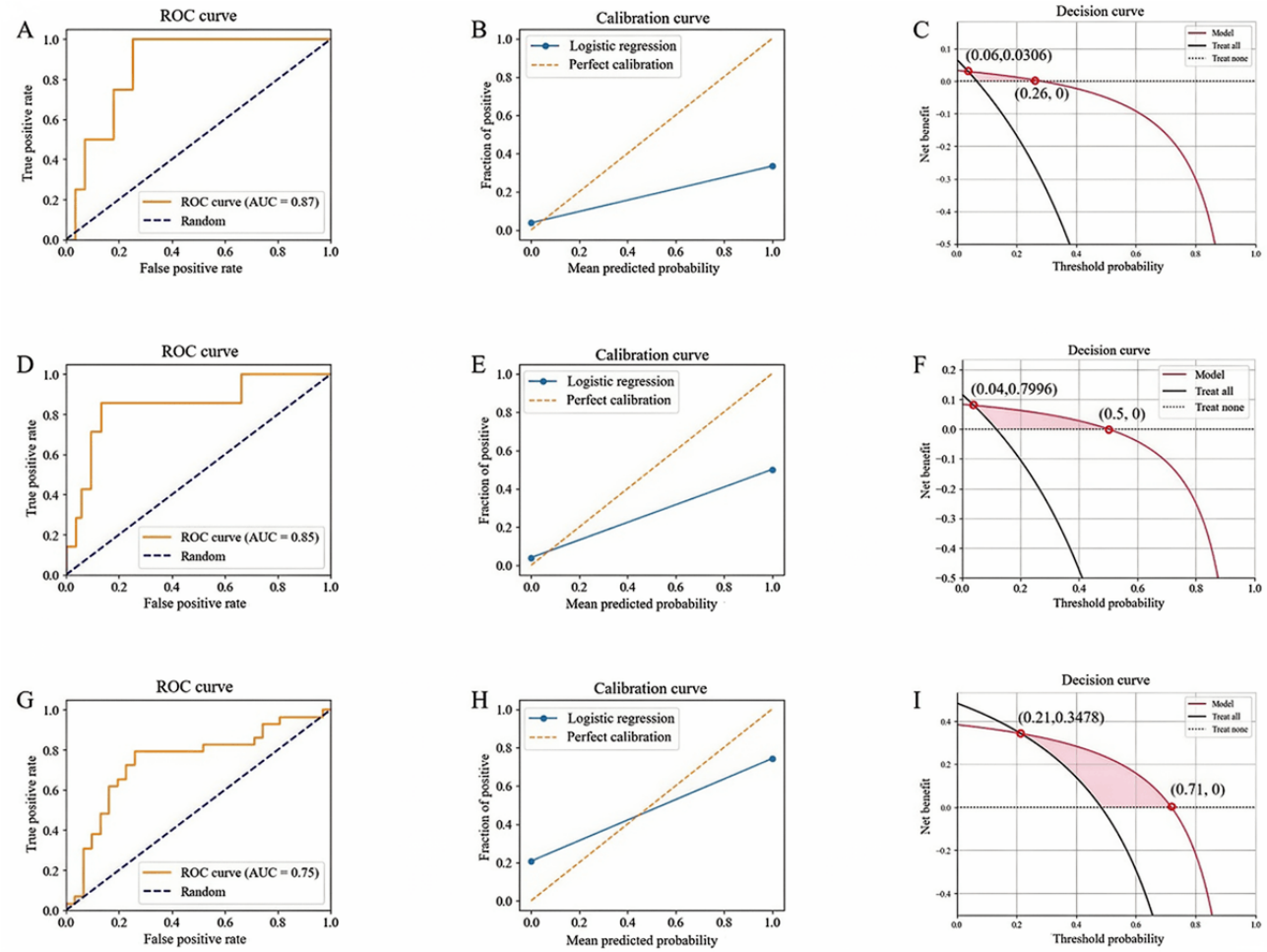


Figure 4. Results of 5-fold cross-validation: Panels A-C depict the ROC curve, calibration curve, and decision curve for 30-day survival, respectively; panels D-F represent those for 180-day survival; and panels G-I denote those for 180-day favorable functional outcome. AUC: area under the curve; ROC: receiver operating characteristic.



XSL•FO
RenderX

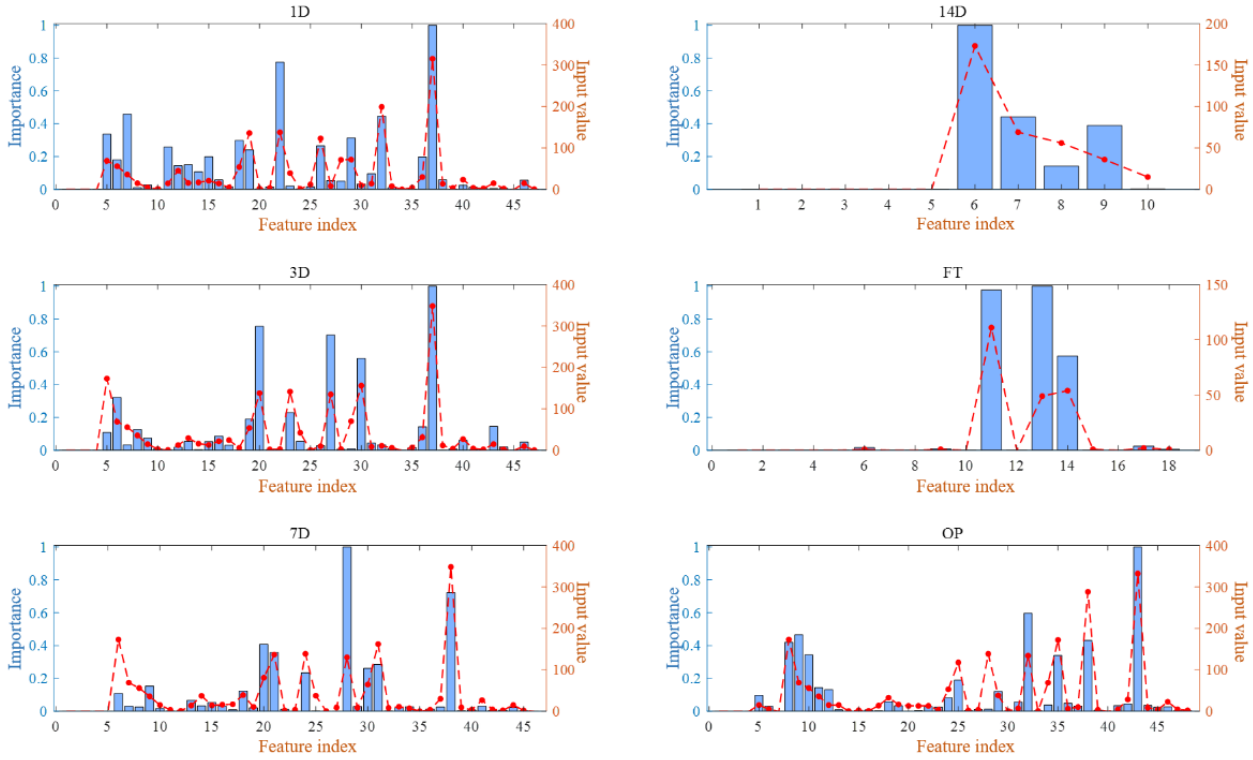
Model Interpretation

In this study, we used deep neural networks to classify clinical data for predicting patients' outcomes. Specifically, we visualized the attribution values calculated by the model alongside the corresponding input feature values in the same diagram to interpret the model's decision-making mechanism.

The attribution values indicated the contribution of each clinical feature to the model's output. The feature importance was relatively similar between postoperative day 1 and day 3, while some changes emerged in the key clinical features affecting predictions by days 7 and 14 (Figure 5).

The horizontal axis in Figure 5 represents the indices of the input features as discrete integers, facilitating the identification of specific features. The left vertical axis denotes the attribution values, displayed with blue bars that illustrate each feature's relative contribution to the output. The right vertical axis corresponds to the original feature values, plotted as a red dashed line to visualize the correspondence between attributions and raw data. This integrated plot helps identify the most influential features on the model's predictions, enabling a more comprehensive understanding of the model's behavior.

Figure 5. Visualization of attribution values of features.



Discussion

Principal Findings

This study reports the construction process and model performance of the MultiStep Transformer model, a novel deep learning model that can dynamically predict short-term and long-term survival and functional outcome in patients with ICH undergoing MIS. The experimental results showed that the model significantly outperformed the mainstream deep learning models (GRU, TCN, and transformer) and traditional clinical scores (GCS, NIHSS, ICH score, and FUNC score) on the 3 prediction tasks of 30-day survival, 180-day survival, and 180-day favorable functional outcome.

In patients undergoing MIS for ICH, neurological functions, vital signs, imaging findings, and laboratory parameters are subjected to continuous changes, generating substantial clinical data. However, temporal data in EHR present challenges, including irregular recording intervals and variable sequence lengths across individuals. Meanwhile, most of the previous studies have conducted predictions at fixed time points, whereas many clinical scenarios necessitate continuous outcome updates with the emergence of new data [22]. Therefore, it is essential to develop a dynamic model to process variably dimensional features. This approach not only addresses the needs of clinical practice but also allows personalized management of patients.

The MultiStep Transformer model incorporates innovative designs to address common challenges in real-world clinical data, such as missing time points, heterogeneous feature dimensions, and inconsistent sampling frequencies.

First, multistage feature prediction addresses data missingness. Through layered stepwise linear mapping, the model not only imputes missing values at certain time points but also enhances temporal dependencies among features, thereby improving the accuracy of predictions.

Second, time point–specific embedding resolves dimensionality inconsistency. Independent linear embedding layers ensure unified representation of heterogeneous input data, strengthening the generalizability of the model.

XSL•FO
RenderX

Third, the self-attention mechanism of the transformer captures long-range dependencies. Compared to traditional recurrent neural networks, the transformer architecture more effectively learns complex nonlinear dependencies and interactions across several time points.

Finally, we visualized feature importance through attribution values. The values revealed which clinical features played a pivotal role in the final predictions. A higher attribution value of a feature suggested a greater influence on the model's decision-making. The MultiStep Transformer model successfully captured the dynamic shifts of key perioperative features during MIS, thereby enhancing the interpretability of the model and its credibility in clinical practice.

Comparison With Prior Work

TCN is a feedforward network architecture that models time series by stacking convolutional layers [23]. Under the constraint of convolutional layers, it can effectively prevent information leakage. Multilayer convolution can also capture local time dependence and short-term dynamic changes in multi–time point clinical data. GRU is a type of recurrent neural network, which uses a gating unit to control the memory and forgetting of information. It has a relatively small number of parameters and high training stability and can model the dependence from local to global. For clinical data with multiple time points, GRU can better extract cross-temporal global features [24]. With the attention mechanism as the core, Transformer can model the relationships between any time points in a sequence in a global scope [25], making it suitable for dealing with long sequences and complex temporal patterns. In addition, it has good parallel computing ability and can effectively use global information.

Compared to the mainstream deep learning models, the MultiStep Transformer model offers full usage and efficient modeling of complex imbalanced time-series data by introducing a multistage feature cycle prediction module, time-sharing embedding mechanism, and transformer encoder structure. Besides, the modular design of the MultiStep Transformer model enables adaptation to flexible inputs, supports dynamic input time points, and is suitable for clinical scenarios with different monitoring frequencies, thereby improving its practicability.

In 30-day survival prediction, GRU, TCN, and transformer achieved 100% precision, indicating rare false positive errors, but their low recall suggests many missed diagnoses. The MultiStep Transformer exhibited a significant improvement in recall and F_1-score, suggesting a more accurate identification ability while avoiding excessive conservatism. The performance of all models decreased with the prolongation of prediction time, but the MultiStep Transformer still showed a high F_1-score in 180-day survival prediction, with greatly improved recall. These findings imply that it is more sensitive in long-term prediction and not overly conservative. In the prediction of a favorable functional outcome at 180 days, all models performed worse compared to their performance in survival prediction, highlighting the challenges in predicting functional outcomes. The MultiStep Transformer was the only model with a recall of more than 80%, indicating that it was more advantageous in identifying patients with a greater chance of recovery.

Traditional scoring systems are widely used in clinical practice due to their simplicity and have demonstrated considerable value in predicting the outcomes of ICH. For predicting mortality, the ICH score achieved AUROCs of 0.83-0.84, while the FUNC score showed AUROCs of 0.79-0.83. In predicting functional outcomes, the ICH score attained AUROCs of 0.77-0.82, compared to 0.76-0.78 for the FUNC score [26]. When predicting death or severe disability, the NIHSS score yielded an AUROC of 0.796, the GCS score achieved an AUROC of 0.650, and the ICH score revealed an AUROC of 0.674 [27]. By comparison, the model developed in this study demonstrated superior predictive performance. For the prediction of 30-day survival, 180-day survival, and 180-day favorable functional outcome, the model achieved AUROCs of 0.87 (95% CI 0.82-0.92), 0.85 (95% CI 0.77-0.93), and 0.75 (95% CI 0.72-0.78), respectively. Meanwhile, calibration curves and DCA confirmed that this model has excellent calibration performance and definite clinical benefit. Furthermore, the inherent limitations of traditional scoring systems are becoming increasingly apparent. McCracken et al. indicated that nearly two decades after its introduction, the predictive accuracy of the ICH score may no longer be sufficient for contemporary clinical practice, necessitating adjustments and optimization [28]. More importantly, the clinical guideline "Guidelines for Neuroprognostication in Critically Ill Adults with Intracerebral Hemorrhage" emphasizes that no single score or clinical variable should be used as the sole basis for prognostic judgment at 3 months or beyond [29]. Accordingly, factors such as preoperative functional status, surgical complications, and intervention timing should be comprehensively considered when developing assessment tools for patients undergoing surgery [30]. The dynamic prognostic model developed in this study aligns with this philosophy. From a clinical perspective, accurate prediction of long-term functional outcomes is crucial for rehabilitation planning and patient communication. For predicting 180-day functional outcomes, the most challenging task, our model achieved an F_1-score of 81.42%, significantly surpassing traditional scoring systems (maximum 71.56%), suggesting its substantial potential in guiding rehabilitation strategies.

Limitations

There are some limitations to this study. First, this study was a single-center retrospective study with a small sample size, which constitutes the most critical methodological limitation. Population heterogeneity, treatment variability, and fluctuations in data quality may potentially compromise model performance. Although our dataset, spanning over ten years and including medical records from both local and nonlocal patients, confers a certain degree of representativeness, and despite rigorous cross-validation for model performance evaluation, single-center data are inevitably influenced by institutional clinical protocols and inherent biases of retrospective data collection. This poses a distinct challenge for dynamic models, whose generalizability is limited by institutional variations in feature distributions and missing patterns. Second, the puncture needle used in MIS leads to artifacts in CT images, which will affect the accurate calculation of hematoma volume. Third, the model only predicted binary outcomes, such as survival or death, favorable

XSL•FO
RenderX

prognosis, and unfavorable prognosis, but could not predict the specific mRS score. The prediction ability of the functional outcome was not comprehensive enough. Finally, we collected clinical data at a few time points, which may have affected the performance of the model. In the future, more clinical data at more time points can help train the model more effectively.

Future Work

All features incorporated into this model are derived from routine clinical practice in MIS for ICH, eliminating the need for additional data collection efforts. Meanwhile, the model supports continuous integration of the latest data and dynamic updating of prediction outcomes, which can precisely match the time-varying characteristics of ICH progression and the clinical decision-making needs of MIS. Leveraging these advantages, the ideal application of the dynamic prediction model developed in this study is to be embedded in the EHR system, thereby providing clinicians with real-time individualized prognostic predictions to support accurate decision-making. However, its implementation faces obstacles such as technical integration and workflow adaptation. Future work will advance in multiple directions. First, more time-point data will be incorporated to improve the model's performance, and external validation of the model will be conducted in multicenter, prospective cohorts. This is a key step to evaluate its real-world generalizability and promote clinical translation. Second, standardized interfaces will be developed to achieve seamless integration of the model with different EHR systems and ensure real-time data updates. Third, the actual impact of the model on clinical decision-making and final clinical outcomes will be evaluated in real clinical settings, and a replicable workflow solution will be formulated.

Conclusions

Based on clinical data, the MultiStep Transformer model uses deep learning methods to achieve the goal of dynamic, comprehensive, and accurate prediction. It can stably output multiple prediction results in a single prognostic model at the same time. This study explored the individualized management of patients with ICH undergoing MIS, providing a new method for predicting short-term and long-term survival and functional outcomes and bringing a reference for the diagnosis and treatment of ICH in clinical practice.

Data Availability

The datasets used during the current study are available from the corresponding author upon reasonable request.

Funding

This work was supported by the Major Program (JD) of Hubei Province (grant number 2023BAA005); Fundamental Research Funds for the Central Universities (grant number YCJJ20251401); the National Natural Science Foundation of China (grant numbers 92148206, 82571528); the Key Research and Development Program of Wuhan (grant number 2024020702030123); the Huazhong University of Science and Technology (grant number 2024JCYJ044); and the Research Fund of Tongji Hospital (grant numbers 2022ZHFY01, AI2024B03).

Authors' Contributions

JW and JS performed investigation, data curation, formal analysis, validation, and wrote the original draft. QY, DC, CP, and YT contributed to conceptualization. YS performed investigation. PZ and ZT handled project administration, supervision, and funding acquisition, and contributed to writing—review and editing. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Baseline Characteristics of Patients Classified by 30-day Survival or Mortality.
[DOCX File , 24 KB - medinform_v14i1e86327_app1.docx]

Multimedia Appendix 2
Baseline Characteristics of Patients Classified by 180-day Survival or Mortality.
[DOCX File , 24 KB - medinform_v14i1e86327_app2.docx]

Multimedia Appendix 3
Baseline Characteristics of Patients Classified by 180-day Favorable or Unfavorable Functional Outcome.
[DOCX File , 24 KB - medinform_v14i1e86327_app3.docx]

Multimedia Appendix 4
Feature labels.

XSL•FO
RenderX

[DOCX File , 17 KB - medinform_v14i1e86327_app4.docx]

References

1. Qureshi AI, Tuhrim S, Broderick JP, Batjer HH, Hondo H, Hanley DF. Spontaneous intracerebral hemorrhage. N Engl J Med 2001 May 10;344(19):1450-1460 [FREE Full text] [doi: 10.1056/NEJM200105103441907] [Medline: 11346811]
2. Sheth KN. Spontaneous Intracerebral Hemorrhage. N Engl J Med 2022 Oct 27;387(17):1589-1596 [FREE Full text] [doi: 10.1056/NEJMra2201449] [Medline: 36300975]
3. Zille M, Farr TD, Keep RF, Römer C, Xi G, Boltze J. Novel targets, treatments, and advanced models for intracerebral haemorrhage. EBioMedicine 2022 Feb 11;76:103880 [FREE Full text] [doi: 10.1016/j.ebiom.2022.103880] [Medline: 35158309]
4. Duan T, Li L, Yu Y, Li T, Han R, Sun X, et al. Traditional Chinese medicine use in the pathophysiological processes of intracerebral hemorrhage and comparison with conventional therapy. Pharmacol Res 2022 May;179:106200 [FREE Full text] [doi: 10.1016/j.phrs.2022.106200] [Medline: 35367344]
5. Bruno C, Cesta CE, Hjellvik V, Ulrichsen SP, Bjørk MH, Esen B, et al. Corrigendum to Antipsychotic use during pregnancy and risk of specific neurodevelopmental disorders and learning difficulties in children: a multinational cohort study [eClinicalMedicine 70 (2024) 102531/DOI: 10.1016/j.eclinm.2024.102531]. EClinicalMedicine 2025 Feb 28;81:103139 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103139] [Medline: 40104086]
6. Wilting FN, Sondag L, Schreuder FH, Dammers R, Klijn CJ, Boogaarts HD. Surgery for spontaneous supratentorial intracerebral haemorrhage. Cochrane Database Syst Rev 2025 Jul 17;7(7):CD015387 [FREE Full text] [doi: 10.1002/14651858.CD015387.pub2] [Medline: 40673401]
7. Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, STICH investigators. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. Lancet 2005 Jan 29;365(9457):387-397 [FREE Full text] [doi: 10.1016/S0140-6736(05)17826-X] [Medline: 15680453]
8. Mendelow AD, Gregson BA, Rowan EN, Murray GD, Gholkar A, Mitchell PM, STICH II Investigators. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial lobar intracerebral haematomas (STICH II): a randomised trial. Lancet 2013 Aug 03;382(9890):397-408 [FREE Full text] [doi: 10.1016/S0140-6736(13)60986-1] [Medline: 23726393]
9. Hanley DF, Thompson RE, Rosenblum M, Yenokyan G, Lane K, McBee N, MISTIE III Investigators. Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial. Lancet 2019 Mar 09;393(10175):1021-1032 [FREE Full text] [doi: 10.1016/S0140-6736(19)30195-3] [Medline: 30739747]
10. Pradilla G, Ratcliff JJ, Hall AJ, Saville BR, Allen JW, Paulon G, ENRICH trial investigators, ENRICH Trial Investigators. Trial of Early Minimally Invasive Removal of Intracerebral Hemorrhage. N Engl J Med 2024 Apr 11;390(14):1277-1289 [FREE Full text] [doi: 10.1056/NEJMoa2308440] [Medline: 38598795]
11. Hemphill JC, Farrant M, Neill TA. Prospective validation of the ICH Score for 12-month functional outcome. Neurology 2009 Oct 06;73(14):1088-1094 [FREE Full text] [doi: 10.1212/WNL.0b013e3181b8b332] [Medline: 19726752]
12. Hemphill JC, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke 2001 Apr 1;32(4):891-897 [FREE Full text] [doi: 10.1161/01.str.32.4.891] [Medline: 11283388]
13. Gregório T, Pipa S, Cavaleiro P, Atanásio G, Albuquerque I, Chaves PC, et al. Prognostic models for intracerebral hemorrhage: systematic review and meta-analysis. BMC Med Res Methodol 2018 Nov 20;18(1):145 [FREE Full text] [doi: 10.1186/s12874-018-0613-8] [Medline: 30458727]
14. Matsumoto K, Ishihara K, Matsuda K, Tokunaga K, Yamashiro S, Soejima H, et al. Machine Learning-Based Prediction for In-Hospital Mortality After Acute Intracerebral Hemorrhage Using Real-World Clinical and Image Data. J Am Heart Assoc 2024 Dec 17;13(24):e036447 [FREE Full text] [doi: 10.1161/JAHA.124.036447] [Medline: 39655759]
15. Zhao X, Zhou B, Luo Y, Chen L, Zhu L, Chang S, et al. CT-based deep learning model for predicting hospital discharge outcome in spontaneous intracerebral hemorrhage. Eur Radiol 2024 Jul;34(7):4417-4426 [FREE Full text] [doi: 10.1007/s00330-023-10505-6] [Medline: 38127074]
16. Huang X, Wang D, Zhang Q, Ma Y, Li S, Zhao H, et al. Development and Validation of a Clinical-Based Signature to Predict the 90-Day Functional Outcome for Spontaneous Intracerebral Hemorrhage. Front Aging Neurosci 2022 May 09;14:904085 [FREE Full text] [doi: 10.3389/fnagi.2022.904085] [Medline: 35615596]
17. Chen Y, Rivier CA, Mora SA, Torres Lopez V, Payabvash S, Sheth KN, et al. Deep learning survival model predicts outcome after intracerebral hemorrhage from initial CT scan. Eur Stroke J 2025 Mar;10(1):225-235 [FREE Full text] [doi: 10.1177/23969873241260154] [Medline: 38880882]
18. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, et al. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. J Biomed Inform 2021 Mar 16;115:103671 [FREE Full text] [doi: 10.1016/j.jbi.2020.103671] [Medline: 33387683]

XSL•FO
RenderX

19. Hall AN, Weaver B, Liotta E, Maas MB, Faigle R, Mroczek DK, et al. Identifying Modifiable Predictors of Patient Outcomes After Intracerebral Hemorrhage with Machine Learning. Neurocrit Care 2021 Feb;34(1):73-84 [FREE Full text] [doi: 10.1007/s12028-020-00982-8] [Medline: 32385834]
20. Schneider H, Huynh TJ, Demchuk AM, Dowlatshahi D, Rodriguez-Luna D, Silva Y, PREDICT/Sunnybrook ICH CTA Study Group. Combining Spot Sign and Intracerebral Hemorrhage Score to Estimate Functional Outcome: Analysis From the PREDICT Cohort. Stroke 2018 Jun;49(6):1511-1514 [FREE Full text] [doi: 10.1161/STROKEAHA.118.020679] [Medline: 29720436]
21. Zhang Z, Zhang R, Xiao K, Sun X. G4Beacon: An In Vivo G4 Prediction Method Using Chromatin and Sequence Information. Biomolecules 2023 Feb 03;13(2):292 [FREE Full text] [doi: 10.3390/biom13020292] [Medline: 36830661]
22. Bashiri FS, Carey KA, Martin J, Koyner JL, Edelson DP, Gilbert ER, et al. Development and external validation of deep learning clinical prediction models using variable-length time series data. J Am Med Inform Assoc 2024 May 20;31(6):1322-1330 [FREE Full text] [doi: 10.1093/jamia/ocae088] [Medline: 38679906]
23. Zheng Z, Zhang X, Yu J, Guo R, Zhangzhong L. Deep Neural Networks for the Classification of Pure and Impure Strawberry Purees. Sensors (Basel) 2020 Feb 23;20(4):1223 [FREE Full text] [doi: 10.3390/s20041223] [Medline: 32102254]
24. Aihara S, Shibata R, Mizukami R, Sakai T, Shionoya A. Deep Learning-Based Myoelectric Potential Estimation Method for Wheelchair Operation. Sensors (Basel) 2022 Feb 18;22(4):73-84 [FREE Full text] [doi: 10.3390/s22041615] [Medline: 35214514]
25. Dirgová Luptáková I, Kubovčík M, Pospíchal J. Wearable Sensor-Based Human Activity Recognition with Transformer Model. Sensors (Basel) 2022 Mar 01;22(5):1911 [FREE Full text] [doi: 10.3390/s22051911] [Medline: 35271058]
26. Gregório T, Pipa S, Cavaleiro P, Atanásio G, Albuquerque I, Chaves PC, et al. Assessment and Comparison of the Four Most Extensively Validated Prognostic Scales for Intracerebral Hemorrhage: Systematic Review with Meta-analysis. Neurocrit Care 2019 Apr;30(2):449-466 [FREE Full text] [doi: 10.1007/s12028-018-0633-6] [Medline: 30426449]
27. You S, Zheng D, Yoshimura S, Ouyang M, Han Q, Wang X, INTERACT Investigators. Optimum Baseline Clinical Severity Scale Cut Points for Prognosticating Intracerebral Hemorrhage: INTERACT Studies. Stroke 2024 Jan;55(1):139-145 [FREE Full text] [doi: 10.1161/STROKEAHA.123.044538] [Medline: 38018833]
28. McCracken DJ, Lovasik BP, McCracken CE, Frerich JM, McDougal ME, Ratcliff JJ, et al. The Intracerebral Hemorrhage Score: A Self-Fulfilling Prophecy? Neurosurgery 2019 Mar 01;84(3):741-748 [FREE Full text] [doi: 10.1093/neuros/nyy193] [Medline: 29762777]
29. Hwang DY, Kim KS, Muehlschlegel S, Wartenberg KE, Rajajee V, Alexander SA, et al. Guidelines for Neuroprognostication in Critically Ill Adults with Intracerebral Hemorrhage. Neurocrit Care 2024 Apr;40(2):395-414 [FREE Full text] [doi: 10.1007/s12028-023-01854-7] [Medline: 37923968]
30. Zyck S, Du L, Gould G, Latorre JG, Beutler T, Bodman A, et al. Scoping Review and Commentary on Prognostication for Patients with Intracerebral Hemorrhage with Advances in Surgical Techniques. Neurocrit Care 2020 Aug;33(1):256-272 [FREE Full text] [doi: 10.1007/s12028-020-00962-y] [Medline: 32270428]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CT: computer tomography
DCA: decision curve analysis
EHR: electronic health record
ENRICH: Early Minimally Invasive Removal of Intracerebral Hemorrhage
FN: false negative
FP: false positive
FUNC: Functional Outcome in Patients with Primary Intracerebral Hemorrhage
GCS: Glasgow Coma Scale
GRU: gated recurrent unit
ICH: intracerebral hemorrhage
MIS: minimally invasive surgery
MISTIE Ⅲ: Minimally Invasive Surgery Plus Alteplase for Intracerebral Hemorrhage Evacuation Phase 3
mRS: modified Rankin Scale
NIHSS: National Institute of Health Stroke Scale
ROC: receiver operating characteristic
STICH: Surgical Trial in Intracerebral Haemorrhage
STICH II: Surgical Trial in Intracerebral Haemorrhage II
TCN: temporal convolutional network
TN: true negative
TP: true positive

XSL•FO
RenderX

TRIPOD-AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence

Edited by M Focsa; submitted 22.Oct.2025; peer-reviewed by M Chen, K Zhang; comments to author 25.Nov.2025; accepted 20.Dec.2025; published 07.Jan.2026.

Please cite as:
Wang J, Shi J, Ye Q, Chen D, Sun Y, Pan C, Tang Y, Zhang P, Tang Z
Deep Learning for Dynamic Prognostic Prediction in Minimally Invasive Surgery for Intracerebral Hemorrhage: Model Development and Validation Study
JMIR Med Inform 2026;14:e86327
URL: https://medinform.jmir.org/2026/1/e86327
doi:10.2196/86327
PMID:

©Jingxuan Wang, Jian Shi, Qing Ye, Danyang Chen, Yuhao Sun, Chao Pan, Yingxin Tang, Ping Zhang, Zhouping Tang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ranking-Aware Multiple Instance Learning for Histopathology Slide Classification: Development and Validation Study

Ho Heon Kim[1,2*], MS; Gisu Hwang[1*], MS; Won Chan Jeong[1], BS; Young Sin Ko[1,3], MD, PhD

[1]AI R&D Center, Seegene Medical Foundation, Seoul, Republic of Korea

[2]Department of Medical Sciences, Graduate School, The Catholic University of Korea, Seoul, Republic of Korea

[3]Pathology Center, Seegene Medical Foundation, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Young Sin Ko, MD, PhD
AI R&D Center, Seegene Medical Foundation
17F, Seegene Medical Tower, 288 Dapsimni-ro, Dongdaemun-gu
Seoul, 02637
Republic of Korea
Phone: 82 2 2228 9349
Email: noteasy@mf.seegene.com

Abstract

Background: Multiple instance learning (MIL) is widely used for slide-level classification in digital pathology without requiring expert annotations. However, even partial expert annotations offer valuable supervision; few studies have effectively leveraged this information within MIL frameworks.

Objective: This study aims to develop and evaluate a ranking-aware MIL framework, called rank induction, that effectively incorporates partial expert annotations to improve slide-level classification performance under realistic annotation constraints.

Methods: We developed rank induction, a MIL approach that incorporates expert annotations using a pairwise rank loss inspired by RankNet. The method encourages the model to assign higher attention scores to annotated regions than to unannotated ones, guiding it to focus on diagnostically relevant patches. We evaluated rank induction on 2 public datasets (Camelyon16 and DigestPath2019) and an in-house dataset (Seegene Medical Foundation-stomach; SMF-stomach) and tested its robustness under 3 real-world conditions: low-data regimes, coarse within-slide annotations, and sparse slide-level annotations.

Results: Rank induction outperformed existing methodologies, achieving an area under the receiver operating characteristic curve (AUROC) of 0.839 on Camelyon16, 0.995 on DigestPath2019, and 0.875 on SMF-stomach. It remained robust under low-data conditions, maintaining an AUROC of 0.761 with only 60.2% (130/216) of the training data. When using coarse annotations (with 2240-pixel padding), performance slightly declined to 0.823. Remarkably, annotating just 20% (18/89) of the slides was enough to reach near-saturated performance (AUROC of 0.806, vs 0.839 with full annotations).

Conclusions: Incorporating expert annotations through ranking-based supervision improves MIL-based classification. Rank induction remains robust even with limited, coarse, or sparsely available annotations, demonstrating its practicality in real-world scenarios.

(JMIR Med Inform 2026;14:e84417) doi:10.2196/84417

KEYWORDS

multiple instance learning; learning to rank; digital pathology; whole slide image; weakly supervised learning; mixed supervision; data-efficient training

Introduction

With the rise of digital pathology, whole slide images (WSIs) have become essential for computer-assisted diagnosis tasks such as tumor detection, subtype classification, and prognosis prediction [1-3]. Due to their extremely high resolution, WSIs are typically divided into smaller patches before being used in deep learning models. In fully supervised learning settings, this requires patch-level labels—an annotation process that is time consuming and labor intensive due to the scarcity of lesion-containing patches. To reduce the burden of exhaustive labeling, the multiple instance learning (MIL) framework has

XSL•FO
RenderX

emerged as the dominant approach for slide-level prediction. MIL operates in a weakly supervised setting, using only slide-level labels and eliminating the need for detailed patch annotations. Despite this advantage, MIL models still require large datasets to achieve generalizable performance due to a weak supervision signal [4]. In data-constrained settings, MIL training often becomes unstable, leading to a significant reduction in predictive accuracy. To address the limitations of weak supervision, recent studies can be broadly categorized based on the strength of supervision into three main streams: (1) slide-level approaches relying solely on slide labels, (2) pseudolabel-based methods that infer patch-level labels, and (3) expert-guided methods leveraging direct annotations.

Slide-level approaches operate solely with slide-level supervision but aim to improve performance through semisupervised learning, self-supervised learning, or multiresolution architectures [5-12]. For instance, dual-stream multiple instance learning (DS-MIL) introduced a dual-stream architecture that models relationships between high- and low-attention instances to mitigate attention collapse and improve feature discrimination [11]. Hierarchical image pyramid transformer (HIPT) leveraged hierarchical self-supervised vision transformers to capture multiscale contextual information across gigapixel WSIs, demonstrating that strong representation learning can alleviate weak supervision limitations [12]. However, because these approaches still rely only on slide-level labels, the supervision signal remains weak and cannot fully guide the model toward diagnostically relevant regions.

Pseudolabel-based methods generate inferred patch-level labels from attention maps to strengthen the weak supervision signal. A prominent data-efficient training using pseudolabel is clustering-constrained attention multiple instance learning (CLAM), which selects high- and low-attention instances within each slide to form pseudopositive or pseudonegative sets and trains an auxiliary instance classifier [13]. Iterative multiple instance learning proposed an iterative framework that refines both instance and slide representations by retraining the feature extractor with attention-derived pseudolabels, though its performance depends on the stability of the initial MIL attention [14]. In general, pseudolabeling techniques aim to fill the gap in patch-level labels by generating inferred annotations [13-18]. Despite their utility, pseudolabels are inherently noisy, as they are inferred rather than observed. Consequently, MIL pipelines that solely rely on pseudolabels often suffer from degraded performance.

In contrast, expert-guided methods offer a promising way to improve both performance and supervision quality by directly integrating expert annotations. Nevertheless, this approach remains relatively underexplored compared to other categories. A notable study described the attention induction model, which uses pathologist-drawn lesion annotations to guide model attention [19]. However, by enforcing attention scores proportional to the size of the annotated regions, this method inherently biases the model toward larger areas, increasing the risk of overlooking small but clinically critical diagnostic features.

To overcome this limitation, we propose rank induction, a novel MIL framework that introduces a pathology-informed inductive bias. Rather than constraining attention based on region size, rank induction assumes that annotated regions should carry higher diagnostic importance than unannotated ones. By learning relative ranks between patches, our method provides finer-grained supervision, guiding the model to focus on diagnostically relevant areas even when annotations are sparse or training data are limited.

Methods

Problem Definition

We extend the standard MIL framework by incorporating patch-level annotations from pathologists as explicit instance labels, either across all training slides or only a subset. In a typical MIL setting, each WSI is divided into K patches $\{x_1,\ldots,x_K\}$, and only a slide-level label, indicating disease presence, is provided. A slide is labeled positive if at least 1 patch contains a lesion and negative if all patches are normal (for simplicity, we assume a fixed K across slides).

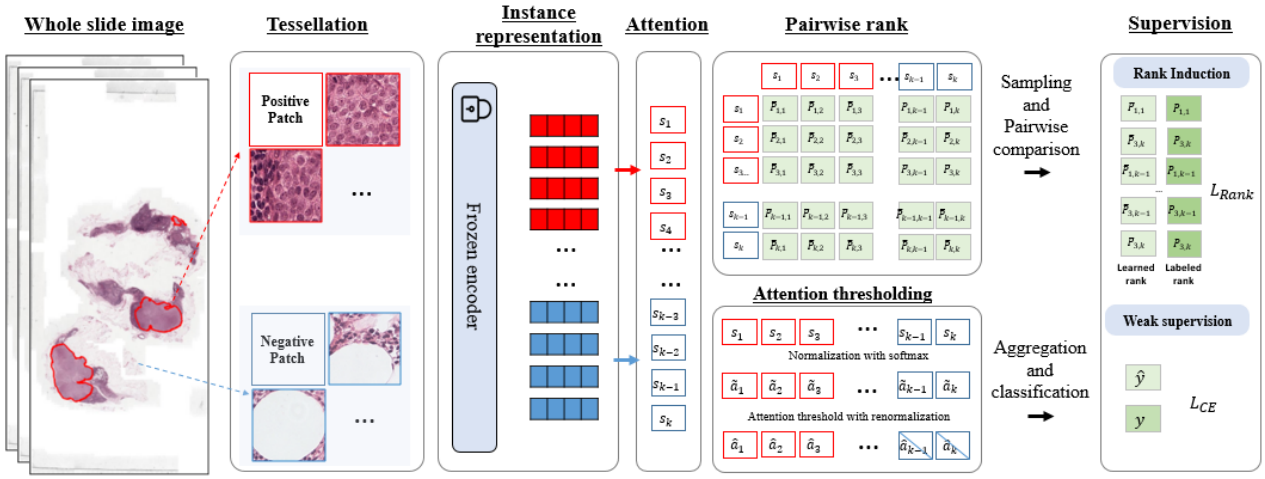
In our extended setting, we additionally use binary labels $\{y_1,\ldots,y_k\}$ at the patch level, where $y_k \in \{0,1\}$, $y_k=1$, if a patch overlaps with a pathologist-annotated region, and $y_k=0$ otherwise. The core inductive bias we introduce is simple but clinically grounded: patches overlapping annotated regions should receive higher attention than unannotated ones. This reflects the assumption that tumor-confirmed regions are more relevant to slide-level diagnosis than those confirmed to be normal.

Rank Induction

As illustrated in Figure 1, rank induction consists of 2 key components: pairwise rank comparison derived from rich annotation supervision and attention thresholding to filter out residual noninformative patches. To implement our ranking-aware attention mechanism, we first define the computation of patch-wise attention. Let z_k denote the feature vector of patch x_k extracted by a backbone feature encoder. We compute a raw attention score s_k from z_k using an attention layer and normalize it using softmax: . Inspired by RankNet [20], rank induction introduces a ranking constraint that encourages higher attention scores for annotated patches than for nonannotated ones, such that $s_i > s_j$, where s_i is the attention score for an annotated patch and s_j for a nonannotated patch.

XSL•FO
RenderX

Figure 1. Overview of rank induction.



To avoid the normalization constraint of softmax ( ), we apply the ranking loss directly on the raw attention scores $s_i \in R$. We define the pairwise ranking preference using a sigmoid function:

Here, *m* is a margin parameter that enforces separation between positive and negative patches and σ is a scaling factor. We exclude same-class pairs (eg, lesion-lesion or normal-normal) and define the valid pair set: $P = \{(i,j)|y_i = 1, y_j = 0\}$. Let be the ground truth ranking label. The rank loss is defined as:

We jointly optimize the slide-level binary cross-entropy loss and the rank loss:

$$L = L_{BCE} + \lambda L_{rank}$$

where λ is a weighting hyperparameter. For slides without annotated lesions (ie, negative slides), we omit the ranking loss and only apply the slide-level loss.

To mitigate slide-level representation dilution, where a few high-attention lesion patches are overwhelmed by numerous low-attention nonlesion patches during the attention-weighted summation, we apply an attention thresholding strategy during training. Specifically, we threshold the normalized attention weights and then renormalize all (including zero-valued) weights as follows:

Here, *T* is a predefined threshold and *K* is the number of patches. The threshold *T* is introduced to prevent the slide-level representation from being dominated by the large number of negative instances when the bag size is large. For example, in a slide containing 2000 instances with only 10 lesion patches, even if attention is highly concentrated on those few positives, the aggregate contribution from thousands of near-zero attention weights of negative patches can still overshadow the lesion signal, leading to degraded slide representations. By applying *T*, we effectively suppress these noise-like weights and maintain a more balanced representation across instances.

Datasets

We evaluated rank induction on 2 public datasets (Camelyon16 [21] and DigestPath2019 [22]) and an in-house dataset (Seegene Medical Foundation-stomach; SMF-stomach). Textbox 1 provides a detailed description of the datasets used in this study, with sample count statistics summarized in Table 1, which presents the statistics for the 3 datasets.

XSL•FO
RenderX

Textbox 1. Description of the datasets used in this study.

- Camelyon16: this dataset consists of 399 hematoxylin and eosin-stained whole slide images (WSIs) of lymph node sections for breast cancer metastasis detection. It is originally split into 67.7% (270/399) training slides and 32.3% (129/399) test slides. For model development and hyperparameter tuning, we further split the 270 training WSIs into 216 (80%) training slides and 54 (20%) validation slides (an 8:2 split). Each slide includes a slide-level diagnosis and polygon-based lesion annotations provided by expert pathologists. In this study, slides containing metastasis were labeled as abnormal, while negative slides were labeled as normal.
- DigestPath2019: we used the colonoscopy tissue segmentation dataset from the DigestPath Challenge 2019. The dataset originally includes 660 training images and 212 test images with slide-level labels (benign or malignant). However, because the test set is not publicly available, we used only the 660 labeled WSIs. We allocated 20% (132/660) of the slides as the test set and used the remaining 80% (528/660) for training and validation. Specifically, we split the 528 slides into 422 (79.9%) for training and 106 (20.1%) for validation. Each WSI is a JPEG image with an average resolution of approximately 5000×5000 pixels. For consistency, we mapped malignant samples to the abnormal class and benign samples to the normal class.
- SMF-stomach (in-house): we constructed an in-house gastric pathology dataset at Seegene Medical Foundation (SMF). Slide preparation and digitization followed standard operating procedures described by Kim et al [23]. A distinct feature of this dataset, as illustrated in Multimedia Appendix 1, is that at least 3 consecutive tissue sections are placed on a single glass slide; however, pixel-level annotations were not exhaustively applied to every section. A single board-certified pathologist reviewed the slides and provided polygon-based annotations only for representative regions; thus, unannotated tissue sections may coexist within the same WSI. Following the diagnostic criteria of Ko et al [24], lesions were originally categorized as malignant, dysplasia, negative, and uncategorized. We excluded the uncategorized class due to its high heterogeneity. For the binary classification task, we mapped malignant and dysplasia to the abnormal class and negative to the normal class. The final dataset comprised 209 slides, allocated as 120 (57.4%) for training, 30 (14.4%) for validation, and 59 (28.2%) for testing.

Table 1. Dataset statistics (whole slide images).

Dataset	Train		Validation		Test	
	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
Camelyon16	89	127	22	32	49	80
DigestPath2019	160	262	40	66	50	82
SMF-stomach[a]	81	39	20	10	39	20

[a]SMF-stomach: Seegene Medical Foundation-stomach.

Data Preprocessing

We applied the same preprocessing pipeline to all datasets. First, we performed tissue-background segmentation using the multilevel Otsu algorithm. Each WSI was converted to grayscale and thresholded to generate a binary mask that separated tissue from the white background. We retained only the foreground tissue regions and discarded blank or irrelevant areas.

Next, we tessellated each foreground region into nonoverlapping patches of size 224×224 pixels. For Camelyon16 and SMF-stomach, we extracted patches at 20× magnification, whereas for DigestPath2019, we used the original image resolution.

After tessellation, we computed the tissue coverage ratio for each patch and discarded patches containing less than 5% tissue, as these were considered mostly empty. Finally, we used a ResNet-50 [24] pretrained on ImageNet-1k [25] to extract features from each patch. Specifically, we used the output of the third residual block and applied an adaptive average pooling layer to obtain a 2048-dimensional feature vector.

Statistical Methods

To evaluate rank induction, we designed a series of experiments targeting performance, data efficiency, and annotation use. Our evaluation included 6 different experiment types, which are discussed in the following subsections.

Model Comparison

We trained and evaluated 7 different MIL models on 3 datasets: the baseline attention-based MIL (AB-MIL), the attention induction method, 2 variants of CLAM (single-branch [SB]; multibranch [MB]), DS-MIL, HIPT, and the proposed rank induction method.

Low-Data Regime

To test data efficiency, we trained models on subsets of the training data at 20%, 40%, 60%, 80%, and 100% of the full set. For each fraction, we trained all 7 models (AB-MIL, attention induction, CLAM-SB, CLAM-MB, DS-MIL, HIPT, and rank induction) from scratch and evaluated them on a common held-out test set.

Annotation Granularity

To simulate real-world coarse annotations, we degraded annotation precision by expanding the original lesion masks on Camelyon16 slides. At 40× magnification, we added symmetric paddings of 0, 448, 896, 1344, 1792, and 2240 pixels to each polygon annotation. For instance, a 448-pixel padding at 40× resolution adds 1 patch-width around the lesion, which corresponds to a 224-pixel patch at 20× magnification—potentially including normal tissue and simulating noisy annotations.

XSL•FO
RenderX

Subset-Annotated Slide Setting

In real-world clinical settings, not all slides can be exhaustively annotated. We tested rank induction under partial supervision by varying the proportion of positive slides with patch-level annotations (5%, 10%, 20%, 40%, 60%, and 80%) in the Camelyon16 training set (111 positive slides in total). The remaining positive slides were treated with slide-level labels only. Negative slides, by definition, had no lesion annotations.

Ablation Test

To assess the contribution of attention thresholding, we performed an ablation study with threshold values $T \in \{0,0.25,0.5,1,2,5\}$. The case of $T=0$ served as the reference condition (ie, without attention thresholding). We compared WSI classification performance and the cumulative attention weight assigned to nonlesion patches on the test set to examine how different thresholds affect model behavior.

Attention Localization

To assess interpretability, we compared how different MIL models spatially focus on tumor regions in Camelyon16. This is not a dense pixel-wise segmentation map; rather, it is a ranking over discrete patches, with scores defined for all patches in the slide [13]. Because no established gold standard protocol exists for evaluating such attention maps, we quantified spatial alignment between attention and expert lesion annotations using 3 metrics: intersection over union (IoU), Dice score, and pointing game [25]. For each WSI, we ranked all patches by attention weight and then selected only the top α% relative to the number of tumor-annotated patches in that slide ($\alpha \in \{1, 5, 10, 20, 50, 75, 100\}$). To ensure that the ideal case could always achieve a score of 1, IoU and Dice score were computed by sampling up to the number of tumor-annotated patches, whereas the pointing game was evaluated by sampling from all patches in the slide. We then computed IoU and Dice score between the selected high-attention patches and the tumor-annotated patches, averaged over positive test slides. We also measured the pointing game by calculating a hit rate, defined as whether at least 1 of the top α highest attention patches fell within the annotated tumor region (classified as a hit). For all malignant test slides, localization accuracy was defined as hits divided by the sum of hits and misses.

For model evaluation, we computed area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and accuracy for slide-level classification. All results were averaged over a 10-fold Monte Carlo cross-validation. We assessed statistical significance using the Mann-Whitney *U* test, with *P* values <.05 considered statistically significant.

Experiment Details

We implemented the models in Python using the PyTorch deep learning framework. WSI data preprocessing was carried out with OpenSlide, OpenCV, and Pillow libraries. Visualization of attention maps and result plots was generated using Seaborn and Matplotlib libraries (specific version information for each library can be found in the GitHub repository). The training and experiments were run on a computing system with a pair of NVIDIA A100 graphics processing units (80 GB memory each) and an Intel Xeon Gold 6338N central processing unit (128 cores, 2.20 GHz) with 512 GB of RAM. Additionally, because the original implementation of attention induction was not publicly available, we reimplemented the method based on the descriptions provided in the paper. Our implementation is available to facilitate reproducibility [26].

For optimization, we used the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. We trained each model for up to 200 epochs using an early stopping criterion: if the validation loss did not improve for 7 consecutive epochs (after an initial 20-epoch warmup period), training was stopped. In practice, most models converged well before 200 epochs under this criterion. For all MIL benchmarks, the ResNet-50 backbone was frozen during training. Crucially, we used a frozen ResNet-50 backbone for all MIL benchmarks. Although methods such as DS-MIL and HIPT typically leverage self-supervised pretraining or hierarchical encoders, we used ResNet-50 as the feature extractor for this experiment. This ensured a direct comparison of the MIL aggregation strategies themselves, isolating their performance from the variations in pretrained feature representations.

We set the rank loss hyperparameters as follows: the scaling factor $\sigma=1$ and margin $m=1$ in the pairwise probability formula, and the rank loss weight $\lambda=1$. In each training iteration, to limit memory use and ensure computational efficiency when computing the rank loss, we randomly sampled up to 1024 positive and 1024 negative patches per slide. The attention threshold *T* was set to 1.

Ethical Considerations

This study used both publicly available deidentified datasets (Camelyon16 and DigestPath2019) and a private clinical dataset (SMF-stomach). The Camelyon16 and DigestPath2019 datasets consist of fully deidentified WSIs, publicly released for research under their respective challenge licenses; therefore, no institutional review board (IRB) approval or informed consent was required for their use. For the SMF-stomach dataset, this study was approved by the IRB of the Seegene Medical Foundation (SMF-IRB-2024-015). Given the anonymous and deidentified nature of the retrospective pathological images, informed consent was not required.

Results

Model Performance

We compared rank induction against 6 representative MIL benchmarks: AB-MIL, attention induction, CLAM, DS-MIL, and HIPT, regarding slide-level classification performance. On Camelyon16, rank induction achieved an AUROC of 0.839 (SD 0.050) and an AUPRC of 0.850 (SD 0.042), outperforming all baseline methods with consistently low variance (Table 2). We observed similar performance trends on the DigestPath2019 dataset. Although all models performed well due to the dataset's relative simplicity, rank induction achieved the best results (AUROC=0.995; AUPRC=0.993) with stable variance. On the SMF-stomach dataset, rank induction demonstrated the best performance, achieving an AUROC of 0.875 (SD 0.008) and an AUPRC of 0.937 (SD 0.006). Due to space constraints, the

XSL•FO
RenderX

full accuracy comparison for all datasets is reported in Multimedia Appendix 2.

Table 2. Comparison of model performance across methods on 2 public datasets and an in-house dataset.

Model	Camelyon16		DigestPath2019		SMF[a]-stomach	
	AUROC[b], mean (SD)	AUPRC[c], mean (SD)	AUROC, mean (SD)	AUPRC, mean (SD)	AUROC, mean (SD)	AUPRC, mean (SD)
AB-MIL[d]	0.741 (0.146)	0.730 (0.183)	0.993 (0.003)	0.990 (0.005)	0.864 (0.031)	0.928 (0.016)
Attention induction	0.742 (0.142)	0.727 (0.179)	0.994 (0.002)	0.990 (0.003)	0.868 (0.031)[e]	0.930 (0.016)[e]
CLAM-SB[f]	0.732 (0.137)	0.700 (0.179)	0.977 (0.019)	0.965 (0.026)	0.837 (0.023)	0.919 (0.013)
CLAM-MB[g]	0.794 (0.128)[e]	0.767 (0.168)[e]	0.976 (0.019)	0.963 (0.026)	0.838 (0.032)	0.916 (0.025)
DS-MIL[h]	0.690 (0.170)	0.666 (0.223)	0.995 (0.004)[e]	0.993 (0.006)[e]	0.839 (0.031)	0.915 (0.016)
HIPT[i]	0.483 (0.103)	0.416 (0.136)	0.962 (0.058)	0.947 (0.091)	0.676 (0.064)	0.751 (0.060)
Rank induction	*0.839 (0. 050)* [j,k, l,m, o ,p]	*0.850 (0. 042)* [j,k, l,m, o ,p]	*0.995 (0. 002)* [j,k, l,m,n,p]	*0.993 (0. 002)* [j,k, l,m,n, p]	*0.875 (0. 008)* [j,m, n,o,p]	*0.937 (0. 006)* [j,m, n,o,p]

[a]SMF: Seegene Medical Foundation.

[b]AUROC: area under the receiver operating characteristic curve.

[c]AUPRC: area under the precision-recall curve.

[d]AB-MIL: attention-based multiple instance learning.

[e]Second-best result.

[f]CLAM-SB: clustering-constrained attention multiple instance learning, single-branch.

[g]CLAM-MB: clustering-constrained attention multiple instance learning, multibranch.

[h]DS-MIL: dual-stream multiple instance learning.

[i]HIPT: hierarchical image pyramid transformer.

[j]Italicization indicates the best result.

[k]P<.05 vs attention-based multiple instance learning.

[l]P<.05 vs attention induction.

[m]P<.05 vs clustering-constrained attention multiple instance learning, single-branch.

[n]P<.05 vs clustering-constrained attention multiple instance learning, multibranch.

[o]P<.05 vs dual-stream multiple instance learning.

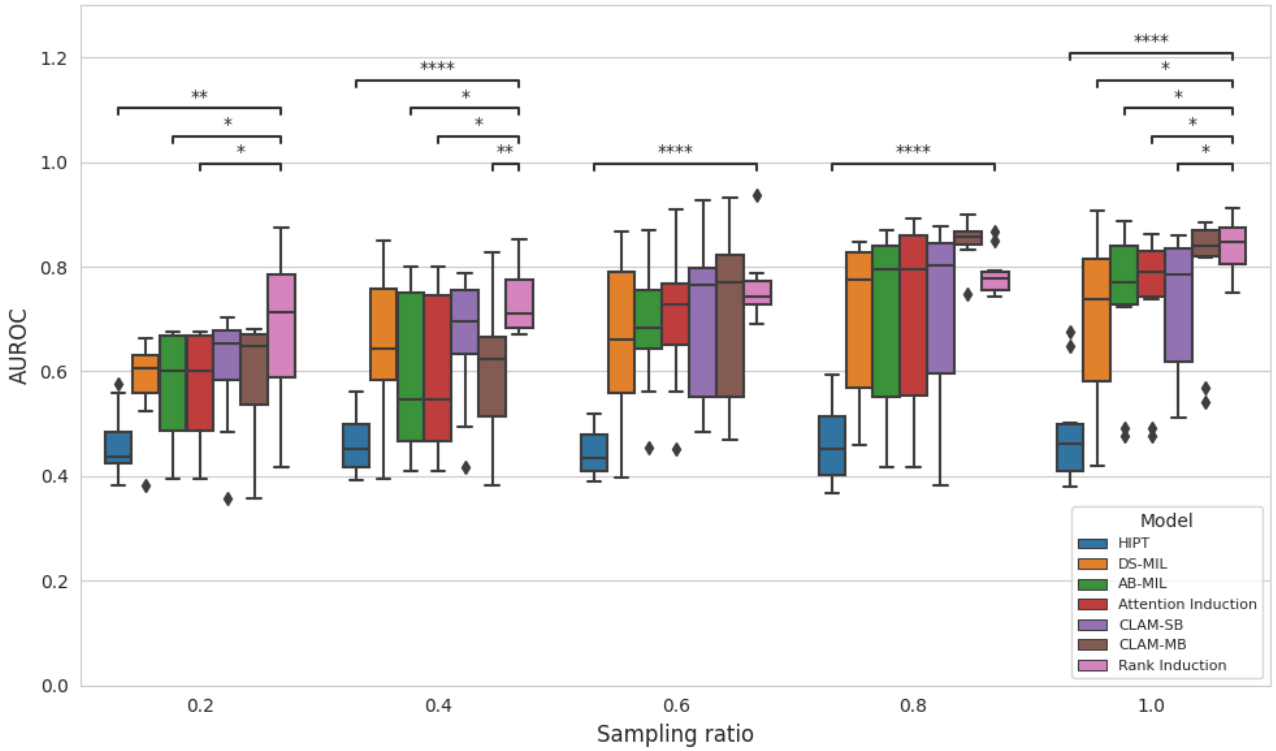
[p]P<.05 vs hierarchical image pyramid transformer.

Data-Efficient Learning With Expert Annotation

To investigate data-efficient training with expert annotation under low-data regimes reflecting real-world constraints, we assessed model performance on Camelyon16 by varying the proportion of training data (Figure 2). With only 20% of the slides, all methods achieved AUROC scores in the range of 0.4 to 0.6, and statistically significant differences were observed between rank induction and the other benchmarks. Among them, rank induction achieved the highest AUROC. At 40% and 60% of the training data size, rank induction significantly outperformed other MIL methods. When trained with 80% of the data, rank induction achieved an AUROC of 0.786 (SD 0.042), surpassing HIPT (AUROC=0.464, SD 0.076), DS-MIL (AUROC=0.709, SD 0.154), AB-MIL (AUROC=0.707, SD 0.186), attention induction (AUROC=0.714, SD 0.193), CLAM-SB (AUROC=0.721, SD 0.180), and CLAM-MB (AUROC=0.847, SD 0.042).

XSL•FO
RenderX

Figure 2. Area under the receiver operating characteristic curve (AUROC) performance across different training set sampling ratios on Camelyon16. AB-MIL: attention-based multiple instance learning; DS-MIL: dual-stream multiple instance learning; CLAM-MB: clustering-constrained attention multiple instance learning, multibranch; CLAM-SB: clustering-constrained attention multiple instance learning, single-branch; HIPT: hierarchical image pyramid transformer. Asterisks indicate statistical significance based on the Mann-Whitney U test: *P<.05, **P<.001, ***P<.0001, and ****P<.00001.

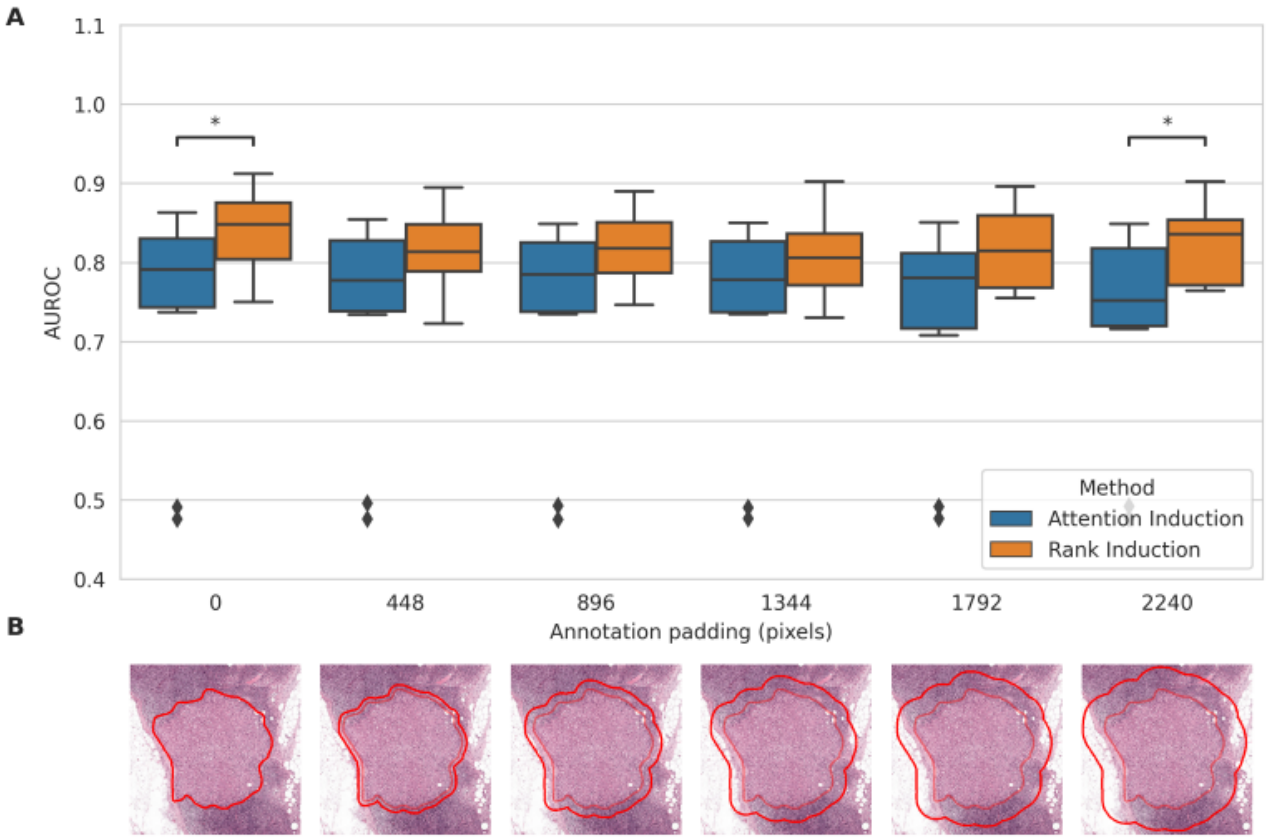


Annotation Granularity

We evaluated model robustness to noisy or imprecise annotations by progressively padding the pathologist-drawn tumor regions on Camelyon16 to simulate coarse lesion markings. Both rank induction and attention induction were trained under each noise condition. With the original annotations (0-pixel padding), rank induction achieved an AUROC of 0.839 (SD 0.050). Even at the highest padding level, its performance declined only slightly to an AUROC of 0.823 (SD 0.050; Figure 3). In comparison, attention induction dropped from an AUROC of 0.742 (SD 0.142) to 0.723 (SD 0.134) as annotation noise increased. Across all levels of annotation coarseness, rank induction consistently outperformed attention induction, although the differences were not statistically significant at some padding levels.

XSL•FO
RenderX

Figure 3. Annotation granularity experiment on Camelyon16. (A) Comparison of slide-level classification performance (area under the receiver operating characteristic curve [AUROC]) between attention induction and rank induction. (B) Representative visualization of the artificially expanded annotations used in the experiment. The thin inner red line indicates the original precise lesion annotation provided by the pathologist (ground truth), while the bold outer red line represents the artificially expanded annotation boundary simulating coarser supervision granularity.

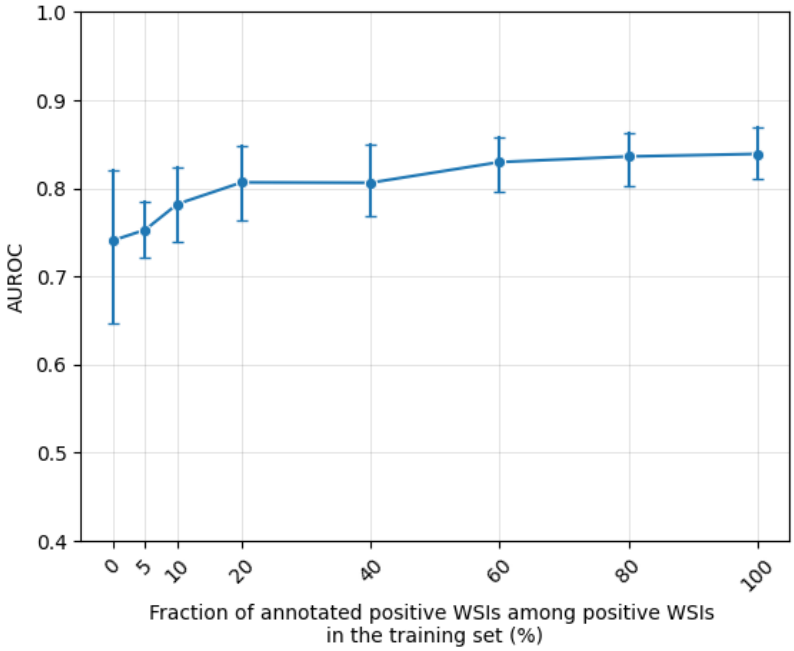


Fraction of Expert Annotation

To simulate real-world scenarios in which annotating all WSIs is infeasible, we evaluated model performance by varying the proportion of annotated positive WSIs in the training set. As a baseline (0% annotation), AB-MIL achieved an average AUROC of 0.741 (SD 0.146; Figure 4). With only 10% (9/89) of positive WSIs annotated in the train dataset, rank induction boosted the performance to an average AUROC of 0.782 (SD 0.070). Increasing the annotation coverage to 20% (18/89) raised the AUROC to 0.807 (SD 0.070), while using all available positive WSIs (N=89) yielded a final AUROC of 0.839 (SD 0.050). Notably, rank induction achieved near-saturated performance with 20% of the annotated data, demonstrating high data efficiency. Moreover, the introduction of even minimal annotation significantly reduced performance variance compared to the unannotated baseline. The performance curve began to plateau beyond approximately 20% annotation coverage, indicating diminishing returns in AUROC gains, while maintaining low variability.

XSL•FO
RenderX

Figure 4. Impact of annotation fraction on area under the receiver operating characteristic curve (AUROC). WSI: whole slide image.



Ablation Test

The impact of varying the attention threshold *T* on model performance and attention distribution is presented in Table 3. As *T* increased from 0 to 2, both AUROC and AUPRC improved steadily, accompanied by a marked reduction in the average sum of attention assigned to nonlesion patches.

Table 3. Effect of attention threshold *T* on whole slide image (WSI) classification performance and sum of attention weight for nonlesion patches.

Threshold	Test dataset (n=129)		Positive WSI in the test dataset (n=49; sum of nonlesion attention), mean (SD)
	AUROC[a]	AUPRC[b]	
0	0.830	0.846	0.465 (0.420)
0.25	0.853	0.878	0.411 (0.434)
0.5	0.869	0.881	0.362 (0.423)
1	0.894	0.905	0.269 (0.387)
2	0.916	0.922	0.178 (0.354)
5	0.909	0.914	0.105 (0.339)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

With no thresholding (*T*=0) or low thresholds (*T*<1, where attention distributions remain nearly uniform), the AUROC remained relatively low. This suggests that weakly attended nonlesion patches diluted the slide-level representation. The best performance was observed at *T*=2 (AUROC=0.916; AUPRC=0.922), indicating that moderate thresholding effectively suppresses noise-like, low-value attention weights while preserving discriminative focus. Excessively high thresholds (eg, *T*=5) yielded a slight decline, reflecting an optimal balance near *T*=2.

Attention Localization and Qualitative Analysis

To evaluate the effect of rank induction on attention localization, we examined whether the attention weights were concentrated within expert-annotated regions. Across all 3 localization metrics, rank induction consistently demonstrated greater focus on the annotated areas compared to other MIL methods. At α=5, AB-MIL, attention induction, and rank induction achieved IoU scores of 0.029, 0.023, and 0.031 and Dice scores of 0.056, 0.045, and 0.059, respectively. When α increased to 75, the performance gap between rank induction and the baseline models became more pronounced (Figures 5A and 5B). A similar trend was observed in pointing game accuracy (Figure 5C). Qualitatively, both AB-MIL and attention induction produced noisy attention maps that highlighted not only tumor regions but also large areas of normal tissue, resulting in frequent false positives. In contrast, rank induction generated sharp and well-localized attention maps, accurately delineating lesion areas with minimal off-target activation—closely matching the ground truth annotations. This distinct separation between tumor and normal regions explains rank induction's

XSL•FO
RenderX

superior performance and enhances the interpretability of its predictions (Figure 6).

Figure 5. Attention localization performance of attention-based multiple instance learning (AB-MIL), attention induction, and rank induction. (A) Intersection over union (IoU) score according to the top α% of patches. (B) Dice score according to the top α% of patches. (C) Average pointing game score. WSI: whole slide image.

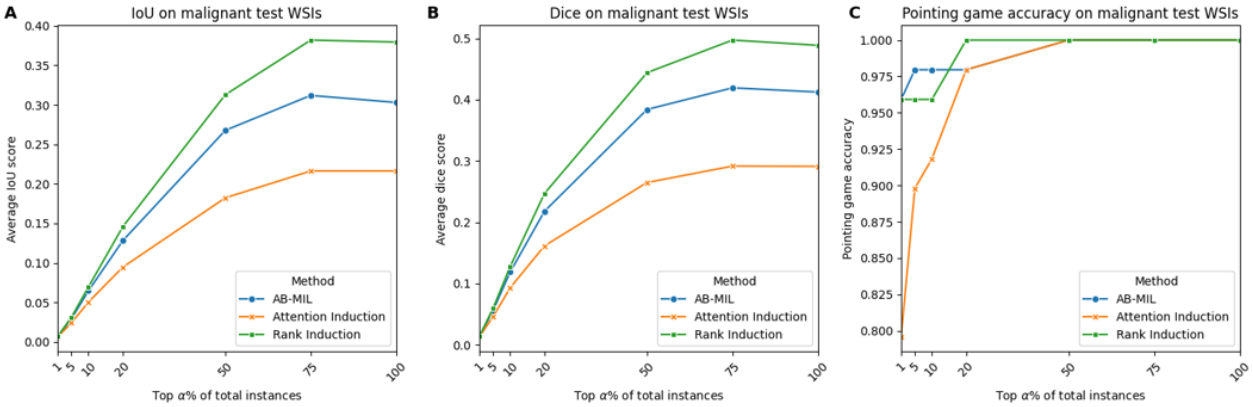
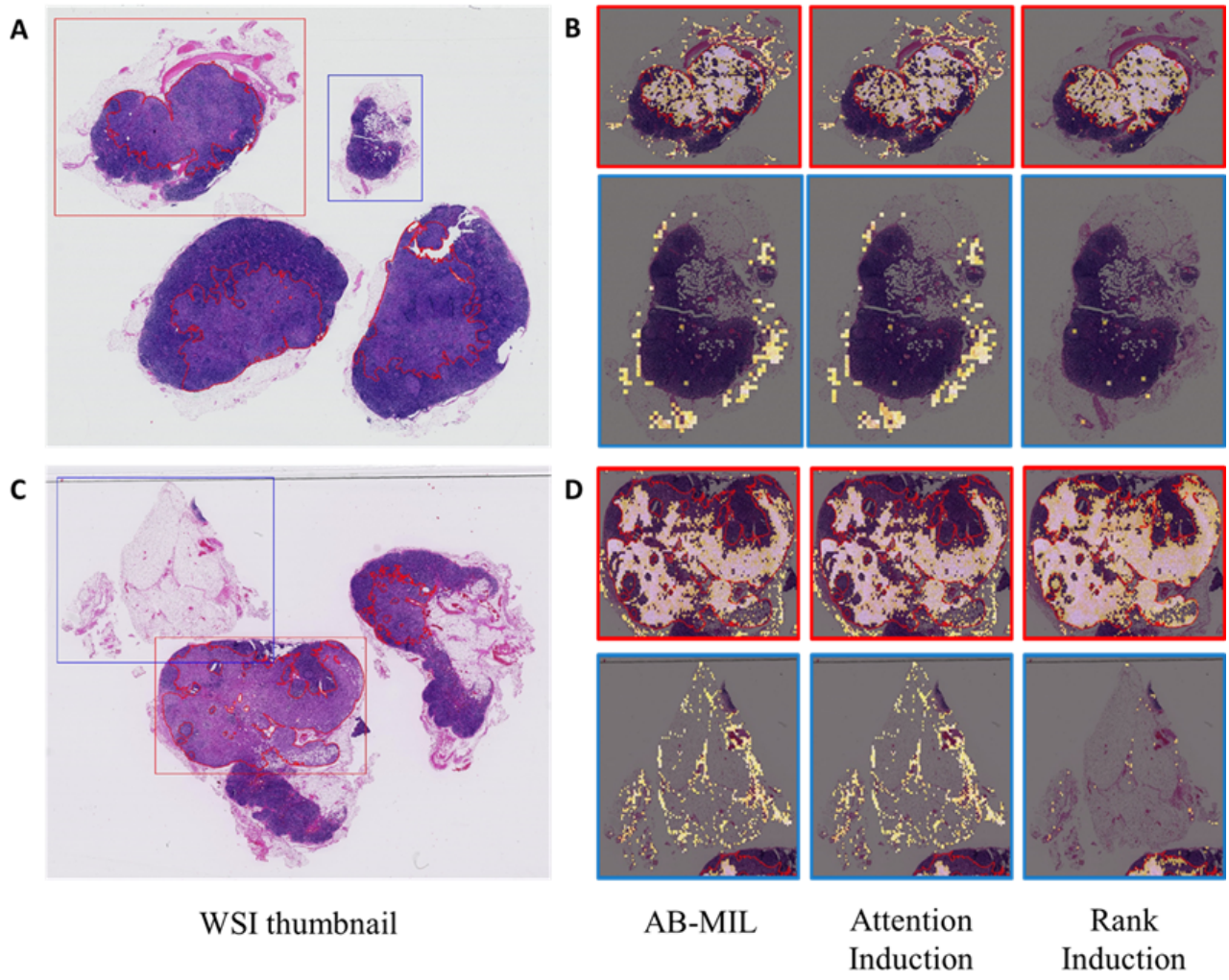


Figure 6. Qualitative visualization of patch selection for attention-based multiple instance learning (AB-MIL), attention induction, and rank induction. (A) A positive WSI thumbnail. (B) Attention distributions from each model in the lesion (red box) and nonlesion (blue box) regions in (A). (C) Another positive WSI thumbnail. (D) Attention distributions from each model in the lesion and nonlesion regions in (C).



XSL•FO
RenderX

Discussion

Principal Findings

In this work, we present rank induction, the first method to integrate MIL with expert annotations using pairwise rank constraints. By enforcing a ranking loss that encourages higher attention scores for annotated patches than for nonannotated ones, our approach effectively strengthens the weak supervision signal inherent to MIL frameworks.

We validated rank induction through comprehensive experiments on 2 public datasets and 1 in-house dataset, covering 5 key scenarios: full-data classification, low-data regimes, varying annotation granularity, subset annotation settings, and qualitative visualization. Across all settings, rank induction consistently outperformed baseline methods, including pseudo-labeling approaches and attention induction, while producing more faithful and interpretable attention heat maps.

Notably, the proposed ranking mechanism is model-agnostic and can be used as a versatile module within other MIL architectures. As demonstrated in Multimedia Appendix 3, integrating rank induction into existing methods such as DS-MIL and HIPT consistently improved their performance. In our main experiments, these baseline models (DS-MIL and HIPT) exhibited lower performance than that reported in their original studies. This discrepancy arises because we used a standard ImageNet-pretrained ResNet-50 backbone across all models to ensure a fair comparison of the MIL aggregation strategies themselves, rather than relying on the self-supervised pretraining typically required by those methods. However, even under this constrained setting, incorporating rank induction significantly boosted their classification accuracy, demonstrating the robustness and generalizability of our ranking-based supervision.

Ultimately, these performance improvements stem from bridging weak and strong supervision through informative pairwise constraints that approximate expert guidance. Rank induction is grounded in a simple yet clinically intuitive assumption: models should prioritize annotated regions over nonannotated ones. This inductive bias aligns naturally with domain knowledge in pathology, establishing rank induction as an effective and practical solution for real-world applications.

Real-World Application With Expert Annotations

The pairwise ranking approach in rank induction addresses key challenges associated with expert annotations in real-world settings. In practice, tumor regions often exhibit irregular and heterogeneous morphology with poorly defined boundaries, making precise annotations difficult. Consequently, annotations are frequently coarse or incomplete. Despite this, rank induction maintained robust classification performance even with noisy annotations (Figure 3). This resilience likely stems from the pairwise training mechanism: although some comparisons may involve incorrectly labeled patches, most pairwise relationships remain valid, enabling the model to suppress noise and learn effectively.

Moreover, due to time and resource limitations, only a subset of positive WSIs can typically be annotated in clinical workflows. In such scenarios, extracting maximum value from limited expert input is critical. Unlike attention induction, which relies on assigning exact attention weights to individual patches, rank induction leverages a larger number of patch-pair comparisons from the valid pair set P, which grows quadratically with the number of annotated patches. This design allows the model to achieve near-saturated performance even with partial annotations (Figure 4).

Although pseudolabeling offers an alternative when expert annotations are unavailable, as seen in methods such as CLAM, it introduces uncertainty by relying on the model's own predictions. These pseudolabels are inherently noisier than expert supervision and can lead to cascading errors. In contrast, rank induction avoids this risk by grounding supervision in verified expert annotations, leading to more stable and accurate slide-level predictions, especially in low-data or noisy-label conditions.

Comparison of Fully Supervised Learning and Rank Induction

In scenarios where exhaustive patch-level annotations are available for all WSIs, fully supervised learning remains a strong baseline. Using the same training settings as our MIL models, a patch-level convolutional neural network with a fine-tuned ResNet-50 backbone and max-pooling aggregation achieved an AUROC of 0.892 on Camelyon16. This result outperformed the AUROC of 0.839 achieved by rank induction, which utilizes a frozen backbone and MIL-based aggregation. This result illustrates the upper bound of fully supervised performance given complete expert annotations and domain-specific fine-tuning (Multimedia Appendix 4).

However, in realistic clinical environments, exhaustive annotation is rarely feasible. When only 10% (9/89) of positive WSIs were annotated in the train dataset, the fully supervised model trained on this subset achieved an AUROC of 0.729. In contrast, rank induction, using the same 10% of patch-level annotations and leveraging slide-level labels from the remaining 90%, achieved a higher AUROC of 0.775. The key difference lies in how each method handles unannotated slides. Fully supervised models are limited to the annotated subset and cannot benefit from weak labels. Rank induction, by integrating both patch- and slide-level supervision through pairwise ranking constraints, makes effective use of all available data, even with sparse annotations.

These findings highlight a clear trade-off: while fully supervised learning performs best under ideal, fully annotated conditions, rank induction offers a more practical and annotation-efficient alternative. Its ability to generalize from partial supervision makes it especially suitable for real-world digital pathology applications, where annotations are incomplete or expensive.

Limitations

Our study has several limitations. First, although we validated our method on 3 datasets, including 1 in-house dataset (SMF-stomach), the diversity of organs and pathologies remains limited. Specifically, the SMF-stomach dataset was collected from a single institution, which may introduce biases related to specific staining protocols or scanner characteristics. To fully assess cross-domain robustness, future work should include

XSL•FO
RenderX

evaluation on larger, multicenter cohorts with diverse scanners. Nonetheless, the inclusion of Camelyon16 partially addresses this concern, as it contains slides from multiple medical institutions, providing evidence of generalization.

Conclusions

In summary, rank induction provides a practical and effective solution for improving slide-level classification in digital pathology by bridging weak and strong supervision via pairwise ranking constraints. By leveraging only sparse expert annotations, it achieves a strong balance between annotation efficiency and model performance, outperforming traditional MIL and pseudolabeling approaches, particularly in low-data or noisy-label settings.

Although fully supervised learning may yield higher accuracy when exhaustive annotations are available, rank induction offers a more scalable and annotation-efficient alternative for real-world use. Its robustness with limited annotations and its ability to highlight diagnostically relevant regions underscore the potential of rank-based supervision to improve both model interpretability and generalizability across diverse clinical environments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of a partially annotated whole slide image from the Seegene Medical Foundation-stomach dataset.
[PNG File , 511 KB - medinform_v14i1e84417_app1.png]

Multimedia Appendix 2
Comparison of slide-level classification accuracy across different multiple instance learning methods on 3 datasets.
[DOCX File , 30 KB - medinform_v14i1e84417_app2.docx]

Multimedia Appendix 3
Impact of integrating rank induction into existing multiple instance learning architectures (dual-stream multiple instance learning and hierarchical image pyramid transformer).
[DOCX File , 31 KB - medinform_v14i1e84417_app3.docx]

Multimedia Appendix 4
Performance comparison between fully supervised learning and rank induction under varying proportions of annotated positive whole slide images.
[PNG File , 41 KB - medinform_v14i1e84417_app4.png]

References

1. Song AH, Jaume G, Williamson DF, Lu MY, Vaidya A, Miller TR, et al. Artificial intelligence for digital and computational pathology. Nat Rev Bioeng 2023 Oct 02;1(12):930-949. [doi: 10.1038/s44222-023-00096-8]
2. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Cancer Genome Atlas Research Network, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep 2018 Apr 03;23(1):181-93.e7 [FREE Full text] [doi: 10.1016/j.celrep.2018.03.086] [Medline: 29617659]
3. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med 2021 May 14;27(5):775-784. [doi: 10.1038/s41591-021-01343-4] [Medline: 33990804]
4. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019 Aug 15;25(8):1301-1309 [FREE Full text] [doi: 10.1038/s41591-019-0508-1] [Medline: 31308507]
5. Zhang H, Meng Y, Zhao Y, Qiao Y, Yang X, Coupland SE, et al. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 Presented at: CVPR 2022; June 18-24, 2022; New Orleans, LA URL: https://ieeexplore.ieee.org/document/9880112?denied= [doi: 10.1109/cvpr52688.2022.01824]
6. Zhou Y, Lu Y. Multiple instance learning with critical instance for whole slide image classification. In: Proceedings of the IEEE 20th International Symposium on Biomedical Imaging. 2023 Presented at: ISBI 2023; April 18-21, 2023; Cartagena, Colombia URL: https://ieeexplore.ieee.org/abstract/document/10230595/ [doi: 10.1109/isbi53787.2023.10230595]
7. Zhou Y, Lu Y. Deep hierarchical multiple instance learning for whole slide image classification. In: Proceedings of the IEEE 19th International Symposium on Biomedical Imaging. 2022 Presented at: ISBI 2022; March 28-31, 2022; Kolkata, India URL: https://ieeexplore.ieee.org/abstract/document/9761678/ [doi: 10.1109/isbi52829.2022.9761678]

XSL•FO
RenderX

8. Qu L, Li S, Luo X, Liu S, Guo Q, Wang M, et al. Weakly semi-supervised whole slide image classification by two-level cross consistency supervision. ArXiv. Preprint posted online on April 16, 2025 2026 [FREE Full text] [doi: 10.48550/arXiv.2504.12132]
9. Ouyang M, Fu Y, Yan R, Shi S, Ling X, Zhu L, et al. MergeUp-augmented semi-weakly supervised learning for WSI classification. ArXiv. Preprint posted online on August 23, 2024 2026 [FREE Full text] [doi: 10.48550/arXiv.2408.12825]
10. Lu MY, Chen RJ, Wang J, Dillon D, Mahmood F. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. ArXiv. Preprint posted online on October 23, 2019 2026 [FREE Full text] [doi: 10.48550/arXiv.1910.10825]
11. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. ArXiv. Preprint posted online on November 17, 2020 2026 [FREE Full text] [doi: 10.48550/arXiv.2011.08939]
12. Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. ArXiv. Preprint posted online on June 6, 2022 2026 [FREE Full text] [doi: 10.1109/cvpr52688.2022.01567]
13. Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021 Jun 01;5(6):555-570. [doi: 10.1038/s41551-020-00682-w] [Medline: 33649564]
14. Zhou Y, Che S, Lu F, Liu S, Yan Z, Wei J, et al. Iterative multiple instance learning for weakly annotated whole slide image classification. Phys Med Biol 2023 Jul 19;68(15). [doi: 10.1088/1361-6560/acde3f] [Medline: 37311470]
15. Zhou Y, Lu Y. Multiple instance learning with task-specific multi-level features for weakly annotated histopathological image classification. In: Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. 2022 Presented at: ICASSP 2022; May 23-27, 2022; Singapore, Singapore URL: https://ieeexplore.ieee.org/abstract/document/9747121/ [doi: 10.1109/icassp43922.2022.9747121]
16. Myronenko A, Xu Z, Yang D, Roth HR, Xu D. Accounting for Dependencies in Deep Learning Based Multiple Instance Learning for Whole Slide Imaging. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, et al, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham, Switzerland: Springer International Publishing; 2021:329-338.
17. Lerousseau M, Vakalopoulou M, Classe M, Adam J, Battistella E, Carré A, et al. Weakly supervised multiple instance learning histopathological tumor segmentation. In: Martel AL, Abolmaesum P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham, Switzerland: Springer International Publishing; 2020:470-479.
18. Qu L, Luo X, Liu S, Wang M, Song Z. DGMIL: distribution guided multiple instance learning for whole slide image classification. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Cham, Switzerland: Springer; 2022:24-34.
19. Koga R, Yokota T, Arihiro K, Hontani H. Attention induction based on pathologist annotations for improving whole slide pathology image classifier. J Pathol Inform 2025 Jan;16:100413. [doi: 10.1016/j.jpi.2024.100413] [Medline: 39845979]
20. Burges C. From RankNet to LambdaRank to LambdaMART: an overview. Microsoft Research. 2010. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/MSR-TR-2010-82.pdf [accessed 2026-01-11]
21. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, the CAMELYON16 Consortium, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017 Dec 12;318(22):2199-2210. [doi: 10.1001/jama.2017.14585] [Medline: 29234806]
22. Da Q, Huang X, Li Z, Zuo Y, Zhang C, Liu J, et al. DigestPath: a benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Med Image Anal 2022 Aug;80:102485 [FREE Full text] [doi: 10.1016/j.media.2022.102485] [Medline: 35679692]
23. Kim HH, Ko YS, Jeong WC, Yun S, Kim K. Evaluating the robustness of slide-level AI predictions on out-of-focus whole slide images: a retrospective observational study. J Pathol Inform 2025 Nov;19:100518 [FREE Full text] [doi: 10.1016/j.jpi.2025.100518] [Medline: 41146682]
24. Ko YS, Choi YM, Kim M, Park Y, Ashraf M, Quiñones Robles WR, et al. Improving quality control in the routine practice for histopathological interpretation of gastrointestinal endoscopic biopsies using artificial intelligence. PLoS One 2022;17(12):e0278542 [FREE Full text] [doi: 10.1371/journal.pone.0278542] [Medline: 36520777]
25. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. Int J Comput Vis 2017 Dec 23;126:1084-1102. [doi: 10.1007/s11263-017-1059-x]
26. JMIR: Official repository for "Ranking-Aware Multiple Instance Learning for Histopathology Slide Classification: Development and Validation Study". GitHub. URL: https://github.com/SMF-AI/rank_induction_JMIR [accessed 2026-01-20]

Abbreviations

AB-MIL: attention-based multiple instance learning
AUPRC: area under the precision-recall curve

XSL•FO
RenderX

AUROC: area under the receiver operating characteristic curve
CLAM: clustering-constrained attention multiple instance learning
CLAM-MB: clustering-constrained attention multiple instance learning, multibranch
CLAM-SB: clustering-constrained attention multiple instance learning, single-branch
DS-MIL: dual-stream multiple instance learning
HIPT: hierarchical image pyramid transformer
IoU: intersection over union
IRB: institutional review board
MIL: multiple instance learning
SMF-stomach: Seegene Medical Foundation-stomach
WSI: whole slide image

Edited by A Benis; submitted 19.Sep.2025; peer-reviewed by Y Zhou, Y Lu; comments to author 23.Oct.2025; revised version received 14.Dec.2025; accepted 01.Jan.2026; published 04.Feb.2026.

Please cite as:
Kim HH, Hwang G, Jeong WC, Ko YS
Ranking-Aware Multiple Instance Learning for Histopathology Slide Classification: Development and Validation Study
JMIR Med Inform 2026;14:e84417
URL: https://medinform.jmir.org/2026/1/e84417
doi:10.2196/84417
PMID:

©Ho Heon Kim, Gisu Hwang, Won Chan Jeong, Young Sin Ko. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Early Prediction of Delirium in Postcardiac Surgery Patients: Machine Learning Model Development and External Validation

Huixiu Hu[1,2,3*], MMed; Yuxiang Wang[4*], BSc; Houfeng Li[5*], MMed; Qinglai Zang[6,7*], MSc; Jing Huang[8], MMed; Ying Zhang[3], MMed; Jinjing Wu[4], MSc; Long Liu[4], MSc; Zhen Xing[3*], MMed; Yaohua Yu[1,2*], BMed

[1]The Graduate School of Fujian Medical University, Fuzhou, Fujian, China

[2]Department of Anesthesiology, The First Hospital of Putian City, Putian, China

[3]Department of Anesthesiology, The First Affiliated Hospital of Hebei North University, Zhangjiakou, Hebei, China

[4]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, Shanghai, China

[5]Department of Anesthesiology, Changhai Hospital, the Second Military Medical University, Shanghai, China

[6]School of Anesthesiology, Second Military Medical University, Shanghai, China

[7]Information Center, The Second Affiliated Hospital of Naval Medical University, Shanghai, China

[8]Graduate School, Wannan Medical College, Wuhu, Anhui, China

[*]these authors contributed equally

Corresponding Author:
Yaohua Yu, BMed
The Graduate School of Fujian Medical University
No.1 Xuefu North Road, Shangjie Town, Minhou County
Fuzhou, Fujian, 351022
China
Phone: 86 139 5956 1579
Email: yyh.8@163.com

Abstract

Background: Delirium is a frequent postoperative complication among patients who have undergone cardiac surgery and is associated with prolonged hospitalization, cognitive decline, and increased mortality. Early prediction of delirium is therefore critical for initiating timely interventions.

Objective: This study proposes the development and validation of a machine learning–based model to predict postoperative delirium in patients undergoing cardiac surgery during intensive care unit (ICU) care, facilitating the early detection of individuals at high risk of delirium and supporting clinicians in the deployment of targeted preventive strategies.

Methods: This study extracted data on postoperative cardiac surgery patients who remained in the ICU for more than 24 hours from the Medical Information Mart for Intensive Care IV version 2.0 (MIMIC-IV 2.0) database and the eICU Collaborative Research Database (eICU-CRD). The MIMIC-IV 2.0 cohort was randomly divided into a training set and an internal validation set in a 7:3 ratio, whereas the eICU-CRD functioned as an independent validation cohort. We used data from the first 24 hours of ICU monitoring to model the likelihood of delirium over the entire ICU admission period. Delirium was identified by a positive Confusion Assessment Method for the Intensive Care Unit evaluation (ie, score ≥4). We built predictive models by using logistic regression, support vector classifier, extreme gradient boosting (XGB), and random forest classifiers. Their performance was assessed via the area under the receiver operating characteristic curve, accuracy, sensitivity, positive predictive value, negative predictive value, and F_1-score.

Results: The analysis involved 2124 patients from the MIMIC-IV 2.0 database and 2406 from the eICU-CRD. A set of 57variables was selected to construct the predictive models. Among the various machine learning models tested, the XGB model demonstrated the best performance for delirium prediction during internal validation. As for external validation, the model achieved an area under the receiver operating characteristic curve of 0.75, indicating strong discriminatory ability. The most important predictive features identified by the model included hospital length of stay, minimum Glasgow Coma Scale score, mean blood pressure, Sequential Organ Failure Assessment score, weight, urine output, heart rate, and age.

Conclusions: The XGB model with strong predictive capability for ICU delirium after cardiac surgery was developed and externally validated. This model offers essential technical support for building real-time delirium alert systems and enables ongoing risk stratification and evidence-based decision-making within the ICU environment.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e73283) doi:10.2196/73283

KEYWORDS

prediction model; machine learning; cardiac surgery; intensive care unit; delirium; Medical Information Mart for Intensive Care IV version 2.0 database; MIMIC-IV 2.0 database; eICU Collaborative Research Database; eICU-CRD

Introduction

Delirium is an acute neuropsychiatric syndrome commonly associated with encephalopathy, acute cerebral dysfunction, and states of confusion, particularly following surgical procedures [1-4]. In patients undergoing cardiac surgery, the incidence of postoperative delirium has been reported to range from 10% to 40% [5-7]. This condition is linked to a variety of adverse outcomes, including heightened pain perception, depression, cognitive impairment, and increased mortality [8,9]. Currently, the assessment of patients' arousal can be conducted using standardized tools such as the Richmond Agitation-Sedation Scale and the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) to identify different types of delirium with distinct characteristics [10-12]. Despite the existence of standardized instruments, the diagnosis of delirium frequently depends on the patient's subjective assessment of their condition. If the occurrence of delirium in patients could be predicted within a short period, it will substantially reduce the aforementioned risks.

Machine learning (ML), a branch of artificial intelligence, has driven notable progress across numerous domains of health care [13,14]. One such area where ML has shown its potential is in the postoperative surveillance with cardiac surgery, offering more information to predict delirium [15,16]. It has the potential to improve patient health care outcomes [17,18]. Compared to traditional data analysis techniques, ML models can provide more intricate predictions and perform real-time monitoring using objective data from all patients [19-21]. Furthermore, recent research has also used ML to forecast the near-term mortality rates of patients after cardiac operations. For example, Nistal-Nuño [22] constructed an extreme gradient boosting (XGB)–based predictive model to estimate 24-hour postoperative mortality following cardiac surgery. The outcome demonstrated that XGB attained an area under the receiver operating characteristic curve (AUC) of 87.5%, signifying the model's exceptional performance in forecasting intensive care unit (ICU) mortality, notably surpassing other models. Zhang et al [23] assessed an ML model, comparing it to existing severity-of-illness systems to develop a real-time tool for predicting death. However, no correlation-predictive models using ML have been developed for patients who experience delirium after cardiac surgery.

In this study, we created and verified 4 models using ML techniques to anticipate the occurrence of delirium and facilitate identification. Furthermore, we enhanced the interpretability of the results by prioritizing the independent variables according to their predictive significance.

Methods

Ethical Considerations

This study used data from 2 publicly available critical care databases: Medical Information Mart for Intensive Care IV version 2.0 (MIMIC-IV 2.0) database and eICU Collaborative Research Database (eICU-CRD). Both databases were approved by the institutional review boards of the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology. As all data were fully deidentified before release, the requirement for individual informed consent was waived in accordance with the Declaration of Helsinki and applicable regulations. All team members underwent certified training in "Data or Specimens Only Research" to comply with ethical regulations governing dataset access.

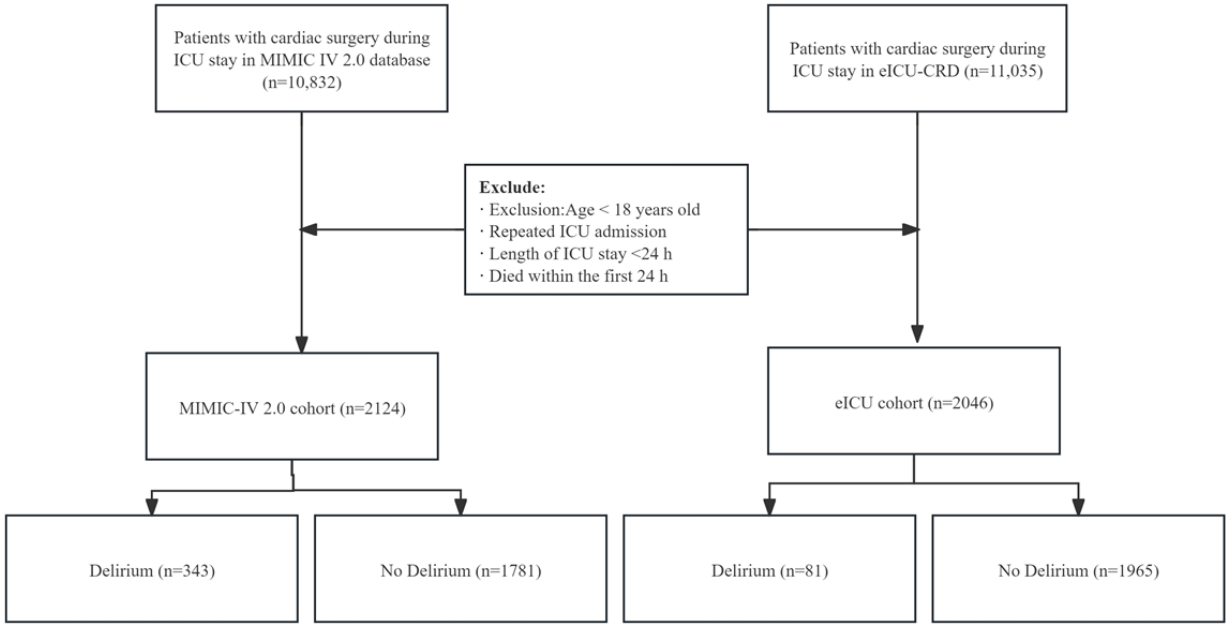
Study Population

This study used 2 publicly available critical care databases to develop and validate a predictive model. The training dataset was sourced from MIMIC-IV 2.0, which includes 76,943 ICU admissions recorded at Beth Israel Deaconess Medical Center (Boston, Massachusetts, United States) between 2008 and 2019 [24]. For external validation, we used the eICU-CRD, which contains deidentified data for over 200,000 patients admitted to 208 US hospitals between 2014 and 2015 [25]. Both databases contain structured clinical data, including demographics, vital signs, laboratory test results, procedures, medications, and outcomes. We included adult patients (aged ≥18 years) who underwent major cardiovascular surgeries, such as coronary artery bypass grafting, heart valve repair or replacement, combined procedures, or other surgeries involving cardiopulmonary bypass.

We applied consistent inclusion and exclusion criteria across both datasets to ensure cohort comparability and data quality. Inclusion criteria required patients to meet the age threshold and have documented cardiovascular surgery. We excluded patients who had ICU stays shorter than 24 hours (to ensure sufficient observation data), missing essential demographic or outcome variables, or delirium recorded within the first 24 hours of ICU admission (to preserve the prediction time window). Although the 2 datasets differ in time range and hospital coverage, we aligned the study population by applying uniform definitions for surgical type and using a standardized 24-hour observation window. Figure 1 summarizes the baseline characteristics of the 2 cohorts.

XSL•FO
RenderX

Figure 1. Schematic representation of the study design. eICU-CRD: eICU Collaborative Research Database; ICU: intensive care unit; MIMIC-IV 2.0: Medical Information Mart for Intensive Care IV version 2.0.



Delirium Assessment

Delirium served as the primary outcome, identified on the basis of a positive CAM-ICU assessment (score ≥4) and consistent diagnostic coding. The observation window was the first 24 hours after ICU admission, during which patient data were collected for modeling. If at least one positive delirium assessment occurred at the time of prediction, the patient was considered delirious.

Data Extraction and Processing

The clinical data were retrieved and extracted using the structured query language, with pgAdmin 4 serving as the administrative platform for PostgreSQL. The prediction model included just clinical and laboratory characteristics that were accessible on the initial day of admission to the ICU, with patients being recognized by their unique ID numbers. The predictors consisted of the following variables— (1) demographics: age, gender, ethnicity and weight; (2) vital signs: heart rate, mean blood pressure, respiratory rate, systolic blood pressure, and temperature; (3) laboratory analysis: hemoglobin level, platelet count, white blood cell count, lactate and urine output; (4) severity scoring: Glasgow Coma Scale (GCS) and Sequential Organ Failure Assessment (SOFA) scores; (5) comorbidities: myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease,diabetes (with control, without control), paraplegia, renal disease, malignant cancer, severe liver disease, and AIDS; (6) medications: opioids, barbiturates, benzodiazepines, acetaminophen, antipsychotics, anticoagulant, antihistamines, diuretics, anesthesia, and anticholinergics; (7) treatment measures: emergency admission, first care unit, last care unit, renal replacement therapy, invasive ventilation, length of ICU stay.

Missing Data Management

Variables exhibiting a missing value rate exceeding 10% were omitted to prevent potential bias. Variables with less than 10% missing values were subjected to multivariable imputations [26].

Data Balance

The dataset showed a marked imbalance, with notably fewer positive delirium cases compared with negative ones, which caused the model to lean toward predicting the majority (negative) class. To mitigate this issue, we used the Synthetic Minority Oversampling Technique to artificially augment the number of positive samples, thereby achieving a more balanced class distribution and enhancing the model's ability to generalize.

Feature Selection

The feature selection process involved using the recursive elimination of features approach of the random forest [27]. This method was used to identify the most optimal combinations of predictive variables. By examining the weight of features and their correlation and after stratified 10-fold cross-validation, final features were selected based on importance scores, correlation analysis, and cross-validation results. This process reduced dimensionality while preserving predictive power.

Model Development and Hyperparameter Tuning

The MIMIC-IV 2.0 dataset (N=2124) was randomly split into a training set (1487/2124, 70%) and a testing set (637/2124, 30%), whereas the eICU-CRD dataset was used as an external validation cohort. We developed prediction models using 4 widely adopted ML algorithms. Logistic regression was implemented for binary classification using maximum likelihood estimation [28]. Random forest, an ensemble learning method, combined multiple decision trees through majority voting to

XSL•FO
RenderX

enhance predictive performance [29]. XGB used a gradient boosting framework to iteratively build strong learners from weak ones [30]. Support vector classifier aimed to find the optimal hyperplane in a high-dimensional space for classification [31]. Bayesian optimization was used to identify optimal hyperparameters for each model, improving training efficiency and performance [32].

Model Performance Evaluation

To comprehensively assess the discriminatory performance of the prediction models, we used the receiver operating characteristic (ROC) as the primary evaluation metric. Additional metrics included accuracy, positive predictive value, negative predictive value, and sensitivity. We also reported the F_1-score, the harmonic mean of precision and sensitivity, to reflect the balance between these two metrics. Together, these indicators were used to evaluate the clinical applicability of each model in stratifying the risk of postoperative delirium among cardiac surgery patients. Shapley Additive Explanations (SHAP) was used to investigate the interpretability of the final predictive model.

Statistical Analysis

All statistical analyses were conducted using Stata 17.0 and SPSS (version 27.0; IBM Corp). Frequencies and percentages were used to summarize categorical variables, with comparisons made via the chi-square test. The distribution of continuous variables was assessed using the Shapiro-Wilk test. Normally distributed data were reported as mean (SD) and compared using independent 2-tailed *t* tests. Skewed data were summarized as median and IQR and analyzed using the Mann-Whitney U or Kruskal-Wallis test, based on group composition. Statistical significance was determined using 2-sided tests, with a threshold of $P<.05$.

Results

Baseline Characteristics

A total of 2124 patients from the MIMIC-IV 2.0 database were included in the final analysis. Among them, 16.1% (343/2124) of cardiac surgery patients were diagnosed with delirium during their hospital stay, occurring after the first day of ICU admission. In the external validation cohort, an analysis was conducted on 2046 cases obtained from the eICU-CRD, of whom 3.81% (81/2046) developed delirium during the same postoperative period, also defined as after the first ICU day. In these patients with delirium, maximum heart rate, minimum mean blood pressure, minimum hemoglobin level, minimum platelet count, maximum white blood cell count, urine output, minimum GCS score, SOFA, and length of ICU stay showed a notable disparity between the two different groups. Tables 1-3 present a concise summary of the comparison of fundamental traits, vital signs, and laboratory analysis between patients with and without delirium. According to the data, patients with delirium are predominantly male, typically older, and have longer hospital stays with higher severity scores upon admission. Additionally, factors such as weight loss, decreased urine output, and decreased mean arterial pressure may all exacerbate the likelihood of delirium in patients.

XSL•FO
RenderX

Table 1. Characteristics of patients and controls from the development dataset for the first 24-hour model cohort: demographics and vital signs.

Patient characteristics	MIMIC-IV 2.0[a] cohort			eICU-CRD[b] cohort		
	No delirium (n=1781)	Delirium (n=343)	*P* value	No delirium (n=1965)	Delirium (n=81)	*P* value
Gender, n (%)			.10			.09
Male	1144.0 (64.2)	195.0 (56.9)		937.0 (49.5)	46 (57)	
Female	637.0 (35.8)	148.0 (43.1)		992.0 50.5)	35 (43)	
Race, n (%)			.80			.62
Asian	47.0 (2.6)	6.0 (1.7)		29.0 (1.5)	6.0 (1.4)	
Black	80.0 (4.5)	19.0 (5.5)		475.0 (24.2)	18.0 (22.5)	
Hispanic	126.0 (7.1)	42.0 (12.2)		1227.0 (62.4)	53.0 (65)	
White	1391.0 (78.1)	219.0 (63.8)		132.0 (6.7)	12.0 (14.8)	
Unknown	29.0 (1.6)	47.0 (13.7)		96.0 (4.9)	10.0 (2.3)	
Other	108.0 (6.1)	10.0 (2.)		6.0 (0.3)	2.0 (2.5)	
Age (y), median (IQR)	70.0 (61.0-79.0)	75.0 (64.0-81.0)	<.001	68.0 (57.0-77.0)	71.0 (61.0-79.8)	<.001
Weight (kg), median (IQR)	83.8 (70.8-96.7)	80.0 (60.0-95.4)	.02	85.2 (70.0-102.0)	84.5 (67.9-103.4)	.31
Vital signs, median (IQR)						
Heart rate min (bpm)	67.0 (60.0-74.0)	68.0 （59.0-77.0）	.12	69.0（59.0-80.0）	70.0（60.0-80.0）	.29
Heart rate max (bpm)	94.0 (86.0-106.0)	97.0 （88.0-110.0）	<.001	101.0（88.0-116.0）	107.0 (92.2-124.5)	<.001
Heart rate mean (bpm)	80.2 (73.3-87.5)	82.7 (75.6-89.3)	<.001	83.6 (75.7-94.4)	86.3 (75.7-97.9)	.007
Mean blood pressure min (mm Hg)	57.0 (53.0-62.0)	55.0 (49.5-60.0)	<.001	62.0 (54.0-72.0)	60.0 (52.0-68.0)	<.001
Mean blood pressure max (mm Hg)	97.0 (89.0-107.0)	95.0 (88.0-108.0)	.33	103.0 (99.0-118.0）	104.0 (91.0-122.0)	.31
Mean blood pressure mean (mm Hg)	74.6 (70.3-79.4)	73.3 (69.3-78.8)	.02	80.0 (76.6-90.4)	78.2 (70.3-88.2)	.02
Respiratory rate min (bpm)	12.0 (10.0-14.0)	12.0 (9.0-14.0)	.75	13.0 (11.0-16.0)	13.0 (11.0-16.0)	.63
Respiratory rate max (bpm)	26.0 (23.0-29.0)	26.0 (23.00-30.0)	.69	27.0 (24.0-32.00)	28.0 (24.0-33.00)	.42
Respiratory rate mean (bpm)	17.8 (16.2-19.6)	18.1 (16.2-20.2)	.23	19.2 (17.1-21.9)	19.2 (16.9-22.9)	.76
Systolic blood pressure min (mm Hg)	93.0 (91.0-95.0)	93.0 (91.00-96.0)	.85	97.0 (95.6-98.4)	97.1 (95.34-98.63)	.001
Systolic blood pressure max (mm Hg)	91.0 (93.0-95.0)	93.0 (91.0-96.0)	.003	92.0 (89.0-94.0)	91.0 (86.3-94.0)	.15
Systolic blood pressure mean (mm Hg)	97.7 (96.5-98.7)	97.9 (96.7-99.0)	.14	100.0 (99.0-100.0)	100.0 (100.0-100.0)	.92
Temperature mean (˚C)	36.7 (36.5-36.9)	36.72 (36.5-37.0)	.48	36.7 (36.6-37.0)	36.8 (36.9-36.5)	.007

[a]MIMIC-IV Medical Information Mart for Intensive Care IV version 2.0.

[b]eICU-CRD: eICU Collaborative Research Database.

XSL•FO
RenderX

Table 2. Characteristics of patients and controls from the development dataset for the first 24-hour model cohort: laboratory test results and comorbidities.

Patient characteristics	MIMIC-IV 2.0[a] cohort			eICU-CRD[b] cohort		
	No delirium (n=1781)	Delirium (n=343)	*P* value	No delirium (n=1965)	Delirium (n=81)	*P* value
Laboratory results, median (IQR)						
Hemoglobin min (g/100 mL)	8.4 (7.5-9.6)	9.0 (7.9-10.3)	<.001	8.4 (7.3-9.4)	9.0 (7.9-10.3)	<.001
Hemoglobin max (g/100 mL)	11.5 (10.0-12.9)	11.4 (10.4-14.2)	.64	11.4 (10.4-12.9)	11.4 (10.4-20.1)	.62
Platelet min (10^9/L)	147.0 (111.0-198.0)	130.0 (95.0-179.0)	<.001	145.0 (109.0-197.0)	128.0 (96.0-178.0)	<.001
Platelet count max (10^9/L)	125.0 (151.0-245.0)	188.0 (149.0-242.0)	.57	128.0 (97.0-235.0)	186.0 (153.0-245.0)	.55
White blood cell min (10^9/L)	8.8 (6.3-11.7)	9.5 (6.9-12.3)	.12	9.2 (6.3-12.0)	9.6 (7.0-12.4)	.14
White blood cell max (10^9/L)	13.1 (10.0-17.3)	14.8 (11.3-19.6)	<.001	14.0 (10.0-17.3)	14.7 (11.4-20.0)	<.001
Lactate min (mmol/L)	1.2（0.9-1.5）	1.2（0.9-1.6）	.21	1.4 (1.0-2.1)	1.3 (0.9-1.9)	.18
Urine output (mL)	1832.0 (1290.0-2617.0)	1575.0 (1002-2321.0)	<.001	1496.0 (796.3-2600.0)	1229.5 (600.0-2056.0)	<.001
Comorbidity, n (%)						
Myocardial infarction	30 (1.7)	3 (0.9)	.37	181 (9.2)	30 (6.8)	.11
Congestive heart failure	724 (40.7)	184 (53.6)	<.001	1251 (63.7)	51 (63)	.78
Peripheral vascular disease	281 (15.8)	84 (24.5)	<.001	14 (0.7)	1.0 (1.1)	.36
Cerebrovascular disease	186 (10.4)	65 (19.0)	<.001	90 (4.6)	6 (7.3)	.02
Chronic pulmonary disease	498 (28)	125 (36.4)	.002	303 (15.4)	11 (13.6)	.35
Renal disease	387 (21.7)	96 (28.0)	.01	321 (16.3)	13 (15.7)	.74
Diabetes with control	1538 (7.7)	30 (8.7)	.53	155 (7.9)	7 (8.5)	.58
Diabetes without control	495 (27.8)	97 (28.3)	.85	503 (30.2)	24 (29.5)	.84
Rheumatic disease	78 (4.4)	23 (6.7)	.06	82.0 (4.6)	5 (6.5)	.06
Peptic ulcer disease	13 (0.7)	7 (2)	.02	9.0 (0.5)	1 (1.5)	.02
Severe liver disease	16 (0.9)	7 (2)	.06	18.0 (1.0)	2 (2.3)	.07
Dementia	11 (0.6)	3 (0.9)	.59	13.0 (0.7)	1 (0.9)	.61
Paraplegia	15 (0.8)	12 (3.5)	<.001	16.0 (0.9)	3 (3.7)	<.001
Malignant cancer	103 (5.8)	16 (4.7)	.41	181.0 (9.2)	5 (6.8)	.11
AIDS	4 (0.2)	5 (3)	.38	1964.00 (99.0)	1 (1.2)	.04

[a]MIMIC-IV Medical Information Mart for Intensive Care IV version 2.0.

[b]eICU-CRD: eICU Collaborative Research Database.

XSL•FO
RenderX

Table 3. Characteristics of patients and controls from the development dataset for the first 24-hour model cohort: score, drug, and treatment measures.

Patient characteristics	MIMIC-IV 2.0[a] cohort			eICU-CRD[b] cohort		
	No delirium (n=1781)	Delirium (n=343)	*P* value	No delirium (n=1965)	Delirium (n=81)	*P* value
Score, median (IQR)						
GCS[c] (min)	14.0 (14.0-15.0)	11.0 (6.0-14.0)	<.001	15.0 (14.0-15.00)	12.0 (8.0-14.0)	<.001
SOFA[d]	4.0 (2.0-7.0)	8.0 (6.0-11.0)	<.001	5.0 (3.0-7.0)	7.0 (5.0-10.0)	<.001
Drug, n (%)						
Acetaminophen	1005 (56.4)	164 (47.8)	.003	1104 (56.2)	41 (50.6)	.054
Anesthesia	38 (2.1)	5 (1.5)	.002	35 (1.9)	1 (1.2)	.02
Anticholinergics	1062 (59.6)	221 (64.4)	.10	387 (19.7)	19 (23.5)	.03
Anticoagulant	519 (29.1)	92 (26.8)	.39	539 (27.4)	36 (44.4)	.41
Antipsychotics	18 (1.0)	11 (3.2)	.001	21 (1.1)	17 (3.9)	<.001
Barbiturates	2.0 (0.1)	343 (100)	.54	3 (0.1)	81 (100)	.54
Benzodiazepines	138 (7.7)	32 (9.3)	.32	142 (8)	35 (9.5)	.34
Diuretics	694 (39)	131 (38.2)	.79	854 (43.5)	29 (35.5)	.008
Opioids	1040 (58.4)	201 (58.6)	.94	981 (49.9)	38 (46.9)	.28
Treatment measures, n (%)						
Emergency admission	1121 (62.9)	231 (67.3)	.12	380 (19.3)	17 (20.9)	.52
First care unit	1724 (96.8)	321 (93.6)	.004	1779 (90.5)	75 (92.6)	.04
Last care unit	42 (2.4)	16 (4.7)	.02	453 (23)	36 (44.4)	<.001
Renal Replacement Therapy	15 (0.8)	11 (3.2)	<.001	719 (36.6)	46 (56.8)	.047
Invasive ventilation	1179 (66.2)	298 (86.9)	<.001	1699 (86.5)	77 (95.1)	<.001
Length of ICU[e] stay, median (IQR)	2.4 (2.0-3.4)	5.9（4.0-10.3）	<.001	3.20 (3.0-5.5)	6.7 (5.3-12.5)	<.001

[a]MIMIC-IV Medical Information Mart for Intensive Care IV version 2.0.

[b]eICU-CRD: eICU Collaborative Research Database.

[c]GCS: Glasgow Coma Scale.

[d]SOFA: Sequential Organ Failure Assessment.

[e]ICU: intensive care unit.

Model Performance Evaluation

Using 4 ML algorithms, we developed predictive models to assess the risk of postoperative delirium in cardiac surgery patients, leveraging electronic health record data for early identification of high-risk individuals. Figure 2 presents ROC curves of all models, allowing a systematic comparison of their discriminative performance. Among the models, XGB demonstrated the best overall predictive performance, achieving the highest AUC for identifying patients at risk of delirium. The random forest classifiers also exhibited strong performance, although slightly lower than that of the XGB model. Notably, both XGB and random forest classifiers maintained high predictive accuracy, indicating good model generalizability. In contrast, support vector classifier and logistic regression models showed substantially lower discriminative power. To further validate the clinical applicability of the XGB model, model performance was assessed using several evaluation metrics,including accuracy, sensitivity, positive predictive value, and negative predictive value (Table 4). Furthermore, the corresponding confusion matrices illustrating these metrics are displayed in Figure 3.

The model was externally validated using the eICU-CRD, a large-scale critical care dataset incorporating records from 208 hospitals, to assess its performance on independent data. The XGB model maintained strong discriminative performance in ROC analysis, with high AUC values confirming its reliability across institutions (Figure 4). The model was further rigorously validated through precision-recall analysis and calibration curves (Figures 5A and B).

XSL•FO
RenderX

Figure 2. Receiver operating characteristic curves of different machine learning algorithms evaluated on the internal validation set. LR: logistic regression; SVC: support vector classifier; RFC: random forest classifier; XGB: extreme gradient boosting.

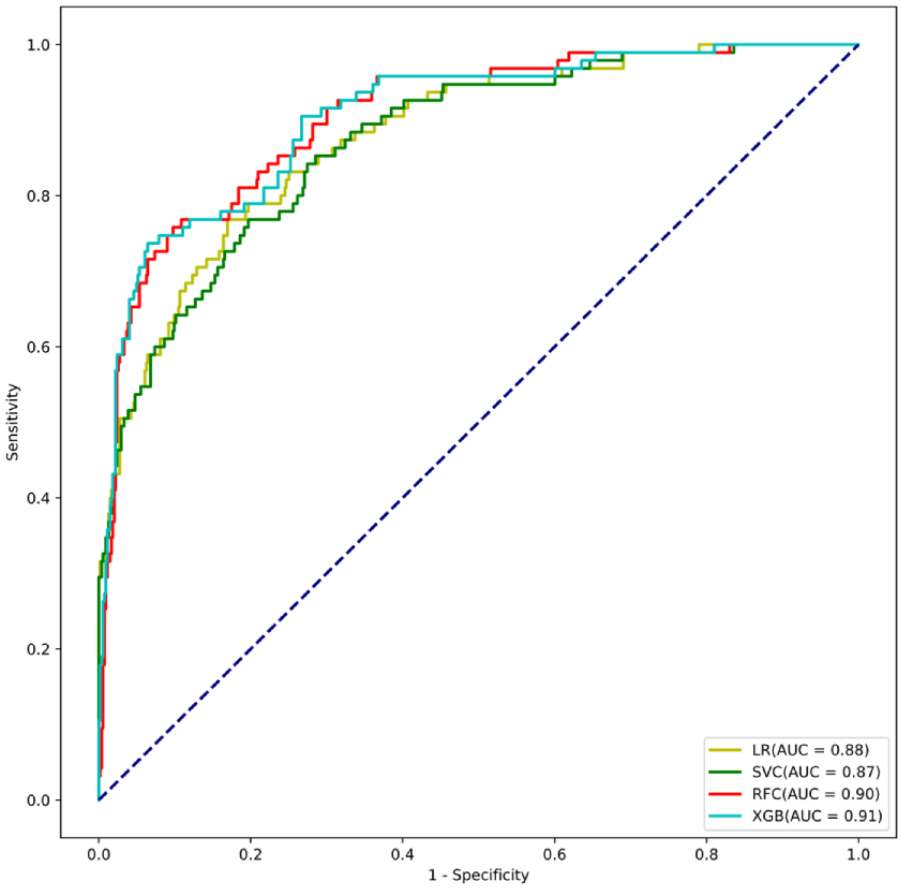


Table 4. Test set evaluation of machine learning model performance.

Model	Accuracy	Sensitivity	AUC[a]	PPV[b]	NPV[c]	F_1-score
LR[d]	0.82	0.73	0.88	0.44	0.95	0.55
XGB[e]	0.83	0.77	0.91	0.47	0.95	0.58
SVC[f]	0.83	0.67	0.87	0.45	0.94	0.54
RFC[g]	0.88	0.75	0.90	0.58	0.95	0.65

[a]AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]LR: logistic regression.

[e]XGB: extreme boosting gradient.

[f]SVC: support vector classifier.

[g]RFC: random forest classifier.

XSL•FO
RenderX

Figure 3. Confusion matrix for binary classification. (A) Logistic regression (LR), (B) extreme gradient boosting (XGB), (C) support vector classifier (SVC), (D) random forest classifier (RFC).

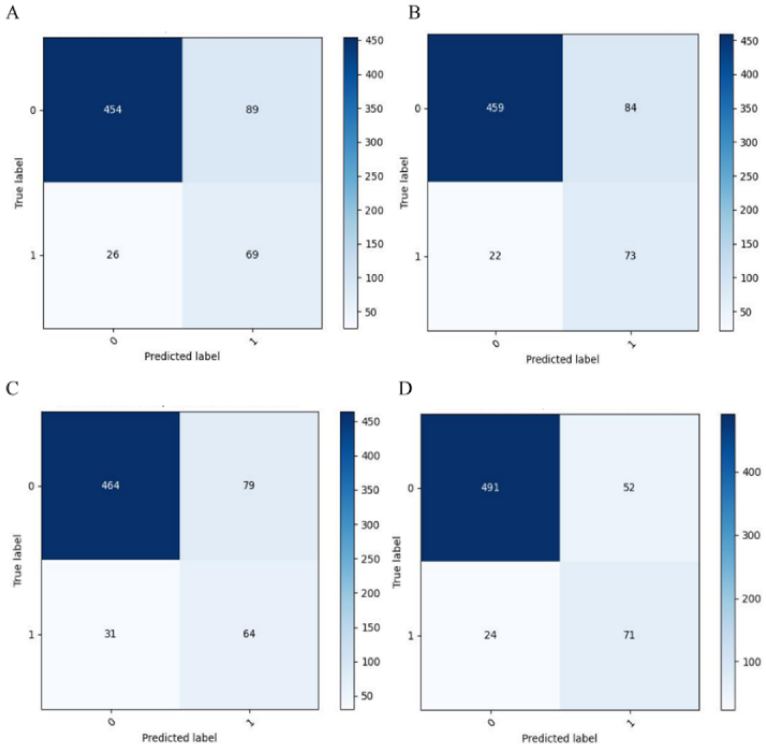
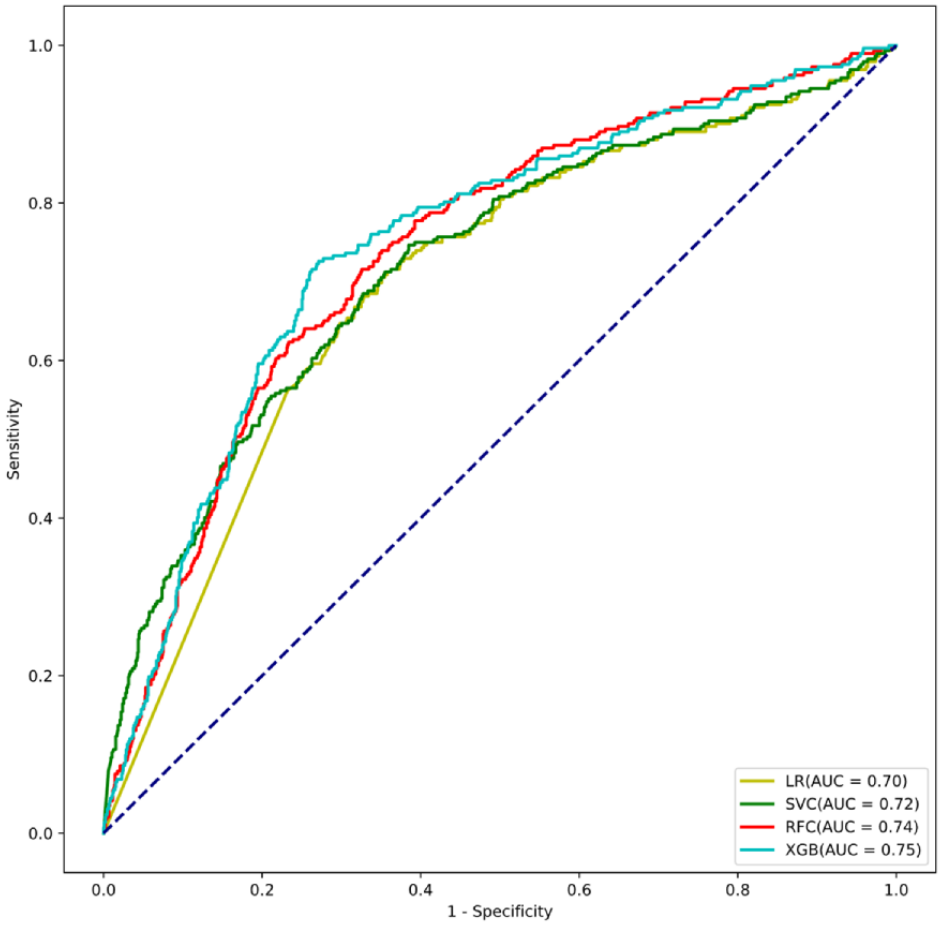
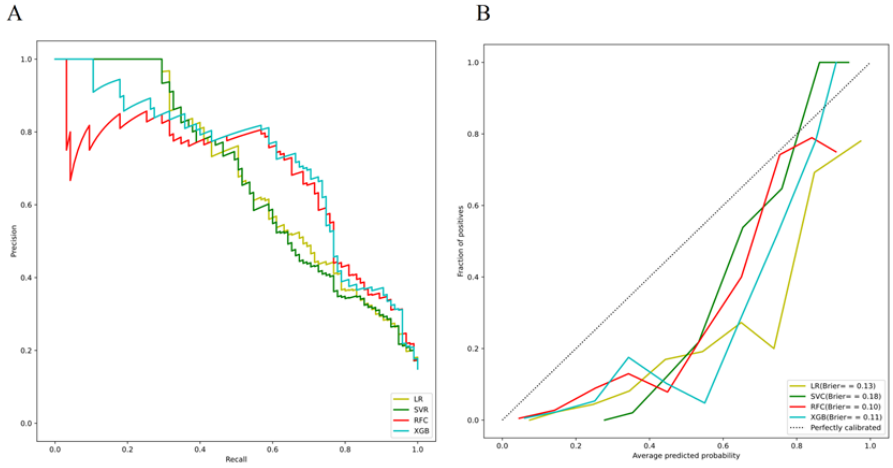


Figure 4. Receiver operating characteristic curves of various machine learning models evaluated on the external validation cohort. LR: logistic regression; RFC: random forest classifier; SVC: support vector classifier; XGB: extreme gradient boosting.



XSL•FO
RenderX

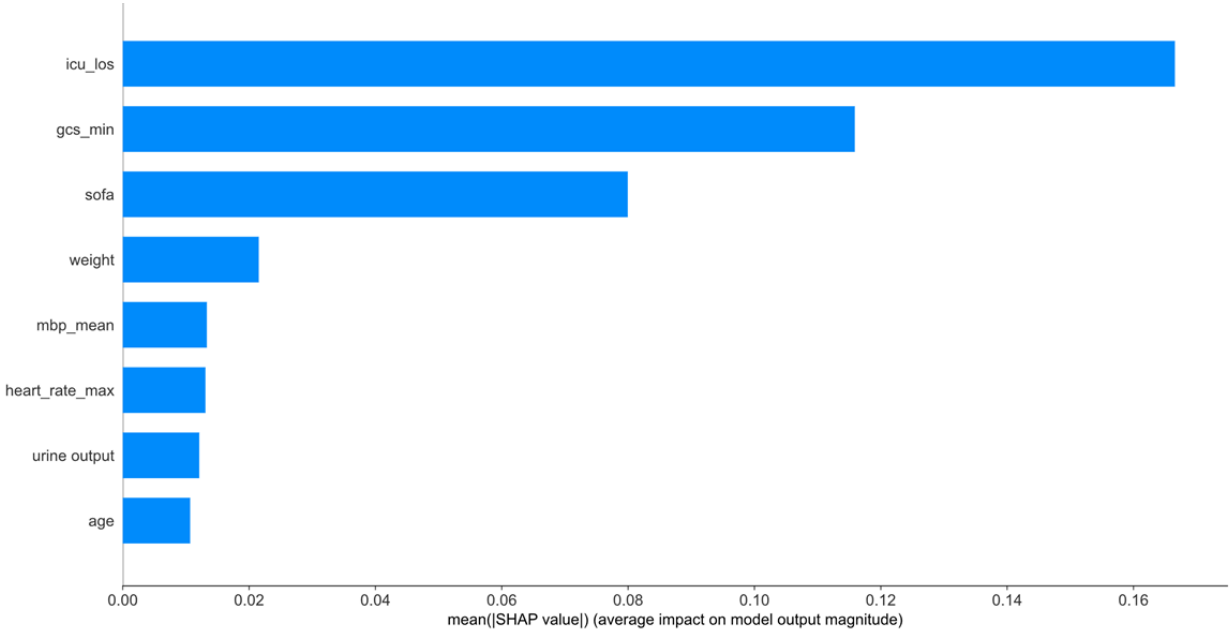
Figure 5. (A) Precision-recall curve. (B) Calibration curves and calculated Brier scores. LR: logistic regression; RFC: random forest classifier; SVC: support vector classifier;XGB: extreme gradient boosting.



Variable Importance

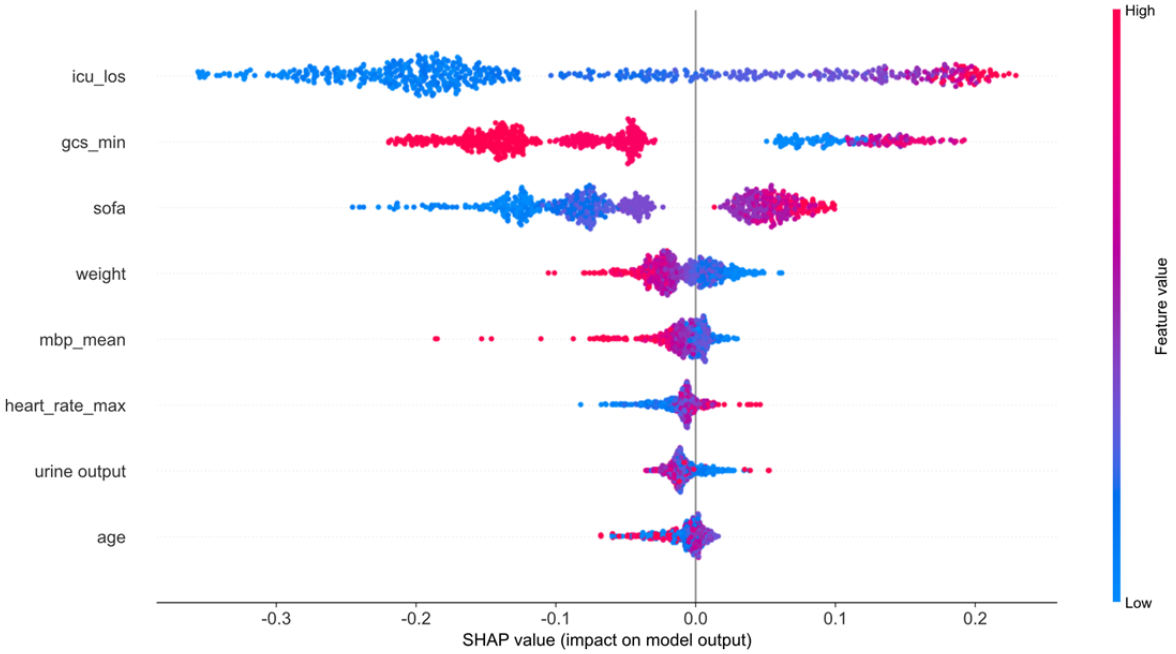
The predictive contribution of each variable to postoperative delirium varied among cardiac surgery ICU patients. The XGB algorithm was used to estimate variable importance and highlight the top predictors influencing model accuracy. The top predictors included length of stay in ICU, lowest GCS score, age, mean blood pressure, SOFA score, weight, heart rate and urine output (Figure 6). To further interpret the influence and directionality of these features on model predictions, we generated SHAP summary plots (Figure 7). These visualizations provide a detailed explanation of how individual features contributed to the model output, offering insights into their relative impact and potential clinical relevance.

Figure 6. Variable contribution rankings estimated using the XGB model. GCS: Glasgow Coma Scale; ICU_los: the length time of ICU; MBP: mean blood pressure; SOFA: Sequential Organ Failure Assessment.



XSL•FO
RenderX

Figure 7. Shapley Additive Explanations (SHAP) summary plots illustrating feature contributions in the extreme gradient boosting (XGB) model. GCS: Glasgow Coma Scale; ICU_los: length of stay in the intensive care unit; MBP: mean blood pressure; SOFA: Sequential Organ Failure Assessment.



Discussion

Principal Findings

In this large-scale retrospective study, we developed and validated an ML model to predict postoperative delirium, a common and burdensome complication after cardiac surgery, with reported incidence rates ranging from 18% to 52% [33-35]. The XGB-based model, trained on the MIMIC-IV 2.0 dataset, demonstrated excellent performance in internal validation (AUC=0.91) and maintained strong discriminative power in external validation using the eICU-CRD dataset (AUC=0.75). Key predictors identified by the model included length of stay in ICU, min GCS score, mean blood pressure, SOFA score, weight, urine output, heart rate, and age. As far as we are aware, this study presents one of the earliest ML models that are developed to predict delirium in a diverse population of cardiac surgery patients. Importantly, our model relies on routinely available clinical data from the first 24 hours of ICU admission, enabling early identification of high-risk patients throughout their ICU stay. This early-warning capability offers a critical window for timely intervention, allowing clinicians to implement targeted strategies at the earliest stages of delirium onset. The model's high temporal relevance and wide applicability support its potential use in guiding personalized care plans and improving patient outcomes.

Comparison With Previous Work

Among existing tools for predicting postoperative cognitive complications, the CAM-ICU score remains the most widely used [36-38]. However, its applicability may be limited in certain surgical contexts or patient subgroups. Some individual populations require more tailored analysis, and it is important to note that CAM-ICU may not capture the full spectrum of delirium severity, potentially compromising its construct validity [12,39,40]. Therefore, balancing feasibility and validity in specific clinical settings is essential. Our ML approach addresses these limitations by offering a data-driven, individualized risk assessment that adapts to patient heterogeneity and enables broader clinical applicability.

Furthermore, this study incorporated SHAP to enhance the interpretability of the XGB model. SHAP provides a transparent and traceable explanation framework for individualized risk prediction of delirium, allowing visualization of how each feature influences the model's output in both direction and magnitude [41,42]. This helps clinicians better understand the rationale behind specific predictions. Through SHAP analysis, we identified a set of clinically relevant predictors closely associated with patient deterioration, including length of stay in ICU, lowest GCS score, age, mean blood pressure, SOFA score, weight, urine output, and heart rate.

The XGB model demonstrated superior predictive performance, its clinical value extends beyond accuracy—it lies in its ability to inform real-time decision-making. Future research should focus on integrating these key features into dynamic clinical monitoring and intervention workflows. For instance, early abnormalities in high-impact variables, such as lowest GCS score, mean blood pressure, or urine output could trigger automated alerts within electronic health record systems, enabling real-time risk stratification. Additionally, individualized SHAP-based explanations could support patient-centered clinical strategies, facilitating a closed-loop "early warning–targeted intervention" model. This approach may ultimately enhance early prevention and improve outcomes in postoperative delirium management.

Strengths and Limitations

This study has several notable strengths. First, model development was based on large scale, real-world data from 2

XSL•FO
RenderX

publicly available critical care databases (MIMIC-IV 2.0 and eICU-CRD), which offer extensive clinical information and robust sample sizes, thereby increasing the trustworthiness and general applicability of our findings. Second, as far as we are aware, this is the earliest known study to apply ML techniques for predicting postoperative delirium in a broad cardiac surgery cohort. This tailored approach enables more accurate risk stratification by accounting for individual variability across surgical subtypes. One notable strength of our model lies in its exclusive use of routinely available clinical data obtained within the first day of ICU stay, eliminating the need for additional or specialized testing. The use of early time-series features to quantify dynamic changes in key physiological indicators allows real-time risk estimation with high clinical practicality and timeliness. From a translational perspective, the model demonstrated a sensitivity of 77% in identifying high-risk patients early, providing a valuable window for initiating preventive interventions. These proactive measures have the potential to reduce ICU length of stay, lower per-patient hospitalization costs, and minimize readmission risk, ultimately interrupting the cascade of adverse outcomes often associated with postoperative delirium. Finally, by focusing on the first day of ICU data, the model creates a strategic opportunity for timely clinical action, supporting the implementation of early, targeted strategies aimed at mitigating symptom progression and reducing complication rates.

Our study has several limitations. First, we acknowledge that our findings may be racially biased against White patients and the applicability to other populations may be limited due to the database being derived from Western countries. Second, the use of open public databases to obtain data may introduce missing data bias, which is unavoidable. Third, as this study is retrospective and observational in nature, selection bias is inevitable. Specifically, the collection of retrospective data depends on existing medical records that may vary in completeness and accuracy due to differences in clinical documentation practices and data entry habits. Such inconsistencies can result in missing or inaccurate key clinical information, which in turn affects the quality of data used for model training. Moreover, retrospective studies lack the ability to actively intervene in or control the research process, making it difficult to eliminate the influence of confounding variables. This limitation weakens the strength of causal inferences drawn from the analysis. Additionally, our model has not yet undergone prospective clinical validation, which limits its current applicability in real-world clinical practice. Prospective validation, ideally conducted under well-controlled trial conditions and standardized workflows, is essential for evaluating the predictive performance and robustness of the model in diverse clinical settings. Without this step, there remains a risk that the model's actual performance may deviate from the retrospective findings, particularly when applied to different health care systems or patient populations.

Conclusions

We constructed and assessed an effective predictive model targeting postoperative delirium in patients admitted to the ICU after cardiac surgery. This model leverages routinely available patient information from the initial 24-hour ICU stay to estimate the risk of delirium throughout the hospitalization period. Given its reliance on readily available early-stage variables, the model has the potential to serve as a practical and accessible risk stratification tool for health care professionals and patients when selecting optimal cardiac treatment strategies.

Acknowledgments

The authors gratefully acknowledge the institutional review boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology for providing authorization to use data from the Medical Information Mart for Intensive Care IV version 2.0 database and eICU Collaborative Research Database.

Funding

The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HH contributed to study design, data extraction, methods, data analysis, and manuscript preparation. YW, QZ, JW, and LL contributed to data interpretation and data analysis. HL, JH, and YZ contributed to data extraction and data interpretation. YY and ZX directed the project. YY is the guarantor of this study.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Li T, Li J, Yuan L, Wu J, Jiang C, Daniels J, RAGA Study Investigators. Effect of regional vs general anesthesia on incidence of postoperative delirium in older patients undergoing hip fracture surgery: the RAGA randomized trial. JAMA 2022 Jan 04;327(1):50-58 [FREE Full text] [doi: 10.1001/jama.2021.22647] [Medline: 34928310]
2. Chen M, Zhang L, Shao M, Du J, Xiao Y, Zhang F, et al. E4BP4 coordinates circadian control of cognition in delirium. Adv Sci (Weinh) 2022 Jun 17;9(23):e2200559 [FREE Full text] [doi: 10.1002/advs.202200559] [Medline: 35713240]
3. Wu YC, Tseng PT, Tu YK, Hsu CY, Liang CS, Yeh TC, et al. Association of delirium response and safety of pharmacological interventions for the management and prevention of delirium: a network meta-analysis. JAMA Psychiatry 2019 May 01;76(5):526-535 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.4365] [Medline: 30810723]
4. Brown KN, Soo A, Faris P, Patten SB, Fiest KM, Stelfox HT. Association between delirium in the intensive care unit and subsequent neuropsychiatric disorders. Crit Care 2020 Jul 31;24(1):476 [FREE Full text] [doi: 10.1186/s13054-020-03193-x] [Medline: 32736572]
5. Tschernatsch M, Juenemann M, Alhaidar F, El Shazly J, Butz M, Meyer M, et al. Epileptic seizure discharges in patients after open chamber cardiac surgery-a prospective prevalence pilot study using continuous electroencephalography. Intensive Care Med 2020 Jul;46(7):1418-1424 [FREE Full text] [doi: 10.1007/s00134-020-06073-8] [Medline: 32405742]
6. Li Q, Shen J, Lv H, Liu Y, Chen Y, Zhou C, et al. Incidence, risk factors, and outcomes in electroencephalographic seizures after mechanical circulatory support: a systematic review and meta-analysis. Front Cardiovasc Med 2022 Aug 03;9:872005 [FREE Full text] [doi: 10.3389/fcvm.2022.872005] [Medline: 35990978]
7. Hawkins RB, Mehaffey JH, Guo A, Charles EJ, Speir AM, Rich JB, Virginia Cardiac Services Quality Initiative. Postoperative atrial fibrillation is associated with increased morbidity and resource utilization after left ventricular assist device placement. J Thorac Cardiovasc Surg 2018 Oct;156(4):1543-9.e4 [FREE Full text] [doi: 10.1016/j.jtcvs.2018.03.169] [Medline: 29801690]
8. Sáez de Asteasu ML, Martínez-Velilla N, Zambom-Ferraresi F, Casas-Herrero Á, Cadore EL, Galbete A, et al. Assessing the impact of physical exercise on cognitive function in older medical patients during acute hospitalization: secondary analysis of a randomized trial. PLoS Med 2019 Jul 05;16(7):e1002852 [FREE Full text] [doi: 10.1371/journal.pmed.1002852] [Medline: 31276501]
9. Fan G, Zhong M, Su W, An Z, Zhu Y, Chen C, et al. Effect of different anesthetic modalities on postoperative delirium in elderly hip fractures: a meta-analysis. Medicine (Baltimore) 2024 Jun 07;103(23):e38418 [FREE Full text] [doi: 10.1097/MD.0000000000038418] [Medline: 38847680]
10. Pun BT, Badenes R, Heras La Calle G, Orun OM, Chen W, Raman R, COVID-19 Intensive Care International Study Group. Prevalence and risk factors for delirium in critically ill patients with COVID-19 (COVID-D): a multicentre cohort study. Lancet Respir Med 2021 Mar;9(3):239-250 [FREE Full text] [doi: 10.1016/S2213-2600(20)30552-X] [Medline: 33428871]
11. Berger M, Terrando N, Smith SK, Browndyke JN, Newman MF, Mathew JP. Neurocognitive function after cardiac surgery: from phenotypes to mechanisms. Anesthesiology 2018 Oct;129(4):829-851 [FREE Full text] [doi: 10.1097/ALN.0000000000002194] [Medline: 29621031]
12. Young M, Holmes N, Kishore K, Marhoon N, Amjad S, Serpa-Neto A, et al. Natural language processing diagnosed behavioral disturbance vs confusion assessment method for the intensive care unit: prevalence, patient characteristics, overlap, and association with treatment and outcome. Intensive Care Med 2022 May;48(5):559-569 [FREE Full text] [doi: 10.1007/s00134-022-06650-z] [Medline: 35322288]
13. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020 Aug 21;3:107 [FREE Full text] [doi: 10.1038/s41746-020-00318-y] [Medline: 32885053]
14. Talpur N, Abdulkadir SJ, Alhussian H, Hasan MH, Aziz N, Bamhdi A. Deep Neuro-Fuzzy System application trends, challenges, and future perspectives: a systematic survey. Artif Intell Rev 2023;56(2):865-913 [FREE Full text] [doi: 10.1007/s10462-022-10188-3] [Medline: 35431395]
15. Deeken F, Sánchez A, Rapp MA, Denkinger M, Brefka S, Spank J, PAWEL Study Group. Outcomes of a delirium prevention program in older persons after elective surgery: a stepped-wedge cluster randomized clinical trial. JAMA Surg 2022 Feb 01;157(2):e216370 [FREE Full text] [doi: 10.1001/jamasurg.2021.6370] [Medline: 34910080]
16. Humeidan ML, Reyes JC, Mavarez-Martinez A, Roeth C, Nguyen CM, Sheridan E, et al. Effect of cognitive prehabilitation on the incidence of postoperative delirium among older adults undergoing major noncardiac surgery: the neurobics randomized clinical trial. JAMA Surg 2021 Feb 01;156(2):148-156 [FREE Full text] [doi: 10.1001/jamasurg.2020.4371] [Medline: 33175114]
17. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA Netw Open 2021 Mar 01;4(3):e212240 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2240] [Medline: 33783520]
18. Heilbroner SP, Few R, Mueller J, Chalwa J, Charest F, Suryadevara S, et al. Predicting cardiac adverse events in patients receiving immune checkpoint inhibitors: a machine learning approach. J Immunother Cancer 2021 Oct;9(10):e002545 [FREE Full text] [doi: 10.1136/jitc-2021-002545] [Medline: 34607896]
19. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev 2022;55(3):1947-1999 [FREE Full text] [doi: 10.1007/s10462-021-10058-4] [Medline: 34393317]

XSL•FO
RenderX

20. Zhong Y, Brooks MM, Kennedy EH, Bodnar LM, Naimi AI. Use of machine learning to estimate the per-protocol effect of low-dose aspirin on pregnancy outcomes: a secondary analysis of a randomized clinical trial. JAMA Netw Open 2022 Mar 01;5(3):e2143414 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.43414] [Medline: 35262718]
21. Sumner SA, Bowen D, Holland K, Zwald ML, Vivolo-Kantor A, Guy GP, et al. Estimating weekly national opioid overdose deaths in near real time using multiple proxy data sources. JAMA Netw Open 2022 Jul 01;5(7):e2223033 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.23033] [Medline: 35862045]
22. Nistal-Nuño B. Machine learning applied to a cardiac surgery recovery unit and to a coronary care unit for mortality prediction. J Clin Monit Comput 2022 Jun;36(3):751-763. [doi: 10.1007/s10877-021-00703-2] [Medline: 33860407]
23. Zhang L, Xu F, Han D, Huang T, Li S, Yin H, et al. Influence of the trajectory of the urine output for 24 h on the occurrence of AKI in patients with sepsis in intensive care unit. J Transl Med 2021 Dec 20;19(1):518 [FREE Full text] [doi: 10.1186/s12967-021-03190-w] [Medline: 34930308]
24. Shi Y, Stornetta DS, Reklow RJ, Sahu A, Wabara Y, Nguyen A, et al. A brainstem peptide system activated at birth protects postnatal breathing. Nature 2021 Jan;589(7842):426-430 [FREE Full text] [doi: 10.1038/s41586-020-2991-4] [Medline: 33268898]
25. Davidson S, Villarroel M, Harford M, Finnegan E, Jorge J, Young D, et al. Day-to-day progression of vital-sign circadian rhythms in the intensive care unit. Crit Care 2021 Apr 22;25(1):156 [FREE Full text] [doi: 10.1186/s13054-021-03574-w] [Medline: 33888129]
26. Murphy MS, Ducharme R, Hawken S, Corsi DJ, Petrcich W, El-Chaâr D, et al. Exposure to intrapartum epidural analgesia and risk of autism spectrum disorder in offspring. JAMA Netw Open 2022 May 02;5(5):e2214273 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.14273] [Medline: 35616937]
27. Proix T, Delgado Saa J, Christen A, Martin S, Pasley BN, Knight RT, et al. Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. Nat Commun 2022 Jan 10;13(1):48 [FREE Full text] [doi: 10.1038/s41467-021-27725-3] [Medline: 35013268]
28. Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg 2021 Jan 14;132(2):365-366 [FREE Full text] [doi: 10.1213/ANE.0000000000005247] [Medline: 33449558]
29. Rhodes JS, Cutler A, Moon KR. Geometry- and accuracy-preserving random forest proximities. IEEE Trans Pattern Anal Mach Intell 2023 Sep;45(9):10947-10959. [doi: 10.1109/TPAMI.2023.3263774] [Medline: 37015125]
30. Li Q, He Y, Pan J. CrossFuse-XGBoost: accurate prediction of the maximum recommended daily dose through multi-feature fusion, cross-validation screening and extreme gradient boosting. Brief Bioinform 2023 Nov 22;25(1):bbad511 [FREE Full text] [doi: 10.1093/bib/bbad511] [Medline: 38216539]
31. Que Z, Lin CJ. One-class SVM probabilistic outputs. IEEE Trans Neural Netw Learn Syst 2025 Apr;36(4):6244-6256. [doi: 10.1109/TNNLS.2024.3395148] [Medline: 38717884]
32. Chen Y, Ou Z, Zhou D, Wu X. Advancements and challenges of artificial intelligence-assisted electroencephalography in epilepsy management. J Clin Med 2025 Jun 16;14(12):4270 [FREE Full text] [doi: 10.3390/jcm14124270] [Medline: 40566015]
33. Wang YY, Yue JR, Xie DM, Carter P, Li QL, Gartaganis SL, et al. Effect of the tailored, family-involved hospital elder life program on postoperative delirium and function in older adults: a randomized clinical trial. JAMA Intern Med 2020 Jan 01;180(1):17-25 [FREE Full text] [doi: 10.1001/jamainternmed.2019.4446] [Medline: 31633738]
34. Brown CH, Neufeld KJ, Tian J, Probert J, LaFlam A, Max L, Cerebral Autoregulation Study Group, et al. Effect of targeting mean arterial pressure during cardiopulmonary bypass by monitoring cerebral autoregulation on postsurgical delirium among older patients: a nested randomized clinical trial. JAMA Surg 2019 Sep 01;154(9):819-826 [FREE Full text] [doi: 10.1001/jamasurg.2019.1163] [Medline: 31116358]
35. Leone M, Einav S, Chiumello D, Constantin JM, De Robertis E, De Abreu MG, Guideline contributors. Noninvasive respiratory support in the hypoxaemic peri-operative/periprocedural patient: a joint ESA/ESICM guideline. Intensive Care Med 2020 Apr;46(4):697-713 [FREE Full text] [doi: 10.1007/s00134-020-05948-0] [Medline: 32157356]
36. Gélinas C, Bérubé M, Puntillo KA, Boitor M, Richard-Lalonde M, Bernard F, et al. Validation of the critical-care pain observation tool-neuro in brain-injured adults in the intensive care unit: a prospective cohort study. Crit Care 2021 Apr 13;25(1):142 [FREE Full text] [doi: 10.1186/s13054-021-03561-1] [Medline: 33849619]
37. Honarmand K, Lalli RS, Priestap F, Chen JL, McIntyre CW, Owen AM, et al. Natural history of cognitive impairment in critical illness survivors. A systematic review. Am J Respir Crit Care Med 2020 Jul 15;202(2):193-201 [FREE Full text] [doi: 10.1164/rccm.201904-0816CI] [Medline: 32078780]
38. Rosas IO, Bräu N, Waters M, Go RC, Malhotra A, Hunter BD, et al. Tocilizumab in patients hospitalised with COVID-19 pneumonia: efficacy, safety, viral clearance, and antibody response from a randomised controlled trial (COVACTA). EClinicalMedicine 2022 May;47:101409 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101409] [Medline: 35475258]
39. Madrigal C, Kim J, Jiang L, Lafo J, Bozzay M, Primack J, et al. Delirium and functional recovery in patients discharged to skilled nursing facilities after hospitalization for heart failure. JAMA Netw Open 2021 Mar 01;4(3):e2037968 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37968] [Medline: 33724390]

XSL•FO
RenderX

40. Saczynski JS, Koethe B, Fick DM, Vo QT, Devlin JW, Marcantonio ER, et al. Cognitive and functional change in skilled nursing facilities: differences by delirium and Alzheimer's disease and related dementias. J Am Geriatr Soc 2024 Nov;72(11):3501-3509. [doi: 10.1111/jgs.19112] [Medline: 39171670]
41. Huang HL, Lee JY, Lo YS, Liu IH, Huang SH, Huang YW, et al. Whole-blood 3-gene signature as a decision aid for rifapentine-based tuberculosis preventive therapy. Clin Infect Dis 2022 Sep 14;75(5):743-752 [FREE Full text] [doi: 10.1093/cid/ciac003] [Medline: 34989801]
42. McKenna M, Shackelford D, Pontes C, Ball B, Nance E. Multiple particle tracking detects changes in brain extracellular matrix and predicts neurodevelopmental age. ACS Nano 2021 May 25;15(5):8559-8573 [FREE Full text] [doi: 10.1021/acsnano.1c00394] [Medline: 33969999]

Abbreviations

AUC: area under the receiver operating characteristic curve
CAM-ICU: Confusion Assessment Method for the Intensive Care Unit
eICU-CRD: eICU Collaborative Research Database
GCS: Glasgow Coma Scale
ICU: intensive care unit
MIMIC-IV 2.0: Medical Information Mart for Intensive Care IV version 2.0
ML: machine learning
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanations
SOFA: Sequential Organ Failure Assessment
XGB: extreme gradient boosting

Edited by J Klann; submitted 01.Mar.2025; peer-reviewed by P Tekkali, A Özel; comments to author 30.Apr.2025; accepted 26.Sep.2025; published 11.Feb.2026.

Please cite as:
Hu H, Wang Y, Li H, Zang Q, Huang J, Zhang Y, Wu J, Liu L, Xing Z, Yu Y
Early Prediction of Delirium in Postcardiac Surgery Patients: Machine Learning Model Development and External Validation
JMIR Med Inform 2026;14:e73283
URL: https://medinform.jmir.org/2026/1/e73283
doi:10.2196/73283
PMID:

©Huixiu Hu, Yuxiang Wang, Houfeng Li, Qinglai Zang, Jing Huang, Ying Zhang, Jinjing Wu, Long Liu, Zhen Xing, Yaohua Yu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ensemble Machine Learning Models for Predicting Patients With High Usage: Model Validation and Economic Impact Analysis

Joshua Kuan Tan[1*], MSc, MBBS; Le Quan[2*], MSc; Hao Yi Tan[1*], MPH, MD; Su-Yen Goh[3*], MBBS; Julian Thumboo[1*], MBBS; Marianne Au[4*], BBus; Yong Mong Bee[5*], MBBS; Jen Hong Tan[2*], PhD

[1]Health Services Research Unit, Singapore General Hospital, Singapore, Singapore

[2]Data Science and Artificial Intelligence Laboratory, Singapore General Hospital, Singapore, Singapore

[3]Future Health Systems Department, Singapore General Hospital, Singapore, Singapore

[4]Group Finance, SingHealth, Singapore, Singapore

[5]Department of Endocrinology, Singapore General Hospital, Singapore, Singapore

[*]all authors contributed equally

Corresponding Author:
Joshua Kuan Tan, MSc, MBBS
Health Services Research Unit
Singapore General Hospital
Outram Road
Singapore, 169608
Singapore
Phone: 65 223322
Email: joshua.tan.kuan@singhealth.com.sg

Abstract

Background: Machine learning models are increasingly used to predict patients at risk of high health care usage for targeted interventions.

Objective: This study aimed to evaluate the predictive performance of multiclass ensemble models across different levels of health care usage and assess their potential application through real-world economic impact analysis.

Methods: A total of 4 previously developed binary classification models (base learners)—boosted trees, multivariate adaptive regression splines, multilayer perceptron, and logistic regression—were extended using a stacking ensemble approach. These base learner models generated individual-level predicted probabilities, which were used as inputs to build multiclass prediction models forecasting usage across defined strata: length of stay (LOS) of <7, 7-13, 14-29, and ≥30 days, and emergency department (ED) visits of <3, 3-4, 5-9, and ≥10 visits. In total, 3 ensemble algorithms were evaluated: random forest, boosted trees, and linear support vector machines. Ensemble models were trained on registry data from 2020-2021 and temporally validated on 2021-2022 data. Performance was assessed using multiclass area under the receiver operating curve, accuracy, and confusion matrix–derived metrics. Economic impact was estimated via Monte Carlo simulations using inpatient billing data, assuming a 20% cost reduction in the following year.

Results: The models were trained on 108,886 patients and validated on 111,004 patients. Among all ensemble configurations, boosted tree models regardless of base learner achieved the highest performance, with multiclass area under the receiver operating curve scores of 0.6877 (95% CI 0.6927-0.7255) for LOS and 0.7601 (95% CI 0.7301-0.7654) for ED visits, and corresponding accuracies of 0.6522 (95% CI 0.6465-0.6579) and 0.7457 (95% CI 0.7405-0.7508), respectively. In the validation set, these models correctly assigned 30.3% of inpatient LOS and 39.8% of ED visits to the correct class, identifying 77% of future inpatient users and 73.9% of future ED users. Economic impact analysis for LOS identified the boosted tree model with logistic regression base learner as dominant, achieving a simulated average cost reduction of SGD $152 million (US $111 million), SGD $2.4 million (US $1.75 million; 1.5%) more than the next best model using a multilayer perceptron base learner.

Conclusions: Ensemble models can effectively predict multilevel health care usage and potentially generate meaningful cost savings when applied to real-world settings. These models may support targeted interventions and guide planning and budgeting in diabetes-related population health programs.

(JMIR Med Inform 2026;14:e77202) doi:10.2196/77202

XSL•FO
RenderX

KEYWORDS

diabetes mellitus; healthcare utilization; population health management, machine learning; artificial intelligence; economic analysis; decision analysis; Monte Carlo simulation

Introduction

Health care expenditure in high-income countries is rising [1], with chronic diseases (specifically diabetes mellitus) being a leading contributor. Diabetes accounts for approximately 15% of health care expenditure in the United States [2], 9% in the European Union [3], and 11% in Singapore [4]. It is a significant cause of cardiovascular disease, kidney failure, eye disease, peripheral vascular disease, and neuropathy, all associated with substantial morbidity, mortality, and higher health care costs.

Health care spending is also highly skewed, with a small proportion of patients incurring most costs [5]. Efforts to identify high-need, high-cost (HNHC) patients—typically the top 5% of usage or cost—have led to predictive models using regression and machine learning [6]. However, the clinical utility of these models remains uncertain, as few studies assess their net benefit compared with not using them [6-8]. Furthermore, the potential to reduce costs among the top decile or 5% of high-cost patients is smaller than commonly assumed [9], possibly due to heterogeneity in patient characteristics and usage patterns, variability in group membership over time, and a small proportion who are amenable to cost-reducing interventions [5,9,10]. Focusing narrowly on the highest users may also overlook many individuals with future health care needs outside this group.

Building on these insights, our previous study expanded the predictive focus beyond traditional HNHC patients by developing binary classification models to predict inpatient and emergency department (ED) usage across various thresholds for patients in the Singapore Health Services Diabetes Registry (SDR), a clinical database of over 130,000 patients with diabetes [11]. In this study, we extended that work by creating separate multiclass ensemble models for inpatient bed days and ED visits. We validated these models, evaluated their relevance for population health program development, and analyzed their potential economic impact using real-world financial data [7,8]. Our goal is to develop clinically actionable predictive models that can be integrated into broader diabetes-related population health initiatives.

Methods

Study Design and Population

This study used data from the SDR spanning 2020-2022, comprising 2 datasets: 2020-2021 and 2021-2022. The 2019-2020 dataset, previously used to develop the binary classification models [11], was excluded from the current analysis to prevent overfitting. Each dataset (2019-2020, 2020-2021, and 2021-2022) was constructed as a cross-sectional snapshot based on a fixed 1-year prediction window. Patients could appear in multiple datasets across years if they continued to meet inclusion criteria for entry into the registry [12]; however, within each dataset, only 1 record per patient per year was used. Multiple visits within the same year were aggregated to derive annual service-level usage metrics (eg, total inpatient bed days and total ED visits).

Predictor Variables

Individual-level predicted probabilities from the binary classification models, which reflect the predicted likelihood of future usage, were used as predictor variables in the corresponding ensemble models (ie, length of stay [LOS] probabilities for the LOS ensemble model and ED probabilities for the ED ensemble model). Additionally, current-year LOS or ED visits were also included as predictors. Full details on predictor variables for the binary models are provided in Table S1 in Multimedia Appendix 1 and our previous publication [11]; briefly, they include age, gender, ethnicity, housing type, comorbidities (eg, hypertension, hyperlipidemia, and chronic kidney disease), diabetes-related complications, and mean glycated hemoglobin.

Outcome Variables

Multiclass ensemble models were developed to predict usage for LOS and ED visits across 4 classes each. For LOS, the classes were <7 days, 7-13 days, 14-29 days, and ≥30 days. For ED visits, the classes were <3 visits, 3-4 visits, 5-9 visits, and ≥10 visits.

Inclusion and Exclusion Criteria

Patients aged 18 years or older with type 2 diabetes mellitus were included. Patients with missing variables were excluded, as data imputation was not performed, and most machine learning algorithms used in this study do not natively support missing values.

Handling Unbalanced Data

Class imbalance negatively affected model performance in previous studies, and random upsampling outperformed the synthetic minority oversampling technique [11]. We applied random upsampling using the ‘upSample’ function from the *caret* package [13] to achieve equal representation across all classes in the training dataset (ie, an equal number of patients across all classes).

Machine Learning Models

Previously developed binary classification models from the multi-institutional SDR were used to predict total inpatient LOS and ED visits over a calendar year [11]. Predictions were made across 3 thresholds in the 2 outcomes: LOS of ≥7, ≥14, and ≥30 days; and ED attendance of ≥3, ≥5, and ≥10 visits.

We used the best-performing binary classification models from our previous work, which were trained using the 2019-2020 dataset as base learners: boosted trees, multivariate adaptive regression splines, multilayer perceptron, and logistic regression [11]. For each base learner type, 3 separate binary models were developed to predict inpatient LOS of ≥7, ≥14, and ≥30 days, and ED attendance of ≥3, ≥5, and ≥10 visits, respectively. The

XSL•FO
RenderX

individual-level predicted probabilities from these binary models were then used as inputs (meta-features) in a stacking ensemble model designed to classify patients into 4 mutually exclusive outcome categories (LOS: <7, 7-13, 14-29, ≥30 days, and ED: <3, 3-4, 5-9, ≥10 visits).

Stacking was selected over voting-based methods for its ability to learn optimal weights for combining base learner outputs, rather than treating all models equally or assigning manual weights [14]. Furthermore, 3 algorithms were tested as meta-learners: random forest, boosted trees, and linear support vector machines (SVMs). All models were developed in R (version 4.3.1; R Core Team) using the *tidymodels* package [15].

The 2020-2021 dataset was used to train and test the ensemble models, with a 75:25 split. Models were then temporally validated on an unseen 2021-2022 dataset. Each dataset represented a patient-year observation with a fixed 1-year prediction horizon. All predictor variables, including inpatient bed days and ED visits, were derived exclusively from the preceding calendar year to predict usage in the subsequent year, ensuring that no future information was used at the time of prediction and avoiding look-ahead bias. While patients could appear in multiple datasets across years, training and validation were temporally separated to ensure that predictions were evaluated on future, unseen patient-year data, consistent with real-world deployment for chronic disease management. To ensure consistency and reproducibility, we used default hyperparameters for all base learners and meta-learners.

Explainability

We did not conduct explainability analysis for the ensemble models, as previous analyses of the base learner models with model-specific variable importance scores, permutation feature importance plots, and partial dependence plots demonstrated consistent and predictable behavior [11]. The ensemble models' transparent construction from base learner outputs rendered additional explainability analysis unnecessary.

Performance Indicators

Model performance was assessed using multiclass area under the receiver operating curve (AUROC; one-vs-rest approach), and accuracy (95% CI) was calculated using the DeLong method (via the pROC package) for AUROC and the confusionMatrix function (from the caret package) for accuracy. Confusion matrices were analyzed, and new metrics specific to our 4-class prediction task were developed for population health programs (Figure 1). Additional metrics, including sensitivity, specificity, positive predictive value, negative predictive value, and balanced accuracy, are detailed in the Multimedia Appendix 1.

These new metrics were designed for population-level interventions using the prediction models. Our priorities were maximizing correct predictions, followed by identifying users, and minimizing unnecessary predictions (ie, patients predicted to have usage but who ultimately did not). Maximizing correct predictions minimizes underpredictions, overpredictions, and missed cases. Increasing identified users enables interventions to reach more patients with future needs, while minimizing unnecessary predictions enhances program efficiency by reducing resources spent on patients unlikely to use services.

Figure 1. Confusion matrix in a 4-class problem and related evaluation variables.

Labels	Inpatient bed days	Emergency Department Visits
0	< 7 days	< 2 visits
1	7 - 13 days	3 - 4 visits
2	14 - 29 days	5 - 9 visits
3	≥ 30 days	≥ 10 visits

Variable	Description	Formula
Total	All patients in the prediction	$= A + B + \cdots + P$
Users	All actual users	$=$ Total $- (A + E + I + M)$
Correct predictions	Ratio of correct predictions (*among patients with actual usage*) to all users	$= \frac{F + K + P}{\text{Utilizers}}$
Underprediction	Ratio of underpredictions (*among patients with actual usage*) to all users	$= \frac{G + H + L}{\text{Utilizers}}$
Overprediction	Ratio of overprediction (*among patients with actual usage*) to all users	$= \frac{J + N + O}{\text{Utilizers}}$
Unnecessary predictions	Ratio of predicted users (*among patients with **no** actual usage*) to all patients with ***no*** usage	$= \frac{E + I + M}{A + E + I + M}$
Missed cases	Ratio of predicted nonusers (*among patients with actual usage*) to all users	$= \frac{B + C + D}{\text{Utilizers}}$
Identified users	Ratio of predicted users (*among patients with actual usage*) to all users	$=$ Correct predictions + Underprediction + Overprediction

XSL•FO
RenderX

Economic Impact Analysis of the Models

We assessed the potential economic impact of the top 5 ensemble models with the largest proportion of correct predictions for inpatient bed days. Using individual-level inpatient cost data, we modeled hypothetical cost reductions to simulate the financial implications of targeted intervention. These simulations assume that accurately predicted patients would receive effective interventions that reduce future inpatient costs. Cost savings were not attributed to the prediction alone, but to downstream programs triggered by the prediction.

In total, 2 key parameters were incorporated—reach and benefits, both factors important to population-level impact. Reach was defined as the proportion of patients within a predicted usage class who would receive the intervention [16], while benefits referred to the hypothetical percentage reduction in future inpatient costs among those reached, attributable to the intervention.

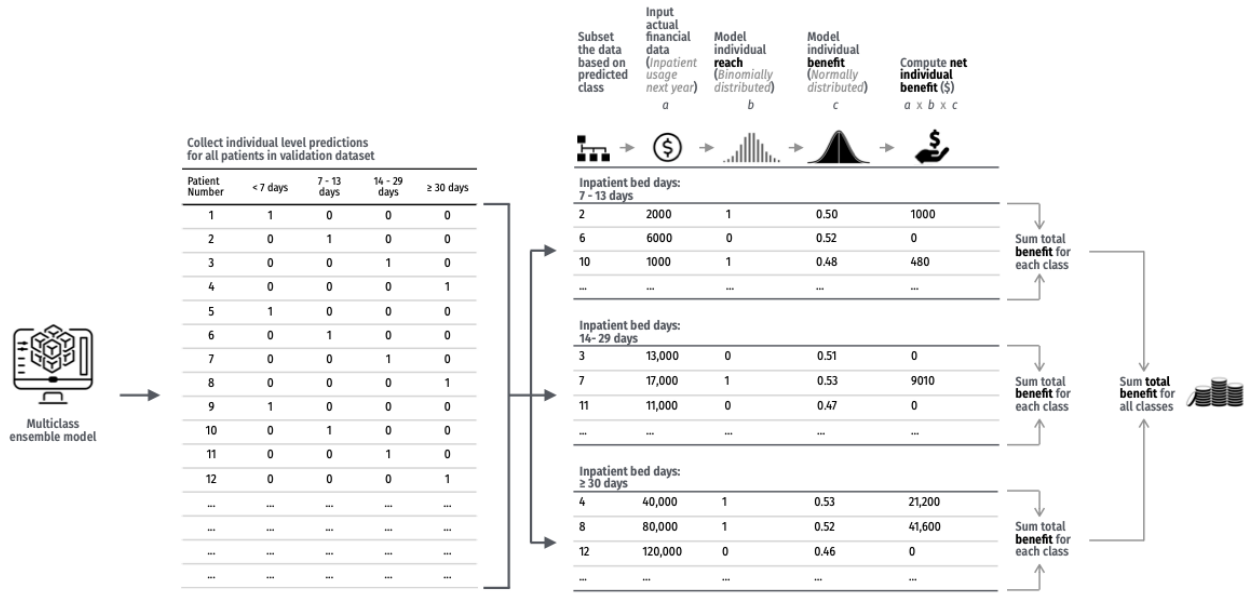
We used total inpatient costs (in Singapore dollars and the equivalent United States dollars [US $], presubsidies) for each calendar year, using the conversion rate of SGD $1=US $1.370, which is accurate as of 1 June 2022. We assessed model performance using fixed effects and Monte Carlo simulations with 10,000 iterations to account for random effects. “Total benefit” was defined as the simulated reduction in inpatient costs in the following year, enabling us to compare models and identify the dominant one. A schematic of the economic impact analysis model is provided in Figure 2, with parameters detailed in Table S2 in Multimedia Appendix 1.

In the fixed effects model, we modeled the benefit at a uniform 50% cost reduction for all identified patients—an optimistic scenario to benchmark maximum potential impact. In contrast, the Monte Carlo simulations modeled a more conservative and realistic benefit at an average cost reduction of 20% across all classes, with stochastic variation applied using a normal distribution.

Real-world programmatic use was further simulated by varying reach across usage tiers: average reach levels were set at 90% for the highest usage class, 70% for the middle class, and 50% for the lowest classes, each with a 5% variance. For example, the reach for the highest users ranged from 85% to 95% across simulations. This tiered approach reflects potential operational realities: patients predicted in the highest usage class are fewer in number and more likely to be prioritized and successfully engaged by targeted interventions. In contrast, lower usage classes include larger populations, making it less feasible to reach all patients in practice. To account for individual variability in intervention uptake, reach levels were also modeled using a binomial probability distribution.

From the Monte Carlo simulations, we identified the dominant model as the one with the highest median cost savings. To explore uncertainty and assess the robustness of this selection, we performed parameter sensitivity analysis stratified by usage class. Specifically, we varied the reach and benefit parameters in 10% intervals (ranging from 10% to 90%), resulting in 81 unique parameter combinations. For each combination, 1000 Monte Carlo runs were conducted to account for stochastic variability. The distribution of cost savings across simulations was analyzed, and results were visualized using the *plotly* package [17] to provide an interactive and detailed view of model performance under different programmatic scenarios.

Figure 2. Schematic of economic impact analysis model.



Reporting Checklist

We reported our study using the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis + Artificial Intelligence (TRIPOD+AI) checklist (Multimedia Appendix 2) [18].

XSL•FO
RenderX

Ethical Considerations

Ethics approval was obtained from the Singapore Health Services Centralized Institutional Review Board before initiating this study (reference 2022/2133). As all participant data were deidentified, a waiver for participant consent was also obtained. All procedures were performed in compliance with relevant laws and institutional guidelines.

Results

Characteristics of the Datasets

The demographics, comorbidities, and usage characteristics of the datasets have been previously described [11]. The 2020-2021 and 2021-2022 datasets included 108,886 of 140,859 (77.3%) patients and 111,004 of 137,584 (80.7%) patients from the total SDR cohort, respectively. Total inpatient costs (presubsidies) per annum are summarized in Table 1, and inpatient LOS distributions are illustrated in Figures S1-S3 in Multimedia Appendix 1.

Inpatient usage showed an increasing trend, with gross inpatient bills (unadjusted for inflation) rising from a nominal value of SGD $808 (US $590) million in 2019 to SGD $1.01 billion (US $737 million) in 2022. The proportion of patients with inpatient usage increased from 46,686 of 134,670 (34.7%) in 2019 to 51,228 of 137,627 (37.2%) in 2022, while the average cost per patient rose from SGD $17,321 to SGD $19,704 (US $12,643 to US $14,382). Total inpatient costs were heavily right-skewed (Table S3 in Multimedia Appendix 1) and followed a Pareto distribution (Figure S1 in Multimedia Appendix 1), with the top 10% of patients accounting for 80% of cumulative health care costs and inpatient bed days annually (Figures S1 and S2 in Multimedia Appendix 1). A strong correlation was observed between gross inpatient bills and inpatient bed days (Figure S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Performance metrics of the top 5 best-performing ensemble models.

Model characteristics	Accuracy (95% CI)	Multiclass AUROC[a] (95% CI)	Correct predictions, %	Underpredictions, %	Overpredictions, %	Unnecessary predictions, %	Identified users, %	Missed cases, %
Best performing ensemble models on test dataset (2021-2022)								
Predicting inpatient bed days category								
Boosted tree with MARS[b] as base learner	0.6522 (0.6465-0.6579)	0.6877 (0.6927-0.7255)	30.5	23.2	24.1	31.3	77.8	22.2
Boosted tree with MLP[c] as base learner	0.6685 (0.6629-0.6741)	0.6832 (0.6875-0.7202)	30.2	20.3	25.5	29.6	76	24
Boosted tree with logistic regression as base learner	0.6495 (0.6438-0.6552)	0.6696 (0.6896-0.722)	30.2	20	26	31.7	76.1	23.9
Boosted tree with boosted tree as base learner	0.6654 (0.6598-0.671)	0.6776 (0.6936-0.7259)	29.7	21.2	24.1	29.9	75.1	24.9
Linear SVM[d] with MLP as base learner	0.7300 (0.7246-0.7352)	0.6758 (0.6757-0.7089)	26.8	19.7	22.6	22.6	69.1	30.9
Predicting emergency department visits category								
Boosted tree with MARS as base learner	0.7457 (0.7405-0.7508)	0.7601 (0.7301-0.7654)	39.5	10.7	24.8	24	75	25
Boosted tree with logistic regression as base learner	0.7205 (0.7151-0.7258)	0.7569 (0.7294-0.7641)	39.3	11	24.9	26.6	75.2	24.8
Boosted tree with MLP as base learner	0.7432 (0.738-0.7484)	0.7557 (0.731-0.7664)	37.9	10.4	25.8	24.2	74	26
Boosted tree with Boosted tree as base learner	0.7425 (0.7372-0.7476)	0.7527 (0.7016-0.7378)	37.1	14.7	18.3	24.2	70.1	29.9
Linear SVM using MARS as base learner	0.7849 (0.7800-0.7898)	0.7722 (0.7316-0.7681)	30.3	6.6	34.8	19.6	71.8	28.2
Best performing ensemble models on validation datasets								
Predicting inpatient bed days category								
Boosted tree with MARS as base learner	0.6405 (0.6377-0.6434)	0.6811 (0.6986-0.714)	30.3	21.1	25.6	32.1	77	23
Boosted tree with boosted tree as base learner	0.6496 (0.6468-0.6524)	0.6828 (0.6966-0.712)	29.9	21.9	24.7	31.1	76.4	23.6
Boosted tree with logistic regression as base learner	0.6357 (0.6329-0.6385)	0.6767 (0.6929-0.7082)	29.4	20.2	27.6	32.6	77.1	22.9
Boosted tree with MLP as base learner	0.6558 (0.6530-0.6586)	0.6782 (0.6957-0.7111)	29.2	19.2	27.4	30.3	75.8	24.2
Linear SVM with MLP as base learner	0.71 (0.7073-0.7127)	0.6742 (0.6851-0.7004)	26.5	19.3	24.5	24	70.4	29.6
Predicting emergency department visits category								
Boosted tree with logistic regression as base learner	0.7053 (0.7026-0.7079)	0.7669 (0.7199-0.7359)	39.8	10.4	23.7	27.9	73.9	26.1
Boosted tree with boosted tree as base learner	0.7317 (0.7290-0.7343)	0.7577 (0.7139-0.73)	39.7	12.5	19	25.1	71.1	28.9
Boosted tree with MLP as base learner	0.7269 (0.7243-0.7295)	0.7563 (0.7158-0.732)	36.9	10.5	23.9	25.5	71.3	28.7
Boosted tree using MARS as base learner	0.7266 (0.7239-0.7292)	0.7664 (0.7169-0.7332)	36.7	10.2	24.7	25.5	71.5	28.5
Linear SVM using MARS as base learner	0.7656 (0.7631-0.7681)	0.7818 (0.7204-0.7368)	30.1	6.7	32.2	21	69	31

[a]AUROC: area under the receiver operating curve.

XSL•FO
RenderX

[b]MARS: multivariate adaptive regression splines.

[c]MLP: multilayer perceptron.

[d]SVM: support vector machine.

Multiclass Model Performance

We developed 12 ensemble models by combining 4 base learner models with 3 ensemble algorithms: random forest, boosted trees, and linear SVM. Detailed performance metrics for all 12 models on the test dataset are provided in Table S4-S6 in Multimedia Appendix 1. We then selected top 5 models for each prediction task based on the highest percentages of correct predictions and identified users (Table 1 and Table S4 in Multimedia Appendix 1), and their performance was temporally validated on the unseen dataset from 2021 to 2022; detailed performance metrics for these models on the test dataset are provided in Tables S7 and S8 in Multimedia Appendix 1.

Boosted tree models outperformed other ensemble algorithms across metrics, including multiclass AUROC, correct predictions, and identified users. While linear SVM models had comparable multiclass AUROC scores to boosted tree models, they showed significantly lower correct predictions, fewer identified users, and more missed cases. Random forest models performed poorly across all metrics.

Boosted tree models demonstrated consistent performance on test and validation datasets, indicating robustness on unseen data (Table 1 and Table S4 in Multimedia Appendix 1). For inpatient bed days, these models correctly assigned approximately 30% of future users to the appropriate category, identified >75% of future users, and generated 30%-32% unnecessary predictions among nonusers. For ED visits, they correctly assigned 36%-39% of future users, identified 70%-75% of users, and generated approximately 25% unnecessary predictions among nonusers.

Economic Impact Analysis of the Models

Results from the fixed effects and Monte Carlo simulations are presented in Table 2 and Figure S4 in Multimedia Appendix 1. The boosted tree ensemble model with logistic regression as the base learner emerged as the dominant model, although other boosted tree models performed similarly. In the fixed effects simulation, the dominant model generated an estimated benefit of SGD $210 (US $153) million. In the Monte Carlo simulations, it produced a median benefit of SGD $152 (US $110) million. Notably, nearly half of the simulations for the dominant model exceeded the 75th percentile of simulations from other models (Table 2 and Figure S4 in Multimedia Appendix 1).

Table 2. Results of fixed effects and Monte Carlo simulations.

Variables	Models				
	Boosted tree with logistic regression as base learner	Boosted tree with MLP[a] as base learner	Boosted tree with MARS[b] as base learner	Boosted tree with boosted tree as base learner	Linear SVM[c] with MLP as base learner
Results for the fixed effects simulation (SGD $)					
Total simulated reduction in inpatient bill	210,220,829	205,184,563	209,535,576	206,000,929	189,308,087
Results for the Monte Carlo simulations (SGD $)					
Mean (SD)	152,794,862 (6,001,058)	150,437,894 (5,816,301)	149,408,731 (5,959,678)	147,699,501 (5,809,931)	136,692,927 (7,365,761)
Median (IQR)	152,854,645 (8,656,193)	150,468,563 (8,399,851)	149,405,404 (8,402,405)	147,756,137 (8,198,816)	136,722,177 (12,202,993)
Range (minimum-maximum)	134,202,838-170,594,658	133,162,981-167,868,627	131,679,740-167,051,047	129,415,870-165,742,786	118,420,401-153,678,680
Q25	148,510,936	146,256,897	145,189,862	143,596,095	130,577,413
Q75	157,167,129	154,656,748	153,592,267	151,794,911	142,780,406

[a]MLP: multilayer perceptron.

[b]MARS: multivariate adaptive regression splines.

[c]SVM: support vector machine.

Parameter Sensitivity Analysis

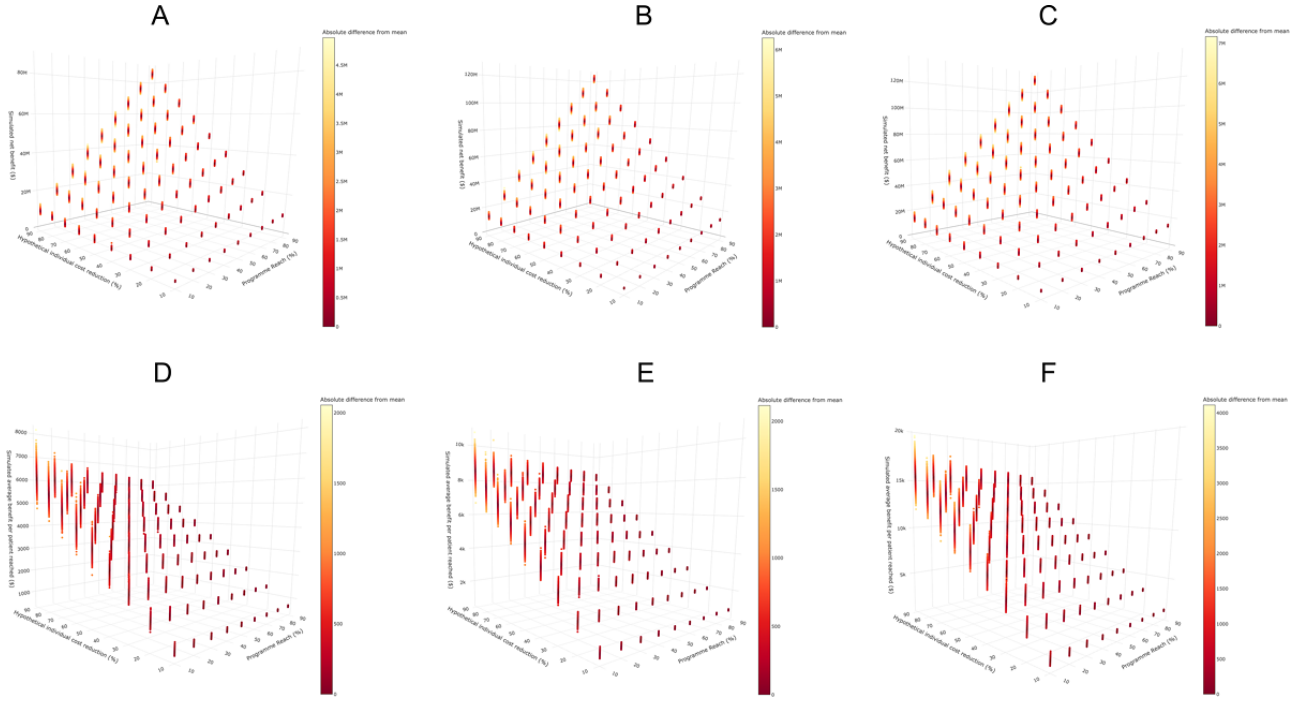
We evaluated the dominant model's performance across various permutations of reach and benefit (Figure 3 and interactive plots available on GitHub [19]). As expected, the net benefit increased with higher individual cost reductions (Figures 3A-3C) and program reach. For predicting LOS of 14-29 days (Figure 3B), with a constant 50% cost reduction, the simulated net benefit ranged from SGD $5.95 to SGD $9.62 (US $4.34 to US $7.02) million at 10% reach, SGD $35.6 to SGD $40.7 (US $26.0 to US $29.7) million at 50% reach, and SGD $66.6 to SGD $70.0 (US $48.6 to US $51.1) million at 90% reach.

With increasing program reach, the range of simulated average benefit per patient narrowed. For LOS of 14-29 days (Figure 3D), with a 50% cost reduction, the average benefit per patient

XSL•FO
RenderX

ranged from SGD $3708 to SGD $5814 (US $2707 to US $4244) at 10% reach and from SGD $4656 to SGD $4389 (US $3399 to US $3204) at 90% reach. The implications of these findings are further discussed in the Discussion section.

Figure 3. 3D scatterplot for Monte Carlo simulations across various permutations of reach and benefit for the boosted tree with logistic regression as base learner ensemble. (A) Simulated net benefit for prediction 7-13 days; (B) simulated net benefit for prediction 14-29 days; (C) simulated net benefit for prediction ≥30 days; (D) simulated average benefit per patient for prediction 7-13 days; (E) simulated average benefit per patient for prediction 14-29 days; and (F) simulated average benefit per patient for prediction ≥30 days. Readers can refer to the interactive plots on GitHub for higher resolution interactive images.



Discussion

Principal Findings

In this study, we expanded on our previous work by developing multiclass ensemble machine learning models to predict future inpatient LOS and ED visits. The models were evaluated using public health program objectives, analyzing predictions through confusion matrices and performance metrics to identify the best-performing models. Economic impact analysis, incorporating inpatient cost data and program-related simulations, identified the boosted tree ensemble algorithm with logistic regression as the base learner as dominant for LOS. Further testing across analytical parameters confirmed its robustness and practical applicability.

While previous studies focused on binary outcomes with limited evaluation of clinical utility [6,20], our study reframes high health care usage as a multiclass problem, enabling more nuanced predictions across different health care events. By incorporating financial data into the economic impact analysis with Monte Carlo simulations, we model the potential benefits of the model, enhancing its relevance for real-world applications [7,8]. This novel approach to economic impact analysis represents a significant contribution to machine learning in health care and the design of resource-efficient health programs.

Advancing High User Prediction Through Multiclass Modeling and Economic Impact Analysis

Our study advances the prediction and targeting of patients with high health care usage by moving beyond binary classification. While many previous studies have focused on predicting binary outcomes—such as identifying HNHC patients in the top 5% or 10% of usage—this strategy has limited practical utility, as most high users do not remain so over time [9,21,22]. Consequently, there is growing recognition of the need to predict risk across multiple levels of usage, enabling stratified interventions tailored to predicted resource needs [9].

To the best of our knowledge, few studies have applied machine learning to predict health care usage across multiple discrete classes. One recent study using Louisiana Department of Health records built multinomial models to predict clinical service usage (ie, number of clinic visits across 5 classes), and found that tree-based ensemble methods, such as extreme gradient boosting and random forest, consistently outperformed logistic regression, neural networks, and naïve Bayes models—likely due to their capacity to model complex nonlinear interactions in heterogeneous data [23]. Another study within the New York City safety net system developed multiple models to predict acute care usage as both binary and continuous outcomes. Their best-performing model, which predicted the number of acute care days as a continuous variable, successfully identified approximately half of the patients who would go on to have very high acute care use [24].

XSL•FO
RenderX

Our study builds on these approaches in several novel ways. First, we applied a stacking ensemble framework to a multiclass prediction problem, enabling stratification across 4 tiers of inpatient bed days and ED visits. This allows for more granular predictions and targeted interventions, unlike previous work focused solely on binary or continuous outcomes. Second, we introduced custom evaluation metrics derived from confusion matrices (Figure 1) that are directly relevant to population health program planning. Third, we conducted a Monte Carlo–based economic impact analysis, incorporating real-world inpatient cost data to estimate potential cost savings under varying assumptions of reach and benefit. This coupling of predictive modeling with economic simulation offers a practical and policy-relevant contribution to the literature on machine learning for health care resource optimization.

Practical Usage of Ensemble Prediction Models for Future Inpatient Bed Days and ED Visits

Our study predicted LOS and ED usage across 4 tiers, rather than focusing on a specific percentile as is the case in many previous studies [6]. This approach enables tailored interventions based on predicted resource use. The distribution of predictions in the confusion matrix (Tables S4 and S5 in Multimedia Appendix 1), with more predictions in the lower usage tier and fewer in the highest usage tier, suggests that low-cost, scalable interventions (eg, clinical flags, EMR risk prompts, or light-touch monitoring) may be suitable for the lower tier (7-13 inpatient days or 3-4 ED visits) [9,10]. For the middle tier (14-29 inpatient days or 5-9 ED visits), moderately scalable interventions, such as programs involving community nurses or care coordinators, may be effective. Higher-cost, intensive interventions are needed for the highest tier (≥30 inpatient days or ≥10 ED visits), although the scalability is limited.

The new model evaluation metrics we developed address the varying costs of misclassification [25]. Models achieving high correct predictions and low unnecessary predictions enhance program efficiency, while those with more identified users allow interventions to reach more patients, although subsequent recalibration of the intervention may be required downstream. Future health programs using these models should assess patient characteristics and clinical needs at enrollment and during follow-up to ensure appropriate intervention assignment. These metrics may also be adapted for other multiclass problems in health care.

Balancing trade-offs between correct and unnecessary predictions is crucial when selecting the best model. Models with higher correct predictions often had more unnecessary predictions (Tables S4-S6 in Multimedia Appendix 1). Similarly, models with more identified users showed higher unnecessary predictions. This trade-off highlights the need to balance these metrics. Although unnecessary predictions may reduce program efficiency, prioritizing correct predictions and identified users led to greater economic benefits, as shown in our economic impact analyses (Table 2 and Figure S4 in Multimedia Appendix 1). Boosted tree ensemble models, which had higher correct predictions, identified users and unnecessary predictions, and significantly outperformed linear SVM ensemble models.

An important feature of our ensemble models is the tiered distribution of predictions (Tables S4-S7 in Multimedia Appendix 1), with the fewest in the highest tier, followed by the middle, and most in the lower tier. This distribution supports efficient resource allocation: low-cost, scalable interventions can be directed to lower tiers, while high-cost, intensive interventions are reserved for the highest tier—ensuring sustainable deployment of resources.

This stratified output also enables a graduated intervention pathway, starting with risk-appropriate outreach and followed by clinical reassessment to adjust intervention intensity. This flexibility is essential given the modest positive predictive value (13%-14%), even in the highest usage group—a reflection of the complexity and unpredictability of health care use, driven by diverse clinical and social factors.

Nevertheless, our best models identified more than 75% of future high users (≥7 inpatient days), offering valuable opportunities for early engagement. Even if further triage is needed, this early identification is a key step in proactive population health management. Future implementations should incorporate qualitative review or clinical vetting to further optimize targeting and minimize risks of overintervention or alert fatigue.

Implications of Economic Impact Analysis and Parameter Sensitivity Analysis

To our knowledge, this is the first study to conduct an economic impact analysis of predictive models for identifying patients at risk of high health care usage. We selected the best-performing ensemble model for inpatient LOS, leveraging inpatient billing data to assess its real-world potential. The boosted tree ensemble algorithm with logistic regression as its base learner emerged as the dominant model (Table 2 and Figure 3), demonstrating superior performance in both fixed effects and Monte Carlo simulations. Notably, the median simulation outcome for the dominant model exceeded the 75th percentile of other models, indicating that its typical performance surpassed even the best-case scenarios (upper quartile) of the alternatives. This highlights the dominant model's consistency and reliability, despite similar AUROC, accuracy, and evaluation metrics across models.

Findings from the parameter sensitivity analysis provide health care planners with insights into the potential cost savings on different combinations of reach and benefit. Our results show that increasing program reach significantly boosts total benefit, while reducing variability in benefit per patient, especially when more high-cost patients (typically those found in the top cost percentiles) are included. This reflects the Pareto distribution of inpatient costs (Figure S1 in Multimedia Appendix 1), where maximizing program reach is essential for capturing high-cost patients and maximizing the overall net benefit.

While the parameter sensitivity analysis does not explicitly incorporate intervention costs, it serves as a useful planning tool by simulating the financial space available for intervention. For example, in a conservative scenario targeting patients predicted to stay 14-29 days, assuming a 30% cost reduction and 50% reach, projected savings reach SGD $21.0 (US $15.3)

XSL•FO
RenderX

million in a year. This estimate can guide budget planning, allowing programs to allocate resources up to this amount while maintaining a positive return on investment. Future iterations of this analysis could incorporate cost parameters and extend to multiyear projections, enhancing its utility for implementation planning.

Limitations

We acknowledge the potential for selection bias arising from the exclusion of patients with missing variables. Complete case analysis was used because data imputation was not performed, and several of the machine learning algorithms applied in this study do not natively support missing values. Nevertheless, a substantial proportion of the registry was retained: 74.6% (100,500/134,670) of eligible patients were included in training the binary base learners, 77.3% (108,886/140,859) in training the stacking ensemble models, and 80.7% (111,004/137,584) in the validation dataset as detailed in our previous study [11]. Patients excluded due to missing data were more likely to have limited interaction with the health care system (eg, no inpatient, ED, or primary care visits), and their exclusion is unlikely to materially affect model performance for predicting higher levels of usage. However, we acknowledge that missingness may be informative, and future work could explore multiple imputation strategies or algorithms that explicitly handle missing data to further assess the robustness and generalizability of the findings.

A methodological limitation is the potential for logical inconsistencies in the predicted probabilities generated by independently trained binary classifiers used as inputs to the stacking ensemble model. For instance, a patient may receive a higher predicted probability for LOS≥14 days than for LOS≥7 days, which violates logical ordering. While these probabilities were not used directly to assign final classes, serving instead as features for a stacking ensemble meta-learner, such inconsistencies could still affect interpretability and calibration. We did not formally assess the prevalence of these inconsistencies in this study. However, we acknowledge this limitation and plan to explore ordinal-consistent modeling strategies in future work, such as classifier chains or ordinal regression frameworks that preserve monotonicity across outcome thresholds.

Additionally, we did not benchmark against simpler baselines, such as heuristic rules (eg, previous year usage) or native multiclass algorithms. Future work could explore these comparisons to better understand the incremental benefits of ensemble architectures in terms of complexity, predictive performance, and operational interpretability.

Another methodological consideration is the absence of hyperparameter tuning, which may have improved absolute model performance. Our objective in this study was to evaluate relative model performance across ensemble strategies and to develop a transparent and reproducible framework. Default hyperparameters were used to ensure consistency and comparability. We acknowledge that in real-world deployment, targeted hyperparameter tuning would likely be performed to further optimize predictive accuracy, and we identify this as a direction for future work.

Relatedly, we did not perform standard decision curve analysis as our multiclass models output discrete class labels corresponding to stratified intervention tiers rather than risk probabilities used for binary decisions. Future work could explore ordinal or multiclass extensions of decision curve analysis that may better reflect stratified care models and potentially enhance decision support performance.

Additionally, in our economic simulations, we also assumed normally distributed variation in intervention benefit; while this was suitable for benchmarking, we acknowledge that future work could explore log-normal or gamma distributions.

A further consideration is the potential impact of pandemic-related disruptions during the study period, which coincided entirely with the COVID-19 pandemic (2020-2022). As the ensemble models were trained and validated on temporally distinct datasets from this period, unmeasured shifts in disease burden and health care delivery—such as deferred elective care, changes in admission criteria, infection control protocols, ED triage processes, and inpatient discharge practices—may have influenced LOS and ED usage patterns. While a detailed exploration of these pandemic-related dynamics is beyond the scope of this study, we acknowledge that they may have affected model performance and limited generalizability to nonpandemic settings. We intend to validate the models using postpandemic data (eg, 2024 onwards) in future work to assess robustness under more stable health care conditions.

Another important limitation is the real-world variability in health care usage among high users; the best models could correctly assign only about 30% of future users to the appropriate category. This variability, likely influenced by regression to the mean and the inherent complexity of the prediction task, highlights the need for additional program measures, such as downstream assessments of predicted patients.

From a pragmatic perspective, not all high users may respond to interventions, and real-world efficacy is likely to vary depending on clinical or social drivers [6]. Our economic impact analyses assumed average cost reductions across usage tiers to facilitate comparative evaluation of ensemble models, but we acknowledge this simplification may not capture the full heterogeneity of intervention effects. Future work could explore stratified modeling based on clinical or social risk profiles when such data becomes available. However, by predicting varying levels of resource use, our models allow for interventions to target patients who may be more amenable to them [9]. This remains an area requiring further research to pinpoint patients with modifiable trajectories [26]. It has been suggested that prediction models should combine quantitative prediction modeling with qualitative assessment for better case identification [6]. In line with this, we plan to explore the use of large language models to analyze unstructured clinical text and better understand social determinants of health [27], supporting a more tailored and prescriptive approach to patient management.

XSL•FO
RenderX

Conclusions

We used binary classification machine learning algorithms to develop ensemble models predicting future health care usage, focusing on inpatient LOS and ED visits. These models were evaluated using metrics relevant to population health programs. Economic and parameter sensitivity analyses of the best-performing models offered valuable insights for designing effective health initiatives. Our ensemble model could be integrated into a diverse decision-making framework, providing tailored recommendations for patients with predicted high health care usage.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation taxonomy (2025), proofreading and editing tasks were delegated to GenAI tools under full human supervision. The GenAI tool used was ChatGPT-4o (OpenAI). However, responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to Singapore Health Services Cluster Data Policy on data sharing restrictions. Data are available from the Singapore Health Services Diabetes Registry Disease Registry Committee for researchers who meet the criteria for access to confidential data. The criteria include institutional review board approval, data use approval, and a research collaboration agreement. The point of contact for the Singapore Health Services Diabetes Registry Disease Registry Committee is Lee Thong Shuen (email: lee.thong.shuen@singhealth.com.sg).

The underlying code for this study is not publicly available for proprietary reasons but may be made available to qualified researchers on reasonable request from the corresponding author.

Authors' Contributions

JKT: Study design, statistical analyses, model development, data interpretation, data visualization, manuscript drafting, and review.
LQ: Study design, model development, data interpretation, and manuscript review
HYT: Data interpretation and manuscript review.
S-YG: Study design, data interpretation, and manuscript review.
JT: Study design, data interpretation, manuscript drafting, and review.
MA: Data collection, data interpretation, manuscript drafting, and review.
YMB: Study design, data interpretation, manuscript drafting, and review.
JHT: Study design, model development, data interpretation, data visualization, and manuscript review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables and figures.
[DOCX File , 14757 KB - medinform_v14i1e77202_app1.docx]

Multimedia Appendix 2
TRIPOD+AI checklist.
[PDF File (Adobe PDF File), 3799 KB - medinform_v14i1e77202_app2.pdf]

References

1. OECD Health Statistics 2023 - OECD. OECD. URL: https://www.oecd.org/health/health-data.htm [accessed 2024-01-19]
2. Parker ED, Lin J, Mahoney T, Ume N, Yang G, Gabbay RA, et al. Economic costs of diabetes in the U.S. in 2022. Diabetes Care 2024;47(1):26-43. [doi: 10.2337/dci23-0085] [Medline: 37909353]

XSL•FO
RenderX

3. MEPs Mobilising for Diabetes Blueprint for Action on Diabetes in the European Union by 2030. MEPs Mobilising for Diabetes. URL: https://www.mepinterestgroupdiabetes.eu/wp-content/uploads/2021/03/MMD-BLUEPRINT-FOR-ACTION-ON-DIABETES.pdf [accessed 2024-09-25]
4. Percentage of spending on mental health, diabetes and ageing versus total healthcare spending in 2018 against amount budgeted for 2019. Ministry of Health Singapore. 2019. URL: https://www.moh.gov.sg/news-highlights/details/percentage-of-spending-on-mental-health-diabetes-and-ageing-versus-total-healthcare-spending-in-2018-against-amount-budgeted-for-2019/ [accessed 2024-09-25]
5. Wammes JJG, van der Wees PJ, Tanke MAC, Westert GP, Jeurissen PPT. Systematic review of high-cost patients' characteristics and healthcare utilisation. BMJ Open 2018 Sep 08;8(9):e023113 [FREE Full text] [doi: 10.1136/bmjopen-2018-023113] [Medline: 30196269]
6. de Ruijter UW, Kaplan ZLR, Bramer WM, Eijkenaar F, Nieboer D, van der Heide A, et al. Prediction models for future high-need high-cost healthcare use: a systematic review. J Gen Intern Med 2022;37(7):1763-1770 [FREE Full text] [doi: 10.1007/s11606-021-07333-z] [Medline: 35018571]
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
8. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35(29):1925-1931 [FREE Full text] [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
9. Madvig P, Pearl R. Managing the Most Expensive Patients. Harv Bus Rev 2020;5(February). URL: https://hbr.org/2020/01/managing-the-most-expensive-patients [accessed 2020-02-05]
10. Crowley C, Perloff J, Stuck A, Mechanic R. Challenges in predicting future high-cost patients for care management interventions. BMC Health Serv Res 2023;23(1):992 [FREE Full text] [doi: 10.1186/s12913-023-09957-9] [Medline: 37710262]
11. Tan JK, Quan L, Salim NNM, Tan JH, Goh S, Thumboo J, et al. Machine learning-based prediction for high health care utilizers by using a multi-institutional diabetes registry: model training and evaluation. JMIR AI 2024;3:e58463 [FREE Full text] [doi: 10.2196/58463] [Medline: 39418089]
12. Lim DYZ, Chia SY, Abdul Kadir H, Mohamed Salim NN, Bee YM. Establishment of the singhealth diabetes registry. Clin Epidemiol 2021;13:215-223 [FREE Full text] [doi: 10.2147/CLEP.S300663] [Medline: 33762850]
13. Max K, Jed W, Steve W, Andre W, Chris K. caret: Classification and Regression Training. URL: https://cran.r-project.org/web/packages/caret/caret.pdf [accessed 2026-01-17]
14. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms 1st Ed. Hall/CRC: Chapman; 2012.
15. Max K, Hadley W. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. URL: https://www.tidymodels.org/ [accessed 2026-01-17]
16. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. MRC Population Health Science Research Network 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
17. Plotly Technologies Inc. Collaborative Data Science. Montreal: Plotly Technologies Inc; 2015.
18. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378 [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
19. Monte-Carlo-Simulations-using-Ensemble-Model-for-Population-Health. Github. URL: https://joshuatankuan.github.io/Monte-Carlo-Simulations-using-Ensemble-Model-for-Population-Health/ [accessed 2026-01-26]
20. Tan JK, Zhang X, Cheng D, Leong IYO, Wong CS, Tey J, et al. Using the johns hopkins ACG case-mix system for population segmentation in a hospital-based adult patient population in singapore. BMJ Open 2023;13(3):e062786 [FREE Full text] [doi: 10.1136/bmjopen-2022-062786] [Medline: 36997258]
21. Ng SH, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019;19(1):452 [FREE Full text] [doi: 10.1186/s12913-019-4239-2] [Medline: 31277649]
22. Johnson TL, Rinehart DJ, Durfee J, Brewer D, Batal H, Blum J, et al. For many patients who use large amounts of health care services, the need is intense yet temporary. Health Aff (Millwood) 2015;34(8):1312-1319. [doi: 10.1377/hlthaff.2014.1186] [Medline: 26240244]
23. Rahman Y, Dua P. A machine learning framework for predicting healthcare utilization and risk factors. Healthcare Analytics 2025;8:100411. [doi: 10.1016/j.health.2025.100411]
24. Li Z, Gogia S, Tatem K, Cooke C, Singer J, Chokshi D, et al. Developing a Model to Predict High Health Care Utilization Among Patients in a New York City Safety Net System. Med Care 2022;61(2):102-108 [FREE Full text] [doi: 10.1097/MLR.0000000000001807] [Medline: 36472326]
25. Theissler A, Thomas M, Burch M, Gerschner F. ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices. Knowledge-Based Systems 2022;247:108651. [doi: 10.1016/j.knosys.2022.108651]

XSL•FO
RenderX

26. Berkman ND, Chang E, Seibert J, Ali R. Characteristics of High-Need, High-Cost Patients : A "Best-Fit" Framework Synthesis. Ann Intern Med 2022;175(12):1728-1742 [FREE Full text] [doi: 10.7326/M21-4562] [Medline: 36343343]
27. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024;7(1):6 [FREE Full text] [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]

Abbreviations

AUROC: area under the receiver operating characteristic curve
ED: emergency department
HNHC: high-need, high-cost
LOS: length of stay
SDR: Singapore Health Services Diabetes Registry
SVM: support vector machine
TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis + Artificial Intelligence

Edited by A Coristine, A Benis; submitted 09.May.2025; peer-reviewed by E Bai, G Lim; comments to author 06.Oct.2025; revised version received 02.Jan.2026; accepted 02.Jan.2026; published 20.Feb.2026.

Please cite as:
Tan JK, Quan L, Tan HY, Goh SY, Thumboo J, Au M, Bee YM, Tan JH
Ensemble Machine Learning Models for Predicting Patients With High Usage: Model Validation and Economic Impact Analysis
JMIR Med Inform 2026;14:e77202
URL: https://medinform.jmir.org/2026/1/e77202
doi:10.2196/77202
PMID:

©Joshua Kuan Tan, Le Quan, Hao Yi Tan, Su-Yen Goh, Julian Thumboo, Marianne Au, Yong Mong Bee, Jen Hong Tan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhanced Magnetic Resonance Imaging–Based Knee Cartilage Segmentation Using a Swin-UNet Conditional Generative Adversarial Network: Development and Validation Study

Jun Young Park[1*], MD, PhD; Ji-Hoon Nam[2,3,4*], MSc; Shakhboz Abdigapporov[3], MSc; Jong-Keun Kim[5], MD, PhD; Yong-Gon Koh[6], MD; Byung Woo Cho[7], MD, PhD; Hyuck Min Kwon[8], MD, PhD; Kwan Kyu Park[8], MD, PhD; Kyoung-Tak Kang[2,3,4], PhD

[1]Department of Orthopaedic Surgery, Yonsei University College of Medicine, Yongin Severance Hospital, Yongin-si, Gyeonggi-do, Republic of Korea
[2]Department of Mechanical Engineering, Yonsei University, Seoul, Seoul, Republic of Korea
[3]Clevion Co., Ltd., Seoul, Republic of Korea
[4]Skyve R&D LAB, Skyve Co., Ltd., Seoul, Republic of Korea
[5]Department of Orthopaedic Surgery, HeungK Hospital, Seoul, Republic of Korea
[6]Department of Orthopaedic Surgery, Joint Reconstruction Center, Yonsei Sarang Hospital, Seoul, Republic of Korea
[7]Department of Orthopaedic Surgery, Yonsei University College of Medicine, Gangnam Severance Hospital, Seoul, Republic of Korea
[8]Department of Orthopaedic Surgery, Yonsei University College of Medicine, Severance Hospital, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Kyoung-Tak Kang, PhD
Department of Mechanical Engineering
Yonsei University
50 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2123 4827
Fax: 82 2 362 2736
Email: tagi1024@gmail.com

Abstract

Background: Accurate segmentation of cartilage from magnetic resonance imaging (MRI) is crucial for the diagnosis and surgical planning of knee osteoarthritis. However, manual segmentation is time-consuming, and conventional computed tomography–based surgical systems are limited by their inability to visualize cartilage.

Objective: This study aimed to develop a clinically targeted deep learning framework, the Swin-UNet conditional generative adversarial network (cGAN), for the automatic segmentation of femoral and tibial cartilage in MRI. We then evaluated its performance against conventional UNet, UNet cGAN, and Swin-UNet baseline models.

Methods: Our dataset comprised 232 knee MRI scans. We conducted quantitative experiments on the proposed Swin-UNet cGAN model and compared the results with those of widely used UNet, UNet cGAN, and Swin-UNet models for femoral and tibial cartilage segmentation, using the Dice similarity coefficient, mean intersection over union, 95th percentile Hausdorff distance, and average symmetric surface distance. All performance metrics were statistically analyzed. In addition, the performance of the Swin-UNet cGAN model was evaluated on an external validation dataset.

Results: The proposed Swin-UNet cGAN achieved the highest mean Dice similarity coefficient and intersection over union scores for both femoral and tibial cartilage segmentation, demonstrating performance statistically comparable to the best-performing baseline (UNet) in the tibia. Regarding distance metrics (average symmetric surface distance and 95th percentile Hausdorff distance), the proposed model significantly outperformed all baselines in the tibia while achieving results comparable to the UNet cGAN in the femur. It also maintained consistently high segmentation performance on both the internal test set and an external validation dataset.

XSL•FO
RenderX

Conclusions: These findings indicate that the proposed Swin-UNet cGAN achieves more accurate knee cartilage segmentation than UNet, UNet cGAN, and Swin-UNet, particularly in terms of boundary accuracy, while maintaining promising generalizability performance across both internal testing and external validation cohorts. This MRI-based deep learning approach addresses critical limitations of computed tomography–based patient-specific instrumentation systems by providing cartilage visualization, potentially improving surgical precision and outcomes in total knee arthroplasty.

(JMIR Med Inform 2026;14:e86155) doi:10.2196/86155

KEYWORDS

cartilage; deep learning; generative adversarial network; knee cartilage segmentation; magnetic resonance imaging; MRI; segmentation; swim-UNet

Introduction

Knee joint osteoarthritis is a progressive degenerative disorder affecting the articular cartilage, underlying subchondral bone, and surrounding periarticular tissues. It presents with pain, stiffness, and reduced mobility, ranking among the leading causes of disability in adults [1,2]. Timely identification and precise quantification of alterations in cartilage structure and composition are essential for optimal management of knee osteoarthritis [3].

An extensive array of imaging modalities is used in clinical and research settings to deepen understanding of osteoarthritis pathogenesis [4]. Magnetic resonance imaging (MRI) offers superior visualization of cartilage structure and is the modality of choice for assessing cartilage damage and morphological alterations. In addition, MRI provides a noninvasive, radiation-free examination and excels at evaluating overall knee pathology, owing to its superior soft-tissue contrast relative to X-ray and computed tomography (CT) [5].

Quantitative analysis of cartilage morphology depends on accurate medical-image segmentation and visualization of the articular cartilage, with subsequent 3D reconstruction enabling measurement of thickness, volume, and surface area [6]. However, this process is complex and time-consuming, as each image slice must be segmented individually using manual segmentation software [7]. In recent years, deep learning (DL)–based artificial intelligence models, particularly convolutional neural network (CNN) architectures, have been adopted as a cutting-edge solution for knee cartilage segmentation and osteoarthritis diagnosis [8]. CNN-based models have shown strong potential for accurately segmenting cartilage and bone across imaging modalities [9].

Toit et al [9] studied a modified UNet algorithm to automatically segment the femoral articular cartilage in 3D ultrasound images of healthy volunteer knees. Wang et al [10] proposed a novel network, Patch Attention UNet, to enhance knee cartilage segmentation by applying an attention mechanism; it was evaluated on multiple datasets. Recently, generative adversarial networks (GANs) have captivated the research community, with several studies demonstrating that GAN-generated delineations can match manual tracings and materially enhance segmentation performance [11]. Gaj et al [11] developed a conditional generative adversarial network (cGAN) model built on the UNet architecture that delivers fully automated, highly accurate segmentation of knee cartilage and meniscus.

Therefore, our objective was to develop a Swin-UNet cGAN, a tailored adversarial framework merging a hierarchical Swin-UNet generator with a pixel-wise CNN discriminator for 2D segmentation of knee joint cartilage, and to evaluate its performance on both an internal test set and an external validation dataset. We compared the segmentation performance of our proposed model with traditional UNet, UNet cGAN, and Swin-UNet models using the Dice similarity coefficient (DSC), mean Intersection over Union (IoU), average symmetric surface distance (ASSD), and 95th percentile Hausdorff distance (HD95).

We hypothesized that the Swin-UNet cGAN model would achieve the highest performance in segmenting femoral and tibial cartilage.

Methods

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of Yonsei Sarang Hospital (IRB number 2025-11-002) and Yongin Severance Hospital (IRB number 9-2025-0189). The requirement for informed consent was waived by the IRB owing to the retrospective nature of this study and the use of deidentified imaging data. All procedures involving human participants were conducted in accordance with institutional ethical standards and approved by the relevant IRBs. All patient data were fully anonymized prior to analysis, with no possibility of individual patient identification. Patients received no compensation for the use of their deidentified MRI data in this research. No identifiable patient information or images are included in this manuscript or its supplementary materials.

Dataset

The internal dataset was collected from Yonsei Sarang Hospital and consisted of 232 MRI scans from patients with knee osteoarthritis scheduled for patient-specific instrument (PSI)–guided total knee arthroplasty (TKA) from October 2023 through December 2024. All internal examinations were performed on a 3.0 T system (uMR 770; United Imaging Healthcare) using a 3D high-resolution T2*-weighted gradient-echo sequence (repetition time [TR]/echo time [TE]=17/9.3 ms; flip angle=25°), with a 20-cm field of view and 1-mm thick sagittal slices oriented relative to the posterior femoral condyles. This imaging protocol is commonly used when fabricating PSI [12]. Each scan comprised 90-110 slices. A medical imaging specialist manually segmented cartilage on every slice. The 232 scans were randomly split at the case level:

XSL•FO
RenderX

a total of 162 (70%) scans were assigned to the training set, 35 (15%) scans to the validation set, and 35 (15%) scans to the testing set. For model development, the original 512 × 512 pixel images and their corresponding masks were down-sampled to 224 × 224 pixels.

For external validation, an independent dataset was obtained from Yongin Severance Hospital, consisting of 25 consecutive PSI-guided TKA cases. All external validation knee MRI examinations were performed on a 3.0 T system (SIGNA Premier; GE Healthcare) with a 20-cm field of view and 2-mm thick sagittal slices oriented relative to the posterior femoral condyles. TR and TE were automatically adjusted by the scanner to optimize image quality, with nominal target values of TR/TE=30/11 ms.

Swin-UNet–Based cGAN Model

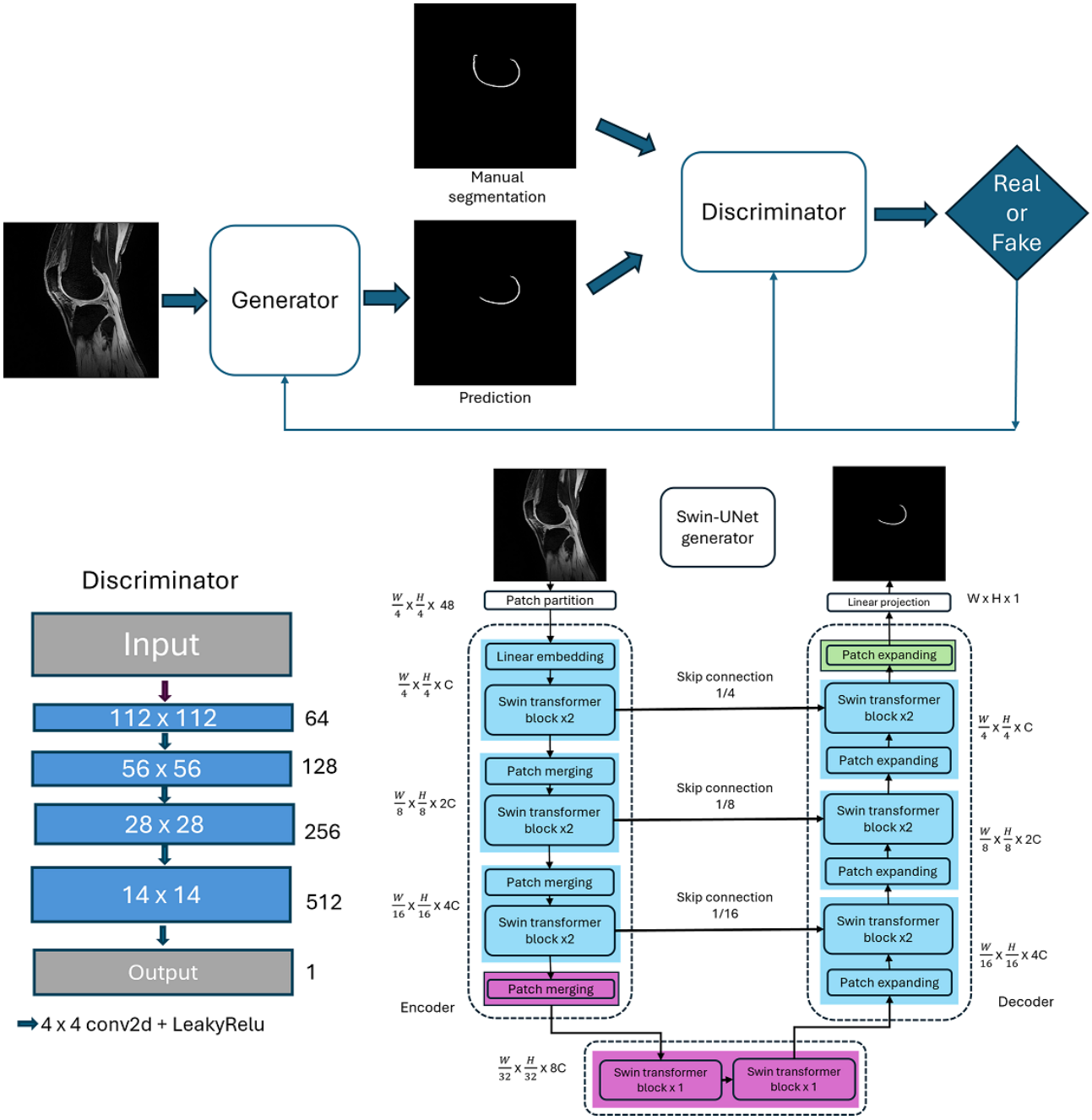
GAN-based networks usually consist of 2 parts: a generator model that produces synthetic samples from noise with a prior distribution and a discriminator model that acts as a binary classifier to differentiate real data from generated synthetic samples. In this study, the widely used cGAN [13] was adopted for MRI knee bone and cartilage segmentation. In this work, the generator is a mapping function from a source domain image into a segmentation mask, rather than mapping noise to real data as in the original GAN model, and predicts a manual segmentation mask. The cGAN-based model uses the discriminator to judge the quality of segmentation given the source image from the dataset. The binary classifier distinguishes the differences between the manual segmentation mask and the synthetic segmentation mask generated by the generator model, designating them as real and fake samples, respectively. This study introduces a Swin-UNet–based cGAN for knee MRI segmentation [14] to create synthetic masks and a PatchGAN discriminator [15] that learns discriminative features to distinguish generated masks from manual annotations.

Generator of the Swin-UNet–Based cGAN Model

The generator of the knee cartilage segmentation model is a hierarchical Swin Transformer architecture integrated into a UNet-like encoder-decoder structure, as shown in Figure 1. The input image size and patch size are set as 224 × 224 and 4, respectively. The C-dimensional (C=32) tokenized inputs with a resolution of H/4 × W/4 are fed into 2 consecutive Swin Transformer blocks (using a 7 × 7 attention window) to perform representation learning, and the feature dimension and resolution remain unchanged. The patch merging layer reduces the number of tokens (down-samples by 2×) and increases the feature dimension to 2× the original dimension. The encoder repeats this procedure 3 times. The patch merging layer in the encoder divides the input patches into 4 parts and concatenates them together. The feature resolution is down-sampled by 2×, and because the concatenate operation increases the feature dimension by 4×, a linear layer is applied on the concatenated features to unify the feature dimension to 2× the original dimension. The bottleneck consists of 2 Swin Transformer blocks and is used to learn deep feature representation. In the bottleneck, the feature resolution and dimension remain unchanged.

XSL•FO
RenderX

Figure 1. Overall architecture of the proposed network.



The decoder has a symmetrical architecture to the encoder, based on the Swin Transformer blocks. The patch expanding layer in the decoder is used to up-sample the extracted deep features. The patch expanding layer reshapes the feature maps of adjacent dimensions into higher-resolution feature maps, up-sampling by 2×, and reduces the feature dimension to half of the original dimension.

Skip connections are used to fuse the multiscale features from the encoder with the up-sampled features. Both shallow and deep features are concatenated together to reduce the loss of spatial information caused by down-sampling.

Discriminator of the Swin-UNet–Based cGAN Model

The discriminator of the proposed model is a PixelGAN-style convolutional binary classifier of "fake" predicted segmented masks and "real" manually segmented masks. The network input has 2 channels (image + mask) and spatial size H × W. The discriminator network architecture consists of four 4 × 4 convolutional layers with LeakyRelu activation function. Normalization is applied to stabilize the training samples, and Dropout2d prevents the discriminator from overpowering the generator in the early stages and encourages robust feature learning. The final layer is a 4 × 4 stride convolution, which produces a grid of real or fake logits. With 224 × 224 input, the discriminator produces a 13 × 13 authenticity map.

In the proposed model, a linearly increasing adversarial weight schedule was used, with λ starting at 0 after pretraining and gradually increasing to 1×10^{-4} over the GAN training epochs. This prevented gradient shocks and was selected as the most stable λ based on validation stability. Prior to selecting λ, larger values of 5×10^{-4} and 1×10^{-3} were empirically explored;

XSL•FO
RenderX

although these values increased the adversarial influence, they led to unstable training.

Loss Function of the Swin-UNet–Based cGAN Model

The proposed knee cartilage segmentation model uses a combination of supervised segmentation loss and adversarial losses. If segmentation performance mainly relies on adversarial loss, the output segmentation mask might not always share the same global structure as the manually segmented mask [16-19]. For traditional segmentation losses, the DSC and cross-entropy can be applied as loss functions for the generator network to improve performance by decreasing the distance between generated "fake" and manually segmented "real" masks. The proposed model uses multiple loss functions during training because the training experiments were conducted in multiple stages: generator model pretraining for segmentation mask predictions and overall training of the developed Swin-UNet cGAN–based knee segmentation model. During both the pretraining and overall model training phases, binary cross-entropy (BCE) with logits loss is adopted as the primary segmentation loss.

where yi is the ground truth mask of the real samples and is the mask predicted by the generator network. For adversarial loss, a hinge loss formulation is applied to improve training stability.

The discriminator network is trained using hinge loss to improve classification of real and fake sample data. For real mask samples, the discriminator is set to output values greater than 1; for fake samples, output are less than –1. The overall loss is computed as the mean of the 2 hinge components. The overall discriminator loss can be written as follows.

In the proposed model training, a gradient penalty was implemented to stabilize training by penalizing deviations in gradient norm from 1. Finally, the generator's overall loss is composed of the total segmentation loss and a weighted adversarial loss.

For the proposed model training, the Adam optimizer was used with a learning rate of 1×10^{-4} for the transformer-based generator and 5×10^{-5} for the patch-based discriminator networks, with a weight decay of 1×10^{-5}. To maximize the prediction results of the segmentation model, the generator was first pretrained for 30 epochs using only the segmentation loss (BCE with logits loss). After pretraining, adversarial training was performed for 40 epochs using a combination of segmentation and adversarial losses. During overall model training, adversarial weight λ_{adv} increased linearly until it reached the final value of 1×10^{-4}. Training of all experimented models was performed using an NVIDIA GeForce RTX 4090 24 GB graphics processing unit.

Evaluation Metrics

The performance of the knee cartilage segmentation models was compared with that of the Swin-UNet segmentation model, the UNet-based GAN segmentation model [11], and the traditional CNN-based UNet model using the DSC, IoU, HD95, and ASSD. The DSC represent the overlap between the predicted and ground truth masks. A DSC of 1 (or 100%) represents perfect overlap, whereas a value of 0 indicates no overlap between the ground truth and predicted masks.

The mean IoU metric is calculated by dividing the area of intersection of the predicted and ground truth masks by the area of the union of 2 masks.

The Hausdorff distance (HD) measures the extent to which 2 subsets of a metric space differ. Because HD is sensitive to outliers, the 95th percentile (HD95) is frequently used instead. By considering the 95th percentile of distances, HD95 provides a more robust measure of boundary conformity, effectively ignoring the most extreme 5% of outliers.

where A and B denote the sets of boundary points for the predicted and ground truth segmentations, respectively; ||a – b|| represents the Euclidean distance between points a and b; and P_{95} denotes the 95th percentile of the accumulated distances.

The ASSD measures the average distance between boundary points, making it highly sensitive to the precise alignment of contours.

where |A| and |B| denote the cardinality of the boundary sets A and B, respectively. All surface metrics (ASSD and HD95) were computed on 224 × 224 masks, the same resolution used for training and inference. Distances were converted to millimeters using the effective in-plane spacing of 0.89 mm/pixel.

Reliability of Manual Segmentation

For consistent and reliable manual labeling of our dataset, DSC and IoU metrics were calculated for intra- and interrater comparisons across femur and tibia segmentations. Two experienced manual data labeling experts, rater A and rater B, segmented the entire dataset. For reliability assessment, MRI data from 30 randomly selected patients were labeled a second time by both raters at a different time of day than the first labeling session. For the final training and test datasets, segmentation masks created by a single expert rater were used for each slice.

XSL•FO
RenderX

Ablation Study

To empirically validate the architectural decisions and hyperparameter configurations of the proposed Swin-UNet cGAN framework, a comprehensive ablation study was designed focusing on 3 critical components. First, model sensitivity to the magnitude of the adversarial loss was evaluated by testing varying adversarial weights ranging from 0.0001 to 0.1 to determine the optimal balance between pixel-wise segmentation accuracy and adversarial shape refinement. Second, to assess the impact of training stability, a constant weight schedule was compared against a linear warm-up strategy, in which adversarial weights increased linearly from zero to the target value over the training epochs. Third, to identify the most effective objective function for segmenting fine cartilage structures within the adversarial framework, standard BCE loss alone was compared with a hybrid objective combining BCE and Dice loss.

Statistical Analysis

Statistical analyses were performed using R (version 4.3.1; R Foundation for Statistical Computing). For multiple comparisons of the results, the normality of the data distribution was first assessed. Data following a normal distribution were analyzed using repeated measures ANOVA, whereas nonnormally distributed data were analyzed using the Friedman test. Post hoc comparisons were conducted using the Bonferroni correction. Inter- and intrarater reliability were evaluated using the DSC. A *P* value of less than .05 was considered statistically significant.

Results

Cartilage Segmentation Results

DSC, mean IoU, ASSD, and HD95 results for femoral and tibial cartilage segmentation are summarized in Table 1.

Table 1. Knee cartilage segmentation results.

Structure and metric	Model, mean (SD)				*P* value
	UNet	UNet cGAN[a]	Swin-UNet	Proposed (Swin-UNet cGAN)	
Femur					<.001
DSC[b] (%)	78.6 (3.1)[c]	80.2 (1.2)[d]	76.5 (4.5)[e]	82 (1.5)[f]	
IoU[g] (%)	69.4 (3.9)[c]	69.9 (1.8)[c]	65.3 (5.2)[e]	71.9 (2.1)[d]	
ASSD[h] (mm)	1.34 (0.32)[e]	0.24 (0.05)[d]	0.58 (0.17)[c]	0.22 (0.08)[d]	
HD95[i] (mm)	5.14 (1.64)[e]	1.25 (0.21)[d]	2.94 (0.96)[c]	1.31 (0.27)[d]	
Tibia					<.001
DSC (%)	79.2 (6.3)[cd]	80 (1.8)[c]	70.5 (8.6)[e]	81 (1.7)[d]	
IoU (%)	69.7 (7.3)[cd]	70.1 (2.5)[c]	59.5 (8.5)[e]	71.1 (2.1)[d]	
ASSD (mm)	1.05 (1.46)[e]	0.32 (0.27)[d]	0.83 (0.83)[c]	0.23 (0.15)[f]	
HD95 (mm)	3.62 (5.62)[e]	1.46 (1.13)[d]	3.33 (2.89)[c]	1.25 (0.53)[f]	

[a]cGAN: conditional generative adversarial network.

[b]DSC: Dice similarity coefficient.

[c-f]Footnotes c, d, e, and f indicate significant differences between models based on post hoc pairwise comparisons (*P*<.05). Superscript letters follow the journal style (alphabetical order by first appearance) and do not represent a rank order of performance. Means sharing the same footnote are not significantly different.

[g]IoU: intersection over union.

[h]ASSD: average symmetric surface distance.

[i]HD95: 95th percentile Hausdorff distance.

For femoral cartilage, the baseline UNet achieved a mean DSC of 78.6% (SD 3.1%) and an IoU of 69.4% (SD 3.9%), with an ASSD of 1.34 (SD 0.32) mm and an HD95 of 5.14 (SD 1.64) mm. Adding adversarial training in UNet cGAN slightly improved overlap (DSC 80.2%, SD 1.2%; IoU 69.9%, SD 1.8%; *P*<.001 vs UNet) and markedly reduced surface errors (ASSD 0.24, SD 0.05 mm; HD95 1.25, SD 0.21 mm; *P*<.001 vs UNet). Swin-UNet without adversarial training underperformed both UNet and UNet cGAN in overlap metrics (DSC 76.5%, SD 4.5%; IoU 65.3%, SD 5.2%) and showed only intermediate surface accuracy (ASSD 0.58, SD 0.17 mm; HD95 2.94, SD 0.96 mm). The proposed Swin-UNet cGAN achieved the best femoral cartilage segmentation, with a DSC of 82% (1.5%) and an IoU of 71.9% (SD 2.1%; *P*<.001 vs all other models), and significantly lower ASSD and HD95 than UNet and Swin-UNet (*P*<.001), while maintaining surface accuracy comparable to UNet cGAN.

For tibial cartilage, the baseline UNet yielded a DSC of 79.2% (SD 6.3%) and an IoU of 69.7% (SD 7.3%), with an ASSD of 1.05 (SD 1.46) mm and an HD95 of 3.62 (SD 5.62) mm. UNet cGAN preserved similar overlap (DSC 80%, SD 1.8%; IoU 70.1%, SD 2.5%) but substantially reduced surface errors (ASSD 0.32, SD 0.27 mm; HD95 1.46, SD 1.13 mm). Swin-UNet again

XSL•FO
RenderX

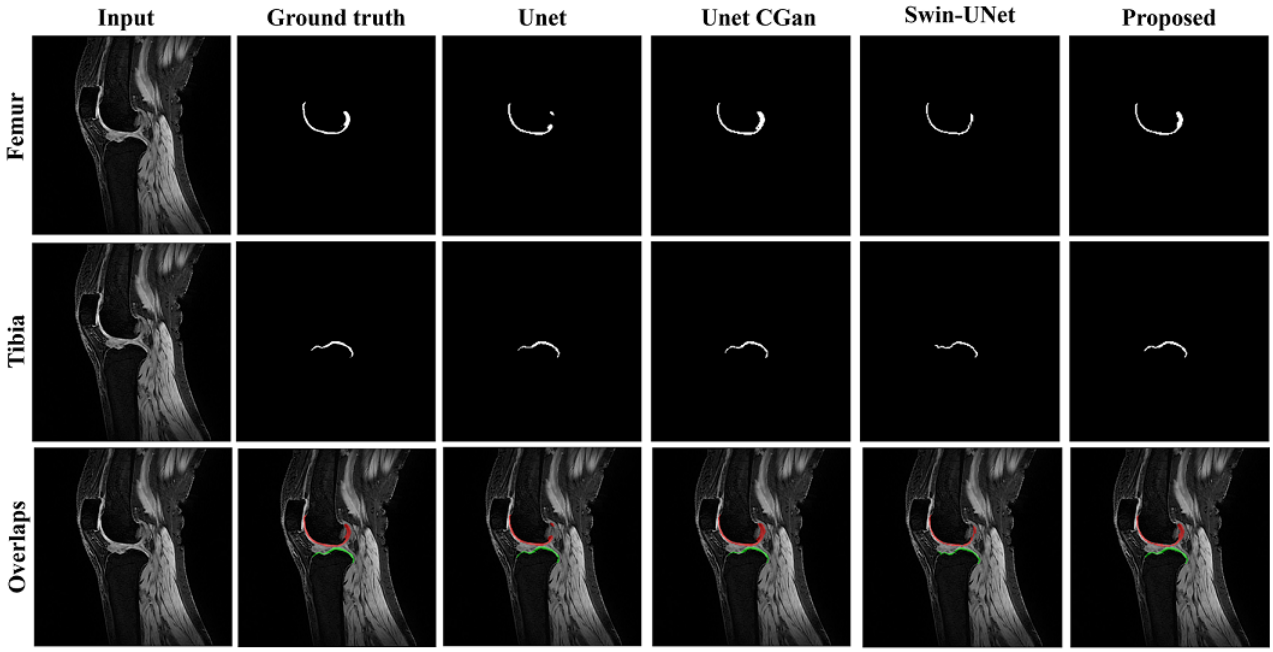
showed the weakest performance (DSC 70.5%, SD 8.6%; IoU 59.5%, SD 8.5%; ASSD 0.83, SD 0.83 mm; HD95 3.33, SD 2.89 mm). The proposed Swin-UNet cGAN yielded the highest overall tibial performance, achieving a DSC of 81% (SD 1.7%) and an IoU of 71.1% (SD 2.1%). While it showed significantly higher overlap than both Swin-UNet and UNet cGAN (P<.001), the difference compared with the standard UNet was not statistically significant. Importantly, it yielded the lowest ASSD (0.23, SD 0.15 mm) and HD95 (1.25, SD 0.53 mm) among all models (P<.001 vs all baselines). Taken together, the progression from UNet to UNet cGAN, Swin-UNet, and Swin-UNet cGAN effectively serves as an ablation series that disentangles the contributions of adversarial training and the Swin-Transformer backbone.

The proposed Swin-UNet cGAN achieved highly efficient inference performance. On an NVIDIA GeForce RTX 4090, the average inference time was 16.9 ms per 2D slice, culminating in a total processing time of 1.68 seconds per patient volume. This corresponds to a throughput of approximately 59 slices per second, demonstrating the model's suitability for near-real-time clinical workflows.

Qualitative Comparison of Segmentation Outputs

Figure 2 illustrates qualitative segmentation outputs of the models on unseen test data. The rows correspond to the femur, tibia, and overlaps (segmentation masks overlaid on the input MRI), respectively. The columns display the Input image, ground truth, and predictions from UNet, UNet cGAN, Swin-UNet, and our proposed Swin-UNet cGAN (Ours). As shown in the "Overlaps" row, the proposed model produces the most visually accurate segmentation for both bone and cartilage compared with the baseline models.

Figure 2. Qualitative comparison results of our proposed model and baseline models on the knee cartilage segmentation dataset.



Reliability of Manual Segmentation

Inter- and intrarater agreement calculations for DSC and IoU scores comparing overlap between labeled masks showed high reliability on the original dataset. For intrarater agreement, DSCs for rater 1 and rater 2 averaged 98.9% and 99%, respectively, across femur and tibia. The IoU scores also were high at 98% and 98.3%, demonstrating excellent internal consistency for labeling MRI data. For interrater agreement, the DSC was 95.5% and the mean IoU was 91.9%, indicating greater than 90% agreement between the 2 raters.

External Validation

Quantitative evaluation results on the external validation dataset are summarized. The proposed model demonstrated consistent segmentation performance across both femoral and tibial structures. For the femur, the model achieved a DSC of 73.1% (SD 6.9%) and an IoU of 65% (SD 6.9%). In terms of boundary delineation accuracy, ASSD was 1.30 (SD 0.70) mm and the HD95 was 5.91 (SD 2.93) mm. For the tibia, the model yielded a DSC of 69.7% (SD 7.4%) and an IoU of 62.6% (SD 7.2%). The distance-based metrics showed similar precision to the femur, with an ASSD of 2.16 (SD 0.94) mm and an HD95 of 7.18 (SD 2.64) mm.

Ablation Study on Model Component

A comprehensive ablation study was conducted to optimize the Swin-UNet cGAN configuration, as summarized in Table 2. First, regarding the adversarial weight, a value of 0.0001 yielded the best performance, whereas increasing the weight led to instability and performance degradation. Second, the linear warm-up strategy outperformed the constant schedule, improving tibial metrics and training stability. Although femoral surface errors were marginally higher with warm-up, this trade-off was negligible given the overall gains. Finally, BCE loss alone outperformed the combined BCE + Dice loss, likely because the discriminator sufficiently enforces global shape constraints, making the additional geometric guidance from Dice loss redundant.

XSL•FO
RenderX

Table 2. Ablation study results of the proposed Swin-UNet conditional generative adversarial network framework analyzing the impact of adversarial weight, scheduling strategies, and segmentation loss functions on segmentation performance.

Component and configuration	Femoral cartilage, mean (SD)				Tibial cartilage, mean (SD)			
	DSC[a] (%)	IoU[b] (%)	ASSD[c] (mm)	HD95[d] (mm)	DSC (%)	IoU (%)	ASSD (mm)	HD95 (mm)
Adversarial weight								
0.0001	81.67 (1.3)	71.52 (1.9)	0.21 (0.05)	1.25 (0.2)	80.55 (1.9)	70.62 (2.3)	0.27 (0.2)	1.35 (0.75)
0.001	80.01 (3.06)	69.58 (3.4)	0.24 (0.05)	1.30 (0.22)	78.63 (2.05)	68.47 (2.5)	0.38 (0.26)	1.64 (0.99)
0.01	76.75 (2.5)	65.53 (3.03)	0.3 (0.05)	1.41 (0.2)	69.61 (3.3)	58.25 (3.8)	0.83 (0.32)	2.9 (1.01)
0.1	55.48 (2.6)	43.02 (2.3)	2.21 (0.53)	3.25 (1.22)	42.92 (3.14)	30.6 (2.8)	3.51 (1.03)	9.79 (3.43)
Scheduling								
Constant	81.67 (1.3)	71.52 (1.9)	0.21 (0.05)	1.25 (0.2)	80.55 (1.9)	70.62 (2.3)	0.27 (0.2)	1.35 (0.75)
Warm-up	82 (1.5)	71.9 (2.1)	0.22 (0.08)	1.31 (0.27)	81 (1.7)	71.1 (2.1)	0.23 (0.15)	1.25 (0.53)
Segmentation loss								
BCE[e]	82 (1.5)	71.9 (2.1)	0.22 (0.08)	1.31 (0.27)	81 (1.7)	71.1 (2.1)	0.23 (0.15)	1.25 (0.53)
BCE + Dice	81.28 (1.5	70.88 (2.2)	0.27 (0.04)	1.43 (0.2)	79.17 (2.28)	68.84 (2.8)	0.39 (0.32)	1.73 (1.3)

[a]DSC: Dice similarity coefficient.

[b]IoU: intersection over union.

[c]ASSD: average symmetric surface distance.

[d]HD95: 95th percentile Hausdorff distance.

[e]BCE: binary cross-entropy.

Discussion

The most important finding of this study is that the proposed Swin-UNet cGAN model provided overall strong knee cartilage segmentation compared with the baseline UNet, UNet cGAN, and Swin-UNet models. Specifically, our model achieved higher DSC and IoU while maintaining tibial mask overlap comparable to UNet, and demonstrated markedly lower ASSD and HD95 while maintaining femoral surface accuracy comparable to UNet cGAN. These findings confirm our hypothesis that the Swin-UNet cGAN model achieves the highest performance in segmenting femoral and tibial cartilage.

Our results demonstrate that combining hierarchical Swin self-attention with adversarial refinement effectively captures the fine-grained pixel information needed for segmentation of knee cartilage. By embedding the Swin-UNet generator into a cGAN framework, we achieve superior cartilage delineation: the hierarchical self-attention mechanism—leveraging windowed self-attention rather than relying solely on skip connections—accurately traces cartilage boundaries and anatomically curved regions. Moreover, the Swin transformer backbone constructs a multiscale feature hierarchy through patch merging and expansion, enabling the network to integrate both detailed structural cues and broader contextual information more effectively than CNN-based encoders such as UNet.

For femoral cartilage, the proposed Swin-UNet cGAN model showed a statistically significant improvement over all baseline models, including UNet, UNet cGAN, and Swin-UNet, with higher DSC and IoU and markedly lower ASSD and HD95 (P<.05), while maintaining comparable surface accuracy to UNet-cGAN. Notably, for tibial cartilage, UNet-cGAN showed a statistically significant improvement over plain UNet, whereas the difference between the proposed Swin-UNet cGAN and UNet, although numerically larger, did not reach statistical significance. This finding is likely related to the relatively large intercase variability of the UNet baseline and the limited sample size, which reduce statistical power to detect small additional gains. Nevertheless, the proposed model consistently achieved the highest mean performance and the lowest ASSD and HD95, indicating the best overall balance between volumetric overlap and surface accuracy. From a clinical perspective, MRI-based PSI relies on accurate 3D surface reconstruction rather than volumetric overlap alone. Therefore, the substantial reductions in ASSD and HD95 achieved by the proposed Swin-UNet cGAN are particularly meaningful, as they indicate more precise cartilage boundaries that are directly relevant for the design and fit of PSI cutting guides.

An interesting finding of this study is that the plain Swin-UNet achieved lower DSC and IoU than UNet for both femoral and tibial cartilage. This may be explained by the inherent structural advantages of convolutional architectures such as UNet, which are highly effective at extracting local image features and can be trained stably even with relatively small datasets. In contrast, transformer-based architectures such as Swin-UNet are designed to capture long-range global context but typically require large-scale training data to fully learn these relationships. Under the limited knee MRI dataset used in this study, plain Swin-UNet appears to have suffered from overfitting and unstable convergence, resulting in lower performance (76.5%) than UNet (79.2%).

XSL•FO
RenderX

In this context, the effectiveness of the proposed Swin-UNet cGAN becomes more evident. As shown in Table 1, although the standalone Swin-UNet performed poorly, its performance markedly improved to 82% when used as the generator within a cGAN framework, surpassing that of UNet (79.2%). This suggests that adversarial learning, through the discriminator, imposes strong anatomical shape constraints on the generator, thereby compensating for the data dependency of the transformer architecture and enabling more accurate cartilage segmentation.

The disproportionate improvement in surface metrics (ASSD and HD95) compared with overlap metrics (DSC and IoU) can be attributed to the fundamental differences in their sensitivity. While DSC and IoU measure volumetric overlap and are relatively insensitive to minor boundary fluctuations, distance-based metrics such as ASSD and HD95 directly quantify contour delineation precision. Adversarial training in the proposed framework explicitly optimizes boundary realism through the discriminator, thereby yielding substantial gains in boundary precision even when volumetric improvements appear modest.

To benchmark our method against the current state of the art, we evaluated nnUNet on the same dataset (Table S1 in Multimedia Appendix 1). While nnUNet provided competitive volumetric overlap, the Swin-UNet cGAN achieved statistically significant improvements in both DSC (P=.02) and IoU (P=.04). Crucially, nnUNet struggled with boundary precision in severe osteoarthritis cases, yielding high HD95 values (2.55 mm and 2.75 mm for femur and tibia, respectively). In contrast, the proposed model markedly reduced surface errors, achieving HD95 values of 1.31 mm and 1.25 mm (P<.001) and significantly lower ASSD (P<.001). This demonstrates that the adversarial framework offers superior robustness in capturing the fine, irregular boundaries of degenerated cartilage compared with the self-configuring UNet approach.

Previous research by Chen et al [20] demonstrated that using Segmentation of Knee Images 2010 (SKI10) data, their model achieved a 1% to 2% DSC improvement over nnUNet, with results of 89% for femur cartilage and 88% for tibia cartilage. Guo et al [21] created a dataset of 700 MRI scans by pooling data from Osteoarthritis Initiative (OAI; n=200), fastMRI (n=175), SKI10 (n=100), and an internal hospital database (n=225). They trained 4 distinct DL models—UNet, UNet++, residual UNet (ResUNet), and transformer UNet (TransUNet)—on this dataset with corresponding labeled cartilage. Among these, TransUNet showed superior performance, achieving DSCs of 82.3% for femur cartilage and 80.3% for tibia cartilage. Further innovation came from Gaj et al [11], who developed a DL model for cartilage and meniscus segmentation by integrating a conditional GAN with UNet. This approach led to notable performance improvements, with femur cartilage segmentation increasing from 84.2% to 89.7% (a greater than 5% gain) compared with traditional UNet. For tibia cartilage, the method increased medial cartilage segmentation from 77% to 86.1% and lateral cartilage from 86.2% to 91.8%, using OAI data. Gaj et al's [11] work clearly illustrates that cGANs can significantly enhance the capabilities of established models like UNet. A quantitative benchmarking table summarizing the datasets, methods, and metrics of these existing studies is presented in Table 3.

XSL•FO
RenderX

Table 3. Comparison of segmentation performance with previous methods.

Study (year) and method	Dataset	Performance
Liu et al [22] (2019)		
SUSAN	SKI10[a] (n=50) and internal (n=120)	• Femur (DSC[b]=81%; ASSD[c]=0.65 mm) • Tibia (DSC=75%; ASSD=0.73 mm)
Gaj et al [11] (2020)		
UNet cGAN[d]	OAI[e] (n=176)	• Femur (DSC=89.7%) • Tibia (DSC=91.8%)
Chen et al [20] (2022)		
3D DNN[f] with adversarial loss	SKI10 (n=100)	• Femur (DSC=89%) • Tibia (DSC=88%)
Guo et al [21] (2024)		
TransUNet[g]	OAI (n=200), fastMRI (n=175), SKI10 (n=100), and internal (n=225)	• Femur (DSC=82.3%) • Tibia (DSC=80.3%)
Wang et al [10] (2025)		
PA-UNet[h] (2D)	OAI-ZIB[i] (n=507)	• Femur (DSC=88.8%; HD95[j]=9.1 mm) • Tibia (DSC=82.7%; HD95=11.7 mm)
PA-UNet[h] (2D)	SKI10 (n=100)	• Femur (DSC=78.2%; HD95=16.4 mm) • Tibia (DSC=71.6%; HD95=18.7 mm)
Proposed framework		
Swin-UNet cGAN	Internal (n=232)	• Femur (DSC=82%; ASSD=0.22 mm; HD95=1.31 mm) • Tibia (DSC=81%; ASSD=0.23 mm; HD95=1.25 mm)

[a]SKI10: Segmentation of Knee Images 2010.

[b]DSC: Dice similarity coefficient.

[c]ASSD: average symmetric surface distance.

[d]cGAN: conditional generative adversarial network.

[e]OAI: Osteoarthritis Initiative.

[f]DNN: deep neural network.

[g]TransUNet: transformer UNet.

[h]PA-UNet: patch attention UNet.

[i]OAI-ZIB: Osteoarthritis Initiative–Zuse Institute Berlin.

[j]HD95: 95th percentile Hausdorff distance.

Beyond spatial overlap assessed by DSC, evaluating boundary precision is crucial for clinical applications such as PSI. Liu et al [22] reported ASSD values of 0.65 mm and 0.73 mm for femoral and tibial cartilage, respectively, using their SUSAN framework. Most recently, Wang et al [10] applied a patch attention UNet (PA-UNet) to the Osteoarthritis Initiative–Zuse Institute Berlin (OAI-ZIB) and SKI10 datasets. Although they achieved competitive DSCs, they reported HD values ranging from 9.1 mm to 18.7 mm, indicating potential challenges in handling outlier boundary pixels.

In this study, the proposed Swin-UNet cGAN demonstrated high surface accuracy, achieving ASSD values of 0.22 mm and 0.23 mm and HD95 values of 1.31 mm and 1.25 mm for femoral and tibial cartilage, respectively. Although direct quantitative comparison is challenging due to differences in datasets, these low error rates suggest that integrating the Swin Transformer with adversarial learning effectively refines segmentation boundaries and may offer superior precision for clinical workflows.

A common limitation observed in prior studies is the use of mixed datasets that combine MRI scans from both healthy individuals and patients without explicitly accounting for osteoarthritis progression. Because knee MRI segmentation is most frequently applied in the context of arthroplasty, we trained our DL model using MRI data from patients diagnosed with Kellgren-Lawrence Grades 3-4 who were already scheduled for TKA. This targeted methodology ensures that our model's findings are highly relevant and directly applicable to real-world clinical scenarios, particularly for patients with advanced osteoarthritis, for whom accurate segmentation is critical for surgical planning and improved patient outcomes.

In addition, a recently published study demonstrated that, compared with training on mixed datasets containing both

XSL•FO
RenderX

healthy participants and patients, training primarily on patient data reduced femoral cartilage DSC from 88.8% to 78.2% and tibial cartilage DSC from 82.7% to 71.6% [10].

In terms of clinical relevance, our results are expected to bring significant advancements not only in quantitative diagnostics during knee arthroscopy but also in computer-assisted surgery, including robotic-assisted TKA. Most contemporary robotic‐assisted procedures are predicated almost entirely on image‐based workflows using CT [23]. However, the patient-specific model is generated exclusively from CT data and includes only the bony anatomy; thus, all registration landmarks must be acquired directly on bone during robot-assisted TKA. Intraoperatively, this requirement necessitates a sharp probe to penetrate the overlying cartilage until the underlying bone surface is reached, ensuring accurate alignment. However, an MRI-derived model inherently preserves the cartilage in its segmentation, facilitating a more straightforward and intuitive registration process.

The clinical impact of incorporating cartilage information has been quantified in several comparative studies [24,25]. Additionally, recent meta-analyses have confirmed that MRI-based PSI demonstrates fewer alignment outliers and superior accuracy compared with CT-based techniques, especially in knees with irregular or worn cartilage [26,27]. The practical implications of the inability of CT to visualize cartilage have led to compensatory measures, such as assuming uniform cartilage thickness or manually removing residual cartilage intraoperatively, potentially compromising surgical precision. In addition, recent studies have shown that CT-based PSI can result in inaccurate alignment in varus deformity knees [28]. According to their study, the authors recommended adding approximately 2 mm of cartilage thickness to the planned lateral tibial resection value in CT-based PSI workflows for patients with varus deformity [28]. However, because cartilage thickness can vary substantially between individuals and across different regions of the joint, artificially increasing the resection thickness by a fixed 2 mm only on the lateral side may introduce design challenges and potential inaccuracies in PSI planning. In contrast, MRI-based PSI directly incorporates patient-specific cartilage thickness, thereby eliminating the need for such heuristic adjustments and potentially overcoming these limitations. Our Swin-UNet cGAN model directly addresses these well-documented limitations by providing accurate, automated segmentation of both femoral and tibial cartilage. This advancement has the potential to eliminate the alignment errors associated with CT-based PSI, representing a significant step toward more precise, reproducible surgical outcomes in TKA.

This study has several limitations. First, we performed segmentation of only the knee joint (distal femur and proximal tibia) cartilage. Future research should expand to include the ligaments and menisci. However, the MRI protocol optimized for PSI fabrication provides a suboptimal evaluation of ligamentous and meniscal structures. Second, the primary training and internal test datasets were acquired on a single 3.0 T MRI system (uMR 770; United Imaging Healthcare) using 1 PSI-optimized 3D T2*-weighted gradient-echo sequence. Although this relatively homogeneous acquisition protocol reduces within-study variability, it may limit the generalizability of our model to other scanner vendors, field strengths, coil configurations, and cartilage-focused imaging protocols beyond those used in this study. External validation was performed on a separate 3.0 T MRI system (SIGNA Premier; GE Healthcare), where segmentation performance was slightly lower than on the internal test set, likely reflecting a degree of domain shift related to differences in slice thickness (2 mm for the external dataset vs 1 mm for the internal dataset), with reduced through-plane resolution and increased partial-volume effects. Nevertheless, the model maintained reasonably high performance on this external cohort, indicating partial robustness to protocol variation. Larger multicenter studies with more heterogeneous MRI protocols and scanner types, together with additional training on datasets acquired with thicker slices and/or explicit domain-adaptation strategies, are needed to further narrow this performance gap and fully establish robustness across diverse clinical environments. Third, the current model uses a downsampled input resolution of 224 × 224 and performs slice-wise 2D segmentation, without explicitly leveraging full 3D spatial context, which is important for volumetric MRI analysis. This resolution choice was primarily driven by the need to manage the high memory complexity of the Swin Transformer backbone within a dual-network GAN framework during training, rather than solely by inference speed constraints. We acknowledge that this downsampling results in an approximate pixel spacing of 0.89 mm, which is relatively coarse compared with thin cartilage structures and may limit the capture of extremely fine details. However, contemporary PSI design workflows for TKA, including most commercial systems, are typically based on sagittal 2D cartilage contours that are subsequently reconstructed into 3D surfaces, making a 2D architecture practically compatible with current clinical and industrial pipelines. While our use of surface-based metrics (ASSD and HD95) provides an indirect assessment of 3D geometric accuracy, future work should extend the framework to 2.5D or 3D architectures with high resolution and directly evaluate interslice consistency and PSI template fit. Fourth, although our proposed model is more advanced than those in previous studies, it exhibited a slightly lower DSC. However, as noted above, this limitation can be attributed to our dataset, which consists of patients (Kellgren-Lawrence grades 3-4) awaiting TKA, where extensive cartilage damage hinders segmentation compared with that in a previous study [11]. Nonetheless, we demonstrated that our Swin-UNet cGAN outperforms the UNet cGAN used in previous studies, even when evaluated on this highly challenging dataset. Future work will explore self-supervised pretraining or advanced augmentation to further improve data efficiency and push performance on small‐sample domains.

The Swin-UNet cGAN integrates hierarchical shifted-window attention into its UNet-style transformer generator and uses a pixel-level CNN discriminator to refine outputs, enabling overall more effective segmentation of fine cartilage structures than the standard UNet, the UNet cGAN, and the Swin-UNet, and this advantage was consistently preserved on an external validation dataset. Clinically, this MRI-based segmentation approach has the potential to address key limitations of current CT-based PSI and robotic-assisted TKA workflows by providing

XSL•FO
RenderX

accurate visualization of femoral and tibial cartilage, enabling more anatomically faithful preoperative models, reducing alignment errors related to unmodeled cartilage thickness, and supporting more intuitive intraoperative registration on the articular surface.

Acknowledgments

During the preparation of this work, the authors used Gemini 3 in order to improve readability and language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional review board restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KTK
Methodology: JHN, SA
Formal analysis: SA
Data curation: JKK, YGK
Writing – original draft: JYP
Writing – review & editing: BWC, HMK
Supervision: KKP, KTK

Conflicts of Interest

KTK is the chief executive officer of Skyve and Clevion Co Ltd, but this did not influence the results of this study. JHN is an employee of Skyve and Clevion Co Ltd, but this also did not affect the results of this study. YGK is a shareholder of Skyve. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1
Quantitative comparison of segmentation performance between nnUNet and the proposed Swin-UNet cGAN.
[DOCX File , 25 KB - medinform_v14i1e86155_app1.docx]

References

1. Koh YG, Lee JA, Kim YS, Lee HY, Kim HJ, Kang KT. Optimal mechanical properties of a scaffold for cartilage regeneration using finite element analysis. J Tissue Eng 2019;10:1-10 [FREE Full text] [doi: 10.1177/2041731419832133] [Medline: 30834102]
2. Koh YG, Lee JA, Lee HY, Kim HJ, Kang KT. Biomechanical evaluation of the effect of mesenchymal stem cells on cartilage regeneration in knee joint osteoarthritis. Appl Sci (Basel) 2019;9(9):1868. [doi: 10.3390/app9091868]
3. Zhang Q, Geng J, Zhang M, Kan T, Wang L, Ai S. Cartilage morphometry and magnetic susceptibility measurement for knee osteoarthritis with automatic cartilage segmentation. Quant Imaging Med Surg 2023;13(6):3508-3521. [doi: 10.21037/qims-22-1245] [Medline: 37284124]
4. Hafezi-Nejad N, Demehri S, Guermazi A, Carrino JA. Osteoarthritis year in review 2017: updates on imaging advancements. Osteoarthritis Cartilage 2018;26(3):341-349 [FREE Full text] [doi: 10.1016/j.joca.2018.01.007] [Medline: 29330100]
5. Blumenkrantz G, Majumdar S. Quantitative magnetic resonance imaging of articular cartilage in osteoarthritis. Eur Cell Mater 2007;13:76-86 [FREE Full text] [doi: 10.22203/ecm.v013a08] [Medline: 17506024]
6. Eck BL, Yang M, Elias JJ, Winalski CS, Altahawi F, Subhas N. Quantitative MRI for evaluation of musculoskeletal disease: cartilage and muscle composition, joint inflammation, and biomechanics in osteoarthritis. Invest Radiol 2023;58(1):60-75 [FREE Full text] [doi: 10.1097/RLI.0000000000000909] [Medline: 36165880]
7. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. Annu Rev Biomed Eng 2000;2(1):315-337. [doi: 10.1146/annurev.bioeng.2.1.315] [Medline: 11701515]
8. Bousson V, Benoist N, Guetat P, Attané G, Salvat C, Perronne L. Application of artificial intelligence to imaging interpretations in the musculoskeletal area: where are we? Where are we going? Joint Bone Spine 2023;90(1):105493. [doi: 10.1016/j.jbspin.2022.105493] [Medline: 36423783]

XSL•FO
RenderX

9. du Toit C, Orlando N, Papernick S, Dima R, Gyacskov I, Fenster A. Automatic femoral articular cartilage segmentation using deep learning in three-dimensional ultrasound images of the knee. Osteoarthr Cartil Open 2022;4(3):100290 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100290] [Medline: 36474947]
10. Wang X, Shi C. Patch attention U-Net for knee cartilage segmentation in magnetic resonance images. Biomed Signal Process Control 2025;106:107754. [doi: 10.1016/j.bspc.2025.107754]
11. Gaj S, Yang M, Nakamura K, Li X. Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks. Magn Reson Med 2020;84(1):437-449. [doi: 10.1002/mrm.28111] [Medline: 31793071]
12. Kwon OR, Kang KT, Son J, Suh DS, Heo DB, Koh YG. Patient-specific instrumentation development in TKA: 1st and 2nd generation designs in comparison with conventional instrumentation. Arch Orthop Trauma Surg 2017;137(1):111-118. [doi: 10.1007/s00402-016-2618-2] [Medline: 28005167]
13. Hamghalam M, Simpson AL. Medical image synthesis via conditional GANs: application to segmenting brain tumours. Comput Biol Med 2024;170:107982 [FREE Full text] [doi: 10.1016/j.compbiomed.2024.107982] [Medline: 38266466]
14. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q. Swin-Unet: Unet-like pure transformer for medical image segmentation. Computer Vision – ECCV 2022 Workshops (Lecture Notes in Computer Science) 2023;13764:205-218 [FREE Full text] [doi: 10.1007/978-3-031-25066-8_9]
15. Demir U, Unal G. Patch-based image inpainting with generative adversarial networks. ArXiv. Preprint posted online on March 20, 2018 2018 [FREE Full text] [doi: 10.48550/arXiv.1803.07422]
16. Xue Y, Xu T, Huang X. Adversarial learning with multi-scale loss for skin lesion segmentation. 2018 Presented at: IEEE 15th international symposium on biomedical imaging (ISBI 2018); April 4-7, 2018; Washington. [doi: 10.1109/isbi.2018.8363707]
17. Jaemin S, Sang J, Kyu-Hwan J. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. ArXiv. Preprint posted online on June 26, 2017 2017 [FREE Full text] [doi: 10.48550/arXiv.1706.09318]
18. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. 2017 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017); July 21-26, 2017; Honolulu. [doi: 10.1109/cvpr.2017.632]
19. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, et al. MedGAN: medical image translation using GANs. Comput Med Imaging Graph 2020;79:101684. [doi: 10.1016/j.compmedimag.2019.101684] [Medline: 31812132]
20. Chen H, Zhao N, Tan T, Kang Y, Sun C, Xie G. Knee bone and cartilage segmentation based on a 3D deep neural network using adversarial loss for prior shape constraint. Front Med (Lausanne) 2022;9:792900 [FREE Full text] [doi: 10.3389/fmed.2022.792900] [Medline: 35669917]
21. Guo J, Yan P, Qin Y, Liu M, Ma Y, Li J. Automated measurement and grading of knee cartilage thickness: a deep learning-based approach. Front Med (Lausanne) 2024;11:1337993 [FREE Full text] [doi: 10.3389/fmed.2024.1337993] [Medline: 38487024]
22. Liu F. SUSAN: Segment unannotated image structure using adversarial network. Magn Reson Med 2019;81(5):3330-3345 [FREE Full text] [doi: 10.1002/mrm.27627] [Medline: 30536427]
23. Hong HT, Koh YG, Cho BW, Kwon HM, Park KK, Kang KT. An Image-Based augmented reality system for achieving accurate bone resection in total knee arthroplasty. Cureus 2024;16(4):e58281 [FREE Full text] [doi: 10.7759/cureus.58281] [Medline: 38752081]
24. Pfitzner T, Abdel MP, von Roth P, Perka C, Hommel H. Small improvements in mechanical axis alignment achieved with MRI versus CT-based patient-specific instruments in TKA: a randomized clinical trial. Clin Orthop Relat Res 2014;472(10):2913-2922 [FREE Full text] [doi: 10.1007/s11999-014-3784-6] [Medline: 25024031]
25. Frye BM, Najim AA, Adams JB, Berend KR, Lombardi AV. MRI is more accurate than CT for patient-specific total knee arthroplasty. Knee 2015;22(6):609-612. [doi: 10.1016/j.knee.2015.02.014] [Medline: 25813758]
26. An VVG, Sivakumar BS, Phan K, Levy YD, Bruce WJM. Accuracy of MRI-based vs CT-based patient-specific instrumentation in total knee arthroplasty: a meta-analysis. J Orthop Sci 2017;22(1):116-120. [doi: 10.1016/j.jos.2016.10.007] [Medline: 27823847]
27. Schotanus MGM, Thijs E, Heijmans M, Vos R, Kort NP. Favourable alignment outcomes with MRI-based patient-specific instruments in total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 2018;26(9):2659-2668. [doi: 10.1007/s00167-017-4637-0] [Medline: 28698929]
28. Kang DG, Kim KI, Bae JK. MRI-based or CT-based patient-specific instrumentation in total knee arthroplasty: how do the two systems compare? Arthroplasty 2020;2(1):1 [FREE Full text] [doi: 10.1186/s42836-019-0020-6] [Medline: 35236432]

Abbreviations

ASSD: average symmetric surface distance
BCE: binary cross-entropy
cGAN: conditional generative adversarial network
CNN: convolutional neural network
CT: computed tomography

XSL•FO
RenderX

DL: deep learning
DSC: Dice similarity coefficient
GAN: generative adversarial network
HD: Hausdorff distance
HD95: 95th percentile Hausdorff distance
IoU: intersection over union
IRB: Institutional Review Board
MRI: magnetic resonance imaging
OAI: Osteoarthritis Initiative
PSI: patient-specific instrument
ResUNet: residual UNet
SKI10: Segmentation of Knee Images 2010
TE: echo time
TKA: total knee arthroplasty
TR: repetition time
TransUNet: transformer UNet

Edited by A Coristine; submitted 20.Oct.2025; peer-reviewed by AM Hasan, YH Kwak, A Humayun; comments to author 17.Nov.2025; revised version received 12.Jan.2026; accepted 30.Jan.2026; published 02.Mar.2026.

Please cite as:
Park JY, Nam JH, Abdigapporov S, Kim JK, Koh YG, Cho BW, Kwon HM, Park KK, Kang KT
Enhanced Magnetic Resonance Imaging–Based Knee Cartilage Segmentation Using a Swin-UNet Conditional Generative Adversarial Network: Development and Validation Study
JMIR Med Inform 2026;14:e86155
URL: https://medinform.jmir.org/2026/1/e86155
doi:10.2196/86155
PMID:

©Jun Young Park, Ji-Hoon Nam, Shakhboz Abdigapporov, Jong-Keun Kim, Yong-Gon Koh, Byung Woo Cho, Hyuck Min Kwon, Kwan Kyu Park, Kyoung-Tak Kang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leveraging Naturalistic Driving Digital Biomarkers for Early Mild Cognitive Impairment Detection: Deep Learning Strategies

Firas Al-Hindawi[1,2], PhD; Teresa Wu[3,4], PhD; Yutong Wen[3,4], MSc; Peter Serhan[5,6], MS; Erica Forzani[6,7], PhD; Francis Tsow[8], PhD; Yonas E Geda[9], MSc, MD

[1]Industrial & Systems Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

[2]Interdisciplinary Research Center for Smart Mobility and Logistics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

[3]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, United States

[4]ASU Mayo Center for Innovative Imaging, Arizona State University, Tempe, AZ, United States

[5]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, United States

[6]Center for Bioelectronics and Biosensors, Biodesign Institute, Arizona State University, Tempe, AZ, United States

[7]School of Engineering for Matter, Transport and Energy, Arizona State University, Tempe, AZ, United States

[8]TF Health Corporation, Phoenix, AZ, United States

[9]Barrow Neurological Institute, Phoenix, AZ, United States

Corresponding Author:
Teresa Wu, PhD
School of Computing and Augmented Intelligence
Arizona State University
Tempe
699 S. Mill Ave.
Tempe, AZ, 85281
United States
Phone: 1 4809655172
Email: teresa.wu@asu.edu

Abstract

Background: Alzheimer disease and related dementias are increasing worldwide, with early detection during the mild cognitive impairment (MCI) stage critical for timely intervention. Driving behavior, which reflects everyday cognitive functioning, has emerged as a promising, noninvasive, and inexpensive digital biomarker when paired with machine learning. However, prior research has often relied on controlled settings, high-level features, or assumptions that fail to capture the sporadic nature of MCI, leaving a gap in modeling naturalistic driving data for robust early detection.

Objective: This study aims to address the limitations of prior work by developing deep learning strategies that leverage driving data collected in a naturalistic setting as digital biomarkers for early detection of MCI.

Methods: Clinically classified participants (8 with MCI and 14 cognitively normal; N=22) drove their personal vehicles under naturalistic conditions for several consecutive days. A total of 3 participants (2 cognitively normal and 1 MCI) withdrew before completing the experiments. In-vehicle sensors recorded GPS, accelerometer, and gyroscope signals, which were segmented into full trips and turning maneuvers. Three modeling strategies were compared: (1) single-view, (2) feature-level fusion, and (3) model-level late fusion. Classification models were trained and evaluated to assess their accuracy, discriminative ability, and participant-level performance.

Results: Models using full-trip data consistently outperformed turn-only inputs, with the best-performing model achieving 78% accuracy and an area under the receiver operating characteristic curve of 77%. Turn-based inputs alone demonstrated limited discriminative power; however, combining them with trip data through late fusion improved performance, though not beyond the full-trip baseline. Participant-level analysis indicated that classification accuracy improved with increased data volume, and trip-wise modeling more effectively captured the episodic nature of MCI than majority-vote aggregation. A frequency-based risk score was proposed as an interpretable and flexible output, enabling practical application in clinical and community settings.

Conclusions: Naturalistic driving behavior offers a scalable and noninvasive approach for early cognitive screening. Deep learning models using full-trip naturalistic driving data show promise for detecting MCI, with fusion strategies providing

XSL•FO
RenderX

supplementary insights. This framework supports proactive, real-world monitoring of cognitive decline, laying the foundation for digital health interventions in dementia prevention.

(JMIR Med Inform 2026;14:e83622) doi:10.2196/83622

KEYWORDS

aging; Alzheimer's disease; mild cognitive impairment; smart driving; machine learning; deep learning; data fusion

Introduction

Alzheimer disease (AD) and AD-related dementias are a growing public health concern. In the United States, cases are projected to reach 13.2 million by 2050, placing a significant strain on health care systems and caregivers [1-6]. Worldwide, over 55 million people are affected, with numbers expected to rise to 152 million by 2050 [5,7-9]. The disease progresses gradually, beginning with an asymptomatic phase, followed by mild cognitive impairment (MCI) and eventually dementia [10]. MCI is a transitional stage between normal aging and dementia, involving cognitive but not functional decline [11,12]. Accurate diagnosis at this stage is vital, as early detection allows timely intervention to slow progression and manage symptoms [13-16].

Identifying MCI early is challenging. Current tools, such as neuropsychological tests or biomarkers such as magnetic resonance imaging and positron emission tomography, are clinic-bound, offer only brief cognitive snapshots, and could be invasive, costly, or impractical for large-scale screening [5,6,17]. These methods often miss early MCI, as markers appear later in disease progression [18], highlighting the need for noninvasive, cost-effective biomarkers that support frequent monitoring in real-life settings [5,6,17,19-22]. Recently, spatial navigation deficits and driving behavior have emerged as potential early biomarkers for preclinical AD [23,24]. Tasks such as driving engage multiple cognitive functions in real-world contexts, and even mild impairments can lead to detectable changes in driving patterns [5,6,19,25]. Research shows that older adults with preclinical AD or MCI may limit their driving range, make more navigation errors, and drive more cautiously than cognitively normal peers [25,26]. These driving changes have been linked to AD biomarkers such as elevated amyloid and tau, even when standard cognitive tests are normal [25,27,28]. Notably, such changes can be detected by machine learning (ML) models using in-vehicle sensor data, as demonstrated in studies such as LongROAD [25] and GPS driving [19]. These studies show that ML can accurately identify individuals with MCI or preclinical AD by analyzing patterns in driving behavior such as speed, acceleration, and route choice, often outperforming conventional tools [5,6,19,25], while showing potential for real-time cognitive monitoring through everyday driving.

Despite encouraging progress, ML-based driving diagnostic systems still face key challenges. Many studies rely on simple, high-level features such as trip frequency and duration [19,25], overlooking nuanced indicators such as turn velocity, acceleration patterns, and path deviations that more directly reflect cognitive function. Furthermore, monthly driving records are treated as independent, even when from the same driver, risking data leakage and inflating performance by allowing driver-specific patterns to appear in both training and testing sets. Although recent work has incorporated richer features such as metabolic rate and ensured proper train-test separation [5,6], those studies were conducted in controlled settings with predefined routes, limiting real-world applicability. Additionally, most approaches rely on summary statistics and classical ML models, missing the potential of more powerful deep learning time-series classification techniques. Finally, conventional evaluation frameworks assume MCI symptoms appear consistently across trips, whereas studies have shown that cognitive performance in MCI can vary markedly from day to day, highlighting its sporadic nature [29-31]. Individuals with MCI may only show impairment in a minority of trips, leading models to misclassify them as cognitively normal, yet even occasional MCI-like trips could still serve as an early indicator of cognitive decline.

To address these limitations, this study proposes a new framework that improves both data representation and model evaluation. It shifts from artificial test settings to naturalistic driving by equipping participants' personal vehicles with tracking devices and allowing several days of normal driving, capturing real-world cognitive function. Instead of treating each participant's data as a uniform whole and relying on majority-based classification, the model adopts a granular, context-aware approach, evaluating each trip and turn individually to detect episodic impairments, consistent with the sporadic nature of MCI. Rather than delivering a binary diagnosis, it identifies subtle changes over time and flags individuals for further evaluation, enabling proactive cognitive health monitoring. The framework aggregates binary trip- and turn-level classifications over time to enable longitudinal monitoring and transforms them into a continuous, participant-level risk score (eg, the proportion of trips flagged as impaired). This frequency-based risk score provides a nonbinary, dynamic measure that reflects gradual cognitive change and supports proactive follow-up rather than a 1-time diagnostic label. Specifically, the framework includes a targeted feature-extraction pipeline that segments time-series data from full trips and cognitively demanding turning maneuvers, producing 2 complementary multivariate inputs. Deep learning architectures are then evaluated under 3 paradigms: (experiment 1) single-view modeling of trips and turns, (experiment 2) feature-level fusion, and (experiment 3) model-level late fusion using dual encoders. This design is guided by 3 key research questions.

For research question 1 (RQ1), effectiveness in naturalistic settings, can driving-based digital-biomarker strategies that showed promise in previous controlled environments [5,6] remain effective amid the noise and variability of day-to-day, real-world driving?

XSL•FO
RenderX

For research question 2 (RQ2), data handling and representation, what preprocessing and feature-extraction pipeline best transforms raw multisensor driving streams into informative inputs for deep learning?

For research question 3 (RQ3), optimal modeling strategy under data constraints, given current dataset limitations, but also anticipating richer data in future studies, which learning paradigm offers the most robust path forward? To answer this, we decompose this question into three comparative subquestions:

1. Single-view baselines: How well can early MCI be detected when trips and turns are modeled independently?
2. Early (data-level) fusion: Does combining trip and turn data at the feature level improve over the best single-view model?
3. Late (model-level) fusion: Can merging latent representations from trip and turn encoders outperform both single-view and early-fusion models?

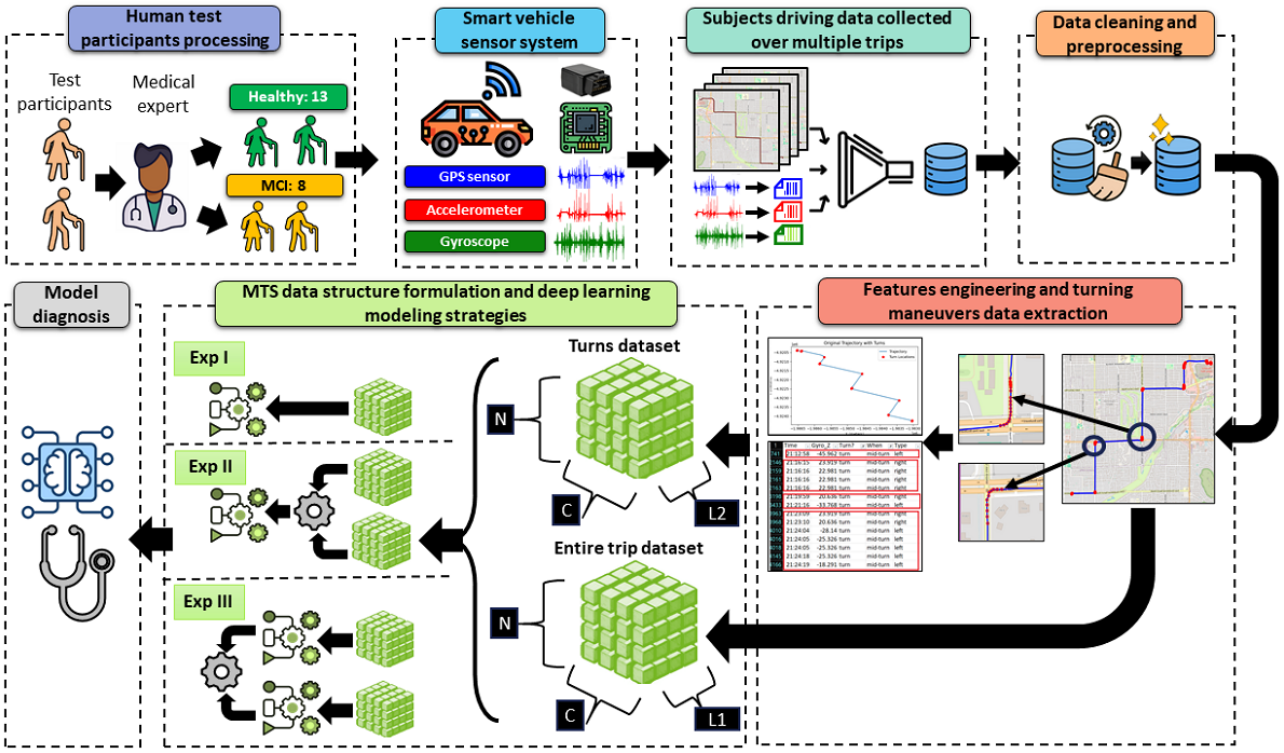
The remainder of this paper is organized as follows. The Methods section describes this study's design, materials, and methods used to develop the framework, and the experimental methodology. The Results section presents the outcomes of the conducted experiments. The Discussion section discusses the findings and concludes the paper.

Methods

Study Design

Figure 1 presents an overview of the proposed methodology, which includes participant recruitment with expert cognitive classification, sensor-based driving data collection, preprocessing, feature engineering, and model training and evaluation. Raw sensor streams were cleaned, synchronized, and segmented into 2 complementary multivariate time-series datasets: full-trip sequences and turning maneuvers, used as inputs for subsequent experiments.

Figure 1. Overview of the adopted methodology, illustrating the sequential steps from participant recruitment to model training and evaluation. N denotes the number of samples (trips or turns), C the number of sensor channels, and L_1 and L_2 the interpolated sequence lengths for the trip and turn time-series inputs, respectively. MCI: mild cognitive impairment; MTS: multivariate time series.



Study Participants

To ensure accurate data interpretation, participant diagnoses were conducted in collaboration with a neurologist specializing in AD. Based on cognitive assessments, 13 participants were classified as cognitively healthy and 8 as having MCI. This grouping enabled structured comparison of driving behavior between cognitively normal and impaired individuals. Table 1 summarizes the demographic and medical diagnoses of the participants.

XSL•FO
RenderX

Table 1. Participant demographics and group classification: this study included 21 older adults (ages 65-85 y), with 11 male and 10 female participants. Medical evaluations classified 13 as cognitively healthy and 8 as having MCI[a].

Participant numbers	Age (y)	Sex	Clinical condition
Participant 1	78	Male	MCI
Participant 2	80	Male	Healthy
Participant 3	70	Male	Healthy
Participant 4[b]	72	Female	Healthy
Participant 5	73	Female	Healthy
Participant 6	74	Male	MCI
Participant 7	77	Female	Healthy
Participant 8	87	Female	Healthy
Participant 9	85	Female	Healthy
Participant 10	82	Female	MCI
Participant 11	71	Female	Healthy
Participant 12	72	Female	MCI
Participant 13	65	Male	MCI
Participant 14[b]	75	Male	Healthy
Participant 15[b]	71	Female	MCI
Participant 16	76	Male	Healthy
Participant 17	75	Female	MCI
Participant 18	83	Male	Healthy
Participant 19	76	Male	MCI
Participant 20	80	Female	Healthy
Participant 21	79	Male	Healthy
Participant 22	75	Male	Healthy

[a]MCI: mild cognitive impairment.

[b]Participants withdrew before completing the experiment, leaving insufficient data for their inclusion in the final analysis.

As shown in the table, this study focused on naturalistic driving performance in adults aged 65-85 years, a group at higher risk for cognitive decline. Further, 21 participants were initially recruited, but 3 participants (participants 4, 14, and 15) withdrew before completing the driving experiment, leaving insufficient data for analysis.

Smart Vehicle Telematics System

Each participant's personal vehicle was fitted with a BitBrew on-board diagnostics (OBD) data logger to continuously record naturalistic driving behavior (Figure 2). This advanced telematics device integrates 3 sensors: GPS, accelerometer, and gyroscope. Together, they capture vehicle location, movement, and dynamic behavior, enabling comprehensive real-world driving analysis.

The GPS module tracks the vehicle's location and speed, recording longitude and latitude coordinates and measuring speed in meters per second (m/s) at a frequency of 1 Hz (1 update per second). The accelerometer captures linear acceleration along 3 axes: longitudinal (forward and backward), lateral (side to side), and vertical (up and down), in m/s^2. It detects events such as harsh braking, rapid acceleration, and sharp turns, operating at 24 Hz for high-resolution motion detection. The gyroscope measures angular velocity in radians per second (rad/s) around 3 axes: pitch (tilting forward and backward), roll (tilting side to side), and yaw (rotational movement around the vertical axis). This sensor operates at 24 Hz and provides data on lane changes, skidding, and rollovers.

Participants drove their vehicles freely with no specific instructions, allowing for natural capture of daily driving behavior. Table 2 summarizes the driving periods for each participant, including start and end dates and total duration. All data from the OBD logger was transmitted to an AWS S3 (Amazon Web Services Simple Storage Service) bucket for cloud-based storage and analysis.

XSL•FO
RenderX

Figure 2. Overview of the data collection system used in the naturalistic driving study. AWS S3: Amazon Web Services Simple Storage Service; OBD: on-board diagnostics.

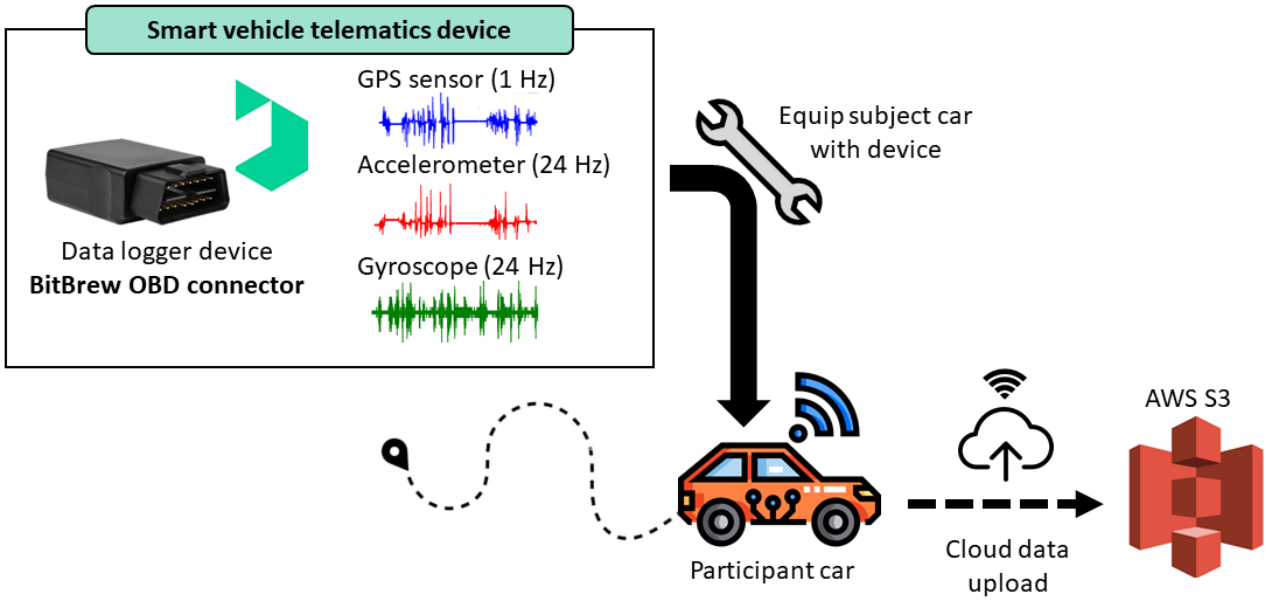


Table 2. Naturalistic driving test periods for study participants, detailing the start and end dates along with the duration (in days) for each observation period.

Participant number	Start date	End date	Duration (d)
Participant 1	March 29, 2023	April 7, 2023	9
Participant 2	April 13, 2023	April 20, 2023	7
Participant 3	May 4, 2023	May 11, 2023	7
Participant 4	—[a]	—	—
Participant 5	May 19, 2023	May 26, 2023	7
Participant 6	May 26, 2023	June 2, 2023	7
Participant 7	June 2, 2023	June 9, 2023	7
Participant 8	June 30, 2023	July 7, 2023	7
Participant 9	July 7, 2023	July 14, 2023	7
Participant 10	July 14, 2023	July 21, 2023	7
Participant 11	July 21, 2023	July 28, 2023	7
Participant 12	August 15, 2023	August 21, 2023	6
Participant 13	September 15, 2023	September 22, 2023	7
Participant 14	—	—	—
Participant 15	—	—	—
Participant 16	August 25, 2023	September 8, 2023	14
Participant 17	August 25, 2023	September 8, 2023	14
Participant 18	September 22, 2023	September 29, 2023	7
Participant 19	September 29, 2023	October 6, 2023	7
Participant 20[b]	October 10, 2023	January 20, 2024	86
Participant 21	October 10, 2023	October 27, 2023	17
Participant 22	October 27, 2023	November 3, 2023	7

[a]Not available.

[b]Participants did the experiment over 2 periods (October 10, 2023, to October 27, 2023, and November 12, 2023, to January 20, 2024).

XSL•FO
RenderX

Naturalistic Driving Data Collection, Cleaning, and Preprocessing

This naturalistic driving study was conducted in Tempe, Arizona, and extended to nearby cities, including Phoenix, Glendale, Chandler, Scottsdale, and Mesa. Participants drove their personal vehicles, each equipped with a BitBrew OBD data logger, which passively recorded real-world driving behavior. The system continuously captured GPS location, accelerometer, and gyroscope data during each session, enabling a comprehensive analysis of driving patterns without imposing specific instructions or restrictions.

Recorded routes spanned a wide range of road types such as residential streets, highways, and urban roads. The dataset included a variable number of trips per participant, with differences in trip length, duration, and driving conditions reflecting the complexity of real-world behavior. However, sensor data was not always consistent across modalities for the same participant due to device malfunctions, intermittent failures, or temporary disconnections. These issues sometimes caused sensors to stop mid-trip, fail to initiate, or record out of sync, resulting in incomplete or misaligned data. To address these challenges, a structured preprocessing pipeline was applied to ensure that only high-quality, reliable, and synchronized driving records were retained for subsequent analysis.

The first step in preprocessing involved segmenting individual trips from the continuous sensor data stream. Timestamps from the accelerometer, GPS, and gyroscope were analyzed to identify gaps between recordings. If the time difference between consecutive points was less than 1 minute, the data was treated as part of the same trip. Gaps exceeding 1 minute marked the start of a new trip. After segmentation, the data were organized into participant-specific folders, with separate subfolders for each sensor and individual CSV files representing distinct trips.

Although segmentation provided structure, several data quality issues emerged. Sensor malfunctions led to missing data, and differences in recording frequency created synchronization challenges. The accelerometer and gyroscope recorded at 24 Hz, while the GPS operated at 1 Hz, resulting in mismatched data lengths. To resolve this, a cleaning process was applied. A mismatch check flagged any case where the number of trips differed across sensors. When discrepancies were found, the system logged debugging information and excluded incomplete trips. Data from all sensors was merged using shared time attributes such as day, month, hour, and minute. Trips with missing values were flagged, stored separately, and excluded from further analysis. Only complete and synchronized trips were retained. Table 3 shows the number of trips kept for each participant after preprocessing.

XSL•FO
RenderX

Table 3. Number of trips retained per participant after data cleaning and synchronization. Due to sensor malfunctions, missing data, and synchronization issues, the total number of trips was reduced.

Participant number	Number of trips recorded by each sensor		
	Accelerometer	GPS	Gyroscope
Participant 1	28	28	28
Participant 2	18	18	18
Participant 3	30	30	30
Participant 4	—[a]	—	—
Participant 5	46	46	46
Participant 6	21	21	21
Participant 7	39	39	39
Participant 8	6	6	6
Participant 9	11	11	11
Participant 10	11	11	11
Participant 11	21	21	21
Participant 12	3	3	3
Participant 13	13	13	13
Participant 14	—	—	—
Participant 15	—	—	—
Participant 16	10	10	10
Participant 17	11	11	11
Participant 18	11	11	11
Participant 19	4	4	4
Participant 20	126	126	126
Participant 21	6	4	6
Participant 22	24	24	24

[a]Not available.

Features, Engineering, and Turning Maneuvers Data Extraction

Turning Maneuvers Data Extraction

Turning a vehicle is a cognitively demanding task involving decision-making, divided attention, working memory, processing speed, and spatial orientation. These functions are often impaired in individuals with MCI and may serve as early biomarkers for AD. Following prior studies [5,6], this work focuses on turning maneuvers, which have shown potential in detecting cognitive decline in standardized controlled experiments. Unlike previous research, this study uses naturalistic driving, where participants follow their routine behavior without instructions. This introduces additional noise into the data, allowing us to test whether turning-related features remain informative under real-world conditions.

Turning detection required converting GPS coordinates from latitude and longitude into Cartesian coordinates for accurate geometric and kinematic analysis. The transformations used are shown in equations 1 and 2. Given a point on the Earth's surface with latitude ∅ and longitude λ, the x and y coordinates were calculated as:

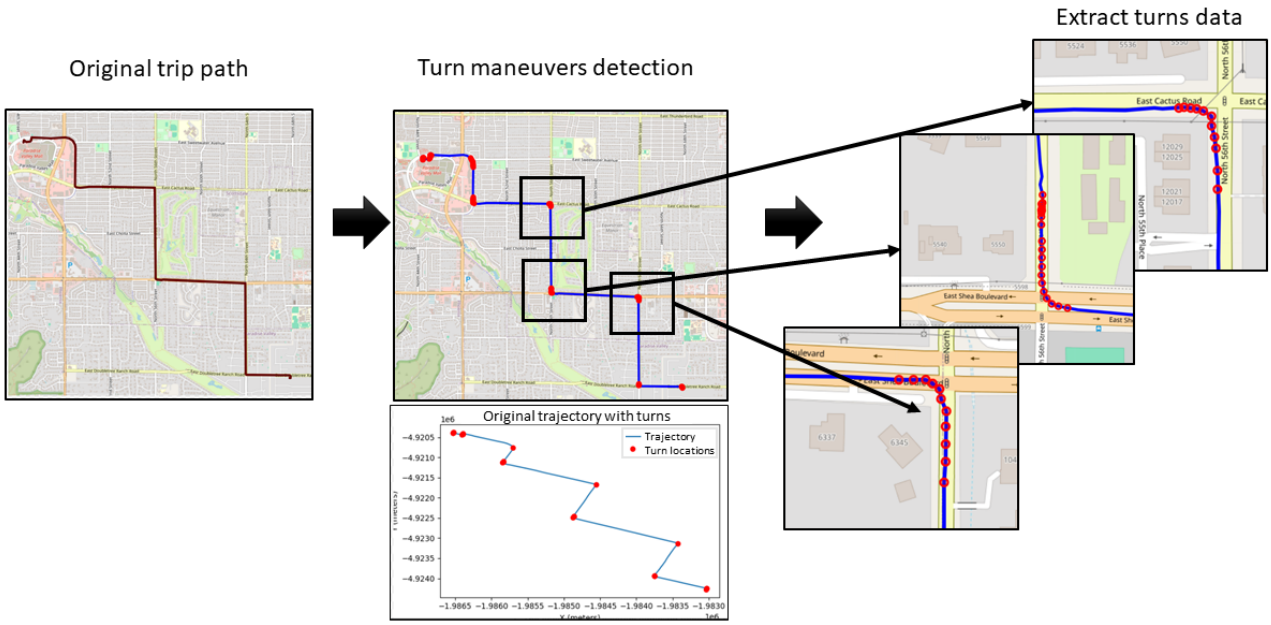
where R is the Earth's radius (approximately 6371 km), and ∅ and λ are expressed in radians.

To identify turning maneuvers, the peak point of each arc path was detected by analyzing the rate of change in x and y rectangular coordinate arrays. By computing the first derivative (dy/dx), the directional changes were captured, with abrupt gradient shifts indicating deviations from straight paths. To filter out gradual curves, various thresholds were tested. It was found that a gradient of 10 degrees (0.174533 radians) reliably captured meaningful turns. However, relying on the gradient alone produced occasional errors. Points where dx=0, resulting in undefined gradients, were discarded. Additionally, turn stamp points with 0 speed, which typically corresponded to stationary periods rather than active maneuvers, were also excluded. Further refinement addressed redundant peak detections within a single turn. In some cases, multiple nearby peaks were

XSL•FO
RenderX

incorrectly classified as separate turns. To resolve this, a smoothing approach was applied. Sequential turn points were merged and replaced by their median index, reducing overlap and ensuring 1 peak per turn event. Once peaks were identified, all surrounding points were extracted by applying a temporal window. Based on empirical observations, a turn maneuver was identified by including all sensor readings occurring 15 seconds before and after the peak point. Accelerometer and gyroscope readings were downsampled to align with the 1 Hz GPS rate by averaging every 24 readings from the higher frequency sensors. Figure 3 illustrates an example of turns extracted from a sample trip.

Figure 3. Illustration of the turn maneuver detection process. The leftmost map represents the original trip path recorded using GPS data. The middle section highlights detected turning maneuvers, marked by red dots, which were identified based on changes in trajectory. The rightmost section zooms in on extracted turn segments, showing detailed turning paths.



Features Engineering (Radial Acceleration)

An additional engineered feature was radial acceleration, which captures how sharply and assertively a driver navigates turns. Higher radial acceleration values may indicate sharper bends or faster speeds, while rapid fluctuations within a maneuver suggest frequent microadjustments in steering and speed. These patterns can reflect reduced motor control and cognitive instability. Radial acceleration was computed by taking the temporal gradient of the gyroscope data along the X, Y, and Z axes. This provided a continuous measure of changes in angular velocity over time, enriching the representation of driver motion. Equations 3-5 show the computation for each axis:

Table 4 summarizes the final set of 10 features used in the input tensor for all models. Each feature represents a specific aspect of motion captured by the GPS, accelerometer, or gyroscope sensors. The Gyro_X_Acc, Gyro_Y_Acc, and Gyro_Z_Acc variables correspond to the angular accelerations derived from the temporal derivatives of the gyroscope signals.

XSL•FO
RenderX

Table 4. Final set of sensor-based features used in the dataset. The table lists the 10 input channels used across all experiments. Each feature captures a different dimension of vehicle motion using raw or derived signals from speed, accelerometer, and gyroscope sensors.

Feature or channel	Description
Speed	Vehicle speed measured in meters per second (m/s)
Acc_X	Linear acceleration along the vehicle's longitudinal axis (forward/backward)
Acc_Y	Linear acceleration along the lateral axis (side-to-side motion)
Acc_Z	Linear acceleration along the vertical axis (up or down motion)
Gyro_X	Angular velocity around the x-axis (roll rate)
Gyro_Y	Angular velocity around the y-axis (pitch rate)
Gyro_Z	Angular velocity around the z-axis (yaw rate or heading change)
Gyro_X_Acc	Angular acceleration around the x-axis, derived from Gyro_X
Gyro_Y_Acc	Angular acceleration around the y-axis, derived from Gyro_Y
Gyro_Z_Acc	Angular acceleration around the z-axis, derived from Gyro_Z

Multivariate Time Series Data Formulation

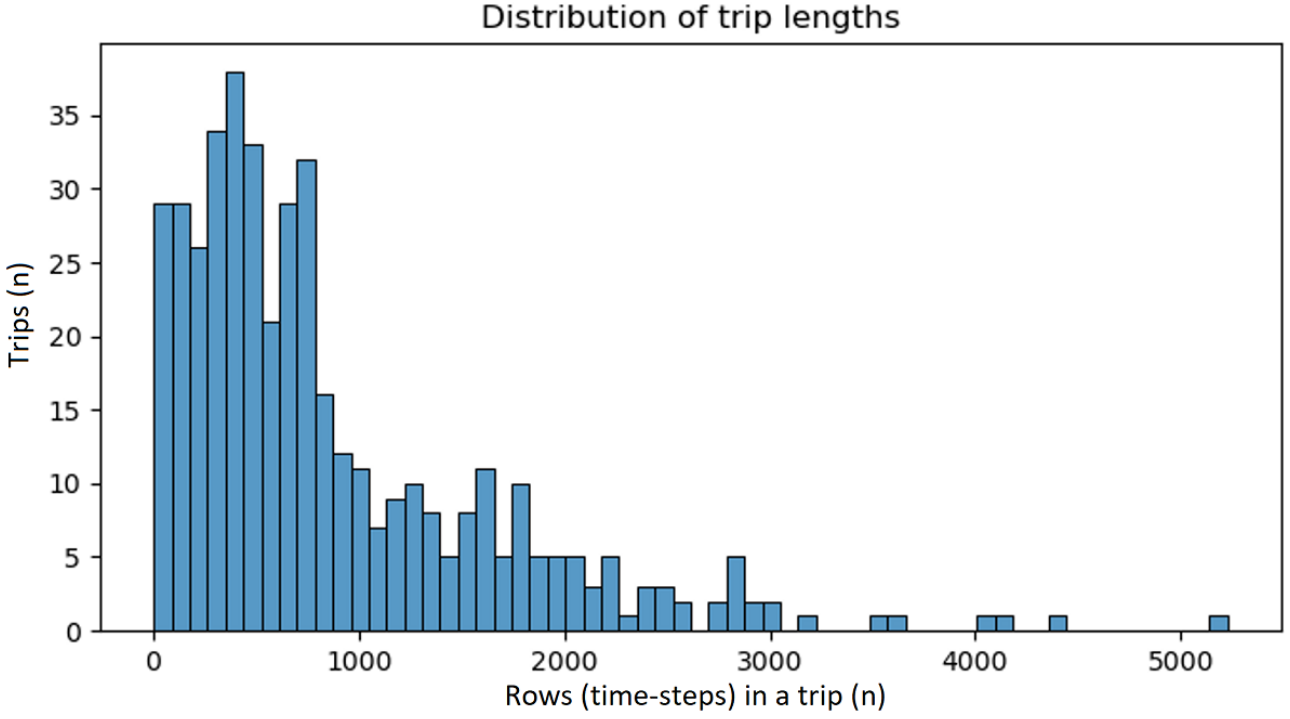
This study used 2 datasets: one capturing entire trip readings and another focused on turning maneuvers. Both were formatted as multivariate time series to preserve the sequential nature of driving behavior. Depending on the dataset, each row represented either a full trip or a turn, and each sensor served as a separate time-dependent variable. To standardize input lengths for model training, sequences were aligned and resampled through interpolation, ensuring uniform temporal resolution across samples.

Time Series Alignment and Interpolation

Trip and turn durations varied widely across drivers, requiring a preprocessing step to standardize time series lengths. Without it, inconsistent sequence sizes would hinder the model's ability to learn temporal patterns. To address this, interpolation was applied to resample all sequences to a fixed number of time points. To select the optimal length, different resolutions were tested by plotting samples and evaluating the balance between data redundancy and information loss. Alignment and interpolation were applied separately to the trip and turn datasets to maintain consistency within each.

For the trip dataset, the distribution of raw sequence lengths (Figure 4) was analyzed, and a target that preserved most trips without excessive compression was selected. This choice was further validated by visually inspecting sensor signals before and after resampling, confirming that key temporal patterns were retained. Based on this process, a fixed length of 1200 points was selected for the trip dataset.

Figure 4. Distribution of trip lengths in the raw dataset. Each bar represents the number of trips with a given sequence length (in time steps). The histogram was used to inform the selection of a standardized interpolation length that balances coverage and temporal resolution.



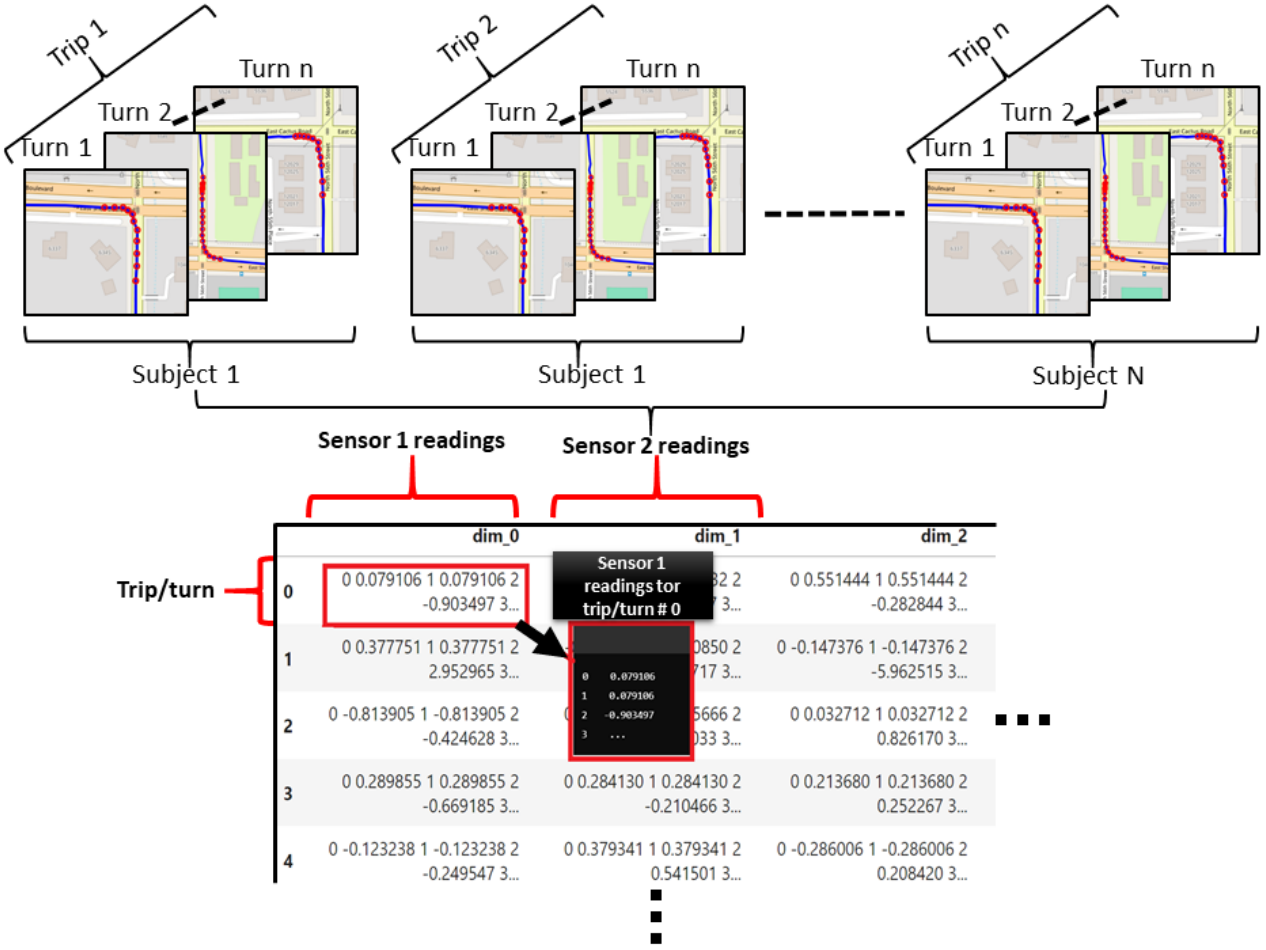
XSL•FO
RenderX

For the turning maneuvers dataset, the extraction process described in Turning Maneuvers Data Extraction produced segments with a standardized length of 31-time steps by centering each around the turn midpoint and including a fixed number of points before and after. In most cases, this window was fully available, making interpolation unnecessary. However, in rare edge cases, such as turns at the very start or end of a trip, the available context was limited, resulting in shorter segments. These were interpolated to ensure consistency across the dataset.

Final Data Representation

After interpolation, all sensor streams were consolidated into a single structured format. Each trip or turn was represented as a 3D tensor with dimensions N × C × L, where N is the number of trips or turns, C is the number of channels (sensors), and L is the standardized sequence length. Each row corresponds to a distinct trip or turn, columns represent the sensor channels (eg, speed), and the entries hold the sequential readings for the interpolated time steps. Figure 5 shows how all sensor data were aligned and merged into this final format.

Figure 5. Representation of the MTS dataset structure used. Each row corresponds to either an entire trip or a turn, depending on the dataset. The dataset is formatted as a 3D tensor with dimensions (trip or turn, channels [sensors], and sensor readings). MTS: multivariate time series.



Experiments

This section describes the experimental setup and methods used to analyze driving behavior and classify participants based on cognitive status. We hypothesize that naturalistic driving contains digital signatures that can support early MCI detection. The central question is how to best process and model this data to maximize predictive performance. To explore this, we designed 3 experiments based on the same source of driving data but varying in input representation and modeling strategy. By decomposing the problem along 2 dimensions, what part of the data to use (entire trip vs turns) and how to architect the model (single vs multiple inputs or models), we aim to identify which configurations most effectively preserve behaviorally relevant information.

Three main architecture families were used: 1D convolutional neural networks (1D-CNNs), recurrent neural networks (RNNs), and temporal convolutional networks (TCNs); because they align well with the characteristics of naturalistic driving data, which are multivariate, sequential, and temporally dependent. 1D-CNNs were selected for their ability to detect short, localized motion patterns such as rapid accelerations or turning dynamics; RNNs were included to capture longer behavioral trends that unfold over an entire trip; and TCNs were chosen for their ability to model both short- and long-range dependencies efficiently and robustly. Variants within each family (eg, lightweight 1D-CNNs and attention-based RNNs) were explored to identify which architectural style best captures early MCI-related temporal signatures in real-world driving behavior.

XSL•FO
RenderX

All models were trained using a unified configuration to ensure comparability across experiments. The loss function was standard cross-entropy with class weighting applied when needed to address class imbalance. Standard scaling normalization was applied where indicated in the tables. Optimization was performed with AdamW using a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. A cosine annealing learning-rate scheduler with T_{max} equal to the number of epochs was used to promote stable convergence. Training was conducted with a batch size of 32 for 200 epochs (reduced from 500 after observing convergence before epoch 200). Dropout was applied only when specified by the architecture variant, as detailed in the appendices.

Table 5 summarizes the details of the experiments, while Figure 6 illustrates the modeling strategies in each.

Table 5. Summary of experimental setups showcasing the data source, feature representation, and modeling paradigm. Sensor channels include speed, linear accelerations (X/Y/Z), angular velocities (X/Y/Z), and derived angular accelerations (X/Y/Z).

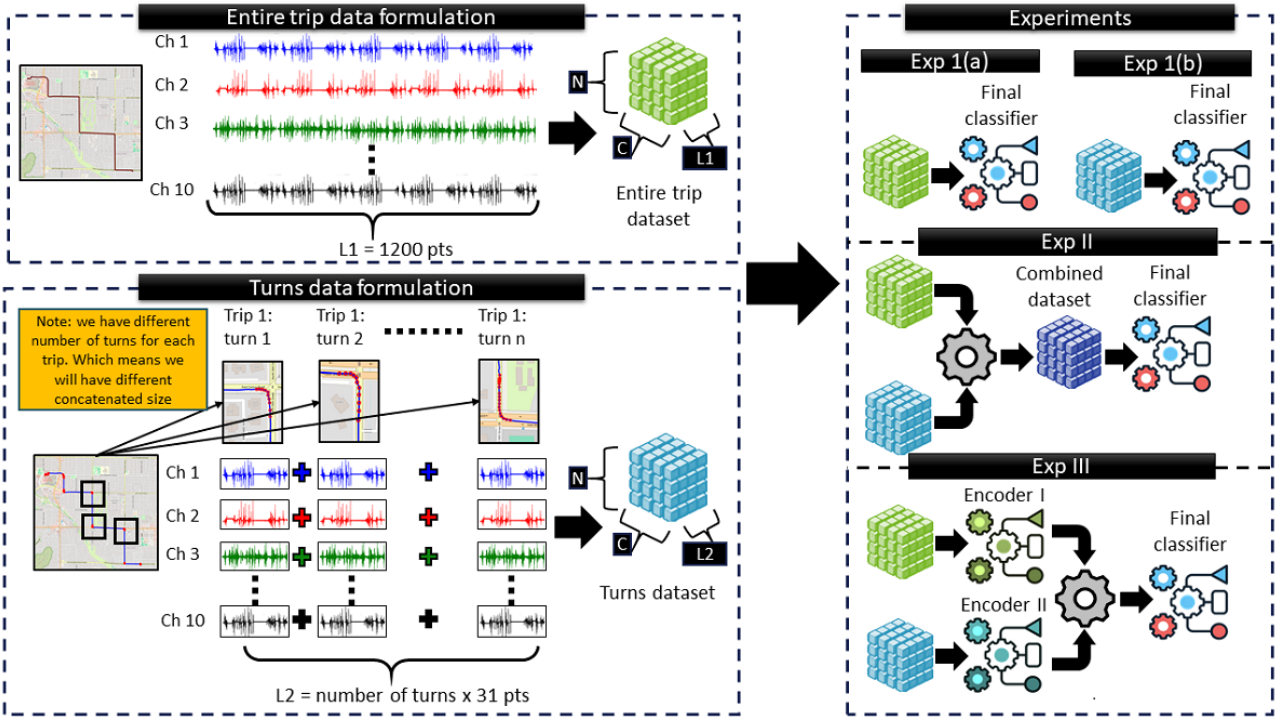
Experiment	Data source	Feature representation (input to model)	Predictive-modeling strategy
1(a)	Entire trip only	Multivariate time series (10 raw sensor channels)	Single-view deep models (1D-CNN[a], RNN[b], and TCN[c] variants)
1(b)	Turning maneuvers only	Multivariate time series (same 10 channels, windowed around turns)	Single-view deep models (1D-CNN, RNN, and TCN variants)
2	Trip and turn signals (concatenated)	Feature-level fusion (20-channel time series)	Single model (1D-CNN, RNN, and TCN) trained on fused input
3	Trip and turn signals (separate)	Two parallel 10-channel time series	Dual encoders with late fusion

[a]1D-CNN: 1D convolutional neural network.

[b]RNN: recurrent neural network.

[c]TCN: temporal convolutional network.

Figure 6. Overview of the data formulation and experimental paradigms. The left side shows the data formulation that produces the 2 multivariate inputs used in this study: the entire-trip sensor readings (green) and concatenated turn-segment readings (blue). The right side summarizes how the produced data are used in our experiment, which consists of 3 modeling paradigms: Exp 1 (single-view models trained on trips or turns separately), Exp 2 (feature-level fusion of the 2 inputs), and Exp 3 (dual encoders with late fusion).



Experiment 1: Single Data Representation With Single Model

This experiment evaluates the effectiveness of using a single type of input, either full trips or turn segments, modeled independently (Figure 6). All input data was treated as a multivariate time series and interpolated to a fixed length using 10 sensor channels. For each data type, various model architectures were tested, including 1D-CNNs, RNNs, and TCNs (details in Multimedia Appendix 1). In the full-trip setting (experiment 1(a)), each trip was used as a standalone sequence to capture broad driving behavior. In contrast, the turn-only setting (experiment 1(b)) focused on concatenated turn segments. This experiment isolated a specific behavior

XSL•FO
RenderX

substructure to examine whether it holds predictive power on its own.

Experiment 2: Combined Input With Single Model (Data Fusion)

In this experiment, feature-level fusion was explored by combining full-trip and turn-based inputs into a single tensor, as illustrated in Figure 6. Sensor readings from 10 channels were extracted for the entire trip and interpolated to a fixed length of 1200 time points (L_1) to ensure consistency. Separately, all turn segments from each trip were concatenated and interpolated to the same length (L_1 =1200), despite variability in the number of turns across trips. After alignment, both sequences were concatenated along the channel dimension, resulting in a unified tensor with shape (N, C=20, and L=1200), where N denotes the number of trips and C includes 10 channels from the trip data and 10 from the turn data. As in experiment 1, a range of model architectures was evaluated, including 1D-CNNs, RNNs, and TCNs, using the fused input. The objective was to assess whether the joint modeling of global and localized driving behavior could improve classification performance. Further architectural details are provided in Multimedia Appendix 2.

Experiment 3: Separate Inputs and Models With Late Fusion (Model Fusion)

In the final experiment, a model fusion approach was adopted. Separate models were trained independently on the trip-level and turn-level data. As shown in Figure 6, a total of 2 distinct datasets were prepared for each trip. In the upper branch, the complete 10-channel sensor stream was resampled to a fixed length of L_1 points and passed to encoder 1, where a latent representation of global driving behavior was learned. In the lower branch, all turning maneuvers for the same trip were concatenated, and the resulting sequence was standardized to L_2 points (which differed from L_1) before being input to encoder 2.

Encoders were selected from 3 architectural families, 1D-CNN, RNN, or TCN (for example, DualTinyFCN, DualGRU, and DualTCN), and were trained with separate weights. The latent vectors from both encoders were combined using a late fusion classifier. This allowed the model to integrate complementary global and local driving features while maintaining the independence of each encoder. Further model details are provided in Multimedia Appendix 3.

Model Training Framework

The training framework for all experiments followed a structured procedure. The process began with data preprocessing, where the turning maneuvers dataset was transformed into a 3D tensor representation with dimensions (N, C, and L). To evaluate the model's performance in distinguishing between cognitively healthy individuals and those with MCI, a 7-fold interparticipant cross-validation strategy was implemented. Given the limited sample size, no independent test set was used. Instead, interparticipant cross-validation was adopted to maximize data use while ensuring that the evaluation remained unbiased. Each fold contained a unique set of participants for testing, with the remaining participants used for training, ensuring that no individual appeared in both sets and thereby eliminating any risk of data leakage across folds.

Stratification was handled at the participant level. Each fold maintained an equal ratio of cognitively normal to MCI participants in the validation set. As the models operated on trip or turn-level data (rather than directly on participants), it was not possible to simultaneously stratify the trip or turn samples and preserve strict participant separation. We therefore prioritized participant-level stratification, as it provides a more reliable measure of generalization to unseen drivers, even though the resulting number of trips or turns per class could become imbalanced across folds. To mitigate any resulting class imbalance, balanced accuracy was reported alongside overall accuracy, area under the receiver operating characteristic curve (AUC), precision, and recall to provide a fairer assessment of performance across both diagnostic groups.

Once the dataset was partitioned, each model was trained on the training folds and evaluated on unseen participants in the validation fold. The average metrics across all 7 folds were used to report the final performance.

Evaluation Metrics

To assess the performance of the ML models, multiple evaluation metrics were used. Given the class imbalance in the dataset, relying solely on accuracy could be misleading. Therefore, we used a combination of accuracy, balanced accuracy, AUC, precision, and recall to provide a comprehensive evaluation of model effectiveness.

Accuracy measures the overall proportion of correctly classified instances, but can be biased when the dataset is imbalanced. Balanced accuracy addresses this issue by computing the average recall across both classes, making it a more reliable metric for evaluating performance in cases where 1 class is underrepresented. AUC provides an indication of the model's ability to distinguish between the 2 classes, with higher values indicating better separability. Precision quantifies the proportion of positive classifications that are actually correct, which is particularly important in minimizing false positives. Recall, on the other hand, measures the model's ability to correctly identify positive instances, ensuring that individuals with MCI are accurately detected. The formulas used for computing these metrics are summarized in Table 6.

XSL•FO
RenderX

Table 6. Evaluation metrics and their mathematical formulations.

Metric	Formula[a]	Description
Accuracy		Proportion of correctly classified instances out of all cases.
Balanced accuracy		Average recall for both classes, accounting for imbalances.
AUC[b]	Computed from the ROC[c] curve	Measures the model's ability to distinguish between classes.
Precision		Proportion of correctly identified positive cases among predicted positives.
Recall (sensitivity)		Proportion of actual positives correctly classified by the model.

[a]FN: false negative, FP: false positive, TN: true negative, TP: true positive.

[b]AUC: area under the receiver operating characteristic curve.

[c]ROC: receiver operating characteristic.

Ethical Considerations

A total of 22 participants were initially recruited, informed about the study, and provided written consent. This study was approved by the Institutional Review Board at Arizona State University (STUDY00006547).

Participant privacy and confidentiality were strictly maintained throughout this study. All data were deidentified using unique participant identification codes, and no personally identifiable information was included in any analyses or publications.

Participants received no financial compensation for their participation.

Results

This section reports the performance of the 3 experimental paradigms, single-view modeling (experiment 1), feature-level data fusion (experiment 2), and model-level fusion (experiment 3), evaluated with 7-fold, participant-stratified cross-validation. For every experiment, we averaged accuracy, balanced accuracy, AUC, precision, recall, and F_1-score across folds.

Experiment 1 Results

Experiment 1 examined whether early MCI signals could be more effectively captured from full-trip data or from isolated turning maneuvers when modeled independently. To provide a performance reference, a nontemporal baseline model was first implemented using summarized trip-level statistics from the raw sensor data. For each trip, the mean and SD of the accelerometer (X, Y, and Z), gyroscope (X, Y, and Z), and speed signals were computed, resulting in 14 static features. These features were used to train a random forest classifier, representing a simple benchmark based on conventional telematics features used in earlier driving and cognitive-monitoring studies [19,25]. Table 7 summarizes the cross-validated results for 21 single-view models. The full-trip experiment (1(a)) consistently outperformed the turn-only experiment across all tested architectures. The best performance was achieved by a lightweight convolutional model (TinyFCN without normalization but with class weighting), reaching 78.3% accuracy, 73.6% balanced accuracy, and an AUC of 76.7%. Other strong performers included deeper convolutional models (InceptionTimeSmall) and attention-enhanced recurrent networks (gated recurrent unit, GRU + Attn), suggesting that both paradigms can leverage the extended temporal context of full trips. Channel normalization was found to make results worse.

XSL•FO
RenderX

Table 7. Cross-validation results for all experiments. Metrics are averaged over 8 folds.

Experiment	Model	Normalize	Weights	Accuracy (%)	BA[a] (%)	AUC[b] (%)	Precision (%)	Recall (%)	F_1-score (%)
Baseline	RF[c]	Yes	—[d]	55	49	52	33	20	25
1(a)	1D-CNN[e]	No	No	73	63	62	57	72	63
1(a)	GRU[f]	No	Yes	73	69[g]	73[g]	63	75	60
1(a)	GRU and attention	No	Yes	77	69	66	64	79	69[g]
1(a)	GRU and multiattention	No	Yes	71	69	65	57	83[h]	64
1(a)	Inception lite	Yes	Yes	59	61	56	47	80	54
1(a)	Inception lite	No	Yes	68	65	60	63	64	53
1(a)	InceptionTimeSmall	Yes	Yes	69	69	69	65[h]	67	58
1(a)	InceptionTimeSmall	No	Yes	78[g]	66	64	69[g]	71	65[h]
1(a)	LSTM[i]	No	Yes	69	69	64	56	78	59
1(a)	ResNet TSC[j]	Yes	Yes	65	66	64	52	82	60
1(a)	ResNet TSC	No	Yes	77[h]	67	65	57	73	63
1(a)	TCN[k]	No	Yes	71	60	59	52	68	57
1(a)	TCN and attention	No	Yes	58	58	62	47	86[g]	56
1(a)	TinyFCN	Yes	No	73	63	62	57	72	63
1(a)	TinyFCN	Yes	Yes	72	66	64	65	73	62
1(a)	TinyFCN	No	Yes	78[g]	74[g]	77[g]	65[h]	66	64
1(a)	Tiny TCN	Yes	Yes	63	64	65	51	76	57
1(a)	Tiny TCN	No	Yes	64	60	58	48	82	57
1(b)	1D-CNN	No	No	63	54	53	44	30	34
1(b)	TinyFCN	Yes	Yes	63	56	58	44	53	46
1(b)	TinyFCN	No	Yes	59	57	59	42	64	47

[a]BA: balanced accuracy.

[b]AUC: area under the receiver operating characteristic curve.

[c]RF: random forest.

[d]Not available.

[e]1D-CNN: 1D convolutional neural network.

[f]GRU: gated recurrent unit.

[g]The second best values.

[h]The best values.

[i]LSTM: long short-term memory.

[j]TSC: time series classification.

[k]TCN: temporal convolutional network.

In contrast, the turn-only experiment 1(b) showed significantly lower performance. The best model in this setting, TinyFCN with normalization, reached only 62.7% accuracy and 56.3% balanced accuracy, a decline of about 16 percentage points compared to its trip-level counterpart. Precision dropped below 45% in most turn-only models, indicating insufficient discriminative power in short maneuver windows. Some architectures (eg, TCN + Attn) achieved high recall (up to 86.4%) by overpredicting the positive class, but at the cost of precision and balanced accuracy. Due to the weak performance of turn-only models, subsequent experiments focused on configurations that included turn data only in combination with full-trip data, either through feature-level fusion (experiment 2) or model-level fusion (experiment 3). This decision avoided further computational investment in a setting that showed limited utility for early AD detection under naturalistic conditions.

Overall, experiment 1 demonstrates that global driving behavior across full trips offers substantially more predictive information than isolated turning maneuvers. Among all architectures,

XSL•FO
RenderX

compact convolutional models such as TinyFCN provided the best balance of performance and efficiency, while recurrent models with attention were also competitive. The performance of full-trip TinyFCN (accuracy ≈78%, AUC ≈77%) serves as the baseline for subsequent fusion experiments.

It is important to note that earlier studies [5,6] identified turn-specific kinematic features as promising biomarkers for MCI. However, those findings were obtained in controlled environments, where participants drove short, predefined routes under standardized conditions. In such settings, turn segments are clearly defined, confounding variables are minimized, and behavioral comparisons are more reliable. In contrast, the present study analyzed turning maneuvers within naturalistic, long-distance trips that varied widely in road conditions, traffic, and route design. As a result, turn segments were shorter, noisier, and occasionally truncated by sensor limitations. These factors contributed to the limited performance of turn-only models and motivated the shift toward combining turn data with richer trip-level context in the remaining experiments.

Experiment 2 Results

In experiment 2, trip and turn signals were concatenated at the feature level, and a single model was trained on the resulting 20-channel input. A total of 4 architectures (TinyFCN, GRU, long short-term memory, and TCN) were evaluated. Their cross-validated performance is reported in Table 8.

Table 8. Cross-validated performance of feature-level data-fusion models (experiment 2). Trip and turn signals were concatenated into a single 20-channel input and evaluated with 7-fold participant-stratified cross-validation.

Experiment	Model	Normalize	Weights	Accuracy (%)	BA[a] (%)	AUC[b] (%)	Precision (%)	Recall (%)	F_1-score (%)
2	TinyFCN	No	Yes	62	59	52	44	86[c]	56[c]
2	GRU[d]	No	Yes	68[e]	63	63[e]	52[e]	62	53
2	LSTM[f]	No	Yes	67	67[c]	64[c]	55[c]	69[e]	55[e]
2	TCN[g]	No	Yes	70[c]	63[e]	63	50	62	53

[a]BA: balanced accuracy

[b]AUC: area under the receiver operating characteristic curve.

[c]The best values.

[d]GRU: gated recurrent unit.

[e]The second best is only underlined.

[f]LSTM: long short-term memory.

[g]TCN: temporal convolutional network.

Overall, performance improved over the turn-only models from experiment 1(b), confirming that global trip context can offset the sparsity of turn data. However, none of the fusion models outperformed the best full-trip baselines from experiment 1(a). The highest accuracy (70%) was achieved by TCN, though its AUC (63%) and balanced accuracy (63%) fell short of the full-trip TinyFCN benchmark. Long short-term memory delivered the highest balanced accuracy (67%), while GRU achieved similar accuracy and recall but lower precision. TinyFCN, which had previously performed best on full trips, dropped to 62% accuracy and 52% AUC, suggesting that feature-level concatenation may impair its ability to model trip dynamics effectively.

These results indicate that early fusion does not offer a reliable method for combining localized and global driving features. While a modest gain in recall was observed, it came at the cost of precision and overall discriminative power. This degradation likely results from the mismatch in temporal characteristics between trips and turns when merged into a single input. These findings support the need for a late-fusion strategy, as explored in experiment 3, where each data type is modeled independently before integration.

Experiment 3 Results

In the final experiment, the trip and turn streams were treated as complementary modalities. Each was processed by a dedicated encoder, and their latent vectors were merged in a late-fusion classifier. Table 9 summarizes the cross-validated results for 7 dual-branch architectures. The dual-encoder strategy outperformed the feature-level fusion models from experiment 2 across all metrics and approached the performance of the best full-trip baselines from experiment 1(a). The highest overall accuracy (75%) was achieved by DualTinyFCN, which also recorded the best AUC (68%) and matched the highest recall (74%). DualTCN with self-attention produced the strongest balanced accuracy (72%) and recall (77%), while maintaining 75% accuracy. DualGRU with a single attention head offered a solid alternative at 73% accuracy and 65% balanced accuracy; its multihead variant improved recall to 84% but at the cost of precision.

XSL•FO
RenderX

Table 9. Cross-validation results for model-level fusion (experiment 3). Each architecture comprises separate trip and turn encoders whose outputs are combined by either concatenation (DualTinyFCN) or a learned gating mechanism (all others).

Experiment	Model	Normalize	Weights	Accuracy (%)	BA[a] (%)	AUC[b] (%)	Precision (%)	Recall (%)	F_1-score (%)
3	Dual GRU[c]	No	Yes	63	62	58	47	67	50
3	Dual GRU Attn	No	Yes	73	65[d]	63	56[d]	76	62[d]
3	Dual GRU Multi-Head	No	Yes	66	63	66	52	84[e]	60
3	Dual LSTM[f]	No	Yes	65	62	64	51	61	50
3	DualTCN	No	Yes	66	60	67[d]	47	79[d]	57
3	DualTCN Attn	No	Yes	75[d]	72[e]	66	57[e]	77	63[e]
3	Dual Tiny FCN[g]	No	Yes	75[e]	65	68[e]	53	74	61

[a]BA: balanced accuracy

[b]AUC: area under the receiver operating characteristic curve.

[c]GRU: gated recurrent unit.

[d]The second best values.

[e]The best values.

[f]LSTM: long short-term memory

[g]FCN: fully convolutional network

Relative to experiment 2, balanced accuracy improved by 6 to 8 percentage points, and AUC increased by about 5 points, suggesting that separate feature extraction for trips and turns reduced the interference observed with early fusion. While none of the dual-branch models surpassed the single-view TinyFCN baseline from experiment 1(a), the performance gap was reduced to within 4 points, with noticeably stronger recall and class balance.

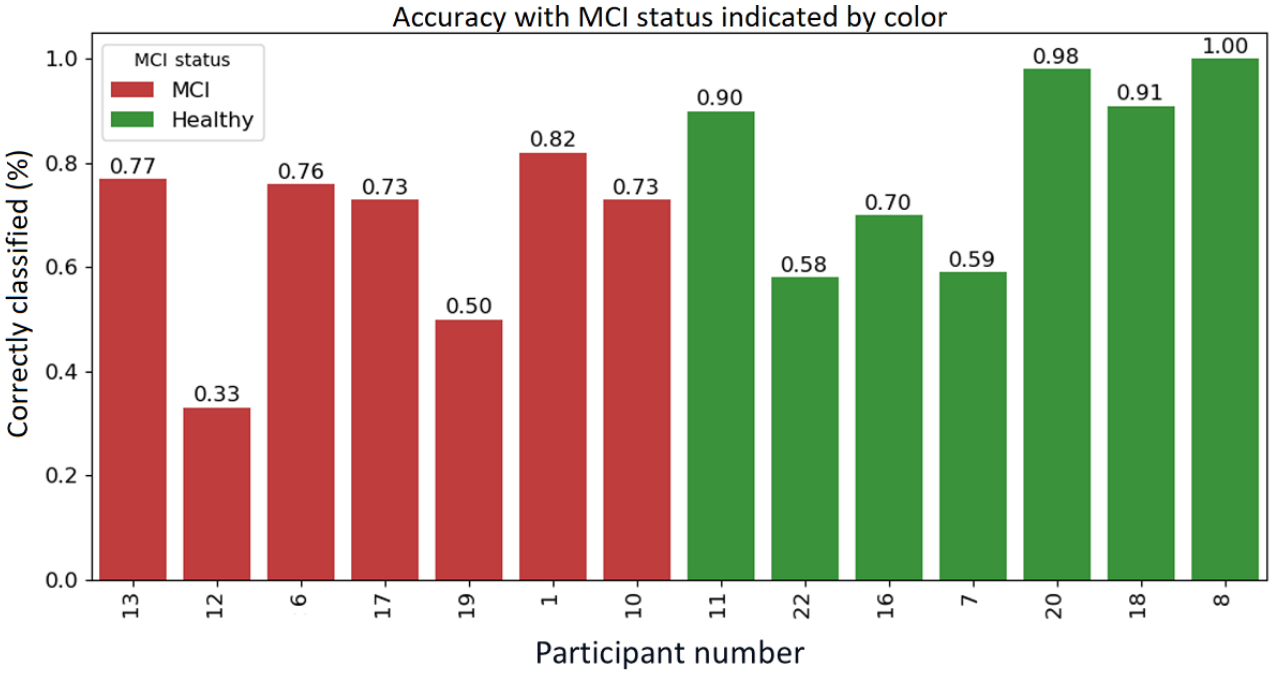
These results indicate that turning behavior carries useful complementary information, which becomes effective when modeled independently before fusion. However, the late-fusion models still fell short of the trip-only baseline, unlike in earlier studies conducted under controlled driving conditions. A likely explanation is the greater behavioral variability in naturalistic turn maneuvers, where factors such as traffic, shallow bends, truncated segments, and GPS inconsistencies introduce noise that diminishes signal quality. To overcome this, 2 improvements are proposed. First, a more selective turn-extraction method could be implemented to filter out low-information maneuvers (eg, shallow curves or stop-sign rolls) and retain only cognitively demanding turns. Second, the use of an adaptive fusion mechanism, such as a trainable gating or context-aware attention layer, could allow the model to dynamically weigh turn features based on context, rather than relying on fixed integration. These enhancements may help future late-fusion models surpass the trip-only benchmark.

Participant-Level Performance of the Best-Performing Model

To further evaluate model effectiveness, participant-level performance was analyzed using the best-performing model from experiment 1(a), which achieved the highest average accuracy (78%). Figure 7 shows the percentage of correctly classified trips per participant, color-coded by diagnostic group.

XSL•FO
RenderX

Figure 7. Per-subject trip classification accuracy of the best performance model, color-coded by cognitive status (green=healthy, red=MCI). MCI: mild cognitive impairment.



This analysis revealed that trips from cognitively healthy participants were classified more accurately than those from the MCI group. Within the MCI cohort, classification accuracy appeared to correlate with the number of available trips. Participants 12 and 19, who contributed only 3 and 4 trips respectively, showed the lowest accuracies (33% and 50%). In contrast, participant 1, also in the MCI group, provided 28 trips and achieved over 80% accuracy, suggesting that greater data volume improves participant-specific classification performance. These findings support the decision to evaluate trips individually rather than applying a majority-vote rule per participant, as done in previous work. As MCI symptoms tend to appear episodically, a majority-vote approach may misclassify participants with sporadic impairments. For instance, participant 12, clinically diagnosed with MCI, was misclassified on 2 of 3 trips. A majority-vote method would incorrectly label this participant as healthy, masking the sporadic but clinically meaningful deviations this framework aims to detect.

The granular trip-wise approach captures the variability that defines mild impairment and provides a more sensitive early warning mechanism than binary participant-level labels. This variability can be expressed as a frequency-based risk score based on the proportion of trips classified as impaired for each driver. Rather than producing a fixed “MCI” or “healthy” label, the score outputs a continuous risk value between 0 and 1, allowing clinicians to define thresholds for follow-up assessment. The choice of such a threshold will have practical implications. A low threshold (eg, flagging drivers if 10% or more of their trips are classified as impaired) increases sensitivity and supports early detection but may lead to more false positives and unnecessary follow-up. A high threshold (eg, 50%) reduces false positives but risks missing individuals with sporadic lapses (precisely the pattern seen in early MCI). As the risk score is continuous, thresholds can be tailored to specific clinical contexts. For example, high sensitivity may be prioritized in population-level screening, while higher precision may be preferred in specialist clinics. By presenting performance curves across thresholds or reporting AUC values alongside confusion matrices at clinically relevant cutoffs, these trade-offs can be made explicit and support informed deployment of the model in practice.

Trip count was also found to influence the stability of those episodic signals. With fewer than 5 trips, a single misclassified journey can flip a driver’s apparent status, explaining the wide swings in accuracy for participants 12 and 19. As the dataset grows (through additional participants, extended monitoring, or data augmentation), the law of large numbers should attenuate this volatility and sharpen the distinction between truly unimpaired drivers and those with early cognitive change. Overall, these results reinforce the importance of context-aware, trip-level evaluation for detecting early MCI. Rather than relying on binary labels, future implementations should adopt frequency-based risk scoring to translate trip-level outputs into longitudinal, clinically actionable indicators.

Discussion

This study demonstrated the potential of naturalistic driving data as a digital biomarker for early MCI detection. Prior limitations, such as controlled test settings, reliance on summary statistics, and rigid evaluation protocols, were addressed through a framework that embraces real-world variability and context-aware analysis, structured around the 3 research questions posed in the Introduction section.

First, the robustness of driving-based biomarkers under naturalistic conditions was assessed (RQ1). Models trained on full-trip data achieved 78% accuracy (AUC 77%) despite variations in routes, traffic, and sensor noise, while turn-only

XSL•FO
RenderX

inputs underperformed by approximately 16 percentage points. These results affirm that early cognitive change is detectable through naturalistic driving; however, short turning maneuvers alone were found to be too noisy to provide reliable indicators.

Second, to address RQ2, a multistage pipeline was developed to transform raw, noisy, multisensor data into structured model inputs. Signals of varying frequencies were aligned, trips segmented, and interpolation applied to ensure consistency. Turn maneuvers were extracted using a targeted method, yielding maneuver-level data while preserving temporal and sensor fidelity for deep learning.

Third, modeling strategies were compared under current data limitations (RQ3). Single-view models on full-trip data remained strongest. While early fusion of trip and turn data failed to outperform this baseline, late fusion using dual encoders significantly improved balanced accuracy and recall, suggesting that turn cues become useful when learned in a dedicated branch. As larger, cleaner turn datasets are acquired, late fusion is expected to surpass trip-only models, especially with mechanisms that emphasize turns only when informative.

In addition, the participant-level analysis revealed that most misclassifications occurred among MCI drivers with few recorded trips, underscoring the effects of behavioral noise and data scarcity. As MCI is marked by sporadic impairments, traditional aggregate predictions may miss subtle signs. By evaluating each trip and maneuver individually, the proposed approach enables detection of episodic lapses, offering a practical early warning system for clinical follow-up.

Future improvements should focus on refining turn extraction through context-aware filtering, enhancing late-fusion models with adaptive gating or attention mechanisms, and enriching the dataset via increased participation and targeted augmentation. Prospective validation and integration of complementary modalities (eg, eye-tracking and in-cabin audio) may yield a more comprehensive behavioral signature of early cognitive decline. Furthermore, given the relatively small dataset, future work could explore the use of transfer learning from models pretrained on large-scale sensor-based time-series datasets. Models trained on motion, activity recognition, or physiological sensing data capture general temporal and kinematic patterns that overlap with vehicle dynamics such as acceleration and rotation. Fine-tuning such pretrained encoders on naturalistic driving data could improve feature extraction, reduce dependence on large labeled cohorts, and enhance generalization to new drivers or sensing platforms. Incorporating transfer learning in this way would be a practical step toward building more data-efficient and scalable cognitive monitoring systems. Collectively, these efforts may further establish naturalistic driving as a viable and nuanced platform for proactive cognitive health monitoring.

Acknowledgments

We want to thank the support of BitBrew Inc (Danlaw Inc)’s colleagues who supported our work by providing the on-board diagnostics. Along this line, we want to express special thanks to Christa Lomazzo, Sudhir Sagi, Priya Gomatham, John Highberg, Mike Carrol, and Larry Dimitrievski for the great product, product support, and collaboration. We also note that TF Health Corporation conducts business under the trade name Breezing Co.

We want to deeply thank the financial and management support of the National Institutes of Health, National Institute of Aging (R43AG078063, R01AG057708, and 2R44AG078063); Wear Tech Center, Partnership for Economic Innovation, Arizona Commerce Authority; and A. J. and Sigismunda Palumbo Charitable Trust. We want to thank the late Ezra Geda, who diligently worked on this project and supported the team with his unforgettable kindness and persistence. Ezra, you are missed, but we will carry forward this project in your honor.

Data Availability

The datasets presented in this paper are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the corresponding author. The underlying code for this study (and training or validation datasets) is not publicly available but may be made available to qualified researchers on reasonable request to the corresponding author.

Funding

This work was supported by the National Institutes of Health (R43AG078063), National Institute of Aging (R01AG057708 and 2R44AG078063), Wear Tech Center, Partnership for Economic Innovation, Arizona Commerce Authority, and Palumbo Charitable Trust.

Authors' Contributions

Conceptualization: FA-H, FT, YEG, TW, EF
Data curation: FA-H, PS
Formal analysis: FA-H, PS
Funding acquisition: TW, EF, FT
Investigation: FA-H, TW
Methodology: FA-H, FT, TW, EF
Project administration: TW, EF

XSL•FO
RenderX

Resources: YEG, FT
Supervision: TW, EF
Validation: FA-H
Visualization: FA-H
Writing - original draft: FA-H, TW, YW, EF, FT, YEG
Writing - review and editing: FA-H, TW, YW, EF, FT, YEG
Software: FA-H
All authors have read and agreed to the published version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experiment 1.
[DOCX File , 15 KB - medinform_v14i1e83622_app1.docx]

Multimedia Appendix 2
Experiment 2.
[DOCX File , 15 KB - medinform_v14i1e83622_app2.docx]

Multimedia Appendix 3
Experiment 3.
[DOCX File , 15 KB - medinform_v14i1e83622_app3.docx]

References

1. Better MA. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]
2. Bennett DA, Schneider JA, Arvanitakis Z, Kelly JF, Aggarwal NT, Shah RC, et al. Neuropathology of older persons without cognitive impairment from two community-based studies. Neurology 2006;66(12):1837-1844. [doi: 10.1212/01.wnl.0000219668.47116.e6] [Medline: 16801647]
3. Knopman DS, Parisi JE, Salviati A, Floriach-Robert M, Boeve BF, Ivnik RJ, et al. Neuropathology of cognitively normal elderly. J Neuropathol Exp Neurol 2003;62(11):1087-1095. [doi: 10.1093/jnen/62.11.1087] [Medline: 14656067]
4. Casey DA, Antimisiaris D, O'Brien J. Drugs for Alzheimer's disease: are they effective? P T 2010;35(4):208-211 [FREE Full text] [Medline: 20498822]
5. Al-Hindawi F, Serhan P, Geda YE, Tsow F, Wu T, Forzani E. LiveDrive AI: a pilot study of a machine learning-powered diagnostic system for real-time, non-invasive detection of mild cognitive impairment. Bioengineering (Basel) 2025;12(1):86 [FREE Full text] [doi: 10.3390/bioengineering12010086] [Medline: 39851360]
6. Serhan P, Victor S, Osorio Perez O, Abi Karam K, Elghoul A, Ransdell M, et al. Smart driving technology for non-invasive detection of age-related cognitive decline. Sensors (Basel) 2024;24(24):8062 [FREE Full text] [doi: 10.3390/s24248062] [Medline: 39771798]
7. Plassman BL, Langa KM, Fisher GG, Heeringa SG, Weir DR, Ofstedal MB, et al. Prevalence of dementia in the United States: the aging, demographics, and memory study. Neuroepidemiology 2007;29(1-2):125-132 [FREE Full text] [doi: 10.1159/000109998] [Medline: 17975326]
8. Shin JH. Dementia epidemiology fact sheet 2022. Ann Rehabil Med 2022;46(2):53-59 [FREE Full text] [doi: 10.5535/arm.22027] [Medline: 35508924]
9. World Alzheimer Report 2019: Attitudes to Dementia. Alzheimer's Disease International. 2019. URL: https://www.alzint.org/resource/world-alzheimer-report-2019/ [accessed 2026-02-05]
10. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol 2014;13(6):614-629. [doi: 10.1016/S1474-4422(14)70090-0] [Medline: 24849862]
11. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7(3):270-279 [FREE Full text] [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
12. Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 2008;41(4):1220-1227 [FREE Full text] [doi: 10.1016/j.neuroimage.2008.03.050] [Medline: 18474436]

XSL•FO
RenderX

13. Crous-Bou M, Minguillón C, Gramunt N, Molinuevo JL. Alzheimer's disease prevention: from risk factors to early intervention. Alzheimers Res Ther 2017;9(1):71 [FREE Full text] [doi: 10.1186/s13195-017-0297-z] [Medline: 28899416]
14. Leifer BP. Early diagnosis of Alzheimer's disease: clinical and economic benefits. J Am Geriatr Soc 2003;51(5 Suppl Dementia):S281-S288. [doi: 10.1046/j.1532-5415.5153.x] [Medline: 12801384]
15. Kishore P, Usha Kumari C, Kumar M, Pavani T. Detection and analysis of Alzheimer's disease using various machine learning algorithms. Mater Today: Proc 2021;45:1502-1508. [doi: 10.1016/j.matpr.2020.07.645]
16. Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 2009;44(4):1415-1422 [FREE Full text] [doi: 10.1016/j.neuroimage.2008.10.031] [Medline: 19027862]
17. Gunes S, Aizawa Y, Sugashi T, Sugimoto M, Rodrigues PP. Biomarkers for Alzheimer's disease in the current state: a narrative review. Int J Mol Sci 2022;23(9):4962 [FREE Full text] [doi: 10.3390/ijms23094962] [Medline: 35563350]
18. Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010;9(1):119-128 [FREE Full text] [doi: 10.1016/S1474-4422(09)70299-6] [Medline: 20083042]
19. Bayat S, Babulal GM, Schindler SE, Fagan AM, Morris JC, Mihailidis A, et al. GPS driving: a digital biomarker for preclinical Alzheimer disease. Alzheimers Res Ther 2021;13(1):115 [FREE Full text] [doi: 10.1186/s13195-021-00852-1] [Medline: 34127064]
20. de la Paz Scribano Parada M, Palau FG, Rodríguez SV, Rincon M, Rico Barroeta MJ, García Rodriguez M, et al. Preclinical cognitive markers of Alzheimer disease and early diagnosis using virtual reality and artificial intelligence: literature review. JMIR Med Inform 2025;13:e62914 [FREE Full text] [doi: 10.2196/62914] [Medline: 39881486]
21. Oh SS, Kang B, Hong D, Kim JI, Jeong H, Song J, et al. A multivariable prediction model for mild cognitive impairment and dementia: algorithm development and validation. JMIR Med Inform 2024;12(1):e59396 [FREE Full text] [doi: 10.2196/59396] [Medline: 39576972]
22. van den Berg RL, van der Landen SM, Keijzer MJ, van Gils AM, van Dam M, Ziesemer KA, et al. Smartphone- and tablet-based tools to assess cognition in individuals with preclinical Alzheimer disease and mild cognitive impairment: scoping review. J Med Internet Res 2025;27:e65297 [FREE Full text] [doi: 10.2196/65297] [Medline: 40424609]
23. Coughlan G, Laczó J, Hort J, Minihane A, Hornberger M. Spatial navigation deficits - overlooked cognitive marker for preclinical Alzheimer disease? Nat Rev Neurol 2018;14(8):496-506. [doi: 10.1038/s41582-018-0031-x] [Medline: 29980763]
24. Allison SL, Fagan AM, Morris JC, Head D. Spatial navigation in preclinical Alzheimer's disease. J Alzheimers Dis 2016;52(1):77-90 [FREE Full text] [doi: 10.3233/JAD-150855] [Medline: 26967209]
25. Di X, Shi R, DiGuiseppi C, Eby DW, Hill LL, Mielenz TJ, et al. Using naturalistic driving data to predict mild cognitive impairment and dementia: preliminary findings from the Longitudinal Research on Aging Drivers (LongROAD) study. Geriatrics (Basel) 2021;6(2):45. [doi: 10.3390/geriatrics6020045] [Medline: 33922735]
26. Bennett JM, Chekaluk E, Batchelor J. Cognitive tests and determining fitness to drive in dementia: a systematic review. J Am Geriatr Soc 2016;64(9):1904-1917. [doi: 10.1111/jgs.14180] [Medline: 27253511]
27. Roe CM, Barco PP, Head DM, Ghoshal N, Selsor N, Babulal GM, et al. Amyloid imaging, cerebrospinal fluid biomarkers predict driving performance among cognitively normal individuals. Alzheimer Dis Assoc Disord 2017;31(1):69-72 [FREE Full text] [doi: 10.1097/WAD.0000000000000154] [Medline: 27128959]
28. Roe CM, Babulal GM, Mishra S, Gordon BA, Stout SH, Ott BR, et al. Tau and amyloid positron emission tomography imaging predict driving performance among older adults with and without preclinical Alzheimer's disease. J Alzheimers Dis 2017;61(2):509-513 [FREE Full text] [doi: 10.3233/JAD-170521] [Medline: 29171997]
29. Cerino ES, Katz MJ, Wang C, Qin J, Gao Q, Hyun J, et al. Variability in cognitive performance on mobile devices is sensitive to mild cognitive impairment: results from the Einstein aging study. Front Digit Health 2021;3:758031 [FREE Full text] [doi: 10.3389/fdgth.2021.758031] [Medline: 34927132]
30. Gorus E, De Raedt R, Lambert M, Lemper J, Mets T. Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. J Geriatr Psychiatry Neurol 2008;21(3):204-218. [doi: 10.1177/0891988708320973] [Medline: 18838743]
31. Zdanowski S, Tieks A, Jeronimus BF, Zuidersma M. Intra-individual variability in cognitive performance can befuddle the study of cognitive impairments and decline. J Alzheimers Dis 2022;85(2):519-525 [FREE Full text] [doi: 10.3233/JAD-210304] [Medline: 34864656]

Abbreviations

1D-CNN: 1D convolutional neural network
AD: Alzheimer disease
AUC: area under the receiver operating characteristic curve
AWS S3: Amazon Web Services Simple Storage Service
GRU: gated recurrent unit
MCI: mild cognitive impairment

XSL•FO
RenderX

ML: machine learning
OBD: on-board diagnostics
RNN: recurrent neural network
TCN: temporal convolutional network

Edited by A Benis; submitted 05.Sep.2025; peer-reviewed by Y Cheng, SB Celebi; comments to author 13.Oct.2025; revised version received 18.Nov.2025; accepted 01.Jan.2026; published 06.Mar.2026.

Please cite as:
Al-Hindawi F, Wu T, Wen Y, Serhan P, Forzani E, Tsow F, Geda YE
Leveraging Naturalistic Driving Digital Biomarkers for Early Mild Cognitive Impairment Detection: Deep Learning Strategies
JMIR Med Inform 2026;14:e83622
URL: https://medinform.jmir.org/2026/1/e83622
doi:10.2196/83622
PMID:

©Firas Al-Hindawi, Teresa Wu, Yutong Wen, Peter Serhan, Erica Forzani, Francis Tsow, Yonas E Geda. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Explainable Machine Learning for Assessing Digital Health Literacy in Older Adults: Validation and Development of a Two-Stage Model Integrating Performance-Based and Self-Assessed Indicators

Choonghee Park[1], MA; Jiyeon Park[2], MA; Seora Kim[1], MA; Ye Seul Bae[3,4], MD, PhD; Jae-Heon Kang[4], MD, PhD; Tae-Min Kim[1,2], MD, PhD; Ji-Won Chun[1,2], PhD

[1]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[2]The Catholic Medical Center Institute for Basic Medical Science, The Catholic University of Korea Catholic Medical Center, Seoul, Republic of Korea

[3]Big Data Research Center, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]Department of Family Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

Corresponding Author:
Ji-Won Chun, PhD
Department of Medical Informatics
College of Medicine
The Catholic University of Korea
222 Banpo-daero, Seocho-gu
Seoul, 06591
Republic of Korea
Phone: 82 3147 9727
Email: jwchun@catholic.ac.kr

Abstract

Background: Digital health literacy (DHL) is the ability to locate, understand, evaluate, and apply health information in digital environments. It is essential for older adults to effectively engage with contemporary health care. However, existing DHL assessments primarily rely on self-reported measures, which are susceptible to subjective bias and often fail to capture actual performance. There is a need for a comprehensive, data-driven approach that integrates objective performance indicators with self-assessments to accurately predict and explain DHL levels in older adults.

Objective: This study develops and validates a machine learning approach to predict DHL levels in older adults by integrating performance-based and self-assessed evaluations.

Methods: We applied a 2-stage methodological framework using 2 independent datasets. In the first stage, to identify performance-based determinants, we assessed actual digital and information comprehension in a separate pilot cohort of 30 older adults (aged 60-74 years). In parallel, to measure self-reported DHL, we conducted an online survey with a distinct group of 1000 older adults (aged 55-74 years) using the Digital Health Literacy Scale and the Korean version of the eHealth Literacy Scale (KeHEALS). Bayesian linear regression was applied to both datasets to identify significant explanatory variables. In the second phase, we trained and validated a binary classification model to predict KeHEALS levels using the survey dataset (n=1000), leveraging the features identified in the first stage. Five machine learning algorithms were evaluated, and the best-performing model was interpreted using Shapley Additive Explanations (SHAP) analysis.

Results: In the pilot performance-based assessment, using a greater number of electronic devices and having higher educational attainment were positively associated with comprehension, whereas alcohol intake showed a negative association. In the self-assessed survey data, key correlates included interest in health-related apps, self-care confidence, age, smoking, alcohol intake, number of devices used, and exercise frequency. Among the machine learning models, categorical boosting demonstrated the most balanced performance (accuracy 0.785, precision 0.769, F1-score 0.765, area under the receiver operating characteristic curve 0.835), outperforming the dummy classifier (accuracy 0.540). SHAP analysis indicated that self-care confidence, health information search, interest in health-related apps, number of electronic devices used, and exercise frequency were the strongest positive contributors to high-DHL predictions, whereas older age and lifestyle factors (alcohol intake, smoking) contributed negatively.

Conclusions: By explicitly integrating performance-based and self-assessed indicators within an explainable machine learning framework, this study demonstrates that DHL in older adults is influenced by both digital engagement and health management

XSL•FO
RenderX

factors. These findings suggest that the proposed framework can serve as a structured approach for evaluating DHL in older adults and inform the design of personalized digital health interventions in clinical and community settings.

(JMIR Med Inform 2026;14:e86171) doi:10.2196/86171

KEYWORDS

digital health literacy; eHealth literacy; digital health care; mHealth; machine learning

Introduction

Background

The rapid advancement of digital technologies and smart devices has transformed everyday life, particularly in the health care domain. Digital health technologies such as electronic health records, telemedicine, and mobile health (mHealth) apps are increasingly recognized as essential tools for enhancing access to health information, supporting self-management, and improving the efficiency and quality of health care delivery [1]. These technologies empower individuals to better manage their health and facilitate personalized, patient-centered care [2]. Furthermore, community-driven digital health systems have demonstrated their potential to establish sustainable and equitable models of care [3].

However, to fully benefit from digital health technologies, individuals must have a sufficient level of digital health literacy (DHL). DHL refers to the ability to locate, understand, evaluate, and apply health-related information in digital environments to make informed decisions and manage health [4]. Higher DHL levels have been associated with improved health outcomes, including better chronic disease management, engagement in preventive behaviors, increased use of health care services, enhanced quality of life, and reduced risk of depression [5-8]. Furthermore, DHL plays a critical role in narrowing the digital divide and promoting equitable access to health resources across diverse populations.

According to the 2024 Digital Information Divide Survey in Korea, the digital informatization level among older adults corresponded to 71.4% of that observed in the total population. While digital access has largely equalized, reaching 95.3% of the general population's level, their digital competencies and utilization indices were relatively lower (55.9% and 75.0%, respectively) [9]. These scores are more likely the result of cohort effects and differences in the availability of educational opportunities, health status, and social and technical support, rather than problems caused by aging [10,11]. Consequently, interventions aimed at enhancing DHL in older populations should be evidence-based and sensitive to the diversity of digital experiences in later life, focusing on modifiable barriers related to access, education, and available support rather than chronological age alone.

The majority of existing studies have relied on self-assessment surveys to measure DHL [12-17]. However, such instruments often fail to capture psychosocial dimensions, such as social support, and may not accurately reflect individuals' actual digital competencies [18]. This underscores the need for a more comprehensive and empirically grounded approach to DHL assessment. Machine learning offers a robust framework for integrating diverse variables, including sociodemographic characteristics, health behaviors, and psychosocial factors, into multidimensional predictive models capable of capturing complex nonlinear interactions [19].

In this study, we employed a Bayesian estimation method to identify key factors associated with DHL-related outcomes, specifically leveraging a wide range of input features, including psychosocial attributes. Subsequently, we developed machine learning models to predict DHL levels and applied Shapley Additive Explanations (SHAP), an explainable artificial intelligence technique, to enhance the interpretability and transparency of the model's outputs [20]. This 2-stage approach aims to provide a more accurate and comprehensive understanding of DHL in older adults and offers practical insights into the design of personalized digital health education and intervention strategies.

Assessment Tool for Measuring Digital Health Literacy: eHealth Literacy Scale

Norman and Skinner's [21] foundational work conceptualized eHealth literacy as a composite of 6 core literacies: traditional literacy, information literacy, media literacy, health literacy, scientific literacy, and computer literacy. This framework is grounded in the "Lily Model," which categorizes skills into analytical and context-specific domains. As digital health technologies have advanced, the original construct of eHealth literacy has broadened and is now more commonly used within the wider notion of DHL.

At the same time, several conceptual and methodological limitations of the eHealth Literacy Scale (eHEALS) have been identified and should be considered when the instrument is used. As a self-report measure, eHEALS assesses subjective confidence rather than objectively measured performance, so scores may diverge from actual digital and health-related skills [22,23]. Factor-analytic studies have also reported inconsistent dimensional structures across populations and languages, indicating that the underlying constructs may vary by cultural and demographic context and that direct score comparisons or universal cutoffs should be interpreted with caution [24]. Nevertheless, eHEALS has been translated and validated in numerous countries, demonstrates consistently high internal consistency, and remains a pragmatic core instrument for assessing perceived eHealth literacy in large-scale surveys and clinical research [24-33].

Using eHEALS, Berkowsky [17] assessed the DHL levels of adults aged 65 and older in California and examined the impact of sociodemographic characteristics and digital usage patterns. Regression analyses indicated that higher educational attainment, more frequent internet use, and engagement in a wider range of online activities (eg, email, online shopping, social media)

XSL•FO
RenderX

were associated with higher DHL levels, whereas basic demographic variables such as age, gender, and income were not significant predictors. These findings suggest that actual digital engagement may be more critical than demographic background in shaping perceived DHL.

Similarly, Kyaw et al [34] investigated DHL among 434 adults aged 65 and older in urban settings, focusing on the influence of social and digital factors. Their findings revealed that older age was negatively associated with DHL, whereas higher income, extended internet use, and active participation in social media showed positive associations. Notably, older adults who used social media exhibited approximately 4-fold higher DHL levels than nonusers, and those using the internet for more than 3 hours per day demonstrated up to a 35-fold increase in DHL levels. These findings suggest that improving DHL among older adults requires not only the provision of digital technologies but also strategies to promote real-world digital engagement and enhance social participation in digital environments.

Assessment Tool for Measuring Digital Health Literacy: Digital Health Literacy Scale

Early definitions of eHealth literacy focused primarily on individuals' abilities to search for and understand health information on the internet. However, today's digital environment increasingly demands interactive engagement through platforms such as social media, mHealth apps, and telemedicine. Consequently, the concept has evolved into DHL, which encompasses not only information seeking but also the skills to evaluate, apply, and interact with digital health content across diverse platforms [35].

In line with this expanded conceptualization, the Digital Health Literacy Scale (DHLS) was developed as a multidimensional tool to comprehensively measure DHL. The DHLS evaluates competencies across 4 domains: use of digital devices, understanding of health information, use and decision of health information, and use intention [36]. The use of digital devices domain assesses individuals' proficiency in operating digital tools and accessing health-related services. The understanding of health information domain measures the ability to comprehend online content, such as medication instructions and test results. The use and decision of health information domain evaluates both the application of health information in decision-making and the critical judgment of its credibility. The use intention domain captures users' motivation and willingness to engage with digital health platforms and services. It was specifically designed to reflect the behavioral, cognitive, and motivational characteristics commonly observed in older adults, particularly those vulnerable to the digital divide. By independently assessing both digital and health literacy and integrating them into a unified DHL construct, the DHLS facilitates a comprehensive and multidimensional assessment of digital health literacy. This approach helps identify differences in digital health literacy across various skill domains, from basic digital operations to advanced critical thinking, providing actionable insights into the specific competencies that require support among older adults.

Machine Learning Approach for Assessing Digital Health Literacy

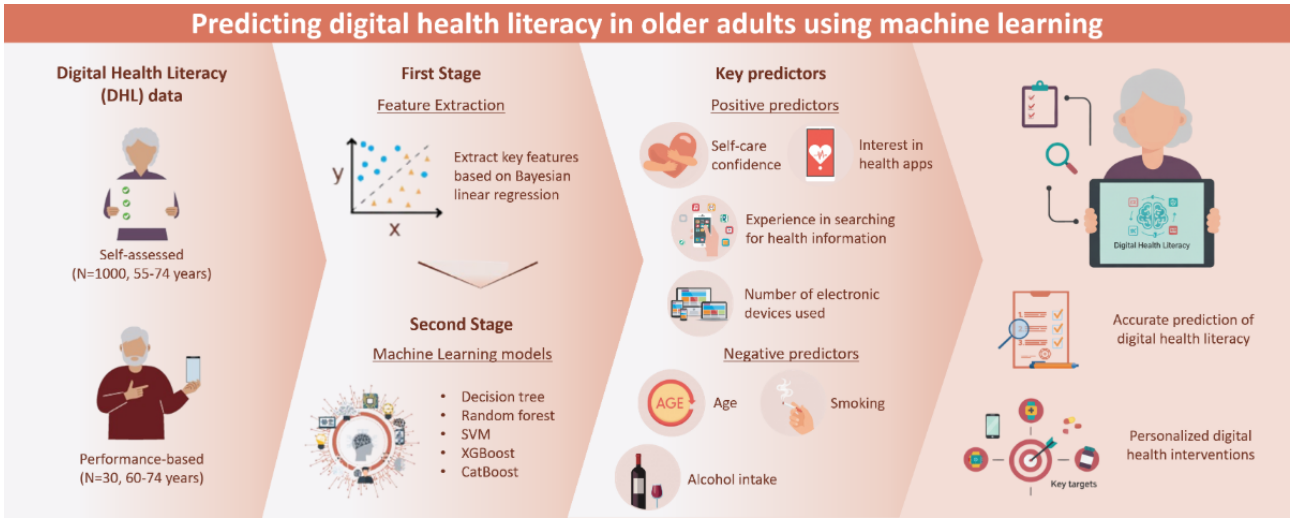
Given what we discussed so far, eHEALS and DHLS are widely used self-report measures for evaluating digital health literacy. Most studies employing these tools use univariate analyses to examine the effects of individual variables independently [37-39]. However, because self-assessment instruments are inherently grounded in subjective perceptions, they have substantial limitations in capturing actual performance and real-world information behaviors. This concern aligns with extensive evidence from cognitive and behavioral research demonstrating that self-assessments are often systematically biased and display only modest associations with objective indicators of competence [40]. Moreover, they fail to account for interactions among variables or evaluate the relative importance of each factor [41,42]. These limitations emphasize the need to develop data-driven approaches that use diverse individual-level characteristics to accurately predict DHL.

To address these limitations, this study adopted a comprehensive 2-stage machine learning framework that integrates performance-based and self-assessed indicators of DHL. Machine learning is a statistical approach that predicts and classifies individual conditions by learning complex, nonlinear patterns from multidimensional data [43]. Research suggests that machine learning algorithms, including Extreme Gradient Boosting (XGBoost), random forest, and logistic regression, can identify key sociodemographic factors influencing general health literacy [44]. Existing studies have demonstrated the effectiveness of machine learning in uncovering intricate variable interactions and enhancing predictive accuracy [45-47], reflecting a broader shift toward objective data-driven assessment in health literacy research.

This study integrated performance-based and self-reported assessments to identify key features associated with DHL. Based on these significant features, a prediction model was developed using a machine learning algorithm. Furthermore, we used SHAP to interpret the contribution of each feature, thereby proposing an explainable artificial intelligence approach for medical apps. This approach aims to effectively predict DHL in older adults and provide a foundation for developing personalized intervention strategies (Figure 1).

XSL•FO
RenderX

Figure 1. Graphical abstract.

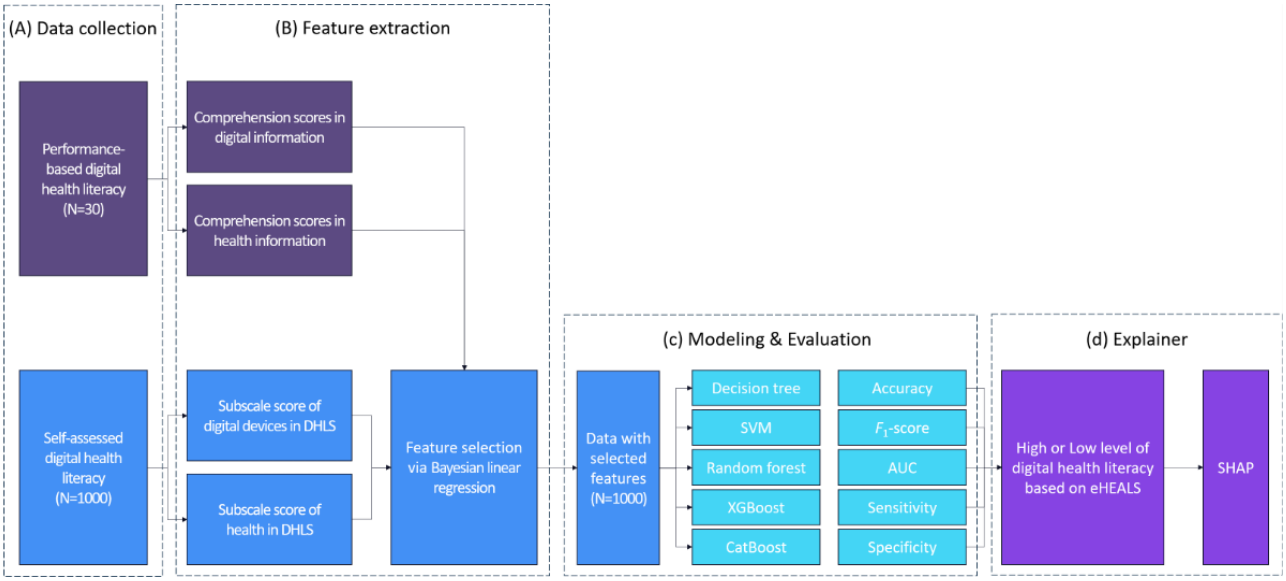


Methods

Overview

In this study, we designed a 2-stage machine learning framework integrating self-assessed (n=1000) and performance-based (n=30) evaluations (Figure 2). In the initial phase, using a performance-based task, digital literacy and health care information scores were obtained (Figure 2A). Subsequently, Bayesian linear regression analysis was performed to identify explanatory variables significantly associated with DHL (Figure 2B). The performance-based evaluation was employed to enhance the explanatory power of the variables and was excluded from the training phase of the DHL predictive model. In the second phase, using a dataset of 1000 self-assessments and the characteristics identified in the initial phase, we constructed prediction models to classify DHL levels (high vs low; Figure 2C). We conducted comparative evaluations of multiple machine learning models, including decision tree, support vector machine (SVM), random forest, XGBoost, and categorical boosting (CatBoost). In addition, SHAP analysis was used to quantitatively assess the impact of significant variables on DHL predictions, thereby enhancing the interpretation of the model's predicted results (Figure 2D).

Figure 2. Overview of the framework used to predict digital health literacy level. AUC: area under the receiver operating characteristic; CatBoost: categorical boosting; DHLS: Digital Health Literacy Scale; eHEALS: eHealth Literacy Scale; SHAP: Shapley Additive Explanations; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.



Performance-Based Digital Health Literacy

Participants

To evaluate performance-based DHL, we recruited a total of 30 participants aged 60-74 (mean age 69.1 years). Participants were enrolled using a nonprobability snowball sampling strategy facilitated through local community networks. To ensure data validity and minimize potential bias, intermediaries who arranged the initial contact were excluded from participation. Trained research staff (CP and SK) conducted a brief telephone screening to confirm eligibility and then scheduled a home visit. During the visits, researchers provided comprehensive information regarding the study's purpose, procedures, anonymity, data usage, and participants' rights. Written

XSL•FO
RenderX

informed consent was obtained before administration of the behavioral tasks. Participation was voluntary, and all participants were informed that they could withdraw from the study at any time without penalty.

To examine differences in performance-based DHL, participants were divided into 2 groups using the eHEALS cutoff score as the criterion: low DHL (n=15) and high DHL (n=15), based on their Korean version of the eHealth Literacy Scale (KeHEALS) scores, as shown in Table 1 [48]. The 2 groups showed no significant differences in age (mean age: low DHL=68.7 years, high DHL=69.4 years; *P*=.19) or gender (male/female: low DHL=2/13, high DHL=5/10; *P*=.34). In addition, no significant difference was found between the groups in years of education (*P*=.49).

Table 1. Comparison of performance-based sociodemographic variables by DHL[a] level based on the KeHEALS[b] score.

Characteristics	Groups n (%)		Total n (%)
	Low	High	
Gender			
Male	2 (29)	5 (71)	7 (100)
Female	13 (57)	10 (43)	23 (100)
Age (years)			
60-64	1 (33)	2 (67)	3 (100)
65-69	9 (64)	5 (36)	14 (100)
70-74	5 (38)	8 (62)	13 (100)
Years of education			
Elementary school	3 (60)	2 (40)	5 (100)
Middle school	3 (75)	1 (25)	4 (100)
High school	8 (57)	6 (43)	14 (100)
Community college	0 (0.0)	1 (100)	1 (100)
University or higher	1 (17)	5 (83)	6 (100)

[a]DHL: digital health literacy.

[b]KeHEALS: Korean version of the eHealth Literacy Scale.

Measurements

Participants engaged in content reading focused on digital literacy and health care information, followed by an identification task designed to assess comprehension. Each reading selection comprised 2 or 3 paragraphs and was displayed on a Samsung Galaxy A15 smartphone equipped with a 163.9-mm display (1080 × 2340 pixels). The smartphone was used as the experimental tool to deliver all visual materials for task performance in a controlled viewing environment. During the identification task, participants were asked to indicate whether 10 digital literacy sentences and 10 health care information sentences derived from the readings were true or false. The average time required to complete the assessment did not exceed 30 minutes. To identify key explanatory variables, scores from this identification task were subsequently used in the initial stage of the DHL framework. Before the assessment, demographic information (gender, age, and years of education), psychosocial factors (living arrangements), digital-related factors (interest in health-related apps, health information search, and number of digital devices used), and health-related factors (alcohol intake, smoking, and exercise frequency) were collected from each participant. Interest in health-related apps was measured using a 5-point Likert scale (1=strongly disagree to 5=strongly agree), capturing the degree of interest in using health-related apps in digital environments. Health information search was assessed using a binary (yes/no) item asking whether participants had personally searched for health-related information.

To evaluate participants' overall level of digital health literacy, the KeHEALS was administered. KeHEALS is an 8-item self-report instrument designed to measure individuals' perceived ability to locate, evaluate, and apply electronic health information, demonstrating high internal consistency (Cronbach α=.89) and good test-retest reliability over a 2-week interval (intraclass correlation 0.80). Each item was rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), yielding a total score between 8 and 40, with higher scores indicating greater perceived digital health literacy. Participants were classified into 2 groups using a sample-specific threshold based on the median total eHEALS score of 29 in the entire sample: those with a total score greater than 29 were categorized as having a high level of digital health literacy, whereas those with a score of 29 or below were categorized as having a low level. This classification facilitated subsequent analyses by enabling group comparisons according to participants' perceived competencies in engaging with digital health information.

XSL•FO
RenderX

Self-Assessed Digital Health Literacy

Participants

To evaluate self-assessed DHL for the predictive modeling phase, survey data were collected from 1000 adults aged 55-74 (mean age 63.6 years). All participants were recruited through an online survey platform operated by a survey agency. Recruitment was facilitated via email and website postings. To ensure a representative sample of the general population, stratified quota sampling was employed based on gender and age groups. Each respondent was assigned a unique access link to ensure single participation and eliminate the possibility of repeated submissions. Given the prospective and anonymous nature of this large-scale survey, the requirement for written informed consent was waived by the Institutional Review Board. Participants were first presented with information on the initial survey screen that provided comprehensive details about the study objectives and procedures. Proceeding to the questionnaire was considered to indicate informed willingness to participate.

To examine group-level differences in DHL, participants were classified into 2 groups based on KeHEALS scores: low DHL (n=563) and high DHL (n=437; Table 2). The high-DHL group had a greater proportion of participants in the younger age range of 55-59 (148/437, 33.9%), whereas the low-DHL group included more participants in the 65-69 (242/563, 43.0%) and 70-74 (80/563, 14.2%) age ranges. Age distribution also differed between the groups (mean age: low DHL=64.2 years, high DHL=62.8 years). Differences were likewise observed in gender distribution (male/female: low DHL=257/306, high DHL=243/194). The proportion of males was higher in the high-DHL group (243/437, 55.6%) than in the low-DHL group (257/563, 45.6%), whereas females constituted the majority in the low-DHL group (306/563, 54.4%). Regarding years of education, a marked difference was found between the 2 groups. In the high-DHL group, more than half of the participants (235/437, 53.8%) had obtained a university degree or higher, whereas the low-DHL group had a greater proportion of high school graduates (271/563, 48.1%) and individuals with less than a middle school education.

Table 2. Comparison of self-assessed sociodemographic variables by DHL[a] level based on the KeHEALS[b] score.

Characteristics	Groups, n (%)		Total, n (%)
	Low	High	
Gender			
Male	257 (51.4)	243 (48.6)	500 (100)
Female	306 (61.2)	194 (38.8)	500 (100)
Age (years)			
55-59	134 (47.5)	148 (52.5)	282 (100)
60-64	107 (57.5)	79 (42.5)	186 (100)
65-69	242 (57.9)	176 (42.1)	418 (100)
70-74	80 (70.2)	34 (29.8)	114 (100)
Education			
Elementary school	5 (83.3)	1 (16.7)	6 (100)
Middle school	45 (93.8)	3 (6.3)	48 (100)
High school	271 (66.4)	137 (33.6)	408 (100)
Community college	62 (50.4)	61 (49.6)	123 (100)
University or higher	180 (43.4)	235 (56.6)	415 (100)

[a]DHL: digital health literacy.

[b]KeHEALS: Korean version of the eHealth Literacy Scale.

Measurements

DHL levels were assessed using the DHLS [36] and KeHEALS [48]. The DHLS consists of 4 subscales: use of digital devices (10 items), understanding of health information (5 items), use and decision of health information (5 items), and use intention (5 items). Participants completed the DHLS questionnaire using a 5-point Likert-type response scale (1=strongly disagree, 5=strongly agree). To determine key explanatory variables related to digital literacy and health care information in older adults, we focused on 3 subscales: use of digital devices, understanding of health information, and use and decision of health information. Based on a sample-specific threshold corresponding to the median total eHEALS score of 29, eHEALS was dichotomized into 2 groups and used as a predictive variable in the machine learning model. Furthermore, we collected demographic information (gender, age, and years of education), psychosocial factors (living arrangements and social support), digital-related factors (interest in health-related apps, health information search, and number of digital devices used), and health-related factors (alcohol intake, smoking, exercise frequency, and self-care confidence). Self-care confidence was assessed as an indicator of health empowerment in older adults using a single item derived from the Korean

XSL•FO
RenderX

Health Empowerment Scale, which measures confidence in making health management decisions [49]. This item was rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree).

Statistical Analysis

Feature Extraction Based on Bayesian

To identify the most influential factors associated with DHL, we applied Bayesian linear regression to both performance- and survey-based datasets. This approach improves modeling efficiency while enhancing interpretability by incorporating prior-based regularization and probabilistic estimation of regression coefficients, thereby allowing direct assessment of variable importance and uncertainty [50]. Each model was specified separately for each observation unit as follows:

where Y_i is the standardized score of the ith participant; X_{ij} is the value of the jth feature for the ith participant; β_0 is the intercept; β_j is the regression coefficient for each feature; ε_i is the error term; and p' is the total number of features considered.

The likelihood for the observed data was specified as follows:

We assigned weakly informative priors to the model parameters to ensure that posterior inferences were primarily driven by the observed data. Specifically, the intercept term β_0 was assigned a prior distribution $N(0, 5)^2$, and each regression coefficient β_j received an independent prior $N(0, 1^2)$ for j=1, ..., p', reflecting an implicit L2 regularization effect. The SD of the error term σ was assigned a HalfCauchy(2) prior.

Posterior distributions were approximated using Markov chain Monte Carlo sampling. For each model, 4 parallel chains were run, each with 2000 tuning steps followed by 2000 sampling steps. Convergence was assessed by monitoring divergence diagnostics throughout the sampling process.

Feature selection was conducted using a Bayesian inference framework. The dataset was first split into training and test sets, and all feature selection procedures were performed exclusively on the training set. In the performance-based dataset, we separately identified predictors of comprehension scores for digital information and health information. In the self-assessed dataset, feature selection was performed for each DHLS subscale, namely, use of digital devices and health information understanding and utilization. To prevent data leakage during the Markov chain Monte Carlo sampling procedure used for feature selection, all modeling steps were conducted exclusively on the training dataset. Variables were retained only if the 95% highest density interval (HDI) of their posterior regression coefficients β did not include 0, indicating a statistically significant positive or negative association with the outcome variable. These statistically validated features were then used exclusively in subsequent stages to construct the final predictive model for DHL.

Modeling

In this study, final predictive modeling was performed using the self-assessed DHL dataset comprising 1000 participants. The goal of the model was to accurately classify participants into 2 distinct groups: high DHL and low DHL. Before developing the final predictive model, a feature selection process was conducted. This process aimed to identify features strongly associated with both performance-based and self-assessed DHL measures. The resulting feature set was designed to comprehensively represent the various aspects of DHL.

Using this selected feature set, 5 machine learning models—decision tree, random forest, XGBoost, CatBoost, and SVM—were implemented to predict DHL levels [51-55]. During data preprocessing, we applied a structured workflow to handle distinct variable types and used mixed encoding strategies to reflect the measurement properties of each predictor. Nominal categorical variables were encoded using OneHotEncoder. Ordinal variables were generally processed with OrdinalEncoder to preserve their inherent ranking. However, for selected lifestyle variables such as alcohol intake, we applied OneHotEncoder to capture category-specific associations across frequency levels that may not be adequately represented by a single ordered scale. These transformations were integrated into a unified preprocessing pipeline and applied consistently across all data splits, ensuring uniform feature representation and minimizing the risk of data leakage.

For model training and testing, the preprocessed dataset was split into a training set (n=800) and a test set (n=200). The hyperparameters of each model were optimized using the Optuna framework [56], with 10-fold cross-validation applied to the training set to maximize accuracy. Following hyperparameter optimization, to assess generalization performance, each final model was trained on the entire training set and evaluated on the test set.

Performance Evaluation

Model performance in predicting DHL levels was quantitatively evaluated using standard metrics for binary classification, including accuracy, precision, sensitivity, specificity, F_1-score, and area under the receiver operating characteristic curve (ROC AUC). Accuracy indicates the overall proportion of correctly classified cases. Precision reflects the proportion of true high-DHL cases among those predicted to have high DHL. The F_1-score, which is the harmonic mean of precision and recall, provides a balanced metric that is particularly useful under class imbalance. The ROC AUC assesses the model's ability to distinguish between high- and low-DHL cases across varying decision thresholds. Collectively, these metrics offer a comprehensive evaluation of model performance.

Furthermore, to address the "black box" nature of machine learning and ensure clinical transparency, we used the SHAP framework. SHAP is a model-agnostic method based on cooperative game theory that assigns each feature an importance value for a particular prediction. It calculates the Shapley value, which represents the average marginal contribution of a feature across all possible coalitions of features [20]. Unlike traditional feature-importance metrics that provide only a global ranking,

XSL•FO
RenderX

SHAP offers both global and local interpretability. It allows us to quantify not only the magnitude but also the direction of each feature's impact on the model's output.

In the context of our model for predicting digital health literacy, SHAP provides a granular explanation of how specific demographic, interest in digital health care, health management, and psychosocial features influence the classification of an individual into the high-DHL or low-DHL group. This methodology enables visualization of individual and global feature effects, allowing identification of key determinants of DHL and verification of whether the model's learned patterns are consistent with established clinical research on digital health.

Ethical Consideration

This study was reviewed and approved by the Institutional Review Board of the Catholic University of Korea, College of Medicine (survey study approval number MIRB 20230825-007; behavioral task study approval number MIRB 20241014-003), and was conducted in accordance with institutional and national guidelines for research involving human participants. For the web-based survey, an online written information sheet described the study purpose, procedures, minimal-risk nature, anonymity, data use, and participants' rights. Data were collected only from individuals who voluntarily agreed to participate after reading this information, with explicit notice that they could withdraw from the survey at any time without penalty. For the behavioral task study, participants received a written information sheet and provided written informed consent before participation. In the survey study, no personally identifiable information was collected. In the behavioral task study, personal information was collected solely for compensation purposes, stored separately from research data, and destroyed immediately after compensation. In both studies, no personal identifiers were used for analysis or reporting; all analytic datasets were anonymized, encrypted, and stored on secure institutional servers accessible only to the research team. Participants who completed the web-based survey received 500 points (approximately US $0.35) redeemable within the online platform, and participants in the behavioral task study received KRW 30,000 in cash (approximately US $20) as compensation for their time. No identifiable images of participants were included in the manuscript or its multimedia appendix.

Results

Post Hoc Age-Group Comparison of Selected Features

To verify the validity of applying selected features from a small cohort of individuals aged 60 years and older to the 55-59 age group included in the final model during the feature-extraction phase, we divided participants into 2 age groups and performed statistical comparisons of the selected features. The comparison revealed significant demographic and behavioral differences in several variables (Table 3). Modest but statistically significant age-related differences were observed for interest in health-related apps (P=.03), health information search (P=.02), smoking (P=.004), and alcohol intake (P=.007). However, no significant differences were observed in the number of electronic devices used (P=.41), self-care confidence (P=.19), or exercise frequency (P=.86).

XSL•FO
RenderX

Table 3. Comparison of self-assessed digital and health-related factors by age groups.

Comparison	Age groups (years), n (%)		Chi-square (*df*)	Cramer *V*	*P* value
	55-59 (n=282)	60-74 (n=718)			
Interest in health-related apps			10.388 (4)	0.102	.03
Strongly disagree	14 (5.0)	76 (10.6)			
Disagree	46 (16.3)	140 (19.5)			
Neither agree nor disagree	75 (26.6)	169 (23.5)			
Agree	86 (30.5)	196 (27.3)			
Strongly agree	61 (21.6)	137 (19.1)			
Health info search			5.617 (1)	0.075	.02
Yes	224 (79.4)	518 (72.1)			
No	58 (20.6)	200 (27.9)			
Number of electronic devices used			2.902 (3)	0.054	.41
1	100 (35.5)	278 (38.7)			
2	111 (39.4)	291 (40.5)			
3	58 (20.6)	116 (16.2)			
4	13 (4.6)	33 (4.6)			
Self-care confidence			6.095 (4)	0.078	.19
Strongly disagree	0 (0.0)	5 (0.7)			
Disagree	23 (8.2)	58 (8.1)			
Neither agree nor disagree	101 (35.8)	273 (38.0)			
Agree	151 (53.5)	347 (48.3)			
Strongly agree	7 (2.5)	35 (4.9)			
Exercise frequency per week			1.314 (4)	0.036	.859
0	38 (13.5)	81 (11.3)			
1	104 (36.9)	259 (36.1)			
2	63 (22.3)	165 (23.0)			
3	49 (17.4)	133 (18.5)			
4	28 (9.9)	80 (11.1)			
Smoking			13.192 (3)	0.115	.004
Never smoke	158 (56.0)	371 (51.7)			
Former smoker	66 (23.4)	244 (34.0)			
Occasional smoker	11 (3.9)	23 (3.2)			
Heavy smoker	47 (16.7)	80 (11.1)			
Alcohol intake			17.904 (6)	0.133	.007
Never drink	26 (9.2)	110 (15.3)			
Abstinence in the past year	33 (11.7)	120 (16.7)			
Less than once a month	52 (18.4)	147 (20.5)			
About once a month	38 (13.5)	90 (12.5)			
2-4 times a month	84 (29.8)	152 (21.2)			
2-3 times a week	36 (12.8)	69 (9.6)			
4 or more times a week	13 (4.6)	30 (4.2)			

XSL•FO
RenderX

Feature Selection From Performance-Based Digital Health Literacy

Feature selection using Bayesian linear regression identified 3 features that were statistically associated with digital information comprehension. The number of electronic devices used showed a strong positive association with comprehension scores (β=.59, 95% HDI 0.21-0.93). By contrast, 2 variables related to alcohol intake were negatively associated with comprehension. Specifically, alcohol intake once a month (β=–1.48, 95% HDI –2.43 to 0.49) and alcohol intake 2-4 times a month (β=–1.05, 95% HDI –1.76 to 0.35) were associated with lower performance in digital literacy tasks.

For health care information comprehension, 2 features were statistically significant. Years of education showed a positive association with comprehension performance (β=.48, 95% HDI 0.10-0.88), indicating that higher educational attainment contributes to better understanding of health-related information. Conversely, alcohol intake 2-4 times a month was again associated with reduced performance (β=–1.03, 95% HDI –1.96 to 0.16). Finally, the features selected from the performance-based DHL were the number of electronic devices used, alcohol intake, and years of education. These selected variables were subsequently used in the second-stage predictive modeling of DHL levels.

Feature Selection From Self-Assessed Digital Health Literacy

The significant predictors associated with 2 DHLS subscales, namely, use of digital devices and health information understanding and utilization, were identified using Bayesian linear regression.

For the digital device utilization subscale, 12 features were statistically significant. Interest in health-related apps was the strongest positive predictor (β=.44, 95% HDI 0.39-0.48). Moreover, several health behavior variables showed positive associations, including abstinence from alcohol intake in the past year (β=.31, 95% HDI 0.09-0.51), alcohol intake about once a month (β=.37, 95% HDI 0.15-0.57), alcohol intake 2-4 times a month (β=.31, 95% HDI 0.11-0.49), alcohol intake 2-3 times a week (β=.26, 95% HDI 0.04-0.52), and being a nonsmoker (β=.19, 95% HDI 0.001-0.37). In addition, self-care confidence (β=.13, 95% HDI 0.03-0.26) was positively associated with digital device use. By contrast, several variables demonstrated significant negative associations with digital device use, including age 70-74 years (β=–.31, 95% HDI –0.50 to 0.11), age 65-69 years (β=–.17, 95% HDI –0.31 to 0.05), number of electronic devices used (β=–.19, 95% HDI –0.26 to 0.13), exercise frequency (β=–.11, 95% HDI –0.17 to 0.05), and years of education (β=–.07, 95% HDI –0.14 to 0.01). Distributions of digital literacy scores by feature values are presented in Multimedia Appendix 1.

For the subscale on health information understanding and utilization, 7 predictors showed statistically significant associations. Interest in health-related apps emerged as the strongest positive predictor (β=.34, 95% HDI 0.27-0.38). Lack of health information search (β=.17, 95% HDI 0.02-0.32), abstinence from alcohol intake in the past year (β=.28, 95% HDI 0.07-0.55), and former smoking status (β=.34, 95% HDI 0.03-0.69) also showed positive associations with health information understanding and utilization. By contrast, negative associations were observed for the number of electronic devices used (β=–.14, 95% HDI –0.21 to 0.07), age 70-74 years (β=–.29, 95% HDI –0.54 to 0.09), and age 65-69 years (β=–.25, 95% HDI –0.41 to 0.11). These variables, including digital information interest and health behavior factors selected from the self-assessed DHL, were subsequently used as input features in the modeling phase.

Predictive Modeling of Digital Health Literacy

The performance of the 5 machine learning models was evaluated and compared using multiple classification metrics, including accuracy, precision, and F_1-score. The detailed results are summarized in Table 4, and the corresponding ROC curves are shown in Figure 3.

To establish a performance baseline given the class distribution, a dummy classifier was evaluated, yielding an accuracy of 0.540. Relative to this baseline, all machine learning models demonstrated improved predictive performance. Among the 5 models, SVM showed strong results, achieving an accuracy of 0.775, a precision of 0.758, and an F_1-score of 0.754. By contrast, the decision tree model exhibited the lowest performance across all metrics, with an accuracy of 0.715, a precision of 0.661, and an F_1-score of 0.716. In addition to standard metrics, ROC curve analysis was conducted to assess the discriminative capability of each model. XGBoost and SVM achieved the highest AUC values (0.850 and 0.841, respectively). Despite these results, CatBoost demonstrated the most balanced and superior overall performance, with an accuracy of 0.785, a precision of 0.769, and an F_1-score of 0.765. Compared with the dummy classifier, CatBoost improved accuracy by 0.245, which was statistically significant (P<.001). This consistently strong performance across primary classification metrics constituted the principal rationale for selecting CatBoost for subsequent SHAP-based post hoc interpretability analysis. These findings indicate that, for DHL prediction, the CatBoost model offers robust and reliable classification performance among the algorithms tested.

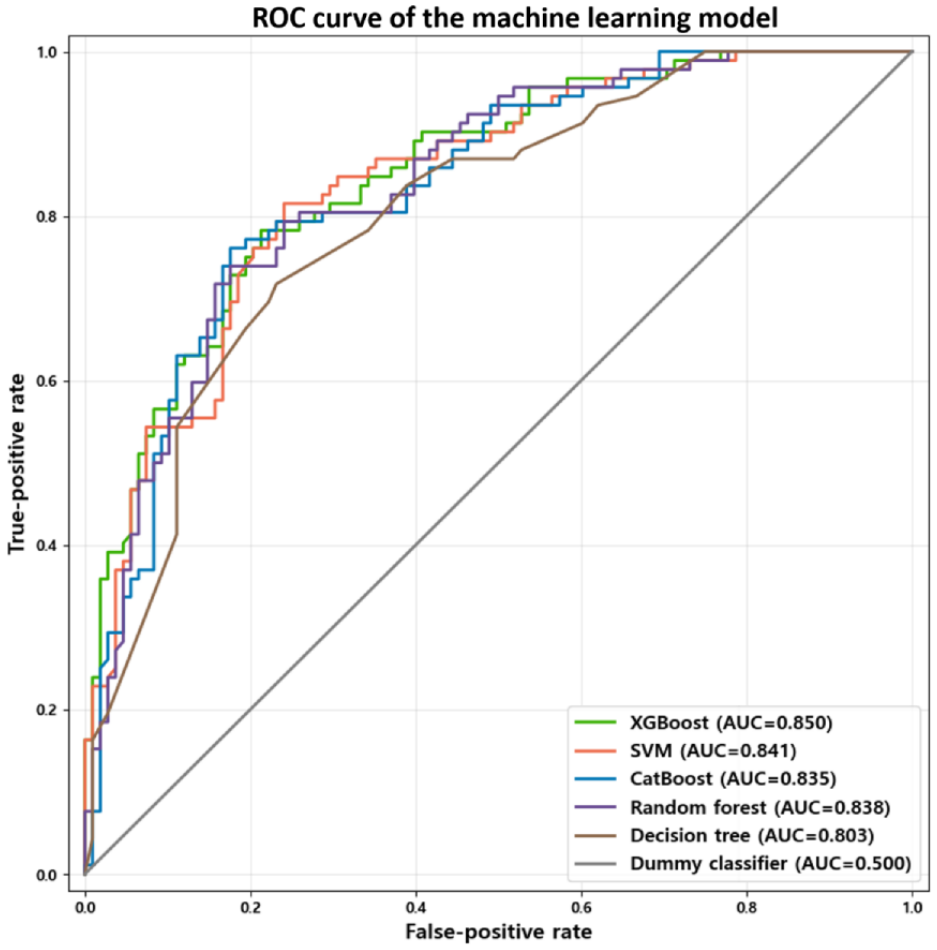
XSL•FO
RenderX

Table 4. Performance metrics of machine learning models.[a]

Model	Accuracy	Precision	Specificity	Sensitivity	F_1-score	Area under the receiver operating characteristic
Dummy classifier	0.540	0.000	1.000	0.000	0.000	0.500
Decision tree	0.715	0.661	0.657	*0.783*	0.716	0.803
Support vector machine	0.775	0.758	0.796	0.750	0.754	0.841
Random forest	0.780	*0.793*	0.843	0.707	0.747	0.838
Extreme gradient boosting	0.780	0.767	0.806	0.750	0.758	*0.850*
Categorical boosting	*0.785*	0.769	*0.806*	0.761	*0.765*	0.835

[a]Values displayed in italics represent the highest performance achieved by each metric when comparing the performance of the 5 machine learning models.

Figure 3. Accuracy and receiver operating characteristic (ROC) curve. AUC: area under the receiver operating characteristic; CatBoost: categorical boosting; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.



Model Interpretation Using SHAP Analysis

The SHAP analysis was conducted to identify the features that contributed most substantially to the model's predictions of DHL levels. Figure 4A presents the mean absolute SHAP values for each feature, indicating their relative importance within the model. The 5 most influential features were self-care confidence, lack of health information search, interest in health-related apps, number of electronic devices used, and exercise frequency. These variables played a central role in distinguishing between high- and low-DHL classifications.

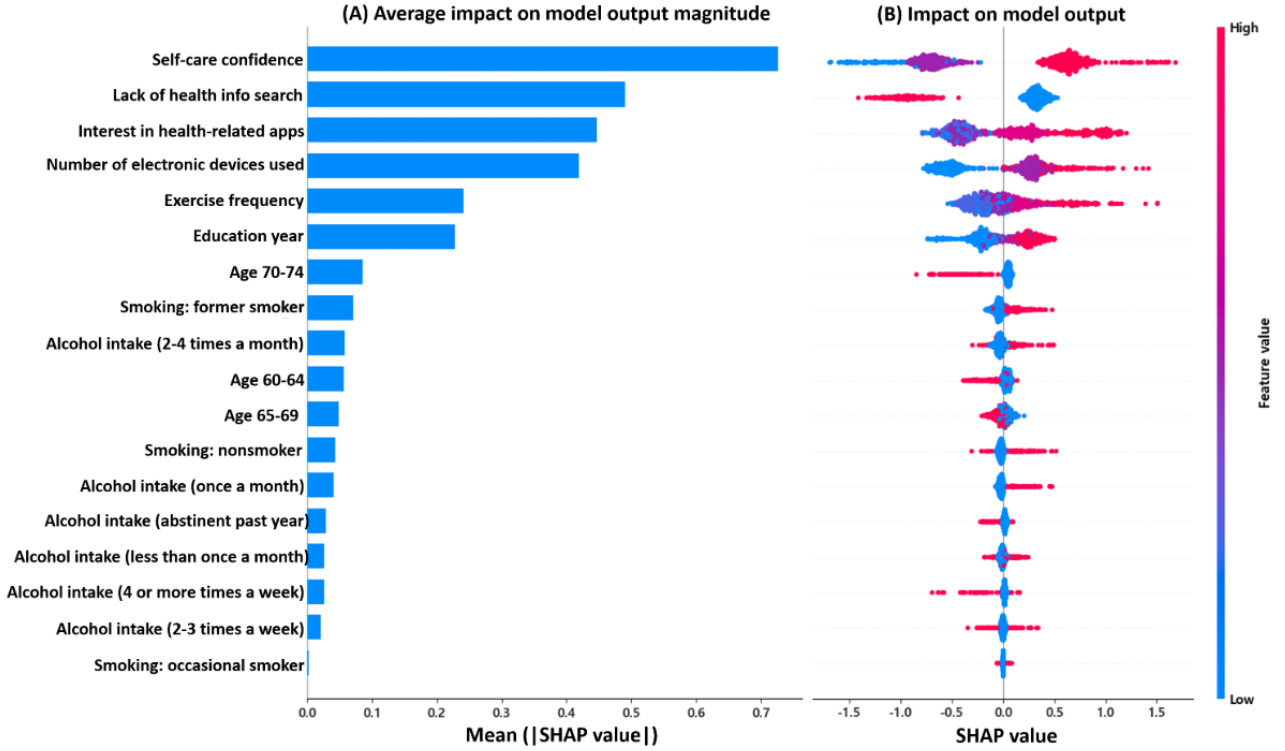
Figure 4B presents a summary plot illustrating the distribution of SHAP values across different feature values. Self-care confidence emerged as the most influential variable, showing the greatest separation between high- and low-DHL predictions. Higher levels of self-care confidence were associated with positive SHAP values, increasing the likelihood of being classified as high DHL. By contrast, lower self-care confidence levels (represented by blue markers) corresponded to negative SHAP values, contributing to low-DHL predictions.

The binary feature representing lack of health information search, derived from one-hot encoding, also demonstrated a

XSL•FO
RenderX

strong effect. Participants who had searched for health information (feature value=0, shown in blue) tended to have positive SHAP values, whereas those who had not (feature value=1, shown in red) exhibited negative SHAP values, indicating a higher likelihood of being classified in the low-DHL group. Similarly, interest in health-related apps and the number of electronic devices used were positively associated with high-DHL predictions. Participants with a greater interest in health-related apps and more frequent use of digital devices showed consistently positive SHAP values, reinforcing their roles as indicators of digital engagement. SHAP values for other contributing factors, including exercise frequency and education level, were relatively small. Age-related features also had comparatively low influence on model predictions; however, the 70-74 age group consistently appeared in association with lower DHL levels, contributing negatively to the prediction outcomes. Overall, the SHAP analysis provided interpretable evidence of how specific behavioral, psychosocial, and demographic features influenced DHL classification, reinforcing both the explainability and practical utility of the CatBoost model in real-world digital health contexts.

Figure 4. Feature importance based on mean Shapley Additive Explanations (SHAP) values according to the contribution of the categorical boosting model to its estimation. (A) Average impact on model output magnitude; (B) Impact on model output.



Discussion

Principal Findings

Digital health literacy, defined as the ability to search for, understand, and apply health-related information in digital environments [4], has emerged as a core competency for effectively navigating modern medical and health care services. As health care delivery models evolve from static information provision to dynamic, user-centered platforms, precise assessment of DHL—particularly among older adults—has become increasingly critical. Accurate measurement of DHL is essential for designing tailored interventions that promote equitable access to health care resources and reduce health disparities.

In this study, we developed an explainable machine learning–based predictive framework to evaluate DHL in a large sample of older adults (n=1000). The framework integrates both performance-based and self-assessed DHL measures and incorporates a range of psychosocial and behavioral variables to identify the most influential predictors contributing to individual differences in digital health literacy.

Selected Features From Performance-Based Digital Health Literacy

To classify participants into 2 DHL-level groups, explainable features were selected based on the intrinsic evaluation criteria of the Bayesian linear regression model. The analysis revealed that, within the performance-based assessment, key predictors included alcohol intake, number of electronic devices used, and years of education. Among these, the number of electronic devices used emerged as the primary positive predictor of digital information comprehension. This finding is consistent with previous studies showing that older adults who use a wider variety of digital devices are more familiar with online health information and demonstrate stronger skills in information search and utilization [57].

Similarly, years of education positively influenced the comprehension of health information. This finding aligns with previous research identifying educational attainment as a key

XSL•FO
RenderX

social determinant of health literacy. Higher educational attainment is generally associated with a greater capacity to comprehend, critically evaluate, and utilize online health information [16,58]. Individuals with higher levels of education are relatively more advantaged in discerning the credibility of health information and translating it into effective communication with health care providers or self-management behaviors, which in turn facilitates greater acceptance and utilization of digital health services [59,60]. A comparative study involving older adults in Korea and the United States likewise identified educational attainment as a consistent predictor of eHealth literacy, while noting that its impact may be moderated by structural and cultural factors within the digital environment [15]. Collectively, these findings underscore the importance of designing digital health interventions that account for educational disparities and incorporate tailored capacity-building strategies for individuals with lower levels of formal education.

By contrast, participants with higher levels of alcohol consumption exhibited significantly lower performance in both digital and health information comprehension, suggesting that lifestyle factors such as alcohol intake may impair information-processing abilities and negatively affect DHL. Moreover, a previous longitudinal study reported that sustained alcohol consumption is associated with cognitive decline, particularly among older women [61]. Taken together, these results indicate that diversity in digital device usage and higher educational attainment contribute meaningfully to actual information comprehension, and that lifestyle behavior management should be considered a critical component of DHL intervention programs.

Selected Features From Self-Assessed Digital Health Literacy

Self-assessed digital health literacy was primarily influenced by interest in health-related apps, age, alcohol intake, number of electronic devices used, and years of education. Additional contributing factors included self-care confidence, smoking status, and exercise frequency. Specifically, perceived competence in using digital devices was associated with greater confidence in health care decision-making, whereas the ability to understand and apply health information was linked to participants' prior health information–seeking behavior.

Among these variables, interest in health-related apps emerged as the most prominent positive predictor of both DHL subscales. Furthermore, a previous study reported that individuals who frequently engage with health care apps and perceive them as useful or reliable tend to demonstrate stronger competencies in locating, evaluating, and applying health information [57]. Interest in health-related apps could reflect more than mere preference, as it represents a motivational factor that enhances digital engagement. Voluntary use of such apps may foster critical components of DHL, including information searching, selection, appraisal, and application. Therefore, interventions designed to improve DHL should extend beyond digital tools and unidirectional health messaging.

In this study, we found that the older the age group, the lower their tendency to evaluate their understanding of and ability to use health information. It has been reported that older adults display a reduced ability to adopt digital technology because of cognitive decline [62]. Moreover, previous studies have found that older adults living alone face both cognitive and emotional barriers when adapting to new technologies [63-67]. Addressing these barriers requires tailored educational strategies that combine practical digital training with emotional support.

The number of electronic devices used and years of education proved to be negatively associated with self-assessed DHL. This finding suggests that greater exposure to digital devices or higher educational attainment does not necessarily translate into higher self-assessed DHL. The strength and nature of the relationship between education and digital health literacy may vary depending on the sociocultural context and the health care and digital infrastructure [68]. In addition, individuals with greater actual abilities may apply stricter self-assessment criteria, reflecting more conservative or critical self-assessment tendencies. A prior finding indicated that older adults may experience increased cognitive load when processing complex health-related information, leading to reduced confidence and lower self-ratings [69].

A similar pattern was observed for health information search. In the feature selection stage, lack of health information search was positively associated with the health information understanding and utilization subscale in the self-assessed survey. By contrast, in the model interpretation stage, the same indicator showed negative SHAP values, thereby decreasing the model's predicted likelihood of high DHL. Individuals who do not search for health information may positively assess their understanding of health information even though they lack it, whereas health information seekers might provide cautious self-assessments despite possessing substantial digital health literacy skills. It is known that self-reported responses do not represent a direct readout of an individual's internal state but are rather the product of complex cognitive processes such as comprehension, retrieval, judgment, and response selection, potentially leading to biased or distorted outcomes [70]. These findings emphasize the constraints of self-reported measures and the imperative to validate subjective data with objective performance metrics [71].

In addition, alcohol intake showed a clear divergence by measurement modality. In the self-assessed digital device utilization subscale, alcohol-related categories were positively associated with utilization scores, including abstinence in the past year, alcohol intake about once a month, and alcohol intake 2-4 times a month. By contrast, in the performance-based DHL tasks, alcohol intake once a month and alcohol intake 2-4 times a month were associated with lower task performance. Prior research in older adults indicates that internet use does not consistently correspond to healthier behaviors and that associations may vary by socioeconomic and lifestyle profiles [72]. A population-based study using CHARLS (The China Health and Retirement Longitudinal Study) 2020 data reported that internet use was associated with a lower likelihood of nondrinking, and this association was partially mediated by social participation [73]. This suggests that digital access may increase opportunities for social activities in which alcohol intake is culturally embedded. Taken together,

XSL•FO
RenderX

performance-based DHL tasks are more closely tied to the cognitive and informational demands of digital health activities and may be more sensitive to alcohol-related decrements. By contrast, self-assessed digital device utilization may reflect a broader lifestyle context rather than objective capability alone.

Development of the Prediction Models

The first stage of the DHL framework identified variables with high explanatory power for both performance-based and self-assessed measures. In the second stage, these selected variables were used to train models predicting DHL levels based on KeHEALS. The model performance evaluation demonstrated that the CatBoost algorithm consistently outperformed the alternative approaches, achieving comparatively high accuracy, precision, and F_1-scores in predicting DHL among older adults (Table 3). Compared with a conventional single decision tree, CatBoost, which effectively handles categorical variables, proved particularly effective for DHL prediction.

To enhance interpretability, a SHAP analysis was conducted on the CatBoost model, providing empirical evidence of the specific factors influencing DHL (Figure 4). The most influential predictors included self-care confidence, lack of health information search, interest in health-related apps, number of electronic devices used, exercise frequency, and years of education. This confirms that the framework successfully integrates both digital and health literacy–related features into a unified predictive model. Among the key predictors, self-care confidence emerged as the single most important predictor, highlighting the strong association between an individual's confidence in making health-related decisions and their DHL level. This finding aligns with prior research reporting that higher eHealth literacy is significantly associated with greater self-efficacy in older adults [74]. This evidence underscores the potential of DHL-focused interventions to enhance broader health management competencies by strengthening decision-making confidence. Furthermore, consistent with the results of the self-assessed DHL, interest in health-related apps reflected an individual's proactive attitude toward information seeking and app utilization, reinforcing the importance of psychosocial factors alongside technical competencies. Accordingly, interventions aimed at improving DHL should combine technical skill development with psychosocial support to enhance motivation for and engagement with digital health tools [75].

In contrast to the preceding factors, a greater number of electronic devices used, higher exercise frequency, and higher educational attainment were paradoxically associated with lower self-assessed scores for use of digital devices and health information understanding and utilization. These results may reflect the subjective biases inherent in self-reports. Self-report measures are constrained by their susceptibility to respondents' memory and cognitive processes, social desirability and consistency motives, and instrument design features such as item wording, response scales, and question order [76-78]. Alcohol intake showed a similar discrepancy. In the self-assessed digital device utilization subscale, abstinence in the past year and infrequent drinking were positively associated with utilization scores. By contrast, similar frequencies of alcohol intake were associated with lower performance in the task-based DHL assessment. This pattern suggests that self-assessed utilization may reflect a broader socioeconomic and lifestyle context, whereas performance-based tasks reflect the cognitive and informational demands required for digital health activities.

Therefore, a multimethod approach that integrates both self-assessments and performance-based evaluations is essential for a comprehensive understanding of DHL, rather than relying solely on self-report data. When variables derived from both measures were incorporated into the machine learning model, SHAP analysis indicated that these factors positively contributed to the classification of the high-DHL group, corroborating the results of the performance-based assessment. This demonstrates that reliance on self-reported data alone may lead to biased inferences or the overlooking of important patterns. By combining multifaceted indicators with explainable machine learning techniques, our study's approach enables more precise identification of the determinants of DHL in older adults and provides practical evidence to inform the design of individualized digital health interventions.

KeHEALS is a broadly used and validated assessment tool, although it yields only a single metric for digital literacy. Therefore, it remains uncertain whether individual vulnerabilities stem from fundamental digital device use, comprehension and appraisal of health information, or demographic and psychosocial factors. By using the detailed subscales evaluated within the DHLS, we can elucidate why a particular individual falls into the low-DHL category according to multidimensional criteria. When these domain-informed features are used to predict KeHEALS levels and the resulting predictions are subsequently examined using SHAP, KeHEALS scores can be analytically decomposed into discrete contributions originating from specific digital and health domains, along with additional contextual factors. This study highlights the limitations of single-score screening instruments and advances personalized interventions by offering explanations of feature importance through SHAP analysis.

This study leverages both performance- and survey-based datasets to bridge the gap between self-reported and actual abilities. Using Bayesian regression, the variables explaining each dimension were identified and integrated into the final DHL prediction model. SHAP analysis further validated the consistency between variable importance in the feature-selection stage and the variables' impact on model learning. This dual-phase approach enabled both ante hoc interpretability—by incorporating transparency into the model design from the outset—and post hoc interpretability, addressing the "black box" issue while retaining explanatory power. Finally, the DHL framework offers practical guidance for assessing users' DHL before implementing digital health interventions. This can be particularly valuable in clinical contexts where evaluation and interpretability are critical, thereby supporting the design of tailored and effective health care strategies for older adults.

Limitations

This study has several limitations that should be noted. First, the performance-based features were identified from a pilot

XSL•FO
RenderX

cohort aged 60-74, whereas the predictive model was trained on a wider age range (55-74 years). Our post hoc comparison revealed that several selected variables differed significantly between the 55-59 and 60-74 age groups, suggesting meaningful demographic and behavioral heterogeneity across age groups. Therefore, the generalizability of the feature-selection process from the small pilot cohort to the younger subgroup may be limited. Future research should employ larger and more diverse samples to enhance the robustness and external validity of the performance-based assessments. Second, because all data collected focused exclusively on older adults in Korea, caution should be exercised when extrapolating the results to populations from different cultural and demographic contexts. Third, the behavioral task was conducted using a single Android smartphone, and all participants were Android users. Thus, familiarity with the Android interface may have been reflected in the digital information comprehension scores and may not fully capture device proficiency among iOS users or across heterogeneous device environments. Fourth, there are inherent limitations related to the measurement scales and variable-processing strategies employed in this study. While offering interpretive simplicity, the binary classification of DHL levels may lead to different outcomes if alternative KeHEALS score thresholds are applied. Furthermore, self-care confidence, which emerged as the primary driver in our predictive model, was assessed using a single-item measure derived from the Korean Health Empowerment Scale. As single-item measures typically exhibit lower reliability and less comprehensive construct coverage than multiitem scales, the high predictive importance of this variable should be interpreted with caution. Additionally, our feature-importance rankings may have been influenced by the mixed encoding strategies used during data preprocessing. Specifically, the use of one-hot encoding for certain lifestyle factors, such as alcohol intake, partitioned these variables into multiple binary features. This approach may have diluted their aggregate predictive power in the SHAP analysis, potentially leading to an artificial deflation of their importance compared with variables maintained as single ordinal features. Finally, actual DHL is likely to exist along a continuous and multifaceted spectrum. Features that demonstrated unexpected associations in the self-assessed DHL analysis, such as the number of electronic devices used, exercise frequency, and years of education, may reflect the inherent limitations of self-report measures or complex interaction effects with other variables. These observations underscore the need for further in-depth analyses to disentangle these relationships.

Conclusions

DHL is a key competency for older adults to effectively navigate modern health care. This study established an explainable machine learning framework that integrates performance-based and self-assessed measures to evaluate DHL. Our findings indicate that using a wider variety of electronic devices and having higher educational attainment were positively associated with performance-based DHL, whereas higher alcohol intake was linked to poorer comprehension. Given the limited sample size of the performance-based pilot, these associations warrant caution and require validation in larger cohorts. In the self-assessed domain, DHL levels were primarily related to interest in health-related apps, age, and self-care confidence. In the predictive modeling stage, the CatBoost model demonstrated superior performance, and SHAP analysis identified self-care confidence and interest in health-related apps as key contributors to high DHL. Overall, by combining performance-based and self-assessed measures, this framework offers a structured and interpretable approach to assessing DHL in older adults and can inform the design of tailored digital health interventions in clinical and community settings.

Acknowledgments

This research was supported by the Basic Medical Science Facilitation Program through the Catholic Medical Center of the Catholic University of Korea, funded by the Catholic Education Foundation.

Funding

This research was supported by the Ministry of Science and ICT of Korea and the Korea Medical Device Development Fund (grant RS-2023-00255145).

Data Availability

The data and code underlying this article will be shared on reasonable request to the corresponding author (JWC).

Authors' Contributions

Conceptualization: JWC (lead), CP (equal)
Data curation: CP
Formal analysis: CP (lead), SK (supporting)
Funding acquisition: JWC
Investigation: YSB, JHK
Methodology: CP
Project administration: JWC, JHK
Validation: JWC, CP

XSL•FO
RenderX

Visualization: CP (lead), JWC (supporting)
Writing – original draft: CP (lead), JP (supporting)
Writing – review & editing: JWC (lead), CP (supporting), TMK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of digital literacy and health literacy scores by values of features.
[DOCX File , 1139 KB - medinform_v14i1e86171_app1.docx]

References

1. Alawiye TR. The impact of digital technology on healthcare delivery and patient outcomes. ETSN 2024;13(02):13-22. [doi: 10.4236/etsn.2024.132002]
2. Lim DSC, Kwok B, Williams P, Koczwara B. The impact of digital technology on self-management in cancer: systematic review. JMIR Cancer 2023 Nov 22;9:e45145 [FREE Full text] [doi: 10.2196/45145] [Medline: 37991831]
3. Botti A, Monda A. Sustainable value co-creation and digital health: the case of Trentino eHealth Ecosystem. Sustainability 2020 Jun 29;12(13):5263. [doi: 10.3390/su12135263]
4. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
5. Nutakor JA, Zhou L, Larnyo E, Addai-Dansoh S, Cui Y. Impact of health information seeking behavior and digital health literacy on self-perceived health and depression symptoms among older adults in the United States. Sci Rep 2024 Dec 28;14(1):31080 [FREE Full text] [doi: 10.1038/s41598-024-82187-z] [Medline: 39730731]
6. Ning L, Zheng Z, Liu M, Gao S, Yang X, Bi J, et al. Profiles of digital health literacy among university students and its association with mental health during the COVID-19 pandemic: a latent profile analysis. Digit Health 2024;10:20552076231224596 [FREE Full text] [doi: 10.1177/20552076231224596] [Medline: 38205038]
7. Liu S, Lu Y, Wang D, He X, Ren W, Kong D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722 [FREE Full text] [doi: 10.3389/fpubh.2023.1200722] [Medline: 37415711]
8. Bíró É, Vincze F, Nagy-Pénzes G, Ádány R. Investigation of the relationship of general and digital health literacy with various health-related outcomes. Front Public Health 2023;11:1229734 [FREE Full text] [doi: 10.3389/fpubh.2023.1229734] [Medline: 37588120]
9. The report on the digital divide. National Information Society Agency. 2024. URL: https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=81623&bcIdx=27832&parentSeq=27832 [accessed 2025-12-04]
10. Hwang M, Kim G, Lee S, Park Y. Digital health literacy and associated factors among older adults living alone in South Korea: a cross-sectional study. Res Community Public Health Nurs 2024 Dec 30;35:389-400. [doi: 10.12799/rcphn.2024.00766]
11. Shi Z, Du X, Li J, Hou R, Sun J, Marohabutr T. Factors influencing digital health literacy among older adults: a scoping review. Front Public Health 2024;12:1447747 [FREE Full text] [doi: 10.3389/fpubh.2024.1447747] [Medline: 39555039]
12. Yang E, Chang SJ, Ryu H, Kim HJ, Jang SJ. Comparing factors associated with eHealth literacy between young and older adults. J Gerontol Nurs 2020 Aug 01;46(8):46-56. [doi: 10.3928/00989134-20200707-02] [Medline: 32936926]
13. Papp-Zipernovszky O, Horváth MD, Schulz PJ, Csabai M. Generation gaps in digital health literacy and their impact on health information seeking behavior and health empowerment in Hungary. Front Public Health 2021;9:635943 [FREE Full text] [doi: 10.3389/fpubh.2021.635943] [Medline: 34055714]
14. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021 Aug;30(8):2235-2243 [FREE Full text] [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
15. Lee OE, Kim D, Beum KA. Factors affecting information and communication technology use and eHealth literacy among older adults in the US and South Korea. Educational Gerontology 2020 Jul 09;46(9):575-586. [doi: 10.1080/03601277.2020.1790162]
16. Cherid C, Baghdadli A, Wall M, Mayo NE, Berry G, Harvey EJ, et al. Current level of technology use, health and eHealth literacy in older Canadians with a recent fracture-a survey in orthopedic clinics. Osteoporos Int 2020 Jul;31(7):1333-1340. [doi: 10.1007/s00198-020-05359-3] [Medline: 32112118]
17. Berkowsky RW. Exploring predictors of eHealth literacy among older adults: findings from the 2020 CALSPEAKS Survey. Gerontol Geriatr Med 2021;7:23337214211064227 [FREE Full text] [doi: 10.1177/23337214211064227] [Medline: 34926723]

XSL•FO
RenderX

18. Quinn S, Bond R, Nugent C. Quantifying health literacy and eHealth literacy using existing instruments and browser-based software for tracking online health information seeking behavior. Computers in Human Behavior 2017 Apr;69:256-267 [FREE Full text] [doi: 10.1016/j.chb.2016.12.032]
19. Komalavalli D, Lopson S, Karthik M. Integrated prediction of mental health and fitness using machine learning models. New York, NY: IEEE; 2024 Presented at: 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS); December 17-18, 2024; Bengaluru, Karnataka, India p. 557-562. [doi: 10.1109/ICICNIS64247.2024.10823203]
20. Lundberg S, Lee S. A unified approach to interpreting model predictions. New York, NY: Curran Associates Inc; 2017 Presented at: The 31st International Conference on Neural Information Processing Systems (NIPS'17); December 4-9, 2017; Long Beach, CA p. 4768-4777. [doi: 10.5555/3295222.3295230]
21. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
22. Thorup CB, Uitto M, Butler-Henderson K, Wamala-Andersson S, Hoffrén-Mikkola M, Schack Thoft D, et al. Choosing the best digital health literacy measure for research: mixed methods study. J Med Internet Res 2025 Apr 08;27:e59807 [FREE Full text] [doi: 10.2196/59807] [Medline: 40198098]
23. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
24. Chaniaud N, Sagnier C, Loup-Escande E. Translation and validation study of the French version of the eHealth Literacy Scale: web-based survey on a student population. JMIR Form Res 2022 Aug 31;6(8):e36777 [FREE Full text] [doi: 10.2196/36777] [Medline: 36044264]
25. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the Simplified Chinese eHealth Literacy Scale: cross-sectional study. J Med Internet Res 2020 Dec 07;22(12):e18613 [FREE Full text] [doi: 10.2196/18613] [Medline: 33284123]
26. van der Vaart R, van Deursen AJ, Drossaert CH, Taal E, van Dijk JA, van de Laar MA. Does the eHealth Literacy Scale (eHEALS) measure what it intends to measure? Validation of a Dutch version of the eHEALS in two adult populations. J Med Internet Res 2011 Nov 09;13(4):e86 [FREE Full text] [doi: 10.2196/jmir.1840] [Medline: 22071338]
27. Petrič G, Atanasova S. Validation of the extended e-health literacy scale: structural validity, construct validity and measurement invariance. BMC Public Health 2024 Jul 25;24(1):1991 [FREE Full text] [doi: 10.1186/s12889-024-19431-8] [Medline: 39054541]
28. Marsall M, Engelmann G, Skoda E, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German Version of the eHealth Literacy Scale. J Med Internet Res 2022 Feb 02;24(2):e28252 [FREE Full text] [doi: 10.2196/28252] [Medline: 35107437]
29. Luz S, Nogueira P, Costa A, Henriques A. Psychometric analysis of the eHealth Literacy Scale in Portuguese older adults (eHEALS-PT24): instrument development and validation. J Med Internet Res 2025 Feb 26;27:e57730 [FREE Full text] [doi: 10.2196/57730] [Medline: 40009846]
30. Lin C, Broström A, Griffiths MD, Pakpour AH. Psychometric evaluation of the Persian eHealth Literacy Scale (eHEALS) among elder Iranians with heart failure. Eval Health Prof 2020 Dec;43(4):222-229. [doi: 10.1177/0163278719827997] [Medline: 30744419]
31. Kim H, Yang E, Ryu H, Kim HJ, Jang SJ, Chang SJ. Psychometric comparisons of measures of eHealth literacy using a sample of Korean older adults. Int J Older People Nurs 2021 May;16(3):e12369. [doi: 10.1111/opn.12369] [Medline: 33527701]
32. Hernández Encuentra E, Robles N, Angulo-Brunet A, Cullen D, Del Arco I. Spanish and Catalan versions of the eHealth Literacy Questionnaire: translation, cross-cultural adaptation, and validation study. J Med Internet Res 2024 May 10;26:e49227 [FREE Full text] [doi: 10.2196/49227] [Medline: 38728072]
33. Chung S, Nahm E. Testing reliability and validity of the eHealth Literacy Scale (eHEALS) for older adults recruited online. Comput Inform Nurs 2015 Apr;33(4):150-156 [FREE Full text] [doi: 10.1097/CIN.0000000000000146] [Medline: 25783223]
34. Kyaw MY, Aung MN, Koyanagi Y, Moolphate S, Aung TNN, Ma HKC, et al. Sociodigital determinants of eHealth literacy and related impact on health outcomes and eHealth use in Korean older adults: community-based cross-sectional survey. JMIR Aging 2024 Aug 13;7:e56061 [FREE Full text] [doi: 10.2196/56061] [Medline: 39140239]
35. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
36. Kim S, Park C, Park S, Kim D, Bae Y, Kang J, et al. Measuring digital health literacy in older adults: development and validation study. J Med Internet Res 2025 Feb 05;27:e65492 [FREE Full text] [doi: 10.2196/65492] [Medline: 39908081]
37. Zhao B, Huang L, Cheng X, Chen T, Li S, Wang X, et al. Digital health literacy and associated factors among internet users from China: a cross-sectional study. BMC Public Health 2024 Mar 27;24(1):908 [FREE Full text] [doi: 10.1186/s12889-024-18324-0] [Medline: 38539176]
38. Liu S, Zhao H, Fu J, Kong D, Zhong Z, Hong Y, et al. Current status and influencing factors of digital health literacy among community-dwelling older adults in Southwest China: a cross-sectional study. BMC Public Health 2022 May 17;22(1):996 [FREE Full text] [doi: 10.1186/s12889-022-13378-4] [Medline: 35581565]

XSL•FO
RenderX

39. Ezeigwe OJ, Nwosu KOS, Afolayan OK, Ojaruega AA, Echere J, Desai M, et al. Technological-based interventions in cancer and factors associated with the use of mobile digital wellness and health apps among cancer information seekers: cross-sectional study. J Med Internet Res 2025 Feb 05;27:e63403. [doi: 10.2196/63403] [Medline: 39909418]
40. Dunning D, Heath C, Suls JM. Flawed self-assessment: implications for health, education, and the workplace. Psychol Sci Public Interest 2004 Dec;5(3):69-106. [doi: 10.1111/j.1529-1006.2004.00018.x] [Medline: 26158995]
41. Vowels MJ. Trying to outrun causality with machine learning: limitations of model explainability techniques for exploratory research. Psychol Methods 2024 Sep 09:met0000699 (forthcoming). [doi: 10.1037/met0000699] [Medline: 39250291]
42. Bonovas S, Piovani D. Simpson's paradox in clinical research: a cautionary tale. J Clin Med 2023 Feb 18;12(4):1633 [FREE Full text] [doi: 10.3390/jcm12041633] [Medline: 36836181]
43. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 07;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
44. İnceoğlu F, Deniz S, Yagin FH. Prediction of effective sociodemographic variables in modeling health literacy: a machine learning approach. Int J Med Inform 2023 Oct;178:105167. [doi: 10.1016/j.ijmedinf.2023.105167] [Medline: 37572386]
45. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
46. Chakraborty C, Bhattacharya M, Pal S, Lee S. From machine learning to deep learning: advances of the recent data-driven paradigm shift in medicine and healthcare. Current Research in Biotechnology 2024;7:100164. [doi: 10.1016/j.crbiot.2023.100164]
47. Alhumaidi NH, Dermawan D, Kamaruzaman HF, Alotaiq N. The use of machine learning for analyzing real-world data in disease prediction and management: systematic review. JMIR Med Inform 2025 Jun 19;13:e68898 [FREE Full text] [doi: 10.2196/68898] [Medline: 40537090]
48. Chang SJ, Yang E, Ryu H, Kim HJ, Yoon JY. Cross-cultural adaptation and validation of the eHealth Literacy Scale in Korea. Korean J Adult Nurs 2018;30(5):504-515. [doi: 10.7475/kjan.2018.30.5.504]
49. Park C, Park Y. Validity and reliability of Korean version of Health Empowerment Scale (K-HES) for older adults. Asian Nurs Res (Korean Soc Nurs Sci) 2013 Sep;7(3):142-148 [FREE Full text] [doi: 10.1016/j.anr.2013.07.004] [Medline: 25030252]
50. Baldwin SA, Larson MJ. An introduction to using Bayesian linear regression with clinical data. Behav Res Ther 2017 Nov;98:58-75. [doi: 10.1016/j.brat.2016.12.016] [Medline: 28081861]
51. Quinlan J. Induction of decision trees. Mach Learn 1986 Mar;1(1):81-106. [doi: 10.1023/A:1022643204877]
52. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. Red Hook, NY: Curran Associates Inc; 2018 Presented at: The 32nd International Conference on Neural Information Processing Systems (NIPS'18); December 3-8, 2018; Montréal, QC, Canada p. 6639-6649. [doi: 10.5555/3327757.3327770]
53. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995 Sep 21;20(3):273-297. [doi: 10.1023/A:1022627411411]
54. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY: Association for Computing Machinery; 2016 Presented at: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16); August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
55. Breiman L. Random forests. Machine Learning 2001 Oct 21;45(1):5-32. [doi: 10.1023/A:1010933404324]
56. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. New York, NY: Association for Computing Machinery; 2019 Presented at: The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19); August 4-8, 2019; Anchorage, AK p. 2623-2631. [doi: 10.1145/3292500.3330701]
57. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
58. Green G. Seniors' eHealth literacy, health and education status and personal health knowledge. Digit Health 2022;8:20552076221089803 [FREE Full text] [doi: 10.1177/20552076221089803] [Medline: 35371531]
59. Seckin G, Hughes S. Patient-reported outcomes in a nationally representative sample of older internet users: cross-sectional survey. JMIR Aging 2021 Nov 24;4(4):e16006 [FREE Full text] [doi: 10.2196/16006] [Medline: 34822340]
60. Price-Haywood EG, Harden-Barrios J, Ulep R, Luo Q. eHealth literacy: patient engagement in identifying strategies to encourage use of patient portals among older adults. Popul Health Manag 2017 Dec;20(6):486-494. [doi: 10.1089/pop.2016.0164] [Medline: 28384076]
61. Han L, Jia J. Alcohol consumption, poor lifestyle choices, and air pollution worsen cognitive function in seniors: a cohort study in China. Environ Sci Pollut Res Int 2022 Apr;29(18):26877-26888. [doi: 10.1007/s11356-021-17891-8] [Medline: 34860343]
62. Wan X, Lighthall NR, Xie R. Consistent and robust predictors of internet use among older adults over time identified by machine learning. Computers in Human Behavior 2022 Dec;137:107413. [doi: 10.1016/j.chb.2022.107413]

XSL•FO
RenderX

63. Charness N, Boot WR. Aging and information technology use. Curr Dir Psychol Sci 2009 Oct 01;18(5):253-258. [doi: 10.1111/j.1467-8721.2009.01647.x]
64. Lee C, Coughlin JF. PERSPECTIVE: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. J Product Innov Manag 2014 Jun 03;32(5):747-759. [doi: 10.1111/jpim.12176]
65. Peek STM, Luijkx KG, Rijnaard MD, Nieboer ME, van der Voort CS, Aarts S, et al. Older adults' reasons for using technology while aging in place. Gerontology 2016;62(2):226-237 [FREE Full text] [doi: 10.1159/000430949] [Medline: 26044243]
66. Piper AM, Brewer R, Cornejo R. Technology learning and use among older adults with late-life vision impairments. Univ Access Inf Soc 2016 Oct 21;16(3):699-711. [doi: 10.1007/s10209-016-0500-1]
67. Schroeder T, Dodds L, Georgiou A, Gewald H, Siette J. Older adults and new technology: mapping review of the factors associated with older adults' intention to adopt digital technologies. JMIR Aging 2023 May 16;6:e44564 [FREE Full text] [doi: 10.2196/44564] [Medline: 37191976]
68. Yuen E, Winter N, Savira F, Huggins CE, Nguyen L, Cooper P, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888 [FREE Full text] [doi: 10.2196/46888] [Medline: 39059006]
69. Vroman KG, Arthanat S, Lysack C. “Who over 65 is online?” Older adults’ dispositions toward information communication technology. Computers in Human Behavior 2015 Feb;43:156-166. [doi: 10.1016/j.chb.2014.10.018]
70. Schwarz N. Self-reports: how the questions shape the answers. American Psychologist 1999 Feb;54(2):93-105. [doi: 10.1037/0003-066x.54.2.93]
71. Kaye S, Lewis I, Freeman J. Comparison of self-report and objective measures of driving behavior and road safety: a systematic review. J Safety Res 2018 Jun;65:141-151. [doi: 10.1016/j.jsr.2018.02.012] [Medline: 29776523]
72. Duplaga M. The association between internet use and health-related outcomes in older adults and the elderly: a cross-sectional study. BMC Med Inform Decis Mak 2021 May 06;21(1):150 [FREE Full text] [doi: 10.1186/s12911-021-01500-2] [Medline: 33957909]
73. Yang X, Chen Y, Qiu T, Zhu X. The mediation effect of social participation in the relationship between internet use and health behaviors among middle-aged and older individuals. Digit Health 2025;11:20552076251346239 [FREE Full text] [doi: 10.1177/20552076251346239] [Medline: 40438191]
74. Park C. Electronic health literacy as a source of self-efficacy among community-dwelling older adults. Clin Gerontol 2024 Jul 01:1-10 [FREE Full text] [doi: 10.1080/07317115.2024.2373894] [Medline: 38949203]
75. Fan S, Jain RC, Kankanhalli MS. A comprehensive picture of factors affecting user willingness to use mobile health applications. ACM Trans Comput Healthcare 2024 Jan 13;5(1):1-31. [doi: 10.1145/3626962]
76. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
77. Huprich SK, Bornstein RF, Schmitt TA. Self-report methodology is insufficient for improving the assessment and classification of axis II personality disorders. J Pers Disord 2011 Oct;25(5):557-570. [doi: 10.1521/pedi.2011.25.5.557] [Medline: 22023295]
78. Haeffel GJ, Howard GS. Self-report: psychology's four-letter word. Am J Psychol 2010;123(2):181-188. [doi: 10.5406/amerjpsyc.123.2.0181] [Medline: 20518434]

Abbreviations

AUC: area under the receiver operating characteristic
CatBoost: categorical boosting
CHARLS: The China Health and Retirement Longitudinal Study
DHL: digital health literacy
DHLS: Digital Health Literacy Scale
eHEALS: eHealth Literacy Scale
HDI: highest density interval
KeHEALS: Korean version of the eHealth Literacy Scale
mHealth: mobile health
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanations
SVM: support vector machine
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by A Coristine; submitted 20.Oct.2025; peer-reviewed by F Fischer; comments to author 17.Nov.2025; revised version received 25.Jan.2026; accepted 26.Jan.2026; published 23.Mar.2026.

Please cite as:
Park C, Park J, Kim S, Bae YS, Kang JH, Kim TM, Chun JW
Explainable Machine Learning for Assessing Digital Health Literacy in Older Adults: Validation and Development of a Two-Stage Model Integrating Performance-Based and Self-Assessed Indicators
JMIR Med Inform 2026;14:e86171
URL: https://medinform.jmir.org/2026/1/e86171
doi:10.2196/86171
PMID:

©Choonghee Park, Jiyeon Park, Seora Kim, Ye Seul Bae, Jae-Heon Kang, Tae-Min Kim, Ji-Won Chun. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Responsible AI for Predicting Delayed Hospital Discharge Among Older Adults: Development and Evaluation Study for Balancing Accuracy, Equity, and Explainability

Somayeh Ghazalbash[1], PhD; Manaf Zargoush[1], PhD; Sara JT Guilcher[2,3], PT, PhD; Kerry Kuluski[4,5], MSW, PhD

[1]Health Policy and Management, DeGroote School of Business, McMaster University, Hamilton, ON, Canada

[2]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

[3]Department of Physical Therapy, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[4]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada

[5]Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Manaf Zargoush, PhD
Health Policy and Management
DeGroote School of Business
McMaster University
1280 Main Street West
Hamilton, ON, L8S 4M4
Canada
Phone: 1 905 525 9140
Email: zargousa@mcmaster.ca

Abstract

Background: Amid growing demands and constrained health care resources, effective hospital bed capacity management is crucial. Delayed hospital discharge, where patients remain in the hospital beyond the need for acute care, strains resources, affects patient outcomes, and reduces system efficiency. Predicting such delays facilitates early interventions to avert them and alleviate burdens on patients, care partners, hospitals, and the broader health care system.

Objective: This study aimed to develop comprehensive predictive analytics for delayed discharges among older adults using explainable machine learning to boost transparency and interpretability, while integrating fairness to mitigate algorithmic biases.

Methods: Leveraging longitudinal data from over 2 decades in Ontario, Canada, we applied extreme gradient boosting and logistic regression models to predict delayed discharges within 90 days post–acute care. Data preprocessing included a 2-year look-back for clinical histories and balanced sampling to address class imbalance. Model performance was assessed via area under the receiver operating characteristic curve, calibration, and clinical utility. Fairness was evaluated across sex, urban or rural residence, and residential instability using several threshold-free metrics. Explainability was examined at the global model level (via partial dependence plots and permutation feature importance) and locally (via Shapley Additive Explanations, breakdown, and ceteris paribus methods), with principal component analysis used to cluster key features for high-risk patients.

Results: The extreme gradient boosting model outperformed logistic regression, achieving an area under the receiver operating characteristic curve of 0.82 on the test set, with acceptable within-group and cross-group ranking fairness across subgroups. Explainability clustering analyses identified functional and cognitive declines (eg, care support needs, dementia, and mobility issues) and regional disparities as primary drivers of high-risk predictions. Bias mitigation improved calibration parity, especially when stratifying by residential instability, underscoring the trade-offs policymakers must weigh between accuracy, fairness, and explainability.

Conclusions: This study demonstrates the potential of responsible artificial intelligence in health care, emphasizing the need to balance predictive accuracy, equity, and interpretability. It uncovers systemic gaps and offers actionable insights for enhanced discharge planning, resource optimization, and equitable care delivery.

(JMIR Med Inform 2026;14:e83244) doi:10.2196/83244

XSL•FO
RenderX

KEYWORDS

delayed discharge; responsible AI; explainability; equity; algorithmic fairness; machine learning; responsible artificial intelligence

Introduction

Delayed hospital discharge (DHD) is defined as staying in hospitals longer than medically necessary. It has become a critical issue due to increasing health care demands under constrained and limited resources. DHD is a prevalent issue in health care systems worldwide, affecting numerous countries, including Canada, the United States, the United Kingdom, Germany, Norway, Sweden, Scotland, Denmark, France, New Zealand, Australia, and South Korea [1]. These delays often stem from unavailable postacute options needed to ensure patients' safe discharge, such as long-term care (LTC), rehabilitation, or other community support services [2,3]. DHD is more common among older adults, especially those with complex health needs due to frailty and multiple chronic conditions [4].

DHD has both direct and indirect impacts on the patients. DHD adversely affects patients and their care partners directly through increased risk of adverse health outcomes, including functional decline, evidenced by a 3%-5% functional deterioration for each day of inactivity [5]. DHD status is also associated with a higher risk of hospital harm and mortality [2]. Patients and care partners often report increased stress, emotional strain, and uncertainty regarding care planning, further compounding the burden of DHD [6].

DHD also has negative effects on the overall efficiency of the health care system. In 2022-2023, DHD patients accounted for 17% of all acute-care bed-days in Canada, with provincial rates ranging from 6.8% to 26% [7]. The occupancy levels of acute-care beds by patients who no longer need such care contribute to system-wide inefficiencies, such as delayed admissions, overcrowding in emergency departments, surgical cancellations, and a domino effect on the entire health care system [2,5,8], leading to hallway medicine [9]. The financial burden associated with DHD is substantial, with hospital-based care for these patients estimated at CAD $730 (US $524.68) to CAD $1200 (US $862.48) per day, significantly exceeding the costs of alternative settings such as LTC (CAD $126 [US $90.56]-CAD $144 [103.50]/day), transitional care units (CAD $155 [US $111.40]/day), or home-based services (CAD $42 [US $30.19]-CAD $118 [US $84.81]/day) [7]. In Canada, inefficiencies related to DHD are estimated to cost the health care system between CAD $5 million (US $3.59 million) and CAD $9 million (US $6.47 million) daily, amounting to billions of dollars annually [10]. Older adults are disproportionately affected by DHD, particularly in Canada, where Ontario, the country's most populous province, reported that over 80% of DHD designations in acute care involve adults aged 65 years and older, with 64% of them older than 75 years [11,12]. Although increasing system capacity, especially in LTC, has been proposed as a potential solution, this strategy requires substantial capital investment and is not practical in the short or medium term [13] and does not address the underlying issues of an undersupply of housing with care models (such as supportive housing or assisted living) and lack of chronic disease management in the community, which could mitigate unnecessary use of hospital care.

One promising alternative solution is to adopt a proactive approach. The risk of DHD can be predicted in advance, enabling health care providers to identify high-risk patients early and implement targeted interventions to reduce its occurrence. Machine learning (ML)–based methodologies are gaining popularity, especially in health care, due to their superior predictive accuracy compared to traditional statistical methods [14] and their effectiveness with longitudinal electronic health records [15,16]. Importantly, identifying a patient at elevated risk of DHD enables the development of upstream, community-based proactive models of care that aim to prevent the clinical and social deterioration leading to hospitalization with a discharge delay, rather than solely improving discharge planning after hospitalization has occurred. Such forward-looking models of care can integrate both health and social care resources, strengthening linkages to primary care and connecting patients and caregivers to essential community supports. These supports may include assistance with instrumental activities of daily living (eg, meal preparation and transportation to medical appointments) as well as fundamental activities of daily living (eg, mobility, eating, and bathing). By addressing these needs proactively, the care model is designed to stabilize older adults within the community, reduce avoidable hospitalizations, and ultimately minimize downstream discharge delays. Timely and accurate risk predictions, therefore, provide critical information for targeting preventive interventions, allocating resources more effectively, and improving patient flow across the continuum of care [17].

Although ML predictions are promising, they present a complex trade-off between accuracy, fairness, and explainability [18]. ML algorithms trained on historical data may unintentionally perpetuate biases [19], reinforcing existing social inequalities and disproportionately affecting marginalized populations [20]. Several studies have shown that automatic algorithms can systematically discriminate against equity-seeking groups, amplifying disparities rather than alleviating them [19,21]. This highlights the need to integrate fairness into predictive models to avoid unjustified disparities in model predictions across groups, thereby ensuring equitable outcomes.

Along with fairness, transparency and explainability are also crucial for ensuring trust and encouraging adoption in clinical settings. Health care professionals not only require accurate and fair predictions but also need clear, interpretable explanations to understand and validate the reasoning behind them, enabling them to integrate artificial intelligence (AI) insights into their decision-making [22]. This has driven the emergence of explainable artificial intelligence (XAI), which addresses concerns about algorithmic opacity by developing models that are both interpretable and accountable. XAI tackles AI's "black box" nature by making predictions interpretable and transparent. This transparency is crucial, especially in health care, where trust as well as shared and informed decision-making are vital

XSL•FO
RenderX

for patients' buy-in and optimizing clinical outcomes and promoting equitable care.

Most of the existing research on DHD has focused on identifying associated patient characteristics and strategies to improve patient flow [3,5,23,24]. Limited studies have investigated predictive upstream approaches to identify patients vulnerable to experiencing delayed discharge. Previous research has applied different methods, ranging from more traditional statistical models, such as logistic regression (LR) [23], to rule-based systems [25] and ML techniques, including neural networks [26]. Such efforts underscore the potential of predictive models to enhance care planning processes and inform decision-making. However, in reviewing the published literature to date, several key gaps emerge. First, the use of ML to predict delayed discharge with large and comprehensive datasets is limited, which impacts the robustness and accuracy of existing models. Second, there is a notable lack of emphasis on responsible AI principles within these models. Addressing these gaps is vital for developing reliable and actionable strategies to effectively prevent and manage delayed discharge. To address these challenges, our study aimed to contribute to the existing literature by (1) developing ML-based predictive models for identifying patients at risk of DHD among older adults, (2) examining the fairness implications of the developed ML-based predictive models to ensure outcome parity among equity-seeking groups, and (3) enhancing the interpretability and explainability of the ML-based predictive models by applying XAI at both global (population-specific) and local (person-specific) levels.

Methods

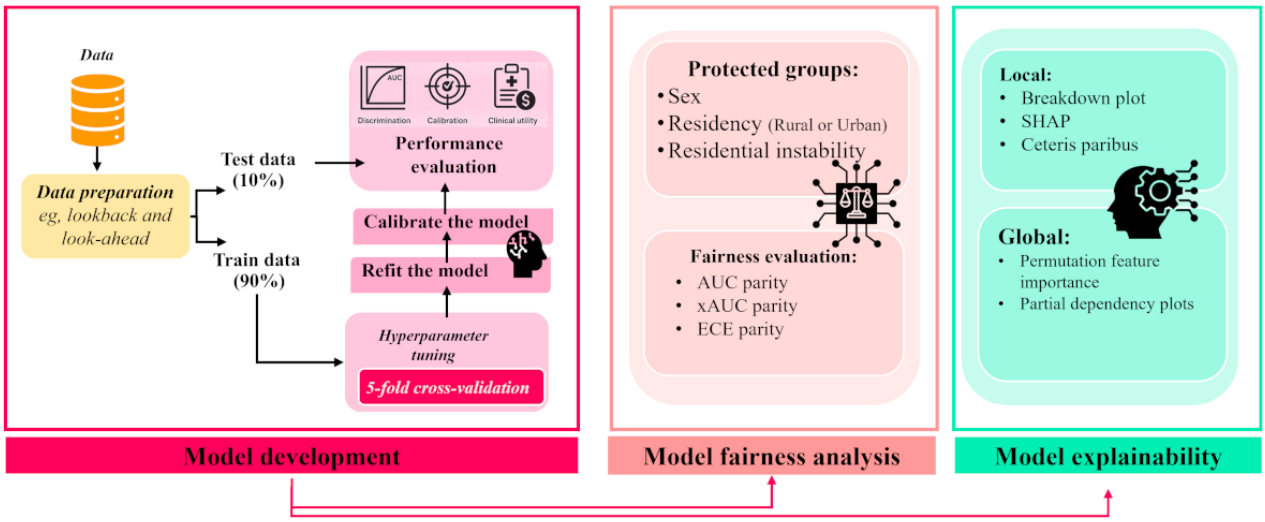
Ethical Considerations

This study received approval from the HiREB (Hamilton Integrated Research Ethics Board) in Ontario (Review Reference: 2025-5472-AP). This study used record-level, coded, and linkable administrative health data held at ICES (Institute for Clinical Evaluative Sciences). Individual informed consent was not required because the study involved the linking together of historical records of personal health information, and consent was impracticable or impossible to obtain. At ICES, all data collections are subject to rigorous privacy and security controls. Participants received no compensation.

Proposed Methodology

We developed a 3-step methodology for this research (Figure 1). The first step involved preparing the data and fine-tuning hyperparameters for the ML algorithms to identify the best prediction model. We used various measures of predictive accuracy to evaluate the model's performance (model development step) and then assessed algorithmic bias across protected groups using multiple fairness metrics (model fairness analysis step). Finally, we examined the model's explainability from both local and global perspectives and used clustering techniques to find important predictive features (model explainability step).

Figure 1. Study methodology framework. AUC: area under the receiver operating characteristic curve; ECE: expected calibration error; SHAP: Shapley Additive Explanation; xAUC: cross-group area under the curve.



Study Population

A retrospective, observational cohort study was conducted using 2 secondary administrative health databases housed at the ICES, including the discharge abstract database and registered persons database. These databases were linked through a unique patient identifier. The cohort included Ontario acute-care admissions from 2004 to 2022, including approximately 1.8 million patients with nearly 12 million observations. Individuals were excluded from the cohort if they were younger than 65 years of age, non-Ontario residents, had missing or invalid data, or were admitted from LTC facilities, resulting in a total of 9 million observations with 1.3 million patients. We excluded admissions from LTC due to their distinct care pathways and discharge patterns. Patients from LTC typically return to their homes after shorter stays, influenced by policies that protect their LTC beds. Including them would obscure the patterns of discharge delays and community transitions central to our research, which aims to support aging in the community.

XSL•FO
RenderX

All variables, including predictors and outcomes, were derived from structured, coded administrative data at ICES, with no manual chart abstraction or unstructured text processing. More detailed information regarding the data source is available in Multimedia Appendix 1.

Data Descriptions

The choice of candidate predictors was based on prior research and data availability, including (1) demographic factors, such as sex and age; (2) socioeconomic factors, including the Ontario Marginalization Index (ON-Marg 2016) [27], income quintile, and rural or urban residency; (3) geographic factors, such as regions; and (4) clinical factors, including patient comorbidities, palliative care, home care, rehabilitation care, fracture, fall, mobility, and disabilities. The primary outcome was DHD within 90 days, reflecting a clinically relevant transition period from the hospital to the next point of care. In Canada, the term DHD refers to the alternate level of care designation, a classification system used in hospitals to identify inpatients who no longer require the full intensity of acute medical services but are not yet ready to be safely discharged home. A patient must be designated as DHD by the most suitable member of the care team, such as a physician, LTC assessor, patient care manager, or discharge planner. The determination of DHD status is a clinical responsibility, based on various indicators of patient readiness and ongoing care needs. These indicators include clinical stability, safety risks, activity tolerance, medication or fluid requirements, diagnostic and therapeutic needs, as well as considerations related to mental health, respiratory care, or palliative care. The DHD designation is formally coded and routinely collected in Canadian hospital administrative data [28]. More detailed information regarding variable selection for model building is also given in Multimedia Appendix 1. To ensure the robustness of our findings, sensitivity analyses were conducted using alternative timeframes of 180 and 365 days, with the results presented in Multimedia Appendix 2.

Data Preprocessing and Preparation

We used a 2-year look-back window to capture patients' clinical history, including prior comorbidities. Further details regarding the construction of these variables can be found in Multimedia Appendix 2. We dichotomized the socioeconomic variables from the ON-Marg (including ethnic concentration, dependency, residential instability, and material deprivation) [27] into binary variables, comparing the most marginalized quintile (Q5) with the other quintiles (Q1-Q4). Missing data were rare (overall <0.3%) and were primarily attributable to incomplete values in the ON-Marg socioeconomic indices. Given the negligible degree of missingness and to avoid introducing additional assumptions through imputation, observations with missing values were removed using a complete-case approach.

Predictive Models

We used supervised ML to predict DHD risk with the extreme gradient boosting (XGB) algorithm. We chose this algorithm due to its superior performance with complex datasets, its ability to generalize, and its robustness against noise and outliers [29,30]. In addition to this ML algorithm, we also included LR as a traditional predictive approach in health care, due to its explanatory power and widespread use. By including LR alongside XGB, we aimed to compare its predictive performance with that of more advanced ML techniques and evaluate the suitability of a widely used model for our specific prediction task. We established the models using the same data structure, preparation, training, and assessment protocols to ensure a fair and meaningful comparison.

Following the standard practice, we used a 2-stage evaluation approach. First, the dataset was split at the patient level into training (90%) and testing (10%) sets to prevent information leakage from repeated visits. The test set, which remained entirely separate from the training process, was exclusively used for the final evaluation of the model's predictive performance. In the second stage, we used 5-fold cross-validation within the training set to optimize hyperparameters and mitigate the risk of overfitting. After the optimal hyperparameters were identified, the model was retrained from scratch on the entire training set. The resulting model was evaluated on the untouched 10% test set, yielding the reported performance metrics. To ensure that the predicted readmission probabilities accurately reflected the true underlying risk, we performed post hoc calibration of our predictive models. Calibration measures the agreement between the predicted probability and the observed (empirical) probability of the outcome class. As the true probability is unknown, proxies such as the proportion of outcomes of interest within a specific group of observations are often used instead of the observed probability [31]. We applied Platt scaling (also known as logistic calibration and sigmoid calibration), a method that transforms model estimates using a trained sigmoid function. It is a univariate LR model with model estimates as independent variables and binary outcomes as dependent variables [32]. The complete pipeline is illustrated in Figure 1.

To ensure a comprehensive and clinically actionable assessment, each predictive model was evaluated across 3 key dimensions: discrimination, calibration, and clinical utility. Discrimination was quantified using the area under the receiver operating characteristic curve (AUC), which measures the model's ability to distinguish patients with and without the outcome. Calibration was assessed using the Brier score and the expected calibration error (ECE), with reliability diagrams used to visualize agreement between predicted probabilities and observed outcomes [33]. The clinical utility was evaluated through decision curve analysis (DCA), which quantifies the clinical net benefit (NB) of our predictive models across various risk thresholds [34]. The NB is calculated as the expected benefit to the cases (ie, true positive rate) minus the expected harm to the controls (ie, false positive rate multiplied by the threshold probability). For the sake of clarity, we used the standard NB, which has a maximum value of 1.0 [35]. In the DCA, our predictive models are compared against two extreme clinical strategies: (1) treating all patients as positive cases, known as the "treat all" strategy, and (2) treating all patients as negative cases, referred to as "treat none" [36]. In the former scenario, the true positive rate is equal to the rate of the high-risk event, π, and the false positive rate is equal to $1–\pi$. In the latter scenario, the NB is equal to zero [37]. Additionally, we conducted a temporal validation to further evaluate the robustness and external validity of our predictive model under

XSL•FO
RenderX

realistic deployment conditions. We partitioned the cohort at the patient level based on the year of each patient's first recorded admission, assigning those with initial admissions before 2014 to the training set and those with initial admissions from 2015 onward to the temporal test set. Data preprocessing, ML model development, and evaluations were performed using R (version 4.2.2; R Foundation). All details regarding the software packages used for each analysis are provided in Multimedia Appendix 3.

Fairness Assessment

We calculated multiple threshold-free fairness metrics for each model to evaluate algorithmic fairness. We used threshold-free fairness metrics to better reflect the application of predictive models in health care, where continuous risk probabilities inform patient prioritization and resource allocation rather than rigid binary classifications. Essentially, in health care settings, risk scores inform care intensity and preventive interventions as ranking tools rather than binary classifiers. Thus, threshold-free measures provide better insights into disparities in utility and resource allocation than threshold-dependent criteria. We evaluated within-group and cross-group ranking fairness using AUC and xAUC (cross-group area under the curve) parity [38], respectively, and examined calibration fairness using ECE parity [39]. AUC parity evaluates the consistency of within-group discriminatory performance, comparing how well each group's positive cases are distinguished from its negatives. xAUC parity measures disparities in misranking errors caused by a predictive score. It quantifies the difference in the probabilities of ranking a random positive case from 1 protected group above a negative one from another group, and vice versa [40]. ECE parity assesses whether a model's predicted probabilities are equally well calibrated across regular and equity-seeking groups [39].

Certain socioeconomic groups, such as females or racial minorities, are often underrepresented in health care data because of biological or nonbiological variation [41,42]. Rural residency is recognized as a distinct barrier to care access, leading to potentially vulnerable or underserved subpopulations [43]. Similarly, residential instability reflects the underlying social vulnerabilities, such as housing insecurity and social isolation [27]. These disparities can introduce or amplify biases in ML predictions and decisions, potentially compromising the quality of care delivered to these equity-seeking groups [44]. To address this, we considered 3 equity-seeking groups in this study: females (vs males), rural (vs urban), and highly marginalized (vs low- to moderate-marginalized) in terms of residential instability. A summary of the prevalence of these equity-seeking groups across the training and test data is provided in Multimedia Appendix 1.

The relative fairness metrics are measured as the ratio of equity measures for the equity-seeking group to those for the regular group in the constructed predictive models. Therefore, a ratio of 1.0 indicates perfect fairness, with any deviation (above or below 1.0) reflecting unfairness toward regular or equity-seeking groups. Finally, to reduce potential algorithmic bias in our model, we used a postprocessing bias mitigation approach based on group-fairness calibration refinement [38]. Separate Platt scaling models were fitted to the predicted probabilities within each subgroup defined by the equity-seeking attributes, ensuring equitable alignment of predicted risks with observed outcomes across groups.

Model Explainability: Local and Global Perspectives

In contrast to traditional models such as LR, ML algorithms often lack interpretability, which makes it difficult to understand how predictors affect outcomes [45]. To tackle this challenge, we applied XAI principles to gain insights into both global and local (ie, patient-specific) explainability. Global explainability, which examines the overall association between predictors and the outcome, was explored using partial dependence plots (PDPs) and permutation feature importance [46]. While PDPs visualize the marginal effects of features on predictions, the permutation feature importance method quantifies how much a model's performance changes when the impact of a feature is removed through resampling or permutation.

Local explainability, which focuses on understanding model predictions for individual cases, was achieved using SHAP (Shapley Additive Explanations) [47], breakdown [48], and ceteris paribus [46]. The SHAP method uses a game-theoretic approach to assess the contribution of each feature to a specific prediction by evaluating all possible combinations of features. The breakdown approach decomposes individual model predictions into interpretable components, attributing the final prediction to the cumulative contributions of specific features. These 2 methods are complementary. Breakdown plots are order-dependent, meaning the importance of features can vary with their sequence, and can be sensitive to feature correlations. In contrast, SHAP provides order-invariant, theoretically consistent attributions, though it does not visualize the incremental path from baseline to final prediction. By using both, we enable cross-validation of local explanations, mitigate order-induced artifacts, and provide a more robust and transparent interpretation of individual-level model decisions [46]. The ceteris paribus approach examines the effect of an explanatory variable while holding the values of all other variables constant. The primary aim is to understand how variations in the values of a variable influence the model's predictions. We conducted the local explanation on a subsample (10%) of test data due to computational constraints for individual predictions.

Local explainability methods generate detailed, instance-specific insights into feature contributions, but analyzing these across numerous predictions can yield a high-dimensional dataset of metrics, including contributions (quantifying how much each feature influences the model's prediction for a given patient), frequencies (capturing how often a particular factor appears among the top 10 features for high-risk patients, thereby indicating its prevalence across such cases), and rankings (denoting the order of importance for each feature, where a lower rank reflects greater consistent influence on predicting risk), which may obscure broader patterns. To address this, we applied principal component analysis (PCA) [49], a well-established unsupervised dimensionality reduction technique that identifies orthogonal components maximizing variance in the data, thereby uncovering latent structures and reducing complexity while preserving key information.

XSL•FO
RenderX

To construct the input dataset for PCA, we first aggregated all local explanation outputs for patients classified as high-risk (predicted probability greater than 0.5). For each predictor, we calculated three summary metrics reflecting its behavior across high-risk individuals: (1) average contribution to the prediction, (2) average rank based on its relative importance, and (3) the prevalence (percentage of high-risk patients for whom the feature appeared among the top contributors). All metrics were standardized using *z* scores to ensure equal weighting across variables. PCA was then applied to the resulting feature-by-metric matrix to reduce dimensionality and identify underlying latent structures. For interpretability, features were positioned within 1 of 4 quadrants (Q1-Q4) according to the signs of their PC1 and PC2 coordinates. To improve conceptual interpretation of the resulting clusters, each feature was categorized into 1 of 3 a priori groups: sociodemographic and geographic characteristics, multimorbidity and health status, and mobility and disability factors. These groups were then examined within each PCA quadrant to identify coherent patterns and clinically meaningful clusters. Additional details about this mapping can be found in Multimedia Appendix 4.

Results

Descriptive Results

Table 1 presents the distribution of sociodemographic, geographical, and clinical characteristics among patients with and without delayed discharge. DHD patients were more likely to be older (median age 84, IQR 79-89 vs 80, 74-85 years), female (201,015/793,276, 57% vs 3,911,541/8,107,313, 48%), and to have lower income (201,015/793,276, 25% in the lowest quintile vs 1,695,425/8,107,313, 21%). DHD patients were more likely to reside in more marginalized areas, particularly concerning residential instability (277,898/793,276, 35% vs 2,251,598/8,107,313, 28%), which reflects factors such as living alone or in nonowned dwellings. Additionally, they showed higher levels of dependency (319,730/793,276, 40% vs 2,993,423/8,107,313, 37%), indicating a greater proportion of seniors and individuals not engaged in the labor force. Furthermore, material deprivation was more prevalent among DHD patients (195,662/793,276, 25% vs 1,688,573/8,107,313, 21%), suggesting they face greater barriers to meeting basic needs such as employment and education [27]. More detailed results of the descriptive analysis of clinical features are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Patient sociodemographic and geographical characteristics.

Characteristic	Overall (N=8,900,589)	Without delayed discharge (N=8,107,313)	With delayed discharge (N=793,276)
Region, n (%)			
Eastern	1,565,790 (18)	1,448,201 (18)	117,589 (15)
Central	2,877,461 (32)	2,610,461 (32)	267,000 (34)
Metropolitan	1,635,504 (18)	1,483,312 (18)	152,192 (19)
Southwestern	1,908,596 (21)	1,748,197 (22)	160,399 (20)
Northern	913,238 (10)	817,142 (10)	96,096 (12)
Income level, n (%)			
Q1-lowest	1,896,440 (21	1,695,425 (21)	201,015 (25)
Q2	1,895,832 (21)	1,722,811 (21)	173,021 (22)
Q3	1,777,314 (20)	1,622,389 (20)	154,925 (20)
Q4	1,665,924 (19)	1,529,430 (19)	136,494 (17)
Q5-highest	1,665,079 (19)	1,537,258 (19)	127,821 (16)
Residency, n (%)			
Urban	7,502,926 (84)	6,817,006 (84)	685,920 (86)
Rural	1,397,663 (16)	1,290,307 (16)	107,356 (14)
Sex, n (%)			
Female	4,362,152 (49)	3,911,541 (48)	450,611 (57)
Male	4,538,437 (51)	4,195,772 (52)	342,665 (43)
Age (years), median (IQR)	80 (75-85)	80 (74-85)	84 (79-89)
ON-Marg[a]: ethnic concentration (Q5 vs others), n (%)	1,446,682 (16)	1,320,691 (16)	125,991 (16)
ON-Marg: dependency (Q5 vs others), n (%)	3,313,153 (37)	2,993,423 (37)	319,730 (40)
ON-Marg: residential instability (Q5 vs others), n (%)	2,529,496 (28)	2,251,598 (28)	277,898 (35)
ON-Marg: material deprivation (Q5 vs others), n (%)	1,884,235 (21)	1,688,573 (21)	195,662 (25)

[a]ON-Marg: Ontario Marginalization Index.

Predictive Performance

Predictive models were evaluated using threshold-free performance metrics on the test set, as shown in Table 2. The results reveal that the XGB model achieved a slightly higher discrimination performance (AUC=0.822) compared to the LR model (AUC=0.799), indicating a modestly superior capacity to rank patients by risk. Both base models showed similar Brier scores (≈0.18), suggesting comparable average squared error between predicted probabilities and observed outcomes. Post hoc Platt scaling significantly improved calibration for both models. The calibrated XGB model demonstrated a better balance of discrimination (AUC=0.822) and reliability (Brier=0.069, ECE=0.138), making it a reliable choice for clinical risk stratification, where both ranking and probability interpretation are essential. The reliability diagrams for the predictive models are displayed in Figure 2.

Table 2. Discrimination and calibration performance measures on the test set.

Models	AUC[a]	Brier score	ECE[b]	Calibration intercept	Calibration slope
LR[c] model	0.799	0.181	0.362	–2.32	0.82
LR + Platt scaling	0.799	0.073	0.144	0.01	1.01
XGB[d] model	0.822	0.175	0.323	–2.26	0.80
XGB + Platt scaling	0.822	0.069	0.138	0.31	1.14

[a]AUC: area under the receiver operating characteristic curve.

[b]ECE: expected calibration error.

[c]LR: logistic regression.

[d]XGB: extreme gradient boosting.

XSL•FO
RenderX

Figure 2. Reliability diagrams for XGB and LR models. LR: logistic regression; XGB: extreme gradient boosting.

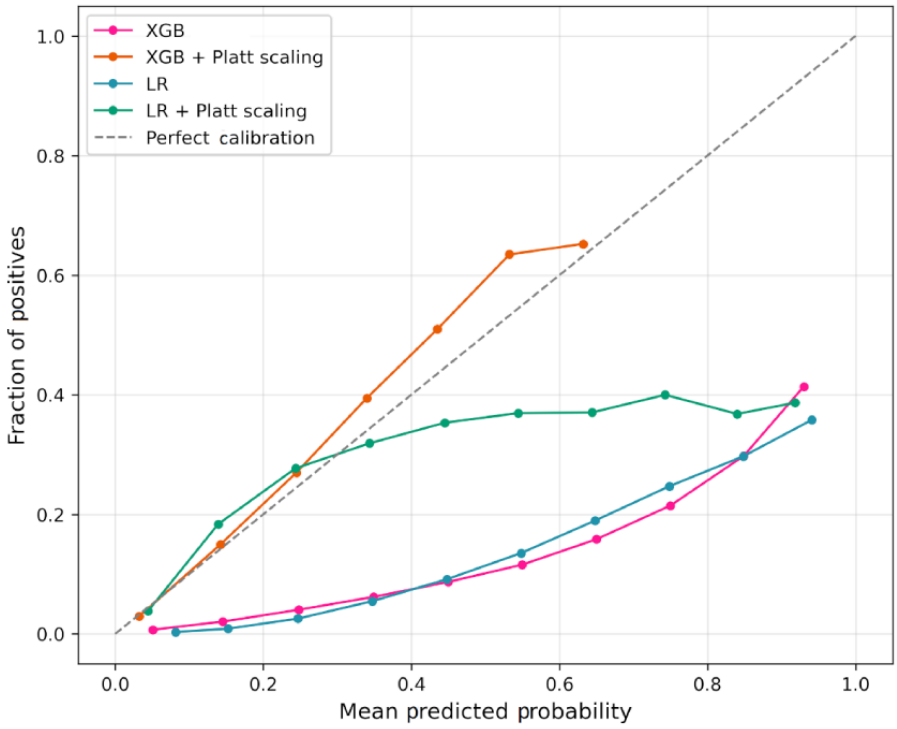


Table 3 summarizes the temporal validation results, including discrimination and multiple calibration indices (Brier score, ECE, calibration-in-the-large, and calibration slope). Overall, both models maintained strong discrimination in the temporal cohort. Post hoc calibration using Platt scaling improved calibration for both models, as reflected by reduced calibration error and calibration slopes closer to 1.0. For the calibrated XGB model, the calibration slope remained near one, and the calibration-in-the-large was relatively stable in the temporal cohort, suggesting limited temporal drift in calibration. In contrast, the calibrated LR model exhibited a larger change in the calibration intercept (from 0.01 to 1.02), consistent with a shift in baseline outcome risk over time and indicating that periodic recalibration may be more important for LR under the temporal distribution shift.

Table 3. Temporally validated models' discrimination and calibration performance.

Models	AUC[a]	Brier score	ECE[b]	Calibration intercept	Calibration slope
Temporal LR[c] model	0.823	0.166	0.346	–2.19	0.89
Temporal LR + Platt scaling	0.823	0.075	0.147	1.02	1.00
Temporal XGB[d] model	0.842	0.162	0.306	–2.14	0.78
Temporal XGB + Platt scaling	0.842	0.069	0.140	6e-10	1.00

[a]AUC: area under the receiver operating characteristic curve.

[b]ECE: expected calibration error.

[c]LR: logistic regression.

[d]XGB: extreme gradient boosting.

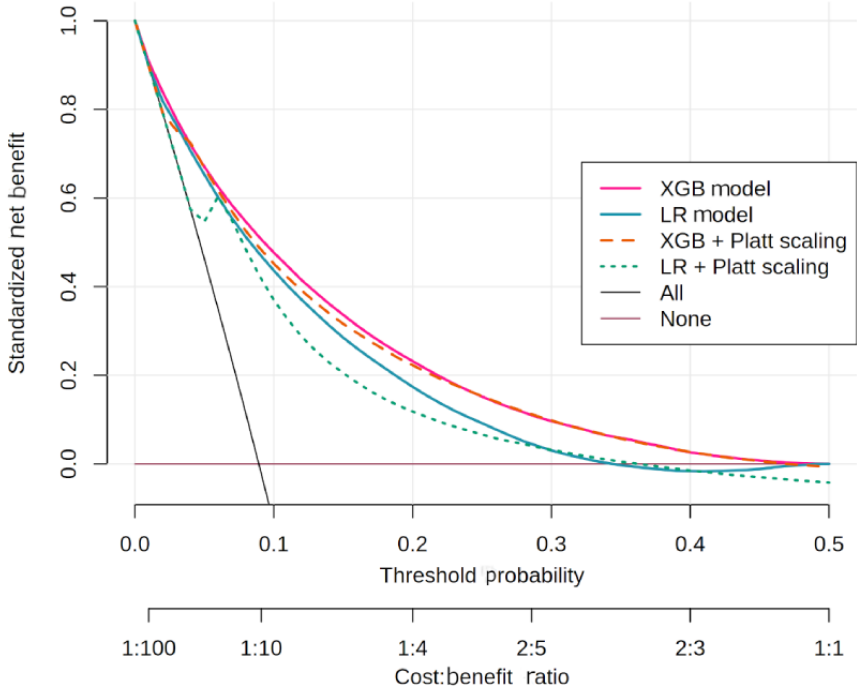
About DCA

Figure 3 presents the clinical utility of prediction models, showing standardized NB across a range of threshold probabilities compared to the default scenarios of "treat all" and "treat none." The XGB and LR models demonstrated broadly similar clinical utility, with XGB consistently yielding slightly higher NB across all thresholds. At low thresholds (10%), both models achieved substantial benefit (NB ≈0.04), indicating a slight advantage over treating all (NB ≈–0.01) or none (NB ≈0). As the threshold increased, the NB of both models gradually declined. However, XGB maintained positive NB up to a threshold probability of approximately 0.5 (cost:benefit ratio ≈1:1), while LR converged toward the "treat none" line near 0.35 (cost:benefit ratio ≈2:5). For instance, at a clinically meaningful threshold of 20% (cost:benefit ratio ≈1:4), where missing a true DHD case is considered 4 times worse than providing an unnecessary intervention, the XGB model retained nearly 50% higher NB (0.020 vs 0.013) compared with LR. This suggests that XGB remains clinically useful even when decision-makers adopt more conservative intervention thresholds, when the cost of unnecessary intervention is comparable to the cost of missing a high-risk case, whereas LR provides meaningful benefit mainly at lower thresholds

XSL•FO
RenderX

(0.05-0.35). Calibration using Platt produced only marginal changes, indicating improved probability reliability (Figure 2) but no substantial gain in clinical utility (Figure 3). More detailed results of the DCA analysis are provided in Multimedia Appendix 5.

Figure 3. Decision curve analysis of XGB and LR models. LR: logistic regression; XGB: extreme gradient boosting.

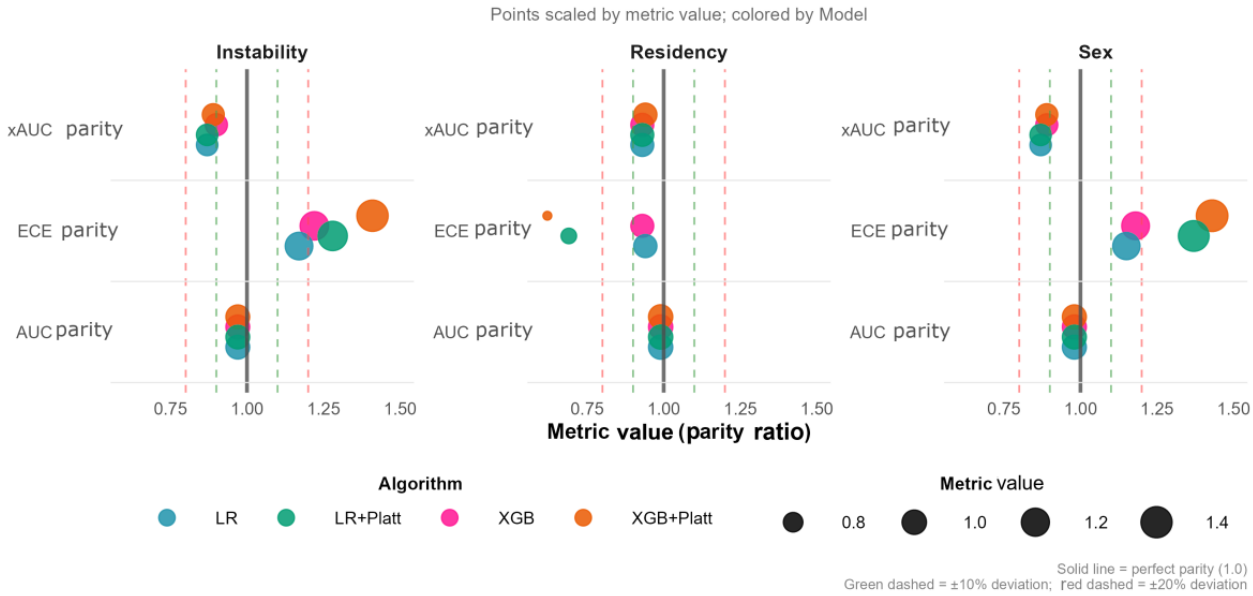


Fairness Analysis

Figure 4 illustrates threshold-free fairness metrics, specifically xAUC parity, ECE parity, and AUC parity, evaluated across 3 equity-seeking groups: females, rural residents, and highly marginalized populations characterized by residential instability. Each point represents a subgroup-specific fairness value, colored by model type and scaled according to its magnitude, with the ideal fairness benchmark of 1.0. Both models demonstrated strong within-group ranking fairness, with AUC parity values consistently approaching 1.0, and robust cross-group ranking fairness, reflected by xAUC parity values generally exceeding 0.9. In contrast, calibration fairness (ECE parity) revealed more substantial disparities, particularly for sex and residential instability, where values often fell outside the ±20% deviation threshold commonly used in the fairness literature [50]. Although Platt scaling improved global calibration performance, its uniform probability adjustment inadvertently magnified calibration differences between groups, thereby reducing ECE parity.

XSL•FO
RenderX

Figure 4. Comparing algorithmic fairness metrics across different predictive models and protected features. The size of each marker represents the magnitude of the fairness metric. The vertical dashed lines at 0.9-1.1 and 0.8-1.2 represent 10% and 20% deviation thresholds from perfect fairness (1.0). AUC: area under the receiver operating characteristic curve; ECE: expected calibration error; LR: logistic regression; xAUC: cross-group area under the curve; XGB: extreme gradient boosting.

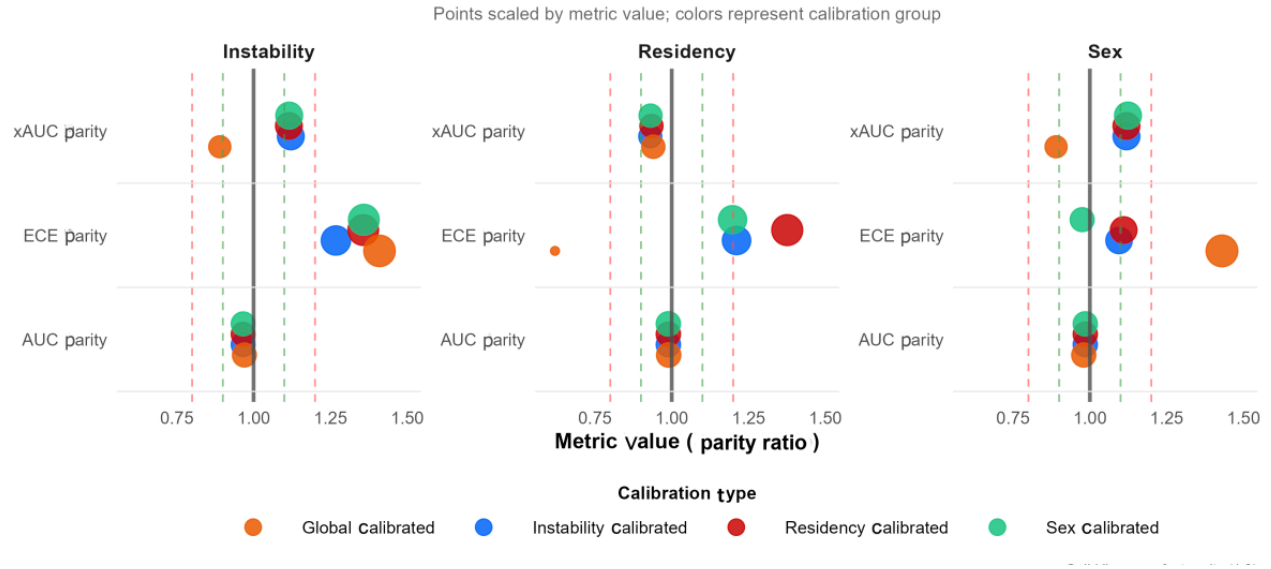


While most metrics remain within a satisfactory level of fairness (less than 20% deviation from perfect parity), prior work has shown that even modest deviations can compound into clinically meaningful inequities for vulnerable populations [21]. Achieving perfect fairness across all criteria simultaneously is inherently difficult due to known fairness-fairness and fairness-accuracy trade-offs [51,52]. To further improve algorithmic fairness and explore its impact on both fairness-fairness and fairness-accuracy trade-offs, we applied postprocessing bias mitigation. This allowed us to assess not only improvements in calibration fairness but also the interactions between fairness criteria and potential impacts on predictive performance.

By implementing the postprocessing bias mitigation described in the section Fairness Assessment, we used group-stratified Platt scaling to enhance calibration fairness (ECE parity) in the XGB model (Figure 5). Among the approaches tested, stratification based on residential instability (depicted by the blue points in Figure 5) proved most effective, yielding substantial improvements in ECE parity across all 3 equity-seeking attributes when compared with global calibration (orange points). This strategy particularly improved calibration parity for sex, bringing ECE parity close to ideal levels (0.975) and well within the ±20% acceptability threshold, followed by moderate improvements in residency and smaller reductions in instability. However, deviations in the latter 2 groups remained slightly above the acceptable range (1.21-1.36). This indicates that residual calibration gaps persist for residency and residential instability even after post hoc mitigation, representing a limitation of group-stratified Platt scaling. Importantly, within-group ranking fairness (AUC parity) remained stable across all calibration schemes (0.966-0.989), confirming that probability rescaling preserved internal risk ordering. Cross-group ranking equity (xAUC parity) showed only minor fluctuations (eg, shifting from approximately 0.9 to 1.1 for instability and sex) yet consistently stayed within accepted bounds.

XSL•FO
RenderX

Figure 5. Comparison of group-based calibration for the XGB before and after post hoc bias mitigation. The size of each marker represents the magnitude of the fairness metric. The vertical dashed lines at 0.9-1.1 and 0.8-1.2 represent 10% and 20% deviation thresholds from perfect fairness (1.0). AUC: area under the receiver operating characteristic curve; ECE: expected calibration error; xAUC: cross-group area under the curve.



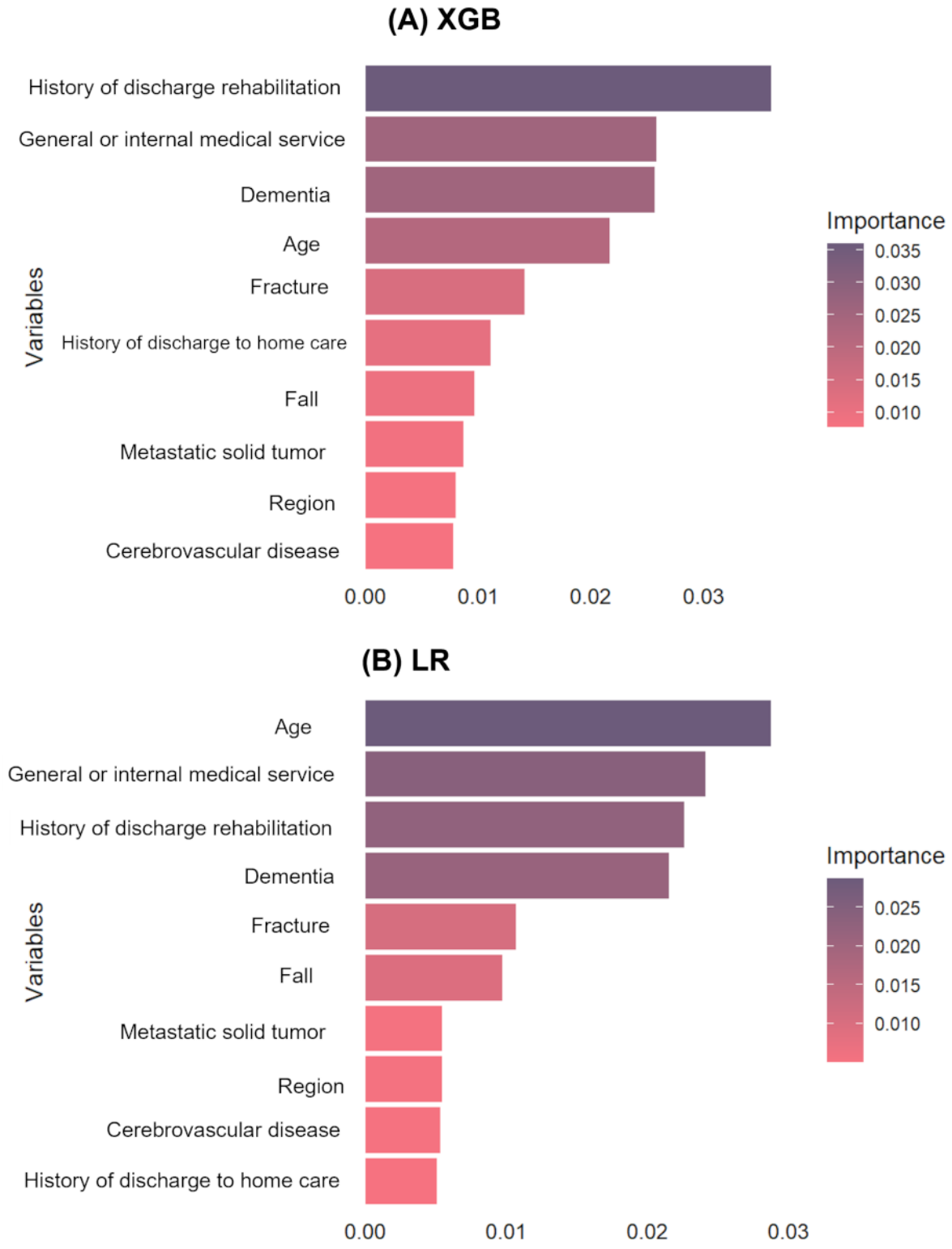
The findings highlight both the promise and limitations of targeted post hoc calibration. Specifically, group-aware recalibration, especially when anchored to the most imbalanced attribute, can significantly mitigate calibration disparities. However, achieving simultaneous improvements across all fairness dimensions remains challenging, reflecting the inherent fairness-fairness trade-off. Moreover, the selection of the stratification attribute introduces a fairness-target trade-off: enhancing calibration for 1 equity-seeking group may unintentionally diminish calibration performance for others. For example, sex-stratified calibration achieved near-perfect parity for sex but performed less favorably for residential instability. Conversely, calibration based on instability improved fairness across all groups, including residency, and in some cases outperformed residency-specific calibration. These patterns illustrate that fairness interventions interact in complex ways; that is, improvements in 1 demographic context can introduce miscalibration in others, highlighting the need for deliberate, context-sensitive prioritization when designing equitable clinical prediction models.

Global Explainability

Figure 6 presents the top 10 most significant predictors, as determined by permutation feature importance scores. Both the XGB and LR models identified several key predictors in common, including the history of needing care support, dementia, fracture, rehabilitation care history, and age. Interestingly, while age was deemed the most important predictor in LR, it held a lower ranking in XGB, suggesting nonlinear models may more effectively capture interactions between variables, thereby reducing the standalone importance of age. Furthermore, the inclusion of regional variables among the top predictors suggests that DHD is not solely driven by patient-level clinical characteristics but is also influenced by geographic or system-level factors.

XSL•FO
RenderX

Figure 6. Top 10 most significant predictors by permutation feature importance scores. (A) Important features by XGB and (B) important features by LR. Hist: history; LR: logistic regression; Med: medical; Rehab: rehabilitation; XGB: extreme gradient boosting.

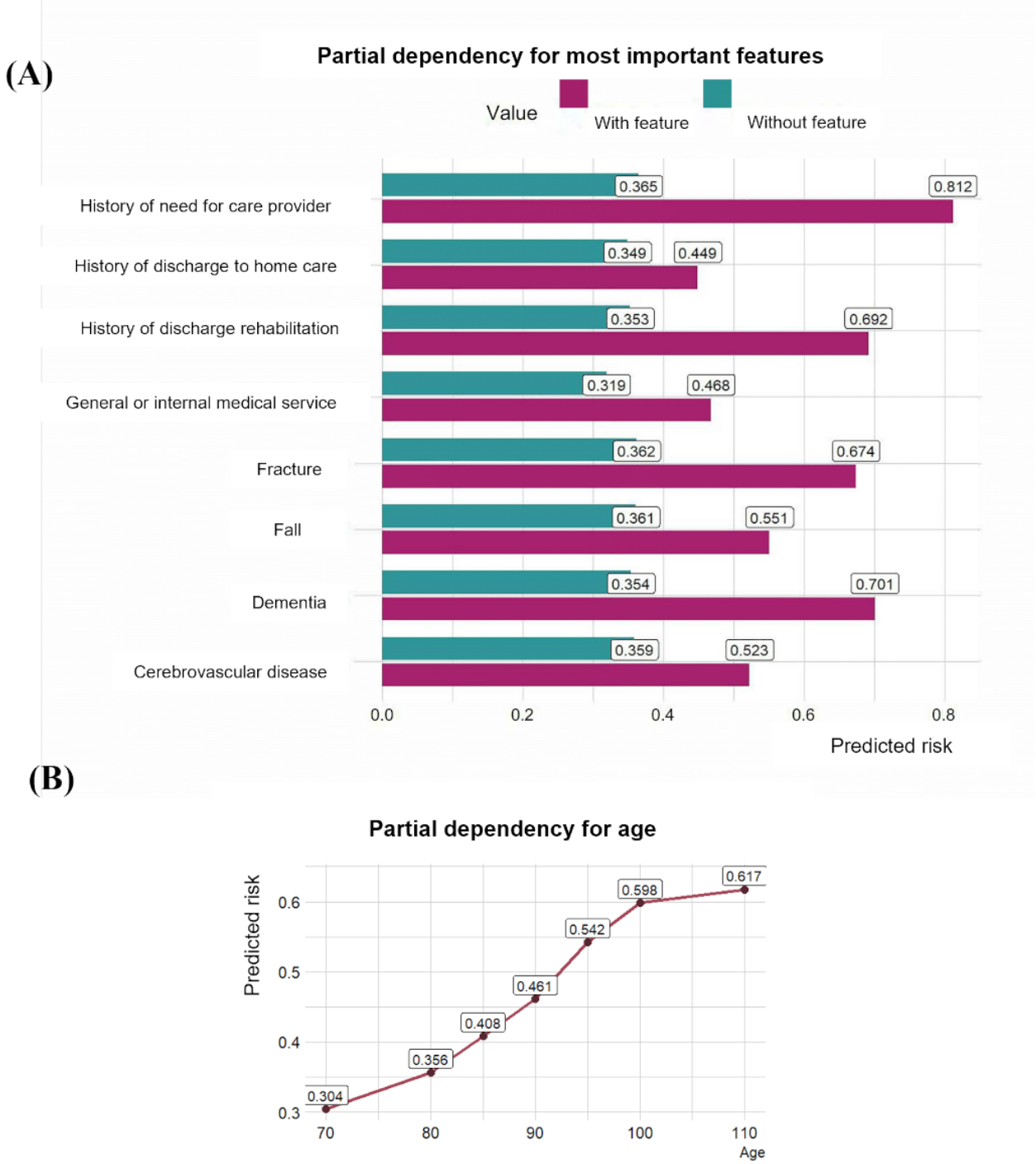


Due to XGB's superior predictive and fairness performances compared to LR, we focused our subsequent analysis on this model for the explainability analysis. Figure 7 shows the PDP for these variables. This figure demonstrates how the predicted risk changes for each predictor when it changes from 0 (with the feature absent) to 1 (with the feature present). As can be seen, the history of need for care support, dementia, and history of discharge to rehabilitation care settings show the most significant changes in predicted risk. These changes are significant enough to raise predicted probabilities from below to above a 0.5 threshold, a commonly used decision point in classification tasks. For instance, patients with dementia exhibit predicted risks exceeding 0.7, underscoring their pivotal role in the risk of DHD. Age also demonstrates a consistent upward trend, with risk increasing steadily from 0.3 at age 70 years to over 0.6 by age 110 years, highlighting age as a critical predictor

XSL•FO
RenderX

at the global level. Additionally, geographic factors appear to influence the predicted risk, with the Northern and Central regions of Ontario associated with the highest risks of delayed discharge (0.43 and 0.39, respectively). In contrast, the Eastern region indicates the lowest risk (0.33). This finding suggests potential geographic disparities in care or health system factors.

Figure 7. Exploring global explainability: partial dependency of top predictors. (A) Clinical factors, (B) demographic factors (age). Hist: history; Med: medical; Rehab: rehabilitation.



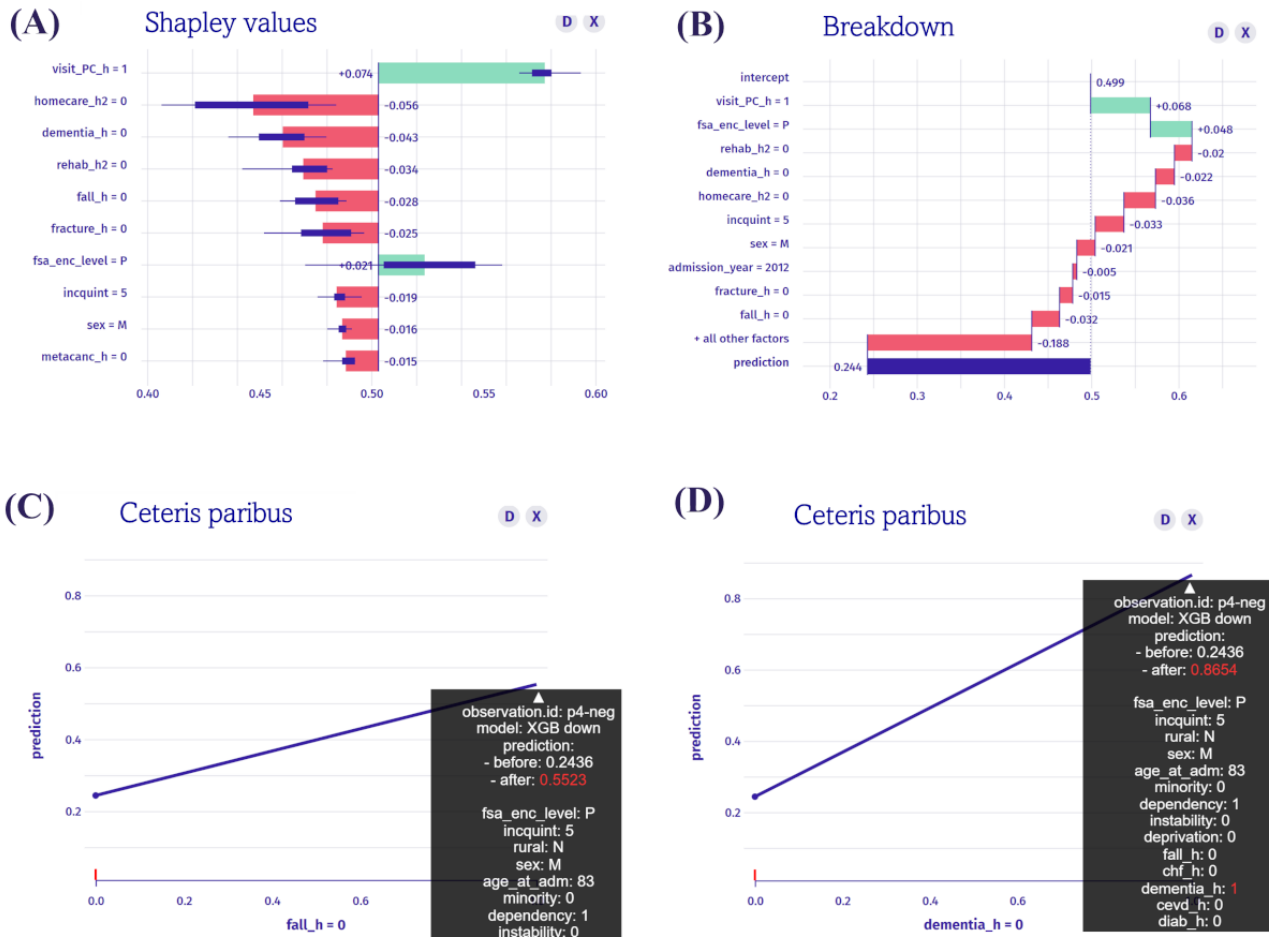
Local Explainability

To illustrate the practical utility of XAI for personalized care planning, we examined a deidentified patient with a predicted DHD risk of 0.24 (Figure 8). While this individual was classified as "low-risk" based on the overall model output, local explainability approaches, including SHAP values, breakdown plots, and ceteris paribus profiles, revealed that the prediction was highly sensitive to minor changes in specific clinical features. The SHAP and breakdown plots showed that the patient's lack of fall history and dementia significantly lowered the predicted risk of DHD. Simulating the presence of either condition using ceteris paribus profiles, while holding all other variables constant, significantly increased the predicted risk

XSL•FO
RenderX

from 0.24 to 0.55 if the same person had a history of falls, and to 0.86 if the person had dementia. This example illustrates how individual-level explainability can uncover a phenomenon we define as "clinical risk instability," a condition in which a patient's health status appears stable until a specific risk factor shifts, resulting in a significant increase in risk. Such underlying vulnerabilities and their critical insights can be achieved through local explainability, whereas global explanations cannot provide such capability.

Figure 8. Local explainability panel for 1 deidentified patient: (A) SHAP values, (B) breakdown plot, (C) ceteris paribus (fall) indicates that introducing a fall history increases the predicted risk from 0.24 to 0.55, while holding other features constant, and (D) ceteris paribus (dementia) reveals a sharper effect, with the risk rising to 0.86 if dementia were present. SHAP: Shapley Additive Explanation; XGB: extreme gradient boosting.



As explained in the Methods section, we also used PCA to cluster features based on their contributions, frequencies, and rankings from SHAP and breakdown explanations for high-risk patients. The resulting biplot is displayed in Figure 9, accounting for 87.8% of the data variance, with 61.5% attributed to PC1 and 26.3% to PC2. PC1 primarily captures the trade-off between high contributions and rankings versus lower frequencies, reflecting features that exert strong but less ubiquitous influence on predictions.

XSL•FO
RenderX

Figure 9. Local interpretations of feature contributions: feature breakdown, clustering, and category distribution. CareDep: history of need for care support; CHF: chronic heart failure; CVD: cerebrovascular disease; Depend: dependency; Depriv: deprivation; Diab: diabetes; GenMed: general medicine service; Hemi/Para: hemiplegia or paraplegia; PallCare: palliative care history; PCA: principal component analysis; PsyMed: psychiatry service; PVD: peripheral vascular disease; Rehab: history of discharge to rehabilitation center; Rheum: rheumatoid; Socio-Demo-Geo: socioeconomic, demographic, and geographic.

In contrast, PC2 emphasizes high contributions paired with frequent appearances, irrespective of ranking, highlighting features that are both impactful and commonly relevant across high-risk cases:

$$PC1 = 0.47 \times contribution + 0.65 \times rank - 0.60 \times percentage$$

$$PC2 = 0.85 \times contribution - 0.15 \times rank + 0.51 \times percentage$$

The coefficients of the variables indicate the magnitude of their influence in the principal component, while their signs reveal whether they exert a positive or negative influence [49]. We categorized the resulting PCA space into 4 regions (Q1-Q4) based on the signs of the coefficients for each feature along the PC1 and PC2 axes. Region Q1, depicted by green points, is characterized by high values on both components and primarily comprises features related to mobility and disability, which account for over 75% of its attributes. The remaining features in Q1 are related to multimorbidity and health status, with no contribution from sociodemographic or geographic factors. This pattern indicates that the loss of independence and the presence of chronic conditions are pivotal in predicting DHD risk. The bar chart confirms that mobility-related issues are significant contributors to DHD.

This PCA-derived clustering provides deeper interpretive value by grouping features with similar profiles, facilitating a holistic understanding beyond individual explanations. For instance, the dominance of mobility and disability in Q1 suggests that interventions targeting physical function could mitigate risks for a substantial portion of high-risk patients, offering policymakers actionable insights to prioritize community-based rehabilitation services that enhance or maintain mobility. Such services, delivered by physiotherapists, occupational therapists, and kinesiologists, include strength and balance training, support for activities of daily living, and personalized exercise programs, and have demonstrated effectiveness in preventing functional decline and alleviating acute care burdens [53]. The history of the need for care support consistently ranked as the most significant contributor in the mobility and disability category and as one of the most impactful features. In the category of multimorbidity, dementia emerged as the most significant factor. Overall, these results reinforce the utility of PCA clustering in distilling complex local data into strategic patterns, bridging microlevel insights with macrolevel policy implications for more equitable and efficient health care delivery. Feature-level PCA outputs, including each predictor's PC1 and PC2 values, quadrant assignment (Q1-Q4), and predefined conceptual categories, are provided in Multimedia Appendix 4 to support transparency and reproducibility of the clustering interpretation.

Discussion

Principal Findings

Our study emphasizes the potential of ML to accurately predict DHD by using extensive longitudinal administrative health data. The XGB model outperformed LR in both predictive accuracy and fairness, supporting the effectiveness of advanced ML techniques in identifying patients at risk for DHD. This predictive capability facilitates proactive interventions, helping to alleviate care bottlenecks and enhance discharge planning.

Explainability analyses revealed that high-risk predictions were predominantly driven by mobility and functional impairments, as well as cognitive impairment factors (eg, dementia). These findings are aligned with previous literature that consistently highlights dementia as one of the most significant risk factors for delayed discharge [54]. PCA clustering of local explanations further highlighted the predominance of mobility and disability features in the high-impact quadrant, accounting for over 75% of attributes there, which shows the critical importance of functional and cognitive declines on high-risk predictions and informs targeted preventive strategies. Local explainability results further shed light on how individual features influence predictions, clarifying what might otherwise remain a "black box." For instance, partial dependency analysis revealed substantial risk shifts associated with dementia and the need for care support, emphasizing their importance in clinical decision-making. We also observed regional patterns in predicted risk, consistent with studies on geographic disparities

XSL•FO
RenderX

with respect to clinical outcomes [55]. By connecting these insights to actionable interventions, such as implementing earlier discharge planning for high-risk groups and conducting routine fear-of-falling screening in primary and community care to trigger fall-prevention referrals [56,57], and ensuring access to postacute rehabilitation for individuals with dementia and those at the risk of falls who are often excluded under current eligibility criteria [58], our study contributes to the growing body of work aiming to make ML tools both interpretable and impactful in real-world health care settings.

Fairness emerged as another important aspect of our analysis. Our results indicate that the models performed fairly across important subgroups, such as sex, rural-urban residency, and residential instability, supporting the broader goal in health care ML to prevent inequities. We highlighted the complex trade-offs that arise not only between fairness and accuracy but also among various aspects of fairness themselves, emphasizing the value of involving decision-makers in the model development process, as resolving these tensions often requires contextual knowledge and consideration of policy goals. This aligns with principles highlighted in a *New England Journal of Medicine* study, specifically regarding the ethical use of AI in health care [59].

Clinical, Managerial, and Policy Implications

DHDs pose a significant challenge to health care systems, straining hospital capacity and compromising patient care. By identifying patients at risk of delayed discharge in advance, our findings enable health administrators to proactively allocate necessary supportive services, such as home care referrals, rehabilitative and therapeutic care, and community support, where they are most needed. This approach promotes a more efficient and proactive health care system, ultimately reducing unnecessary admissions that may lead to prolonged hospital stays. Second, understanding the factors that consistently contribute to an elevated risk of delayed discharge can reveal systemic deficiencies. If patients with specific mobility or cognitive impairments are frequently categorized as high-risk, it may indicate underinvestment in and insufficient publicly funded community-based rehabilitation services. Recognizing these deficiencies can help direct policy initiatives and funding toward addressing them, thereby enhancing system capacity over time. Third, by synthesizing individual-level insights, health system planners can monitor trends within specific population subgroups. Regions consistently associated with higher risk or patient cohorts facing distinct challenges can be targeted through specialized programs, infrastructure investments, or policy reforms. Fourth, leveraging local XAI analyses, such as ceteris paribus and SHAP, to detect clinical risk instability (where a patient's seemingly stable condition rapidly shifts to a high-risk state) empowers clinicians to initiate anticipatory interventions, including cognitive screening, home safety evaluations, or pre-emptive referrals to community services.

At a systems level, this highlights the value of integrating XAI tools not only to explain retrospective predictions but also to inform real-time, forward-looking care strategies that align with the principles of personalized and preventive medicine. Finally, with a clearer understanding of how various factors interact to hinder timely discharge, decision-makers can implement strategic interventions, such as early geriatric assessments, specialized dementia care pathways, and improved coordination with community care providers to prevent hospitalization delays. Instead of responding to existing backlogs, policymakers can take proactive measures to enhance the quality and continuity of care.

Limitations

First, the predictive models are developed using retrospective cohort data, which limits control over data collection and can introduce selection bias. Second, our predictive models only focus on older delayed-discharge patients; therefore, their applicability to other populations is limited. Third, our data may lack confounders, such as social determinants of health (eg, individual-level income, individual-level education, race, housing support, caregiver availability, and capacity), the severity of cognitive or physical impairments, and caregiver support levels. Thus, predictive model performance may improve by incorporating these risk factors. Fourth, PDPs summarize average marginal associations between predictors and model predictions. However, when predictors are correlated, PDP estimates can be influenced by the joint distribution of features and may implicitly rely on extrapolation into combinations of covariate values that are uncommon or unsupported in the observed data, complicating causal or purely marginal interpretation [46,60]. Future work could use accumulated local effects profiles as a robustness check, as accumulated local effects are less sensitive to feature dependence and can provide more stable global effect summaries under correlation. Finally, model performance was evaluated using a large held-out test set and an independent temporally separated cohort, providing an assessment of both internal performance and temporal generalizability. Given the size of the test set, performance estimates from a single evaluation are expected to be stable under the law of large numbers; however, this approach does not provide explicit uncertainty estimates (eg, CIs). Consequently, while XGB consistently demonstrated higher discrimination and improved calibration relative to LR across both the random hold-out and temporal validation settings, we caution against overinterpreting performance differences. Future work could strengthen comparative inference and quantify performance variability through repeated cross-validated evaluation or nested cross-validation, particularly when the objective is to establish statistically robust superiority between competing models.

Conclusions

Our study used an integrated approach that develops predictive models, ensures fairness, and incorporates XAI, offering a forward-looking perspective on the ongoing challenge of delayed discharge. By identifying at-risk patients early and promoting equitable outcomes while illuminating the factors driving our predictions, we established a more robust foundation for prevention-oriented policies. By focusing on accuracy, fairness, and explainability, this study advances the development of ML models that are not only accurate but also ethical and interpretable. These triple considerations are critical for fostering trust in AI-driven decision-making in health care and improving

XSL•FO
RenderX

patient outcomes. This enables health system leaders, policymakers, clinicians, and community partners to implement evidence-based interventions that address the root causes of DHD. Ultimately, this enhances their ability to manage care, reduce delays, optimize resource allocation, and improve the quality and equity of care for older adults at risk of delayed discharge.

Acknowledgments

This study contracted ICES (Institute for Clinical Evaluative Sciences) Data & Analytic Services (DAS) and used deidentified data from the ICES Data Repository, which is managed by ICES with support from its funders and partners: Canada's Strategy for Patient-Oriented Research (SPOR), the Ontario SPOR Support Unit, the Canadian Institutes of Health Research, and the Government of Ontario. The opinions, results, and conclusions reported are those of the authors. No endorsement by ICES or any of its funders or partners is intended or should be inferred.

The generative AI tool Grammarly was used to improve readability, narrative coherence, and grammatical structure. Grammarly was solely used as a writing aid. The scientific content remains the work of the authors.

Data Availability

We are unable to provide a minimal dataset for this study due to privacy, legal, and ethical restrictions, as well as prescribed entity designations. All data used in this study are securely housed at ICES, Ontario, Canada, in coded form and are subject to their privacy, legal, prescribed entity designations, and ethical governance. While legal data-sharing agreements between ICES and data providers (eg, health care organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet prespecified criteria for confidential access. A high-level script skeleton and pseudocode outlining the full analysis pipeline are publicly available [61].

Funding

This research project was funded by the Canadian Institutes for Health Research (186083). KK holds the Dr. Mathias Gysler Research Chair in Patient- and Family-Centred Care, funded by the Trillium Health Partners Foundation. S Guilcher holds a pain scientist salary award from the Leslie Dan Faculty of Pharmacy, University of Toronto. The funder played no role in this study's design, data collection, analysis and interpretation of data, or the writing of this paper.

Authors' Contributions

S Ghazalbash and MZ carried out problem formulation. S Ghazalbash conducted data analytics. S Ghazalbash drafted the initial version of this paper. S Ghazalbash and MZ contributed to the interpretation of results and revised this paper for intellectual content. S Guilcher and KK provided critical insights into the practice and policy implications of the results and contributed to further revisions of this paper. All authors read and approved this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of data sources and cohort characteristics.
[DOCX File , 31 KB - medinform_v14i1e83244_app1.docx]

Multimedia Appendix 2
Data preparation.
[DOCX File , 33 KB - medinform_v14i1e83244_app2.docx]

Multimedia Appendix 3
Technical and package information.
[DOCX File , 19 KB - medinform_v14i1e83244_app3.docx]

Multimedia Appendix 4
Mapping features.
[DOCX File , 20 KB - medinform_v14i1e83244_app4.docx]

Multimedia Appendix 5
Decision curve analysis.

XSL•FO
RenderX

[DOCX File , 17 KB - medinform_v14i1e83244_app5.docx]

References

1. McGilton KS, Vellani S, Babineau J, Bethell J, Bronskill SE, Burr E, et al. Understanding transitional care programs for older adults who experience delayed discharge: a scoping review. BMC Geriatr 2021;21(1):210 [FREE Full text] [doi: 10.1186/s12877-021-02099-9] [Medline: 33781222]
2. Landeiro F, Roberts K, Gray AM, Leal J. Delayed hospital discharges of older patients: a systematic review on prevalence and costs. Gerontologist 2019;59(2):e86-e97. [doi: 10.1093/geront/gnx028] [Medline: 28535285]
3. Walker D, Lead PA. Caring for our aging population and addressing alternate level of care. 2011. URL: https://www.loftcs.org/wp-content/uploads/2019/11/Caring-for-our-Aging-Population-and-Addressing-ALC-report-walker_2011-ontario.pdf [accessed 2026-03-25]
4. Guilcher SJT, Bai YQ, Wodchis WP, Bronskill SE, Rashidian L, Kuluski K. What happened to the patients? Care trajectories for persons with a delayed hospital discharge during wave 1 of COVID-19 in Ontario, Canada; a population-based retrospective cohort study. PLoS One 2024;19(9):e0309155 [FREE Full text] [doi: 10.1371/journal.pone.0309155] [Medline: 39331610]
5. Barnable A, Welsh D, Lundrigan E, Davis C. Analysis of the influencing factors associated with being designated alternate level of care. Home Health Care Manage Practice 2015;27(1):3-12. [doi: 10.1177/1084822314539164]
6. Rojas-García A, Turner S, Pizzo E, Hudson E, Thomas J, Raine R. Impact and experiences of delayed discharge: a mixed-studies systematic review. Health Expect 2018;21(1):41-56 [FREE Full text] [doi: 10.1111/hex.12619] [Medline: 28898930]
7. Alternate level of care in Canada: evidence assessment report. Canadian Journal of Health Technologies. 2025. URL: https://www.cda-amc.ca/sites/default/files/ou-tr/OP0557-ALC_Combined_Report.pdf [accessed 2026-04-07]
8. McCloskey R, Jarrett P, Stewart C, Nicholson P. Alternate level of care patients in hospitals: what does dementia have to do with this? Can Geriatr J 2014;17(3):88-94 [FREE Full text] [doi: 10.5770/cgj.17.106] [Medline: 25232367]
9. Rosalie Wyonch – fixing the ALC patient problem is a triple win for Canadians. 2024. URL: https://www.cdhowe.org/intelligence-memos/rosalie-wyonch-fixing-alc-patient-problem-triple-win-canadians [accessed 2026-03-25]
10. Durante S, Fyie K, Zwicker J, Carpenter T. Confronting the alternate level of care (ALC) crisis with a multifaceted policy lens. SPP Publications 2023;16(1):1-32. [doi: 10.55016/ojs/sppp.v16i1.76748]
11. Alternate level of care leading practices guide. Provincial Geriatrics Leadership Ontario. 2021. URL: https://geriatricsontario.ca/resources/alternate-level-of-care-leading-practices-guide/ [accessed 2026-03-25]
12. No authors. Strategies to reduce alternate level of care: environmental scan [internet]. Ottawa (ON): Can Agency Drugs Technol Health 2024 [FREE Full text] [Medline: 39236207]
13. Bell B. We aren't properly preparing for long-term care demand. Toronto Star. 2019. URL: https://www.thestar.com/opinion/contributors/we-aren-t-properly-preparing-for-long-term-care-demand/article_d4e110bf-f9f8-54f3-8215-40f31d37cabf.html [accessed 2024-09-19]
14. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021;20(2):154-170. [doi: 10.1002/wps.20882]
15. Swinckels L, Bennis FC, Ziesemer KA, Scheerman JFM, Bijwaard H, de Keijzer A, et al. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. J Med Internet Res 2024;26:e48320 [FREE Full text] [doi: 10.2196/48320] [Medline: 39163096]
16. Hunik L, Chaabouni A, van Laarhoven T, Olde Hartman TC, Leijenaar RTH, Cals JWL, et al. Diagnostic prediction models for primary care, based on AI and electronic health records: systematic review. JMIR Med Inform 2025;13:e62862 [FREE Full text] [doi: 10.2196/62862] [Medline: 40845324]
17. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
18. Singh M, Ghalachyan G, Varshney KR, Bryant RE. An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness. arXiv. Preprint posted online on Sep 29, 2021 . [doi: 10.48550/arXiv.2109.14653]
19. Liu M, Ning Y, Teixayavong S, Liu X, Mertens M, Shang Y, et al. A scoping review and evidence gap analysis of clinical AI fairness. NPJ Digit Med 2025;8(1):360 [FREE Full text] [doi: 10.1038/s41746-025-01667-2] [Medline: 40517148]
20. Ghassemi M, Nsoesie EO. In medicine, how do we machine learn anything real? Patterns (N Y) 2022;3(1):100392 [FREE Full text] [doi: 10.1016/j.patter.2021.100392] [Medline: 35079713]
21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
22. Yang CC. Explainable artificial intelligence for predictive modeling in healthcare. J Healthc Inform Res 2022;6(2):228-239 [FREE Full text] [doi: 10.1007/s41666-022-00114-1] [Medline: 35194568]
23. Bai AD, Dai C, Srivastava S, Smith CA, Gill SS. Risk factors, costs and complications of delayed hospital discharge from internal medicine wards at a Canadian academic medical centre: retrospective cohort study. BMC Health Serv Res 2019;19(1):935 [FREE Full text] [doi: 10.1186/s12913-019-4760-3] [Medline: 31801590]

XSL•FO
RenderX

24. Carfagnini QA, Ayanso A, Law MP, Orlando E, Faught BE. What factors increase odds of long-stay delayed discharge in alternate level of care patients? J Am Med Dir Assoc 2023;24(9):1327-1333. [doi: 10.1016/j.jamda.2023.02.104] [Medline: 36996875]
25. Turcotte LA, Daniel I, Hirdes JP. The post-acute delayed discharge risk scale: derivation and validation with Ontario alternate level of care patients in ontario complex continuing care hospitals. J Am Med Dir Assoc 2020;21(4):538-544.e1. [doi: 10.1016/j.jamda.2019.12.022] [Medline: 32089427]
26. Francis NK, Luther A, Salib E, Allanby L, Messenger D, Allison AS, et al. The use of artificial neural networks to predict delayed discharge and readmission in enhanced recovery following laparoscopic colorectal cancer surgery. Tech Coloproctol 2015;19(7):419-428. [doi: 10.1007/s10151-015-1319-0] [Medline: 26084884]
27. Matheson FI, Moloney G, van Ingen T. 2016 Ontario marginalization index. 2018. URL: https://www.publichealthontario.ca/-/media/documents/U/2018/userguide-on-marg.pdf [accessed 2026-03-25]
28. Definitions and guidelines to support ALC designation in acute inpatient care. Canadian Institute for Health Information. 2016. URL: https://www.cihi.ca/sites/default/files/document/acuteinpatientalc-definitionsandguidelines_en.pdf [accessed 2024-09-19]
29. Rahman S, Irfan M, Raza M, Moyeezullah Ghori K, Yaqoob S, Awais M. Performance analysis of boosting classifiers in recognizing activities of daily living. Int J Environ Res Public Health 2020;17(3):1082 [FREE Full text] [doi: 10.3390/ijerph17031082] [Medline: 32046302]
30. Nayak S, Amin A, Reghunath SR, Thunga G, Acharya D, Shivashankara KN, et al. Development of a machine learning-based model for the prediction and progression of diabetic kidney disease: a single centred retrospective study. Int J Med Inform 2024;190:105546 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105546] [Medline: 39003788]
31. Huang Y, Jiang X, Gabriel RA, Ohno-Machado L. Calibrating predictive model estimates in a distributed network of patient data. J Biomed Inform 2021;117:103758 [FREE Full text] [doi: 10.1016/j.jbi.2021.103758] [Medline: 33811986]
32. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020;27(4):621-633 [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
33. Calster BV, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet 2025;7(12):100916 [FREE Full text] [doi: 10.1016/j.landig.2025.100916] [Medline: 41391983]
34. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3:18 [FREE Full text] [doi: 10.1186/s41512-019-0064-7] [Medline: 31592444]
35. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016;34(21):2534-2540 [FREE Full text] [doi: 10.1200/JCO.2015.65.5654] [Medline: 27247223]
36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574 [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
37. Zhang Z, Rousson V, Lee WC, Ferdynus C, Chen M, Qian X, et al. Decision curve analysis: a technical note. Ann Transl Med 2018;6(15):308 [FREE Full text] [doi: 10.21037/atm.2018.07.02] [Medline: 30211196]
38. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform 2021;113:103621 [FREE Full text] [doi: 10.1016/j.jbi.2020.103621] [Medline: 33220494]
39. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. PMLR 2017;70:1321-1330 [FREE Full text]
40. Kallus N, Zhou A. The fairness of risk scores beyond classification: bipartite ranking and the xauc metric. arXiv. Preprint posted online on Jun 1, 2019 . [doi: 10.48550/arXiv.1902.05826]
41. Kallus N, Mao X, Zhou A. Assessing algorithmic fairness with unobserved protected class using data combination. Manage Sci 2022;68(3):1959-1981. [doi: 10.1287/mnsc.2020.3850]
42. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
43. Subedi R, Greenberg TL, Roshanafshar S. Does geography matter in mortality? An analysis of potentially avoidable mortality by remoteness index in Canada. Health Rep 2019;30(5):3-15 [FREE Full text] [doi: 10.25318/82-003-x201900500001-eng] [Medline: 31091331]
44. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
45. Kim K, Yoon Y, Shin S. Explainable prediction of problematic smartphone use among South Korea's children and adolescents using a machine learning approach. Int J Med Inform 2024;186:105441. [doi: 10.1016/j.ijmedinf.2024.105441] [Medline: 38564961]
46. Biecek P, Burzykowski T. Explanatory Model Analysis. New York, NY: Chapman and Hall/CRC; 2021.
47. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NIPS'17: Proc 31st Int Conf Neural Inf Proces Syst 2017:4768-4777. [doi: 10.5555/3295222.3295230]

XSL•FO
RenderX

48. Staniak M, Biecek P. Explanations of model predictions with live and breakDown packages. R J 2018;10(2):395-409. [doi: 10.32614/rj-2018-072]
49. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. Nat Rev Methods Primers 2022;2(1):100. [doi: 10.1038/s43586-022-00184-w]
50. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2022;55(3):1-44. [doi: 10.1145/3494672]
51. Rodolfa KT, Lamba H, Ghani R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nat Mach Intell 2021;3(10):896-904. [doi: 10.1038/s42256-021-00396-x]
52. Dang VN, Cascarano A, Mulder RH, Cecil C, Zuluaga MA, Hernández-González J, et al. Fairness and bias correction in machine learning for depression prediction across four study populations. Sci Rep 2024;14(1):7848 [FREE Full text] [doi: 10.1038/s41598-024-58427-7] [Medline: 38570587]
53. Li F, Goh KS, Yu X, Png GK, Chew THS, Ang GC, et al. Mobility matters: a protocol to improve mobility and reduce length of stay in hospitalised older adults. BMJ Open Qual 2025;14(1):e003084 [FREE Full text] [doi: 10.1136/bmjoq-2024-003084] [Medline: 40086812]
54. Aaltonen M, El Adam S, Martin-Matthews A, Sakamoto M, Strumpf E, McGrail K. Dementia and poor continuity of primary care delay hospital discharge in older adults: a population-based study from 2001 to 2016. J Am Med Dir Assoc 2021;22(7):1484-1492.e3. [doi: 10.1016/j.jamda.2020.11.030] [Medline: 33358723]
55. Wilkie RZ, Ho JY. Life expectancy and geographic variation in mortality: an observational comparison study of six high-income Anglophone countries. BMJ Open 2024;14(7):e079365 [FREE Full text] [doi: 10.1136/bmjopen-2023-079365] [Medline: 39138004]
56. Asai T, Oshima K, Fukumoto Y, Yonezawa Y, Matsuo A, Misu S. The association between fear of falling and occurrence of falls: a one-year cohort study. BMC Geriatr 2022;22(1):393 [FREE Full text] [doi: 10.1186/s12877-022-03018-2] [Medline: 35509040]
57. Yoshikawa A, Ramirez G, Smith ML, Lee S, Ory MG. Systematic review and meta-analysis of fear of falling and fall-related efficacy in a widely disseminated community-based fall prevention program. Arch Gerontol Geriatr 2020;91:104235. [doi: 10.1016/j.archger.2020.104235] [Medline: 32911232]
58. McGilton KS, Vellani S, Zheng N, Wang D, Yeung L, Escrig-Pinol A. Healthcare professionals' perspectives on rehabilitating persons with cognitive impairment. Dementia (London) 2021;20(5):1772-1790 [FREE Full text] [doi: 10.1177/1471301220969615] [Medline: 33222528]
59. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
60. Baniecki H, Parzych D, Biecek P. The grammar of interactive explanatory model analysis. Data Min Knowl Discov 2024;38:2596-2632 [FREE Full text] [doi: 10.1007/s10618-023-00924-w] [Medline: 36818741]
61. alc-delayed-discharge-ml-pipeline-. GitHub. URL: https://github.com/ghazalbs/alc-delayed-discharge-ml-pipeline- [accessed 2026-04-07]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
DCA: decision curve analysis
DHD: delayed hospital discharge
ECE: expected calibration error
HiREB: Hamilton Integrated Research Ethics Board
ICES: Institute for Clinical Evaluative Sciences
LR: logistic regression
LTC: long-term care
ML: machine learning
NB: net benefit
ON-Marg: Ontario Marginalization Index
PDP: partial dependence plot
PCA: principal component analysis
SHAP: Shapley Additive Explanations
XAI: explainable artificial intelligence
xAUC: cross-group area under the curve
XGB: extreme gradient boosting

XSL•FO
RenderX

Edited by A Benis; submitted 30.Aug.2025; peer-reviewed by M Al-Agil, V Moglia; comments to author 07.Oct.2025; revised version received 18.Feb.2026; accepted 06.Mar.2026; published 13.Apr.2026.

Please cite as:
Ghazalbash S, Zargoush M, JT Guilcher S, Kuluski K
Responsible AI for Predicting Delayed Hospital Discharge Among Older Adults: Development and Evaluation Study for Balancing Accuracy, Equity, and Explainability
JMIR Med Inform 2026;14:e83244
URL: https://medinform.jmir.org/2026/1/e83244
doi:10.2196/83244
PMID:41973500

©Somayeh Ghazalbash, Manaf Zargoush, Sara JT Guilcher, Kerry Kuluski. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of Feature Selection Methods in Machine Learning Models of Cancer Information Seeking Among United States Adults: Cross-Sectional Study

Ying Liu[1*], PhD; Kesheng Wang[2,3*], PhD

[1]Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN, United States
[2]Cancer Survivorship Research Center, College of Nursing, University of South Carolina, Columbia, SC, United States
[3]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
[*]all authors contributed equally

Corresponding Author:
Ying Liu, PhD
Department of Biostatistics and Epidemiology, College of Public Health
East Tennessee State University
P.O. Box 70259
1276 Gilbreath Dr.
Johnson City, TN, 37601
United States
Phone: 1 4234396662
Email: liuy09@etsu.edu

Abstract

Background: Feature selection is the process of identifying the most informative and relevant features from a larger set of candidate features in machine learning (ML) models. The Boruta algorithm and the least absolute shrinkage and selection operator (LASSO) are 2 widely used methods.

Objective: This study aimed to (1) compare several feature-selection strategies, including Boruta, LASSO, their intersection, principal component analysis (PCA), and a no–feature-selection baseline, and (2) evaluate ML models to predict cancer information–seeking behavior among US adults.

Methods: Data from 5505 individuals (2630 cancer information seekers and 2875 nonseekers) were selected from the 2022 Health Information National Trends Survey. The Boruta algorithm, LASSO, and PCA were used to perform feature selection of 73 variables. Five ML tools (the support vector machine algorithms, logistic regression [LR], random forest [RF], k-nearest neighbor, and extreme gradient boosting) were applied to develop ML models to predict cancer information–seeking. The area under the receiver operating characteristic curve (AUC) and the DeLong test were used to evaluate and compare the performance of the models. Stepwise LR analysis was performed to estimate the odds ratios and their 95% CIs for the associations of potential variables selected in ML analyses with the outcome.

Results: Overall, 47.8% (2630/5505) of respondents reported seeking cancer information (949/2189, 43.4% of men; 1681/2189, 50.7% of women). RF achieved the highest AUC (0.781) and second-highest accuracy (0.714) using LASSO-selected variables, while the support vector machine with linear kernel and LR models using all 73 features yielded the highest accuracy (0.717). Notably, RF produced comparable AUCs when using Boruta-only features, LASSO-only features, or no feature selection yet (all 73 features); these AUCs were significantly higher than those derived from PCA components or from the 20 PCA-loading–based variables. Stepwise LR confirmed that 19 of the 27 shared variables selected by both Boruta and LASSO were independently associated with information seeking (P<.05). The top predictors included a personal history of cancer, greater worry about developing cancer, a family history of cancer, non-Hispanic White race, higher household income, awareness of genetic testing, viewing health-related videos on social media, interest in cancer screening, being offered access to an online medical record, and knowledge of human papillomavirus.

Conclusions: Boruta and LASSO demonstrated strong and consistent performance in feature selection for predicting cancer information seeking, whereas PCA provided a dimension-reduced yet less predictive alternative. Findings offer actionable insights for tailoring public health communication strategies and improving engagement in cancer information resources among US adults.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e75862) doi:10.2196/75862

KEYWORDS

beliefs; cancer information seeking; data mining; feature selection; health behaviors; knowledge; machine learning; PCA; random forest

Introduction

With the development of science and technology, people seek health information from various sources, including online and offline approaches, for different reasons, such as to understand personal or family members' health conditions and to make decisions on health professionals' recommendations [1]. According to the United States National Cancer Institute (NCI) report, approximately 40.5% of people will have cancer in their lifetime based on 2017-2019 data. Over 2 million people were diagnosed in the United States in 2024 [2]. Patients often actively seek cancer information to better understand their diagnosis, treatment options, and home care. Their family members frequently join in the search for additional information. Healthy people may seek cancer information to enlarge their knowledge and improve their quality of life, such as changing their diet to healthy food and having regular physical exams [3]. Generally, cancer information–seeking behavior may increase knowledge, preventive behaviors, and screening behaviors; moreover, cancer information seekers may be more likely to adopt healthy lifestyle behaviors and get screened for cancer [3-7].

Several sociodemographic factors, such as being female, aged 55-64 years vs 40-44 years, having higher education, identifying as Black or Hispanic, and being married, have been positively associated with cancer information seeking in the US population. Racial disparities and variations by marital status and cancer status have also been reported [5,8-13]. Furthermore, behaviors such as alcohol use and tobacco smoking use, as well as certain chronic conditions (eg, cancer and anxiety), may influence cancer information seeking [3,14-16]. Beliefs about cancer and knowledge of genetic testing have also been associated with cancer information–seeking behavior, though findings in these areas remain inconsistent [13,15,17-22].

Machine learning (ML) and predictive analytics are commonly used in many areas, and they can transform data into useful insights for better understanding and faster decision-making. ML methods can address high-dimensional data, model the etiological and clinical heterogeneity, and translate univariate variable findings into clinically useful multivariate decision support systems [23-27]. Feature selection is a critical step in ML, not only to reduce the dimensionality of the feature space, but also to reveal the most relevant features without losing too much information [28-30]. Feature selection, as a preprocessing stage, is essentially the process of picking some informative and relevant features from a larger collection of features that produce a better characterization of patterns of multiple classes [28,31]. Several feature-selection methods have been used in ML, such as the least absolute shrinkage and selection operator (LASSO) [32-35] and the Boruta algorithm [32,33]. Principal component analysis (PCA) is an unsupervised feature reduction technique that explains the variance-covariance structure of a set of variables through linear combinations known as principal components (PCs) or factors. PCA has been used to reduce the dimensions, but it is not a feature-selection method because all variables remain in each factor [31,36-38].

Numerous studies have investigated cancer information–seeking behaviors [3,8-12,14,19,39]. However, relatively limited research has applied ML approaches to systematically identify key factors associated with cancer information–seeking behaviors. The Health Information National Trends Survey (HINTS) conducted by the NCI is a nationally representative cross-sectional survey of civilian noninstitutionalized adults aged 18 years and older in the United States. HINTS collected comprehensive data on the access to, use of, and needs for health- and cancer-related information, as well as knowledge, perceptions, attitudes, and related health behaviors. It has been widely used to address cancer information seeking [3,8-11,13-20]. Several ML tools, such as logistic regression (LR), support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and extreme gradient boosting (XGBoost), have been used in the classification of binge drinking, e-cigarette use, and severe psychological distress using HINTS data [40-43]. Therefore, this study aimed to (1) compare feature-selection methods—Boruta, LASSO, combination of Boruta and LASSO, and PCA-based methods—and (2) develop ML tools to predict cancer information seeking among US adults using the data from the 2022 Health Information National Trends Survey (HINTS 6).

Methods

Sample

The data for this study were selected from the HINTS 6, which included 6252 respondents. The HINTS is a nationally representative survey administered by the NCI since 2003. The HINTS targets adults aged 18 years or older in the civilian noninstitutionalized population of the United States. The HINTS, sponsored by the NCI, provides a unique opportunity to explore the characteristics of information seekers and nonseekers, as well as the content of information being sought by the public in a nationally representative sample. Data collection for HINTS 6 started on March 7, 2022, and concluded on November 8, 2022. The overall household response rate, based on the next-birthday method, was 28.1%.

Ethical Considerations

The original HINTS 6 data collection by the NCI was designated "exempt research" under 45 CFR 46.104 and approved by the Westat Institutional Review Board on May 10, 2021 (project #6632.03.51), with a subsequent amendment approved on November 24, 2021 (amendment ID #3597). This study is a secondary analysis of HINTS 6, a publicly available, deidentified dataset. In accordance with US federal regulations (45 CFR 46) and institutional policies, secondary analyses of

XSL•FO
RenderX

deidentified, publicly available data do not require additional institutional review board review. Additional details about the informed consent process, incentives, and methodology can be found in the HINTS 6 methodology report [44].

Outcome Variable

Individuals were classified as cancer information seekers if they responded "yes" to the question "Have you ever looked for information about cancer from any source?" and those who responded "no" were classified as nonseekers.

Data Processing of Predictors

A total of 87 predictive variables (including demographic factors, alcohol and tobacco use, health care, medical record, chronic diseases, beliefs about cancer, social media, health and nutrition, etc) were included in the initial analysis. Previous simulation and real data analyses revealed that statistical analysis is likely to be biased if the percentage of missing values is more than 10% [45-47]. Therefore, variables with a missing value rate higher than 10% were removed for further analysis. Finally, 74 variables, including the outcome, were left. After excluding individuals with missing data on the outcome, age, gender, or race, the final sample size was 5505.

Demographic characteristics included gender, age group (18-49 years, 50-64 years, 65-74 years, and 75 years or older), race, education, full-time work (yes or no), income, and health insurance (yes or no). Race was recoded as Hispanic, non-Hispanic White, non-Hispanic Black or African American, non-Hispanic Asian, and other. Education had 4 categories (less than high school, some college, bachelor's degree, and postbaccalaureate degree). The 4 categories of annual income were <US $19,999, US $20,000-US $49,999, US $50,000-US $74,999, and ≥US $75,000. Table 1 lists the demographic variables.

Table 1. Prevalence of cancer information seeking across demographic factors.

Variable	Total, n	Seeking, n	Prevalence, %[a]	*P* value[b]
Gender				
Male	2189	949	43.4	<.001
Female	3316	1681	50.7	
Age group				
18-49 years	1935	895	46.3	.01
50-64 years	1607	791	49.2	
65-74 years	1234	623	50.5	
>75 years	729	321	44.0	
Race				
Non-Hispanic White	3175	1738	54.7	<.001
Non-Hispanic African American	878	314	35.8	
Hispanic	984	371	37.7	
Non-Hispanic Asian	286	120	42.0	
Other	182	87	47.8	
Education[c]				
Less than high school	1300	375	28.8	<.001
Some college	1574	721	45.8	
Bachelor's degree	1550	843	54.4	
Postbaccalaureate degree	1070	687	64.2	
Income (US $)				
<19,999	904	271	30.0	<.001
20,000-49,999	1431	569	39.8	
50,000-74,999	946	488	51.6	
>75,000	2224	1302	58.5	
Overall	5505	2630	47.8	

[a]The prevalence is the ratio of seeking and total.

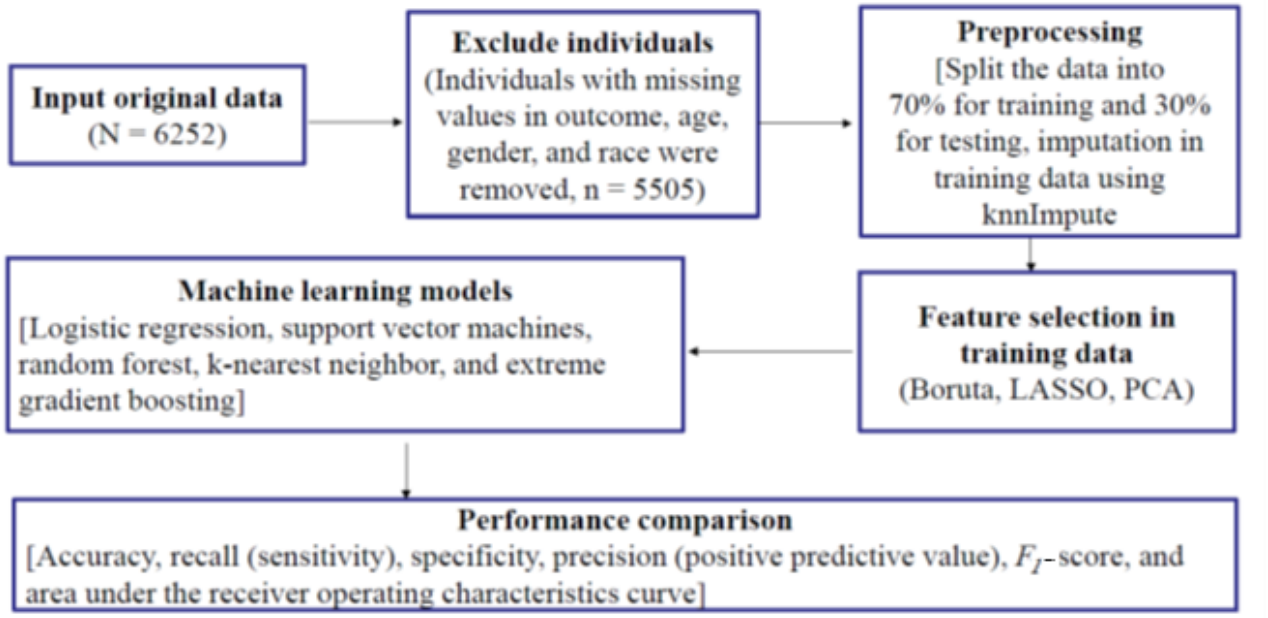
[b]*P* value is based on the chi-square test.

[c]11 participants had missing values in the education variable.

XSL•FO
RenderX

Race was generated for dummy variables. Other predictive variables were binary, ordinal, or continuous. In this dataset, 70% (3854/5505) of the entries were used for training the models, leaving the remaining 30% (1651/5505) for the testing set. The missing values were imputed using the "knnImpute" method in caret based on the training data only [48]. The derived imputation values were subsequently applied to the test data. The full list of 73 predicting variables is listed in Multimedia Appendix 1. Figure 1 shows an overview of the data curation and ML process.

Figure 1. Overview of data curation and machine learning workflow. LASSO: least absolute shrinkage and selection operator; PCA: principal component analysis.



Feature-Selection Methods

Feature selection was performed within the training data. The Boruta algorithm, implemented in R software (version 4.5.2; R Foundation for Statistical Computing) via the "Boruta" package, automatically performs feature selection on a dataset using an RF classifier [49]. The LASSO feature selection was applied using the "glmnet" package in R software [50]. This method regularizes the model by applying a penalty (λ), shrinking the regression coefficients, and reducing some of them to 0. The feature-selection phase occurs after the shrinkage, where non-0 values are selected as model parameters.

The PCA is a dimensionality reduction technique that transforms a set of correlated variables into a smaller set of uncorrelated variables called PCs. Generally, the first PC will be the linear combination of the variables that captures the maximum amount of information in the data and will be correlated with at least some of the observed variables, while the second PC identified accounts for the second-largest amount of variance in the data and is uncorrelated with the first PC, and so on. Eigenvalues indicate the amount of variance explained by each PC. A scree plot was used to visualize eigenvalues. The eigenvalue-one criterion (eigenvalue ≥1) is commonly used to decide how many PCs to retain. Eigenvectors are the weights used to calculate PC scores. The PC score is a linear combination of observed variables weighted by eigenvectors. Assume there are n individuals and k observed predictors included in the PCA, then there are k PCs in total. The equation of the PC score for each individual can be written as:

$$PC_{ij} = b_{1j}x_{i1} + b_{2j}x_{i2} + ... + b_{qj}x_{iq}$$

where

PC_{ij} = the jth PC score for the ith individual, i = 1, 2, ..., n; j = 1, 2, ..., k

b_{kj} = the regression coefficient for observed variable k in $_{the}$ jth PC

X_{ik} = the value of individual i on the observed variable k

q = the number of PCs, q = 1, 2, ..., k

In PCA, the factor loading of a variable represents the correlation between the original variable and a given PC. The loading indicates how much each original variable contributes to a specific PC. Large absolute values of loadings indicate that the corresponding variable has a strong relationship with that particular PC. A factor loading of a variable is considered large if its absolute value exceeds 50%. The PCA was performed using SPSS software (version 31; IBM Corp). For ML analysis, the PC scores for each individual were initially used as predictors. Furthermore, from each of the uncorrelated PCs, 1 variable with the highest loading/correlation coefficient with the PC was chosen for further ML analysis.

ML Methods

A total of 5 ML algorithms were used, including LR, SVM, RF, KNN, and XGBoost. The caret package, incorporating other packages in R, was used for LR, KNN, RF, SVM, and XGBoost [48]. A 10-fold cross-validation approach was applied, and multiple parameters for each algorithm were optimized using a grid search.

XSL•FO
RenderX

For the LR model, the "glmnet" in the caret package was used. In the grid search, we set alpha = 0:1 and lambda = seq(0.001, 1, length = 10).

The SVM algorithm includes linear kernel and radial kernel [51]. In the grid search, we set C = c(0.01, 0.1, 0.2, 0.5, 1, 2)) for the linear kernel; sigma = c(0.05, 0.25, 0.5, 1, 2) and C = c(0.05, 0.25, 0.5, 1, 2)) for the radial kernel.

The RF algorithm randomly selects a subset of variables to construct multiple decision trees (DTs) [52,53]. In the grid search, we set mtry = c(1:15) and ntree = 300, where the mtry parameter refers to the number of variables used in each random tree, while ntree refers to the number of trees that the forest contains. The mtry range (1-15) was chosen to cover and exceed the conventional default ($\sqrt{p} \approx$ 8-9) while limiting overfitting and computational burden.

KNN, a simple ML algorithm based on a clustering algorithm with supervised learning, calculates the average of the numerical target of the k-nearest neighbors [54]. KNN is more suitable for low-dimensional data with a small number of input variables. In the grid search, we set k=1:20.

XGBoost [55] is a supervised ML method for regression and classification tasks similar to the RF classifier. In the grid search, we set the nrounds = c(200,300), max_depth = c(6, 10, 20), colsample_bytree = c(0.5, 1.0), eta = c(0.1, 0.3), gamma= c(0, 0.5), min_child_weight = c(1,2), and subsample = c(0.75, 1.0).

Performance of ML

To evaluate the performance of feature-selection methods, we used several metrics, including accuracy, recall (sensitivity), specificity, precision (positive predictive value), F_1-score, and area under the receiver operating characteristic curve (AUC). The R packages used included "caret," "kernlab," and "ROCR."

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives. The F_1-score is a harmonic mean that combines both recall and precision. The DeLong test was used to compare the statistical differences of the AUC between different models [56].

Statistical Analysis

The categorical variables were presented in their raw values along with the proportions for categorical variables. The chi-square test was used to examine the associations of categorical variables with cancer information seeking across demographic variables. Stepwise LR analysis was performed to estimate the odds ratios (ORs) and their 95% CIs for the associations of potential factors selected in ML analyses with the outcome. All statistical analyses were performed using SPSS software (version 31).

Results

Prevalence of Cancer Information Seeking

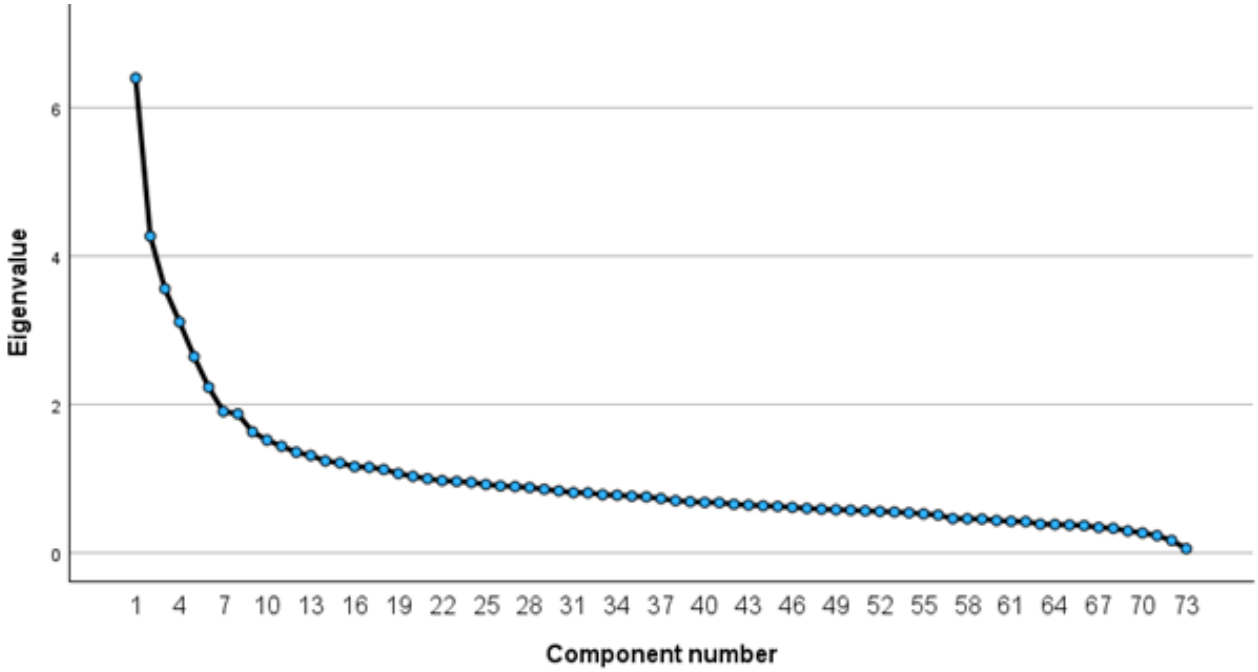
Among the 5505 adult respondents, 2630 were classified as cancer information seekers and 2875 as nonseekers (Table 1). The overall prevalence was 47.8% (2630/5505; 949/2189, 43.4% for men and 1681/2189, 50.7% for women). The prevalence increased with age (895/1935, 46.3%; 791/1607, 49.2%; 623/1234, 50.5% for age groups 18-49, 50-64, and 65-74 years, respectively). The age group of >75 years had a lower prevalence (321/729, 44%). The prevalence was higher in those with higher education and higher income.

Feature Selection

The Boruta algorithm selected 43 variables, and LASSO selected 42 variables related to cancer information seeking (Multimedia Appendix 1). Of these, 27 variables were identified by both methods. The PCA identified 20 uncorrelated PCs with eigenvalues >1 (Figure 2 and Multimedia Appendix 2). From each PC, we selected the variable with the highest loading/correlation (Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Scree plot showing 20 principal components with eigenvalues higher than 1.



ML Performance

The performance statistics are summarized in Multimedia Appendix 4. RF achieved the highest AUC (0.781) and second-highest accuracy (0.714) using 42 LASSO-selected variables, while using all 73 features yielded the highest accuracy (0.717). Using the 27 common variables identified by both Boruta and LASSO, the RF model achieved the highest predictive accuracy (0.711), the same as using SVM with a linear kernel, closely followed by SVM with a radial basis function (RBF) kernel (0.708) and LR (0.708). When using 20 PC scores and selecting 1 variable from each of 20 PCs with the highest loading with the PC, the RF models showed lower accuracy (0.694 and 0.678, respectively). In Figure 3, the mean decrease in accuracy and the mean decrease in Gini metrics from the RF algorithm are shown for the 27 variables identified by LASSO and Boruta. Figure 4 displays the same metrics for the 20 variables derived from each PC. Based on the Gini values in Figure 3, the strongest predictors included cancer-related worry, income, a family history of cancer, interest in cancer screening, meaning and purpose in life, education, more frequent provider visits, having a cancer diagnosis, being offered access to an online medical record, watching health-related videos on social media, and awareness of genetic testing. Plot of mean decrease accuracy and mean decrease Gini values using RF algorithm, XGBoost, gradient boosting machine, LR models, and the 21 factors are illustrated in Figures S1-S4 in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 3. Plot of mean decrease accuracy (left panel) and mean decrease Gini (right panel) values using the random forest algorithm and 27 variables selected by both Boruta and least absolute shrinkage and selection operator.

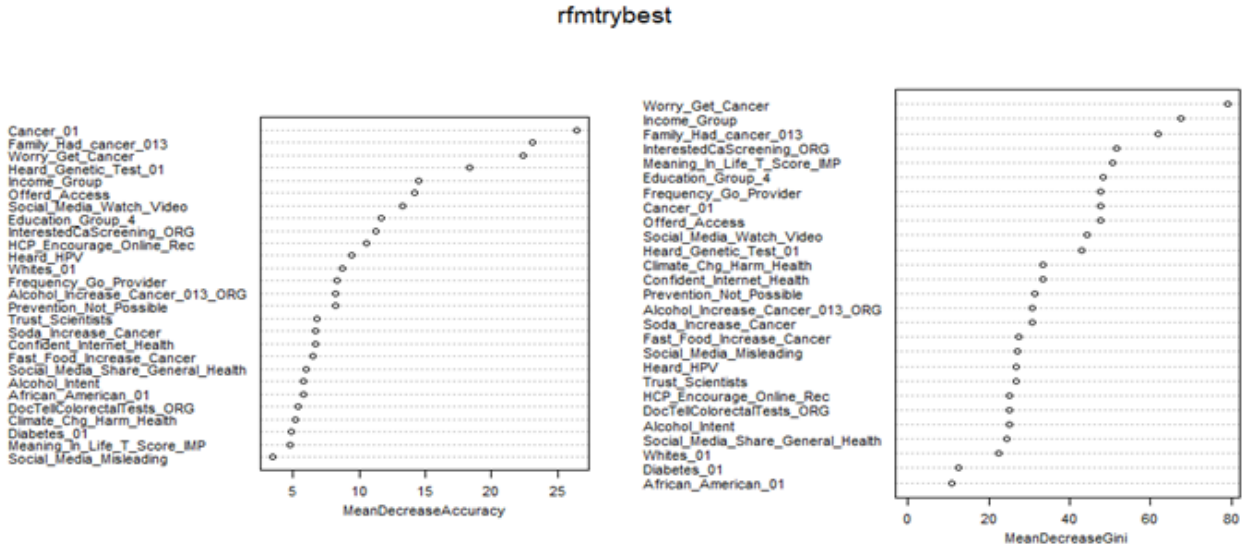
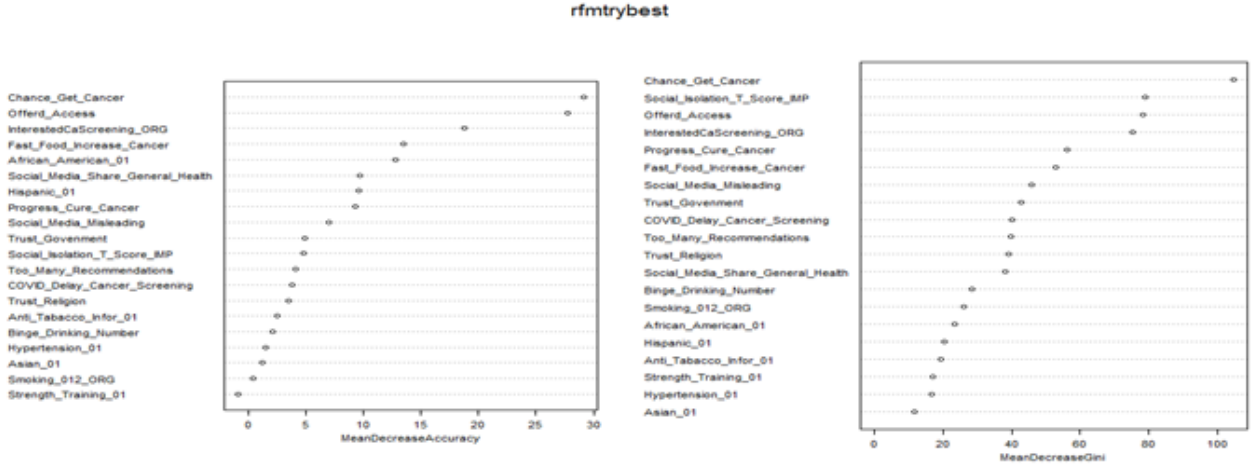


Figure 4. Plot of mean decrease accuracy (left panel) and mean decrease Gini (right panel) values using the random forest algorithm and 20 variables with the highest loading from 20 PCs.



To determine whether the RF model had a significantly higher AUC than other models, the DeLong test was used. The AUC differences between RF and other models, 95% CIs of the AUC difference, and *P* values are illustrated in Table 2. The RF model yielded comparable AUC values with SVM with a linear kernel and LR, except for using PCA-based 20 variables ($P>.05$). In contrast, the AUCs from these RF models were significantly higher than those obtained using SVM with an RBF kernel, KNN, and XGBoost ($P<.05$), except for SVM with an RBF kernel using 20 PC scores and XGBoost using LASSO and PCA-based 20 variables. Figure 5 presents the AUC in the test data for each ML model across the 6 feature-selection methods.

XSL•FO
RenderX

Table 2. Comparison of machine learning models with random forest using the DeLong test of area under the receiver operating characteristic curve (AUC) difference.

Features	SVM[a]_Linear		SVM_RBF[b]		LR[c]		KNN[d]		XGBoost[e]	
	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value
All features	0.0007 (–0.0101 to 0.0115)	.89	0.0509 (0.0359 to 0.0659)	<.001[f]	–0.0012 (–0.0098 to 0.0074)	.78	0.0464 (0.0306 to 0.0622)	<.001[f]	0.0109 (0.0015 to 0.0204)	.02[f]
LASSO[g]	–0.0019 (–0.0113 to 0.0075)	.69	0.0117 (0.0014 to 0.0219)	.03[f]	–0.0044 (–0.0135 to 0.0047)	.34	0.0288 (0.0156 to 0.0420)	<.001[f]	0.0057 (–0.0037 to 0.0152)	.24
Boruta	0.0057 (–0.0037 to 0.0151)	.23	0.0165 (0.0054 to 0.0276)	.004[f]	0.0048 (–0.0045 to 0.0142)	.31	0.0373 (0.0228 to 0.0519)	<.001[f]	0.0090 (0.0002 to 0.0179)	.04[f]
LASSO and Boruta	0.0005 (–0.0074 to 0.0084)	.91	0.0091 (0.0007 to 0.0174)	.04[f]	0.0005 (–0.0078 to 0.0088)	.90	0.0256 (0.0132 to 0.0380)	<.001[f]	0.0182 (0.0089 to 0.0207)	.001[f]
PCA[h] score	–0.0082 (–0.0190 to 0.0026)	.14	0.0032 (–0.0078 to 0.0143)	.57	–0.0089 (–0.0197 to 0.0019)	.11	0.0438 (0.0278 to 0.0599)	<.001[f]	0.0278 (0.0117 to 0.0338)	<.001[f]
PCA (highest loading variable)	0.0189 (0.0084 to 0.0293)	<.001[f]	0.0238 (0.0133 to 0.0342)	<.001[f]	0.0173 (0.0070 to 0.0276)	.001[f]	0.0469 (0.0303 to 0.0635)	<.001[f]	0.0105 (–0.0029 to 0.0238)	.13

[a]SVM: Support vector machine.

[b]RBF: Radial basis function.

[c]LR: Logistic regression.

[d]KNN: k-nearest neighbor.

[e]XGBoost: Extreme gradient boosting.

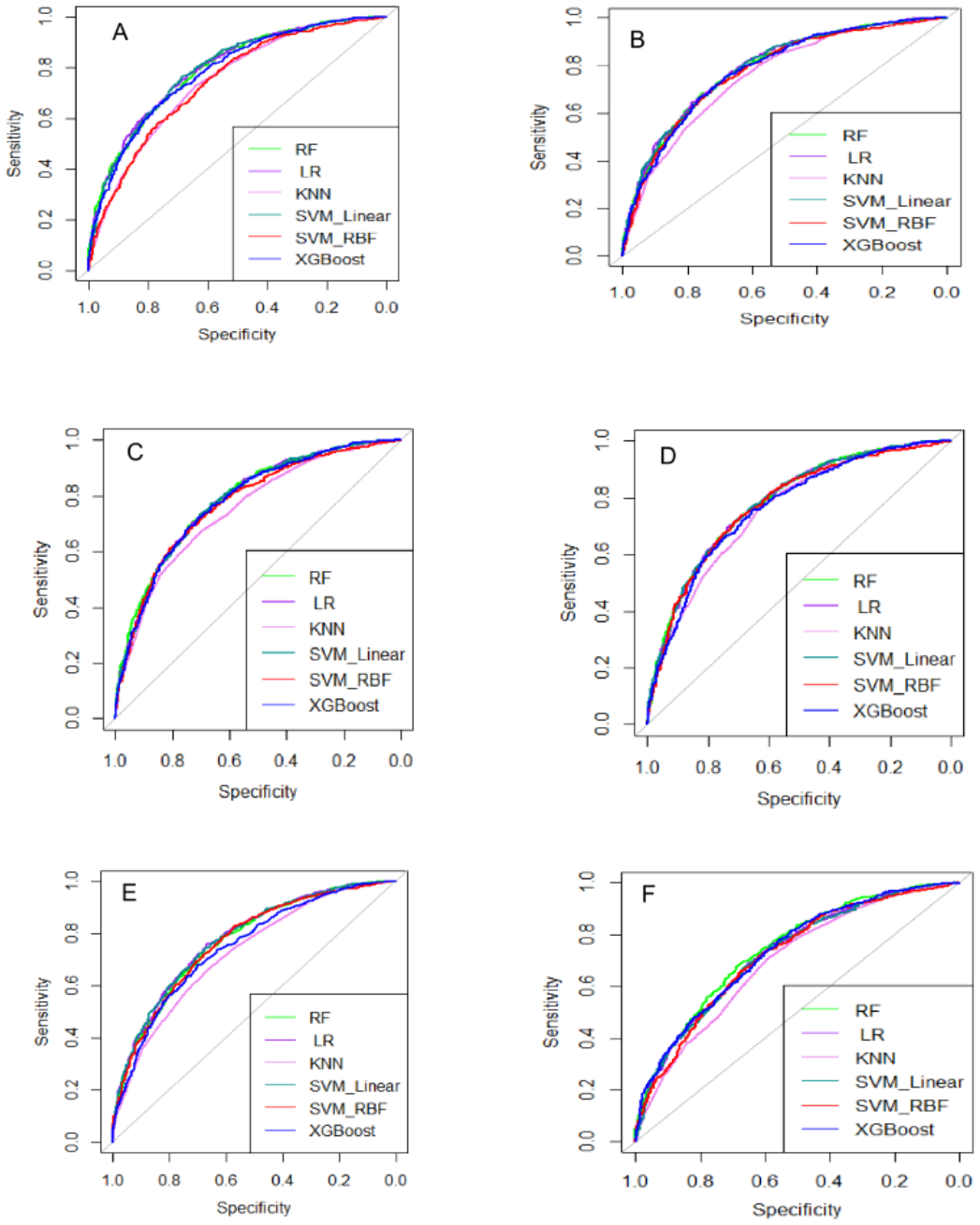
[f]*P*<.05 based on DeLong test.

[g]LASSO: The least absolute shrinkage and selection operator.

[h]PCA: Principal component analysis.

XSL•FO
RenderX

Figure 5. AUC curves in the test data in comparison of RF with other five ML models. (a) All features, (b) LASSO, (c) Boruta, (d) LASSO and Boruta, (e) 20 PCA scores, (f) 20 highest loading variables from 20 PCs.



To evaluate whether differences exist among feature-selection methods, the RF model and the DeLong test were used. The pairwise AUC differences among 6 methods, 95% CIs of the AUC difference, and *P* values are illustrated in Table 3. Boruta, LASSO, and using all 73 features did not show significant differences ($P>.05$), whereas these methods showed higher AUC than methods using PCA score and PCA-based selection of highest loading variables ($P<.05$). The combined method of Boruta and LASSO did not show a significant difference from LASSO-only, whereas it showed lower AUC than methods using all 73 features and Boruta, but higher AUC than methods using PCA score and PCA-based selection of highest loading variables ($P<.05$). Furthermore, using PCA scores had a higher AUC than using 20 PCA-based variables ($P<.05$). AUC in the test data for all ML models across the 6 feature-selection methods are illustrated in Figure 6.

XSL•FO
RenderX

Table 3. Comparison of feature-selection methods using random forest and the DeLong test of area under the receiver operating characteristic curve (AUC) difference.

Features	Boruta		LASSO[a]		Boruta and LASSO		PCA[b] score		PCA (highest loading variable)	
	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value	AUC difference (95% CI)	*P* value
All features	0.0011 (–0.0035 to 0.0057)	.64	0.0054 (–0.00086 to 0.0116)	.09	0.0080 (0.0013 to 0.0146)	.02[c]	0.0230 (0.0109 to 0.0351)	<.001[c]	0.0492 (0.0339 to 0.0644)	<.001[c]
Boruta	—[d]	—	0.0043 (–0.0028 to 0.0114)	.23	0.0069 (0.0012 to 0.0125)	.02[c]	0.0219 (0.0093 to 0.0345)	<.001[c]	0.0481 (0.0315 to 0.0646)	<.001[c]
LASSO	—	—	—	—	0.0026 (–0.0028 to 0.0079)	.35	0.0176 (0.0050 to 0.0303)	.006[c]	0.0438 (0.0274 to 0.0601)	<.001[c]
LASSO and Boruta	—	—	—	—	—	—	0.0150 (0.0018 to 0.0283)	.03[c]	0.0412 (0.0239 to 0.0583)	<.001[c]
PCA score	—	—	—	—	—	—	—	—	0.0262 (0.0081 to 0.0442)	.005[c]

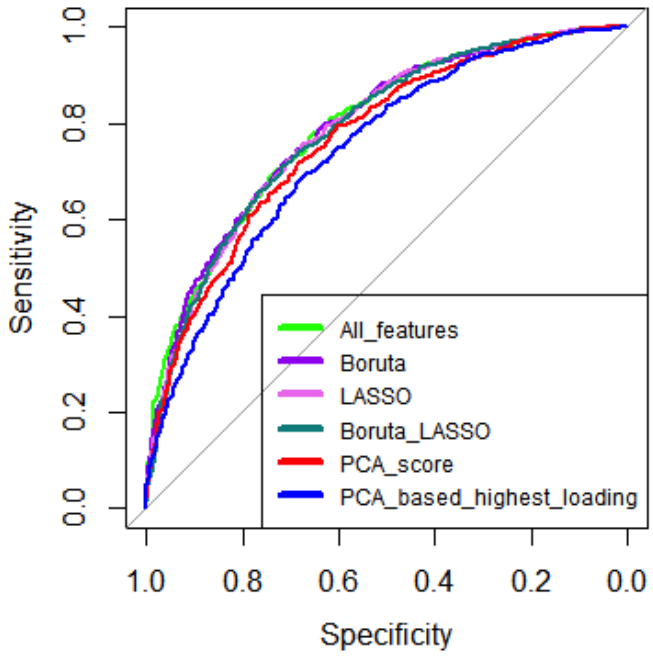
[a]LASSO: least absolute shrinkage and selection operator.

[b]PCA: principal component analysis.

[c]*P*<.05 based on DeLong test.

[d]Not applicable.

Figure 6. Area under the receiver operating characteristic curve in the test data using the random forest model for feature-selection methods. LASSO: least absolute shrinkage and selection operator; PCA: principal component analysis.



Stepwise LR Analysis

For the stepwise LR, we used the single training data, where both the Boruta algorithm and LASSO selected 27 common variables for cancer information seeking. Stepwise LR of the 27 variables identified by both Boruta and LASSO, and further confirmed that 19 variables were associated with cancer information seeking (Table 4). Based on ORs, the top predictors included having a cancer diagnosis (OR 1.46, 95% CI 1.35-1.59), greater worry about developing cancer (OR 1.41, 95% CI 1.31-1.53), a family history of cancer (OR 1.35, 95% CI 1.25-1.47), White (OR 1.25, 95% CI 1.12-1.32), having a higher household income (OR 1.24, 95% CI 1.13-1.35), having heard of genetic testing (OR 1.24, 95% CI 1.14-1.35), watching health-related videos on social media (OR 1.24, 95% CI 1.15-1.34), interest in cancer screening (OR 1.20, 95% CI

XSL•FO
RenderX

1.11-1.29), being offered access to an online medical record (OR 1.17, 95% CI 1.08-1.27), and knowledge of human papillomavirus (HPV; OR 1.17, 95% CI 1.08-1.27).

Table 4. Stepwise logistic regression analyses of 27 variables.

Variable	aOR[a] (95% CI)	*P* value
Race (White; reference=other races)	1.25 (1.12-1.32)	<.001
Education (high school or above; reference=less than high school)	1.16 (1.06-1.26)	.001
Income (1-4; 4=>US $75,000)	1.24 (1.13-1.35)	<.001
Alcohol_increase_cancer (1-4; 4=a lot)	1.16 (1.08-1.26)	<.001
Confident_internet_health resource (1-5; 5=very confident)	1.12 (1.03-1.21)	.007
Diabetes (yes; reference=no)	0.91 (0.84-0.99)	.02
Cancer (yes; reference=no)	1.46 (1.35-1.59)	<.001
Family had cancer (yes; reference=no)	1.35 (1.25-1.47)	<.001
Heard_genetic_test (yes; reference=no)	1.24 (1.14-1.35)	<.001
Heard_HPV (yes; reference=no)	1.17 (1.08-1.27)	<.001
Doctor told colorectal cancer tests (yes; reference=never)	1.12 (1.04-1.21)	.004
Interested in cancer screening (0-4; 4=very)	1.20 (1.11-1.29)	<.001
Prevention of cancer not possible (1-4; 4=strongly agree)	1.16 (1.07-1.25)	<.001
Worry_get_cancer (1-5; 5=extremely)	1.41 (1.31-1.53)	<.001
Frequency_go_provider last year (1-6; 6=10 or more times)	1.12 (1.04-1.21)	.004
Offer_access online medical record (yes; reference=no)	1.17 (1.08-1.27)	<.001
Meaning in life (1-5; 5=a lot)	1.09 (1.01-1.18)	.03
Social_media_misleading health information (1-5; 5=a lot)	1.09 (1.01-1.17)	.03
Social_media_watch_video (1-5; 5=almost everyday)	1.24 (1.15-1.34)	<.001

[a]aOR: adjusted odds ratio

Discussion

Principal Findings

This study compared 3 feature-selection approaches, including Boruta, LASSO, and PCA, and their variations across 5 ML models to predict cancer information–seeking behavior. Both Boruta and LASSO identified the same set of 27 variables, whereas PCA produced 20 uncorrelated PCs. Based on AUC, the RF model emerged as the best-performing algorithm, yielding comparable AUC values when using Boruta-selected features, LASSO-selected features, or no feature selection at all. In addition, a stepwise LR using the 27 variables identified by both Boruta and LASSO confirmed that 19 variables were significantly associated with cancer information seeking (P<.05). The top predictors included having a cancer diagnosis, prior awareness of genetic testing, higher household income, a family history of cancer, being offered access to an online medical record, knowledge of HPV, worry about developing cancer, watching health-related videos on social media, and interest in cancer screening.

Comparison to Prior Work

Feature selection is a critical step in ML to reduce the dimensionality while retaining the most relevant features without excessive information loss [28-30]. Previous studies have used several feature-selection methods, including LASSO, Boruta, and RF [32-35]. In this study, we applied Boruta and LASSO methods and selected 27 overlapping variables (Multimedia Appendix 1 and Figure 3). The Boruta algorithm is a feature-selection method based on RF and considers important variables, but does not account for collinearity while selecting variables, whereas LASSO regression is useful when multicollinearity exists in the model [49,57,58]. The LASSO can be used as a variable selection method because numerous β coefficients that are not strongly associated with the outcome are decreased to 0, which is equivalent to removing those variables from the model. However, the disadvantage of LASSO is that it assumes a more restrictive set of assumptions than RF. For example, the LASSO is a linear model, so anything that matters for your outcome that is not linear in the parameters under estimation is at risk of getting eliminated. Most variables selected by LASSO were also selected by Boruta, which independently confirmed the association between those variables and cancer information–seeking behavior. Previous studies have revealed that combining these methods can lead to more robust feature selection, potentially improving model performance and interpretability [59,60]. In addition, PCA has been used to reduce the dimensions, but it is not a feature-selection method because all variables remain in each PC [31,36-38]; however, PCA can help identify the most important features by examining their contributions (loadings) to the PCs. This study initially conducted data mining using PCA and selected important

XSL•FO
RenderX

uncorrelated PCs with eigenvalues >1.0, and for ML analysis, the PC scores were used as predictors, whereas all variables remained in each PC. Notably, based on each of the uncorrelated PCs, we performed feature selection by choosing 1 variable with the highest loading/correlation coefficient with the PC for ML analysis. In this study, PCA identified 20 uncorrelated PCs, while 20 features were selected from these PCs for the development of ML models. Boruta, LASSO, and the use of all 73 features did not show significant differences. However, these methods showed higher AUC than methods using PCA scores and PCA-based selection of the highest loading variables. The combined method of Boruta-LASSO did not yield a statistically significant improvement in model performance compared with LASSO alone. Moreover, the Boruta-LASSO method demonstrated lower discrimination than models using all 73 features or Boruta alone, but achieved higher AUC than models based on PCA-derived features or PCA-based selection of variables with the highest loadings. In addition, models using PCA scores directly showed better performance than models constructed using a reduced set of 20 variables selected based on PCA loadings.

ML methods can accommodate a large number of predictors and capture complex relationships among variables, making them well-suited for predicting health information–seeking behaviors. For example, 3 ML algorithms such as LR, SVM, and RF were applied in predicting the information-seeking behavior of clinicians using an electronic medical record system [61]. Furthermore, LASSO was used to predict health information–seeking behaviors [62], and elastic net and LASSO models were used in predicting internet health seeking [63]. Another study used 4 algorithms (ie, RF, SVM, Bayes generalized linear model, gradient boosting, and an ensemble of the individual methods) to identify search terms and patterns that correlate with changes in obesity and overweight prevalence across Africa [64]. However, limited research has applied ML approaches to systematically identify key factors associated with cancer information–seeking behaviors. This study compared 5 ML tools using 10-fold cross-validation and tested multiple parameters for each algorithm using a grid search for optimal performance. Using 42 LASSO-selected variables, RF achieved the highest AUC (0.781) and second-highest accuracy (0.714). Using the 27 common variables identified by both Boruta and LASSO, the RF model achieved the highest predictive accuracy (0.711), the same as using SVM with linear kernel, closely followed by SVM with RBF kernel (0.708) and LR (0.708). Previous studies have shown that RF is one of the best ML tools. For example, 1 previous study evaluated 10 ML classifiers, including LR, linear discriminant analysis, naive Bayes, KNN, SVM with RBF kernel, DT, RF, XGBoost, AdaBoost, and artificial neural network (ANN), and found that the RF model and the SVM model showed the best performance [65]. Another study developed 7 ML models (LR, KNN, SVM, DT, RF, XGBoost, and ANN) and found that RF and ANN models with the same AUC outperformed other models [66]. Another study compared LR, KNN, gradient boosting, XGBoost, RF, multilayer perceptron, and SVM for diagnosing breast cancer and found that RF achieved the maximum accuracy of 90.68% [67]. However, 1 recent study compared 6 ML algorithms (XGBoost, LR, SVM, RF, KNN, and DT) in developing prognostic models for patients with alpha-fetoprotein–positive hepatocellular carcinoma, and the XGBoost model performed the best, and RF was the second best [68]. Another recent study compared 5 ML models (SVM, XGBoost, Gaussian naïve Bayes, adaptive boosting, and RF) and found that XGBoost and RF achieved superior predictive performance, as evidenced by higher AUCs [69], whereas another study did not find differences among LR, SVM, and RF [70]. As can be seen, there are studies in the literature on the use of ML algorithms in the diagnosis of different types of cancers, as well as other diseases and conditions. The comparisons of ML tools may show heterogeneity.

Furthermore, although LASSO was our primary feature-selection method due to its advantages in handling multicollinearity, high-dimensional data, and overfitting, we used stepwise regression as a secondary, confirmatory analysis for several reasons. First, LASSO selects variables by shrinking coefficients through penalization, but it does not provide traditional inferential statistics such as SEs, *P* values, or likelihood-based model comparison, which are still expected in many epidemiologic and clinical research settings. Stepwise regression allowed us to evaluate whether LASSO-identified predictors remained significant under a conventional regression framework. Second, stepwise selection served as a sensitivity analysis, enabling assessment of the robustness and stability of the LASSO-identified variable set. Recent methodological papers recommend combining penalized regression with stepwise or likelihood-ratio–based checks when the goal includes both prediction and interpretation [71,72]. Accordingly, stepwise regression was used not as the primary modeling strategy, but as a supplementary sensitivity analysis to evaluate the robustness and interpretability of the variables selected by LASSO.

Patient with cancer often seek information regarding their diagnosis, treatment options, treatment costs, potential side effects, and the implications for daily life and survival. While physicians remain a primary source of such information, patients frequently turn to additional resources such as the internet and books to supplement their understanding [12]. A previous study has identified that various factors influence cancer information–seeking behaviors, including gender, education levels, income, and cancer type [3]. In this study, ML techniques were used to identify key predictors of cancer information–seeking behavior. Furthermore, 18 out of 27 selected variables from LASSO and Boruta were confirmed to be significantly associated with cancer information seeking by a stepwise LR model (Table 4). Consistent with prior research, variables including higher educational attainment, race/ethnicity, personal cancer history, family history of cancer, cancer-related beliefs, knowledge of genetic testing, and awareness of HPV were included in the analysis [3,5,8-15,18-20,22]. Furthermore, this study expanded the existing literature by identifying several additional predictors of cancer information–seeking behavior. These included beliefs that certain nutritional factors (such as red meat and alcohol consumption) and climate change increase cancer risk, engagement in social media activities, having access to online medical records, more frequent provider visits, and interest in cancer screening.

XSL•FO
RenderX

Social media has become a main platform and important resource for people to obtain and exchange health-related information and advice [73,74], and has become a new channel for promoting cancer prevention [75]. For example, 1 study found that young adults with cancer used social media to connect with cancer peers for support [76]. Another study showed that social media could enable the seeking and sharing of breast cancer–related information, and enhance patient education, communication, engagement, and empowerment [77]. Social media platforms may increase access to health information and decision aids [78]. However, while social media can make health information more accessible, the use of social media for health information seeking can also create the risk of harm through exposure to misinformation. Because misinformation perceptions can affect attitudes and behaviors, a better understanding of the public's perceptions of health misinformation on social media and their ability to detect it, as well as possible subgroup differences in such perceptions, is needed [79]. However, misinformation and disinformation on social media have become widespread, which can lead to a lack of trust in health information sources and, in turn, lead to negative health outcomes [80]. In this study, 3 variables related to social media use were selected by both Boruta and LASSO (Multimedia Appendix 1), including social_media_share_general_health (1-5, 5 = almost everyday), social_media_watch_video (1-5, 5 = almost everyday), and social_media_misleading (1-5, 5= a lot). Stepwise LR further confirmed these variables, including social_media_watch_video and social_media_misleading (Table 4). These findings highlight that people involved in social media activities have increased odds of seeking cancer information. Furthermore, it has been shown that most social media users perceive some (46%) or a lot (36%) of false or misleading health information on social media using HINTS 6 data [80]. This study added that most social media users with a high prevalence of false and misleading health information on social media are positively seeking cancer information.

This study further added that beliefs about alcohol use causing cancer were associated with increased odds of cancer information seeking. Alcohol consumption increases the risk of several types of cancer, including liver, esophageal, colorectal, and breast cancer; however, public awareness of the association between alcohol use and cancer remains low and varies by type of alcoholic beverage [81-83]. For example, using the HINTS (2020) data, 1 study found that awareness of the alcohol-cancer link was highest for liquor (31.2%), followed by beer (24.9%) and wine (20.3%). More US adults believed wine (10.3%) decreased cancer risk, compared with beer (2.2%) and liquor (1.7%). Most US adults (>50%) reported not knowing how these beverages affected cancer risk [82]. Another study using the HINTS (2020) data found that 34% of those reporting current alcohol consumption believed that drinking wine decreases or has no effect on cancer risk, compared with 20.8% of those reporting no alcohol consumption [84]. A recent study using several HINTS cycle datasets did not find significant differences in diet-related cancer risk awareness and behaviors between cancer survivors and those without a history of cancer [85]. Among the European Union general population, awareness of the link between alcohol and breast cancer ranged between 10% and 20%, head and neck cancer (15%-25%), colorectal and esophagus cancer (15%-45%), and liver cancer (40%). Awareness was higher among young people and specialized health professions and lower among women (the latter specifically for breast cancer) [83].

Practical Implications

This study identified a set of variables associated with cancer information seeking. We address 2 implications. First, social media users had higher odds of seeking cancer information. Previous studies found that web-based infotainment videos are an effective approach in increasing public understanding about science and health care among web-based health information seekers and are a useful and effective approach in relaying complex health information, motivating interested viewers to seek additional health information, and driving public audiences to credible and reliable sources of information [86]. Furthermore, social media plays a significant role in how people seek and share information about cancer, both for themselves and for others. While social media can be a valuable tool for connecting with support networks and accessing information, it also presents challenges related to misinformation and the potential for information overload. Despite many perceived benefits of social media use among oncology stakeholders, misinformation poses a critical threat to the value of social media for seeking and sharing cancer-related information [87]. It has been suggested that it is necessary for all key stakeholders—including patients and the public, health care providers, researchers, technology companies, and governmental organizations to proactively address the problem of online health misinformation [87].

Furthermore, awareness and beliefs about alcohol and red meat were significantly associated with cancer information seeking. It has been suggested that knowledge about diet-related cancer risks is essential for behavior change; therefore, increasing public knowledge and risk beliefs about the link between alcohol and cancer, particularly among those who consume alcohol, may contribute to declines in the burden of alcohol-related disease in the United States [84]. Further research is warranted to understand these factors better and to develop effective strategies to improve dietary behaviors among cancer survivors [85].

Strengths and Limitations

This study has several notable strengths. First, this study used the most recent HINTS 6 data to examine the prevalence of cancer information seeking. The HINTS data provides unparalleled insights into health information seeking behaviors, social media use, beliefs about alcohol and cancer, etc. Second, we performed feature selection using 2 widely used methods, LASSO and Boruta, to identify common variables across both methods. Third, we inferred PC scores (factor scores) and then used weighted LR analyses to estimate the associations of potential factors and PC scores with colorectal cancer screening. Fourth, we compared 5 ML algorithms and found that the RF model demonstrated outstanding classification performance in predicting cancer information seeking. Moreover, we used stepwise LR analysis to confirm the results from ML techniques.

XSL•FO
RenderX

Despite these strengths, our analysis has some limitations. First, because the HINTS data are cross-sectional, we could only identify correlations rather than causal relationships. Future research could address this limitation by applying a rigorous quasi-experimental method to longitudinal datasets. Second, since information from participants was self-reported, our study may be subject to recall bias as well as social desirability bias. Third, a notable limitation is the relatively low response rate of the HINTS 6. Low response rates raise concerns about nonresponse bias, particularly if nonrespondents differ meaningfully from respondents on key variables such as alcohol use, health information–seeking, or cancer awareness. Although HINTS incorporates sampling weights, replicate weights, and nonresponse adjustments to enhance population representativeness, these statistical corrections cannot fully eliminate bias arising from selective participation. Therefore, caution is warranted when interpreting the generalizability of our findings to all US adults. Fourth, the data were collected in 2022, and the COVID-19 pandemic may have influenced both data collection and results. In addition, this study used only the binary outcome (seeker vs nonseeker). For the seekers, there are still 4 questions in the data such as how much do you agree or disagree: it took a lot of effort to get the information you needed, you felt frustrated during your search for the information, you were concerned about the quality of the information, and the information you found was hard to understand. We did not include these subquestions. Furthermore, because the outcome captured lifetime ("ever") cancer information–seeking, whereas several predictors reflected respondents' current status in 2022, temporal misalignment may exist. Information seeking may have occurred prior to the measurement of current characteristics; therefore, associations should be interpreted as correlational rather than causal.

Conclusion

This study provided the updated prevalence of cancer information seeking among US adults. Furthermore, we performed feature selection and compared 5 ML algorithms for classifying cancer information seeking, and identified that RF was the best performer with the highest AUC. Moreover, PCA proved useful for data mining to reduce the indicators in complex survey data and aid in feature selection. In addition, based on the stepwise regression model, 19 out of 27 selected variables were significantly associated with cancer information seeking. We identified a set of predictive variables for cancer information seeking, such as having cancer, having a family history of cancer, worrying about getting cancer, knowledge of genetic tests and HPV, being offered access to health records, higher income, often watching videos on social media, believing that alcohol consumption increases cancer risk, and frequency of visiting providers. Our findings may benefit researchers, policymakers, and health care providers by increasing public awareness and supporting targeted education on cancer information seeking.

Acknowledgments

The authors would like to thank the National Cancer Institute for providing the data from the 2022 Health Information National Trends Survey.

Data Availabilty

The data that support the findings of this study are openly available at the National Cancer Institute [88].

Funding

No funding source is given for this paper.

Authors' Contributions

YL contributed to writing—original draft, software, methodology, formal analysis, and data curation. KW contributed to writing—original draft, review and editing, software, methodology, formal analysis, and conceptualization.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Feature selection based on LASSO and Boruta. LASSO: least absolute shrinkage and selection operator.
[XLSX File (Microsoft Excel File), 16 KB - medinform_v14i1e75862_app1.xlsx]

Multimedia Appendix 2
Eigenvalues based on PCA. PCA: principal component analysis.
[XLSX File (Microsoft Excel File), 17 KB - medinform_v14i1e75862_app2.xlsx]

Multimedia Appendix 3
Rotated component/loading of variables for 20 factors.
[XLSX File (Microsoft Excel File), 21 KB - medinform_v14i1e75862_app3.xlsx]

XSL•FO
RenderX

Multimedia Appendix 4
Machine learning and comparison of performance.
[DOCX File , 25 KB - medinform_v14i1e75862_app4.docx]

Multimedia Appendix 5
Additional figures.
[DOCX File , 177 KB - medinform_v14i1e75862_app5.docx]

References

1. Weaver JB, Mays D, Weaver SS, Hopkins GL, Eroğlu D, Bernhardt JM. Health information–seeking behaviors, health indicators, and health risks. Am J Public Health 2010;100(8):1520-1525. [doi: 10.2105/ajph.2009.180521]
2. Cancer statistics. National Cancer Institute. 2024. URL: https://www.cancer.gov/about-cancer/understanding/statistics [accessed 2025-03-20]
3. Roach AR, Lykins ELB, Gochett CG, Brechting EH, Graue LO, Andrykowski MA. Differences in cancer information-seeking behavior, preferences, and awareness between cancer survivors and healthy controls: a national, population-based survey. J Cancer Educ 2009;24(1):73-79 [FREE Full text] [doi: 10.1080/08858190802664784] [Medline: 19259869]
4. Vanderpool RC, Huang GC, Mollica M, Gutierrez AI, Maynard CD. Cancer information-seeking in an age of COVID-19: findings from the National Cancer Institute's Cancer Information Service. Health Commun 2021;36(1):89-97. [doi: 10.1080/10410236.2020.1847449] [Medline: 33225770]
5. Kelly B, Hornik R, Romantan A, Schwartz JS, Armstrong K, DeMichele A, et al. Cancer information scanning and seeking in the general population. J Health Commun 2010;15(7):734-753 [FREE Full text] [doi: 10.1080/10810730.2010.514029] [Medline: 21104503]
6. Shim M, Kelly B, Hornik R. Cancer information scanning and seeking behavior is associated with knowledge, lifestyle choices, and screening. J Health Commun 2006;11 Suppl 1:157-172. [doi: 10.1080/10810730600637475] [Medline: 16641081]
7. Niederdeppe J, Hornik RC, Kelly BJ, Frosch DL, Romantan A, Stevens RS, et al. Examining the dimensions of cancer-related information seeking and scanning behavior. Health Commun 2007;22(2):153-167. [doi: 10.1080/10410230701454189] [Medline: 17668995]
8. Mayer DK, Terrin NC, Kreps GL, Menon U, McCance K, Parsons SK, et al. Cancer survivors information seeking behaviors: a comparison of survivors who do and do not seek information about cancer. Patient Educ Couns 2007;65(3):342-350 [FREE Full text] [doi: 10.1016/j.pec.2006.08.015] [Medline: 17029864]
9. Rutten LJF, Squiers L, Hesse B. Cancer-related information seeking: hints from the 2003 Health Information National Trends Survey (HINTS). J Health Commun 2006;11 Suppl 1:147-156. [doi: 10.1080/10810730600637574] [Medline: 16641080]
10. Sacca L, Maroun V, Khoury M. Predictors of high trust and the role of confidence levels in seeking cancer-related information. Inform Health Soc Care 2022;47(1):53-61. [doi: 10.1080/17538157.2021.1925676] [Medline: 34014145]
11. Mahmood A, Kedia S, Ogunsanmi DO, Kabir U, Entwistle C. Patient-centered communication and cancer information-seeking experiences among cancer survivors: a population-based study in the United States. Patient Educ Couns 2025;135:108710. [doi: 10.1016/j.pec.2025.108710] [Medline: 40010060]
12. Nagler RH, Gray SW, Romantan A, Kelly BJ, DeMichele A, Armstrong K, et al. Differences in information seeking among breast, prostate, and colorectal cancer patients: results from a population-based survey. Patient Educ Couns 2010;81 Suppl:S54-S62 [FREE Full text] [doi: 10.1016/j.pec.2010.09.010] [Medline: 20934297]
13. Zhang D, Hu H, Shi Z, Li B. Perceived needs versus predisposing/enabling characteristics in relation to internet cancer information seeking among the US and Chinese public: comparative survey research. J Med Internet Res 2021;23(1):e24733 [FREE Full text] [doi: 10.2196/24733] [Medline: 33427668]
14. Cho B, Lee S, Pan Y, Sharma M, Holland K. Association of cancer information seeking behavior with cigarette smoking and e-cigarette use among U.S. adults by education attainment level: a multi-year cross-sectional analysis from a nationally representative sample in 2017-2020. Prev Med 2023;172:107550. [doi: 10.1016/j.ypmed.2023.107550] [Medline: 37210044]
15. Kaphingst KA, Lachance CR, Condit CM. Beliefs about heritability of cancer and health information seeking and preventive behaviors. J Cancer Educ 2009;24(4):351-356 [FREE Full text] [doi: 10.1080/08858190902876304] [Medline: 19838898]
16. Zhao X, Cai X. The role of risk, efficacy, and anxiety in smokers' cancer information seeking. Health Commun 2009;24(3):259-269. [doi: 10.1080/10410230902805932] [Medline: 19415558]
17. Arora NK, Hesse BW, Rimer BK, Viswanath K, Clayman ML, Croyle RT. Frustrated and confused: the American public rates its cancer-related information-seeking experiences. J Gen Intern Med 2008;23(3):223-228 [FREE Full text] [doi: 10.1007/s11606-007-0406-y] [Medline: 17922166]
18. Agurs-Collins T, Ferrer R, Ottenbacher A, Waters EA, O'Connell ME, Hamilton JG. Public awareness of direct-to-consumer genetic tests: findings from the 2013 U.S. Health Information National Trends Survey. J Cancer Educ 2015;30(4):799-807 [FREE Full text] [doi: 10.1007/s13187-014-0784-x] [Medline: 25600375]

XSL•FO
RenderX

19. Lu L, Liu J, Yuan YC. Cultural differences in cancer information acquisition: cancer risk perceptions, fatalistic beliefs, and worry as predictors of cancer information seeking and avoidance in the U.S. and China. Health Commun 2022;37(11):1442-1451. [doi: 10.1080/10410236.2021.1901422] [Medline: 33752516]
20. Van Stee SK, Yang Q. Online cancer information seeking: applying and extending the comprehensive model of information seeking. Health Commun 2018;33(12):1583-1592. [doi: 10.1080/10410236.2017.1384350] [Medline: 29083231]
21. Kye SY, Yun EH, Park K. Factors related to cancer information scanning and seeking behavior among high school students in Korea. Asian Pac J Cancer Prev 2012;13(4):1439-1445 [FREE Full text] [doi: 10.7314/apjcp.2012.13.4.1439] [Medline: 22799345]
22. Nelissen S, Beullens K, Lemal M, Van den Bulck J. Fear of cancer is associated with cancer information seeking, scanning and avoiding: a cross-sectional study among cancer diagnosed and non-diagnosed individuals. Health Info Libr J 2015;32(2):107-119 [FREE Full text] [doi: 10.1111/hir.12100] [Medline: 25809822]
23. Hurtado M, Siefkas A, Attwood MM, Iqbal Z, Hoffman J. Machine learning applications and advancements in alcohol use disorder: a systematic review. Addict Med 2022:12. [doi: 10.1101/2022.06.06.22276057]
24. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, et al. What is machine learning, artificial neural networks and deep learning?-examples of practical applications in medicine. Diagnostics (Basel) 2023;13(15):2582 [FREE Full text] [doi: 10.3390/diagnostics13152582] [Medline: 37568945]
25. Theodosiou AA, Read RC. Artificial intelligence, machine learning and deep learning: potential resources for the infection clinician. J Infect 2023;87(4):287-294 [FREE Full text] [doi: 10.1016/j.jinf.2023.07.006] [Medline: 37468046]
26. Kumari J, Kumar E, Kumar D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. Arch Comput Methods Eng 2023;30(6):3673-3701 [FREE Full text] [doi: 10.1007/s11831-023-09915-y] [Medline: 37359744]
27. Shastry KA, Vijayakumar V. Deep learning techniques for the effective prediction of Alzheimer's disease: a comprehensive review. Healthcare (Basel) 2022;10(10):1842 [FREE Full text] [doi: 10.3390/healthcare10101842] [Medline: 36292289]
28. Chen X, Kopsaftopoulos F, Wu Q, Ren H, Chang F. Flight state identification of a self-sensing wing via an improved feature selection method and machine learning approaches. Sensors (Basel) 2018;18(5):1379 [FREE Full text] [doi: 10.3390/s18051379] [Medline: 29710832]
29. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PLoS One 2019;14(6):e0218760 [FREE Full text] [doi: 10.1371/journal.pone.0218760] [Medline: 31242238]
30. Cömert Z, Şengür A, Budak Ü, Kocamaz AF. Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models. Health Inf Sci Syst 2019;7(1):17 [FREE Full text] [doi: 10.1007/s13755-019-0079-z] [Medline: 31435480]
31. Raihan-Al-Masud M, Mondal MRH. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. PLoS One 2020;15(2):e0228422 [FREE Full text] [doi: 10.1371/journal.pone.0228422] [Medline: 32027680]
32. Atuegwu NC, Litt MD, Krishnan-Sarin S, Laubenbacher RC, Perez MF, Mortensen EM. E-cigarette use in young adult never cigarette smokers with disabilities: results from the behavioral risk factor surveillance system survey. Int J Environ Res Public Health 2021;18(10):5476 [FREE Full text] [doi: 10.3390/ijerph18105476] [Medline: 34065407]
33. Atuegwu NC, Oncken C, Laubenbacher RC, Perez MF, Mortensen EM. Factors associated with e-cigarette use in U.S. young adult never smokers of conventional cigarettes: a machine learning approach. Int J Environ Res Public Health 2020;17(19):7271 [FREE Full text] [doi: 10.3390/ijerph17197271] [Medline: 33027932]
34. Choi J, Jung HT, Ferrell A, Woo S, Haddad L. Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (hookah) users. J Clin Med 2021;10(5):972 [FREE Full text] [doi: 10.3390/jcm10050972] [Medline: 33801175]
35. Fu R, Shi J, Chaiton M, Leventhal AM, Unger JB, Barrington-Trimis JL. A machine learning approach to identify predictors of frequent vaping and vulnerable Californian youth subgroups. Nicotine Tob Res 2022;24(7):1028-1036 [FREE Full text] [doi: 10.1093/ntr/ntab257] [Medline: 34888698]
36. Castillo-Barnes D, Su L, Ramírez J, Salas-Gonzalez D, Martinez-Murcia FJ, Illan IA, et al. Autosomal dominantly inherited Alzheimer disease: analysis of genetic subgroups by machine learning. Information Fusion 2020;58:153-167. [doi: 10.1016/j.inffus.2020.01.001]
37. Chen C, Qin Y, Cheng J, Gao F, Zhou X. Texture analysis of fat-suppressed T2-weighted magnetic resonance imaging and use of machine learning to discriminate nasal and paranasal sinus small round malignant cell tumors. Front Oncol 2021;11:701289 [FREE Full text] [doi: 10.3389/fonc.2021.701289] [Medline: 34966664]
38. Grazioli G, Martin RW, Butts CT. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. Front Mol Biosci 2019;6:42 [FREE Full text] [doi: 10.3389/fmolb.2019.00042] [Medline: 31245383]
39. Rex DK. Colonoscopy remains an important option for primary screening for colorectal cancer. Dig Dis Sci 2025;70(5):1595-1605. [doi: 10.1007/s10620-024-08760-8] [Medline: 39666212]

XSL•FO
RenderX

40. Zhang X, Ren H, Gao L, Shia BC, Chen MC, Ye L, et al. Identifying the predictors of severe psychological distress by auto-machine learning methods. Inform Med Unlocked 2023;39:101258 [FREE Full text] [doi: 10.1016/j.imu.2023.101258] [Medline: 37152204]
41. Chen Y, Liu X, Gao L, Zhu M, Shia B, Chen M, et al. Using the H2O automatic machine learning algorithms to identify predictors of web-based medical record nonuse among patients in a data-rich environment: mixed methods study. JMIR Med Inform 2023;11:e41576 [FREE Full text] [doi: 10.2196/41576] [Medline: 37335616]
42. Huang X, Dai Z, Wang K, Luo X. Machine learning-based prediction of binge drinking among adults in the United State: analysis of the 2022 Health Information National Trends Survey. 2024 Presented at: Proceedings of the 2024 9th International Conference on Mathematics and Artificial Intelligence; 2024, May 10-12; Beijing, China. [doi: 10.1145/3670085.3670090]
43. Fang W, Liu Y, Xu C, Luo X, Wang K. Feature selection and machine learning approaches in prediction of current e-cigarette use among U.S. adults in 2022. Int J Environ Res Public Health 2024;21(11):1474 [FREE Full text] [doi: 10.3390/ijerph21111474] [Medline: 39595741]
44. Health Information National Trends Survey HINTS 6 methodology report. Westat. 2023. URL: https://hints.cancer.gov/docs/methodologyreports/HINTS_6_MethodologyReport.pdf [accessed 2025-08-31]
45. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol 2017;17(1):162 [FREE Full text] [doi: 10.1186/s12874-017-0442-1] [Medline: 29207961]
46. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7(2):147-177. [Medline: 12090408]
47. Shi J, Fu R, Hamilton H, Chaiton M. A machine learning approach to predict e-cigarette use and dependence among Ontario youth. Health Promot Chronic Dis Prev Can 2022;42(1):21-28 [FREE Full text] [doi: 10.24095/hpcdp.42.1.04] [Medline: 35044141]
48. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28(5):1-26. [doi: 10.18637/jss.v028.i05]
49. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw 2010;36(11):1-13. [doi: 10.18637/jss.v036.i11]
50. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22. [doi: 10.18637/jss.v033.i01]
51. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273-297. [doi: 10.1007/BF00994018]
52. Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, et al. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 2017;151:147-160. [doi: 10.1016/j.catena.2016.11.032]
53. Kesler SR, Rao A, Blayney DW, Oakley-Girvan IA, Karuturi M, Palesh O. Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning. Front Hum Neurosci 2017;11:555 [FREE Full text] [doi: 10.3389/fnhum.2017.00555] [Medline: 29187817]
54. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med 2016;4(11):218-218 [FREE Full text] [doi: 10.21037/atm.2016.03.37] [Medline: 27386492]
55. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco California USA; 2016, August 13-17; San Francisco, California p. 785-794. [doi: 10.1145/2939672.2939785]
56. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-845. [Medline: 3203132]
57. Kaneko H. Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables. Heliyon 2021;7(6):e07356 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e07356] [Medline: 34195450]
58. Wang S, Nan B, Rosset S, Zhu J. Random lasso. Ann Appl Stat 2011;5(1):468-485. [doi: 10.1214/10-aoas377]
59. Fu Y, Zhao J, Wang Y. Lasso regression and Boruta algorithm to explore the relationship between neutrophil percentage to albumin ratio and asthma: results from the NHANES 2001 to 2018. Clin Exp Med 2025;25(1):149. [doi: 10.1007/s10238-025-01701-3] [Medline: 40347409]
60. Huang J, Liu W. Comparison of machine learning models for predicting stroke risk in hypertensive patients: lasso regression model, random forest model, Boruta algorithm model, and Boruta algorithm combined with lasso regression model. Medicine (Baltimore) 2025;104(22):e42690 [FREE Full text] [doi: 10.1097/MD.0000000000042690] [Medline: 40441184]
61. King AJ, Cooper GF, Hochheiser H, Clermont G, Hauskrecht M, Visweswaran S. Using machine learning to predict the information seeking behavior of clinicians using an electronic medical record system. AMIA Annu Symp Proc 2018;2018:673-682 [FREE Full text] [Medline: 30815109]
62. Comulada WS, Step M, Fletcher JB, Tanner AE, Dowshen NL, Arayasirikul S, Special Projects Of National Significance Social Media Initiative Study Group. Predictors of internet health information-seeking behaviors among young adults living with HIV across the United States: longitudinal observational study. J Med Internet Res 2020;22(11):e18309 [FREE Full text] [doi: 10.2196/18309] [Medline: 33136057]

XSL•FO
RenderX

63. Comulada WS, Goldbeck C, Almirol E, Gunn HJ, Ocasio MA, Fernández MI, Adolescent Medicine Trials Network (ATN) CARES Team. Using machine learning to predict young people's internet health and social service information seeking. Prev Sci 2021;22(8):1173-1184 [FREE Full text] [doi: 10.1007/s11121-021-01255-2] [Medline: 33974226]
64. Oladeji O, Zhang C, Moradi T, Tarapore D, Stokes AC, Marivate V, et al. Monitoring information-seeking patterns and obesity prevalence in Africa with internet search data: observational study. JMIR Public Health Surveill 2021;7(4):e24348 [FREE Full text] [doi: 10.2196/24348] [Medline: 33913815]
65. Zhao H, Su Y, Wang M, Lyu Z, Xu P, Jiao Y, et al. The machine learning model for distinguishing pathological subtypes of non-small cell lung cancer. Front Oncol 2022;12:875761 [FREE Full text] [doi: 10.3389/fonc.2022.875761] [Medline: 35692759]
66. Hu P, Li Y, Liu Y, Guo G, Gao X, Su Z, et al. Comparison of conventional logistic regression and machine learning methods for predicting delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage: a multicentric observational cohort study. Front Aging Neurosci 2022;14:857521 [FREE Full text] [doi: 10.3389/fnagi.2022.857521] [Medline: 35783143]
67. Hassan M, Hassan M, Yasmin F, Khan M, Zaman S, Galibuzzaman, et al. A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction. Decis Anal J 2023 Jun;7:100245. [doi: 10.1016/j.dajour.2023.100245]
68. Dong R, Fu N, Siriwardane EMD, Hu J. Generative design of inorganic compounds using deep diffusion language models. J Phys Chem A 2024;128(29):5980-5989. [doi: 10.1021/acs.jpca.4c00083] [Medline: 39008628]
69. Li Y, Hui S. Regression analysis and validation of risk factors for upper limb dysfunction following modified radical mastectomy for breast cancer patients. Am J Transl Res 2025;17(4):2614-2628. [doi: 10.62347/CZYA6232] [Medline: 40385071]
70. Ma W, Hou C, Yang M, Wei Y, Mao J, Guan L, et al. Different MRI-based radiomics machine learning models to predict CD3+ tumor-infiltrating lymphocytes in rectal cancer. Front Oncol 2025;15:1509207 [FREE Full text] [doi: 10.3389/fonc.2025.1509207] [Medline: 40356764]
71. Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom J 2018;60(3):431-449 [FREE Full text] [doi: 10.1002/bimj.201700067] [Medline: 29292533]
72. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54(8):774-781. [doi: 10.1016/s0895-4356(01)00341-9] [Medline: 11470385]
73. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
74. Zheng F, Wang K. The impact of social media on guideline-concordant cervical cancer-screening: insights from a national survey. Public Health 2023;223:50-56. [doi: 10.1016/j.puhe.2023.07.025] [Medline: 37598576]
75. Han CJ, Lee YJ, Demiris G. Interventions using social media for cancer prevention and management: a systematic review. Cancer Nurs 2018;41(6):E19-E31 [FREE Full text] [doi: 10.1097/NCC.0000000000000534] [Medline: 28753192]
76. Lazard AJ, Meernik C, Collins MKR, Vereen RN, Benedict C, Valle CG, et al. Social media use for cancer support among young adults with cancer. J Adolesc Young Adult Oncol 2023;12(5):674-684. [doi: 10.1089/jayao.2023.0025] [Medline: 37257189]
77. Aristokleous I, Karakatsanis A, Masannat YA, Kastora SL. The role of social media in breast cancer care and survivorship: a narrative review. Breast Care (Basel) 2023;18(3):193-199 [FREE Full text] [doi: 10.1159/000531136] [Medline: 37404835]
78. Johnson AR, Longfellow GA, Lee CN, Ormseth B, Skolnick GB, Politi MC, et al. Social media as a platform for cancer care decision-making among women: internet survey-based study on trust, engagement, and preferences. JMIR Cancer 2025;11:e64724 [FREE Full text] [doi: 10.2196/64724] [Medline: 40053770]
79. Gaysynsky A, Senft Everson N, Heley K, Chou WS. Perceptions of health misinformation on social media: cross-sectional survey study. JMIR Infodemiol 2024;4:e51127 [FREE Full text] [doi: 10.2196/51127] [Medline: 38687591]
80. Stimpson JP, Park S, Pruitt SL, Ortega AN. Variation in trust in cancer information sources by perceptions of social media health mis- and disinformation and by race and ethnicity among adults in the United States: cross-sectional study. JMIR Cancer 2024;10:e54162 [FREE Full text] [doi: 10.2196/54162] [Medline: 38717800]
81. Gapstur SM, Bandera EV, Jernigan DH, LoConte NK, Southwell BG, Vasiliou V, et al. Alcohol and cancer: existing knowledge and evidence gaps across the cancer continuum. Cancer Epidemiol Biomarkers Prev 2022;31(1):5-10 [FREE Full text] [doi: 10.1158/1055-9965.EPI-21-0934] [Medline: 34728469]
82. Seidenberg AB, Wiseman KP, Klein WMP. Do beliefs about alcohol and cancer risk vary by alcoholic beverage type and heart disease risk beliefs? Cancer Epidemiol Biomarkers Prev 2023;32(1):46-53 [FREE Full text] [doi: 10.1158/1055-9965.EPI-22-0420] [Medline: 36453075]
83. Kokole D, Ferreira-Borges C, Galea G, Tran A, Rehm J, Neufeld M. Public awareness of the alcohol-cancer link in the EU and UK: a scoping review. Eur J Public Health 2023;33(6):1128-1147 [FREE Full text] [doi: 10.1093/eurpub/ckad141] [Medline: 37802887]
84. Rohde JA, Klein WMP, D'Angelo H. Alcohol and cancer risk beliefs as correlates of alcohol consumption status. Am J Prev Med 2023;65(6):1181-1183. [doi: 10.1016/j.amepre.2023.06.012] [Medline: 37364661]

XSL•FO
RenderX

85. Mavadiya H, Lu Y. Diet-related awareness and behaviours in cancer survivors compared with non-cancer individuals: a pooled analysis of the HINTS study. Public Health Nutr 2025;28(1):e102. [doi: 10.1017/S1368980025100505] [Medline: 40457752]
86. Sommers J, Dizon DS, Lewis MA, Stone E, Andreoli R, Henderson V. Assessing health information seeking behaviors among targeted social media users using an infotainment video about a cancer clinical trial: population-based descriptive study. JMIR Cancer 2025;11:e56098 [FREE Full text] [doi: 10.2196/56098] [Medline: 40029972]
87. Loeb S, Langford AT, Bragg MA, Sherman R, Chan JM. Cancer misinformation on social media. CA Cancer J Clin 2024;74(5):453-464 [FREE Full text] [doi: 10.3322/caac.21857] [Medline: 38896503]
88. National Cancer Institute. URL: https://hints.cancer.gov/data/default.aspx [accessed 2026-03-25]

Abbreviations

ANN: artificial neural network
AUC: area under the receiver operating characteristic curve
DT: decision tree
HINTS 6: 2022 Health Information National Trends Survey
HINTS: Health Information National Trends Survey
HPV: human papillomavirus
KNN: k-nearest neighbor
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
ML: machine learning
NCI: National Cancer Institute
OR: odds ratio
PC: principal component
PCA: principal component analysis
RBF: radial basis function
RF: random forest
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by A Coristine; submitted 11.Apr.2025; peer-reviewed by X Liu, S Mohanadas, V Alemede, KA Obisesan Olawuni, RT Potla, MB Patel, S Chowdhury; comments to author 02.Jun.2025; revised version received 09.Mar.2026; accepted 12.Mar.2026; published 20.Apr.2026.

Please cite as:
Liu Y, Wang K
Comparison of Feature Selection Methods in Machine Learning Models of Cancer Information Seeking Among United States Adults: Cross-Sectional Study
JMIR Med Inform 2026;14:e75862
URL: https://medinform.jmir.org/2026/1/e75862
doi:10.2196/75862
PMID:34898427

©Ying Liu, Kesheng Wang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Predictive Accuracy of Mood Symptoms Using Wearable Devices and Machine Learning in Bipolar Disorder

P Keerthana Nair*, MBBS; Priyanka Renita D'Souza*, MBBS, MD

Department of Psychiatry, Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Manipal, India

*all authors contributed equally

Corresponding Author:
Priyanka Renita D'Souza, MBBS, MD
Department of Psychiatry, Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Manipal, India

Related Article:

https://medinform.jmir.org/2025/1/e66277

(JMIR Med Inform 2026;14:e92172) doi:10.2196/92172

KEYWORDS

machine learning; wearable device; bipolar disorder; mood disorder; digital mental health

We read the paper "Using Wearable Device and Machine Learning to Predict Mood Symptoms in Bipolar Disorder: Development and Usability Study" by Wu et al [1] with great interest. The integration of wearable data with machine learning to predict mood symptoms is clinically relevant and holds significant potential for real-world applications. However, we would like to highlight specific methodological aspects and possible future directions that could significantly enhance clinical reliability.

Our primary observation concerns the use of weekly mood assessments with backfilling. The authors acknowledged that their labeling method relied on self-rated questionnaires with backfilled mood symptoms for the preceding 7 days, and the participants may not complete these promptly as symptoms emerged, which could affect the accuracy of mood labels. In line with this, we appreciate the intent to minimize patient burden, but using retrospective entry to accurately capture daily mood fluctuations may not be reliable in the context of bipolar disorder. A patient feeling particularly low at the time of filling the weekly scale may unknowingly project that state onto the entire week, and that single net rating may not reflect the preceding days. This may create a risk of labeling daily physiological data with an inaccurate mood state, especially as the wearable data (sleep, activity, heart rate) are collected daily. To address this, future study designs may incorporate ecological momentary assessment, similar to the studies by Schwartz et al [2] and Li et al [3] that used digital biomarkers in studies on mood disorders. Implementing a quick visual check-in once or twice a day, for instance, a 5-point emoji scale with an optional short note, would reduce recall bias and ensure that daily wearable data are mapped against actual daily mood.

Furthermore, physiological markers such as average heart rate and total sleep hours are nonspecific and easily influenced by external stressors, including daily conflicts, work pressure, or life events. Without accounting for these confounding factors, the model risks interpreting situational stress as mood signals. As noted by the authors, heart rate is an index of sympathetic and parasympathetic influences of the autonomic nervous system in both stress and resting states, highlighting the overlap between stress response and mood states. Therefore, including a brief daily measure in future models to capture perceived daily stress would help in differentiating reactive distress from endogenous mood changes.

In conclusion, the integration of wearable biosensors is a promising development. However, we believe that a shift in study design from retrospective weekly aggregates to real-time, context-sensitive monitoring would further maximize the clinical utility of these models.

Acknowledgments

We used the generative artificial intelligence tool Gemini to make minor edits to the manuscript to improve language clarity, which were further reviewed and revised by the authors. The authors take full responsibility for the content of the publication.

Funding

The authors declare that no financial support was received for this work.

XSL•FO
RenderX

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: PKN, PRD
Formal analysis: PKN
Writing – original draft: PKN
Writing – review & editing: PRD
Supervision: PRD
Both authors approved the final manuscript.

Conflicts of Interest

None declared.

Editorial Notice

The corresponding author of "Using Wearable Device and Machine Learning to Predict Mood Symptoms in Bipolar Disorder: Development and Usability Study" was given the opportunity to respond to this letter but did not respond.

References

1. Wu CT, Hsieh MH, Chen IM, et al. Using wearable device and machine learning to predict mood symptoms in bipolar disorder: development and usability study. JMIR Med Inform 2025 Sep 16;13(1):e66277. [doi: 10.2196/66277] [Medline: 40957006]
2. Schwartz S, Schultz S, Reider A, Saunders EFH. Daily mood monitoring of symptoms using smartphones in bipolar disorder: a pilot study assessing the feasibility of ecological momentary assessment. J Affect Disord 2016 Feb;191:88-93. [doi: 10.1016/j.jad.2015.11.013] [Medline: 26655117]
3. Li H, Mukherjee D, Krishnamurthy VB, et al. Use of ecological momentary assessment to detect variability in mood, sleep and stress in bipolar disorder. BMC Res Notes 2019 Dec 4;12(1):791. [doi: 10.1186/s13104-019-4834-7] [Medline: 31801608]

Edited by A Coristine; submitted 26.Jan.2026; this is a non–peer-reviewed article;accepted 27.Feb.2026; published 12.Mar.2026.

Please cite as:
Nair PK, D'Souza PR
Enhancing Predictive Accuracy of Mood Symptoms Using Wearable Devices and Machine Learning in Bipolar Disorder
JMIR Med Inform 2026;14:e92172
URL: https://medinform.jmir.org/2026/1/e92172
doi:10.2196/92172

© P Keerthana Nair, Priyanka Renita D'Souza. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study

Jundan Chen*, PhD; Zhihe Mao*, BA

School of Physical Education, Hunan University of Arts and Science, 3150 Dongting Road, Changde, Hunan, China

*all authors contributed equally

Corresponding Author:
Zhihe Mao, BA
School of Physical Education, Hunan University of Arts and Science, 3150 Dongting Road, Changde, Hunan, China

Related Article:

Correction of: https://medinform.jmir.org/2026/1/e80163

Abstract

(JMIR Med Inform 2026;14:e92491) doi:10.2196/92491

In “Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study” [1], the authors noted two errors.

The author list has been corrected from the following:

Zhihe Mao, PhD; Jundan Chen, PhD

The author list now reads:

Jundan Chen, PhD; Zhihe Mao, BA

Author ZM’s contact information has been corrected to the following:

Phone: 86 13789072016

Fax: 86 07367283047

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Chen J, Mao Z. Comprehensive pediatric health risk stratification using an AI-driven framework in children aged 2 to 8 years: design and validation study. JMIR Med Inform 2026 Jan 26;14:e80163. [doi: 10.2196/80163] [Medline: 41587023]

Submitted 30.Jan.2026; this is a non–peer-reviewed article;accepted 05.Feb.2026; published 27.Feb.2026.

Please cite as:
Chen J, Mao Z
Correction: Comprehensive Pediatric Health Risk Stratification Using an AI-Driven Framework in Children Aged 2 to 8 Years: Design and Validation Study
JMIR Med Inform 2026;14:e92491
URL: https://medinform.jmir.org/2026/1/e92491
doi:10.2196/92491

© Jundan Chen Jr, Zhihe Mao Jr. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation

He Ayu Xu[1*], PhD; Bernard Maccari[2*], MSc; Hervé Guillain[3], MD, DrPH; Julien Herzen[2], PhD; Fabio Agri[4,5*], MBA, MD; Jean Louis Raisaro[1*], PhD

[1]Biomedical Data Science Center, Lausanne University Hospital, Lausanne, Switzerland

[2]Unit8 SA, Lausanne, Switzerland

[3]Public Health Solutions Ltd, Promasens, Switzerland

[4]Department of Administration and Finance, Lausanne University Hospital, Lausanne, Switzerland

[5]Department of Visceral Surgery, Lausanne University Hospital, Lausanne, Switzerland

[*]these authors contributed equally

Corresponding Author:
He Ayu Xu, PhD
Biomedical Data Science Center
Lausanne University Hospital
CHUV, Centre hospitalier universitaire vaudois Rue du Bugnon 21
Lausanne, 1011
Switzerland
Phone: 41 0795566886
Email: he.xu@chuv.ch

Related Article:

Correction of: https://medinform.jmir.org/2023/1/e38150

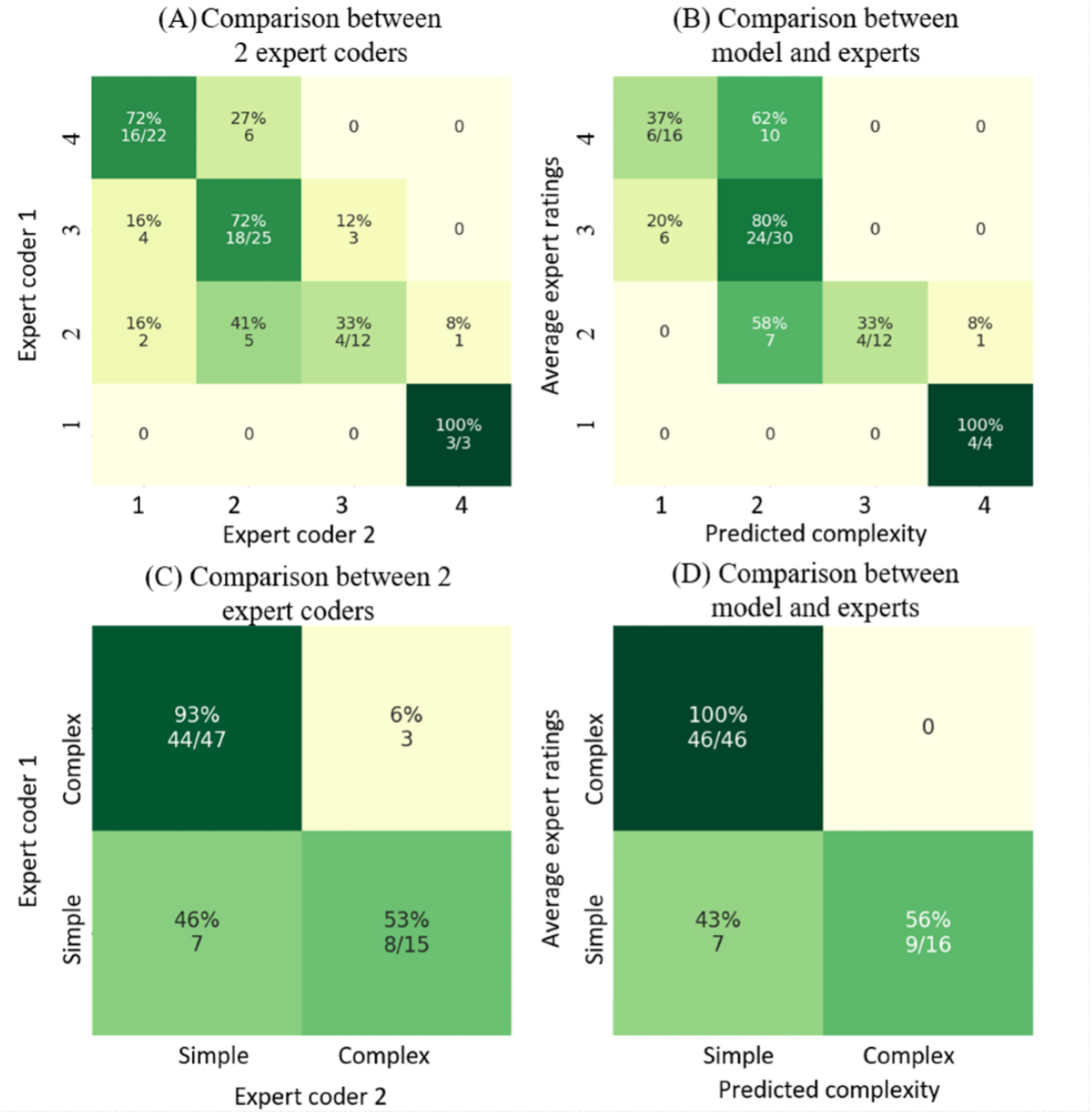
(JMIR Med Inform 2026;14:e94047) doi:10.2196/94047

In “An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation” [1] the authors made two corrections. The y-axis labels in Figures 9 and 10 were misprinted and did not correspond to the intended values shown in the original submission. The underlying data and conclusions of the paper remained unchanged.

The first figure originally appeared as follows (Figure 1):

XSL•FO
RenderX

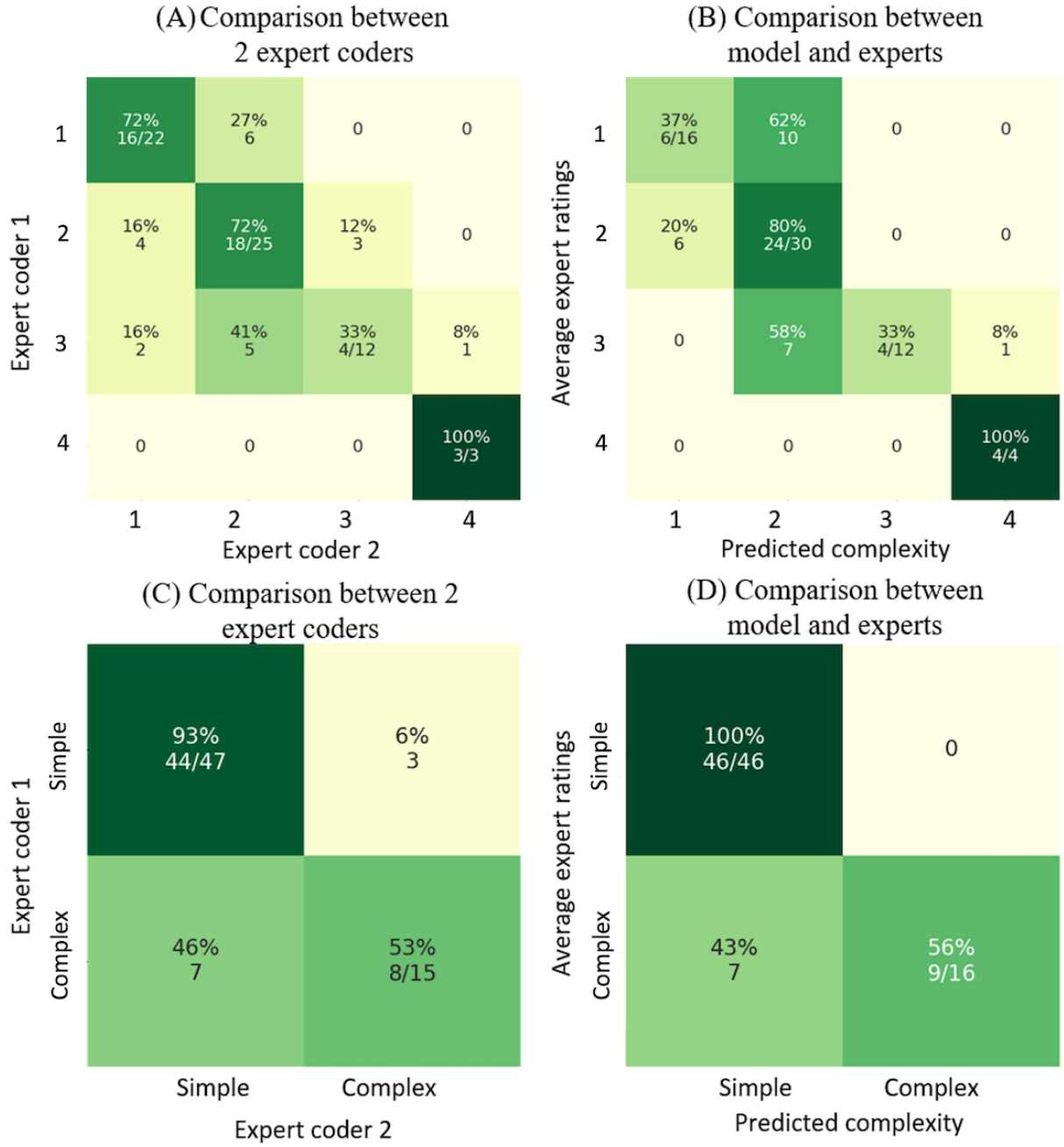
Figure 1. Original Figure 9, (A) The complexity rating comparison between 2 expert coders on the gold-standard set. (B) The comparison between the validation model's predictions and average expert ratings on the gold-standard set. (C) The comparison between 2 expert coders' ratings on the gold-standard set when grouping into simple (complexity 1 and 2) and complex (complexity 3 and 4) cases. (D) The comparison between average expert ratings and the validation model's predictions on the gold-standard set when grouping into simple and complex cases. The average expert ratings are rounded up to the next largest integer.



This figure has been revised to the following (Figure 2):

XSL•FO
RenderX

Figure 2. Revised Figure 9, (A) The complexity rating comparison between 2 expert coders on the gold-standard set. (B) The comparison between the validation model's predictions and average expert ratings on the gold-standard set. (C) The comparison between 2 expert coders' ratings on the gold-standard set when grouping into simple (complexity 1 and 2) and complex (complexity 3 and 4) cases. (D) The comparison between average expert ratings and the validation model's predictions on the gold-standard set when grouping into simple and complex cases. The average expert ratings are rounded up to the next largest integer.

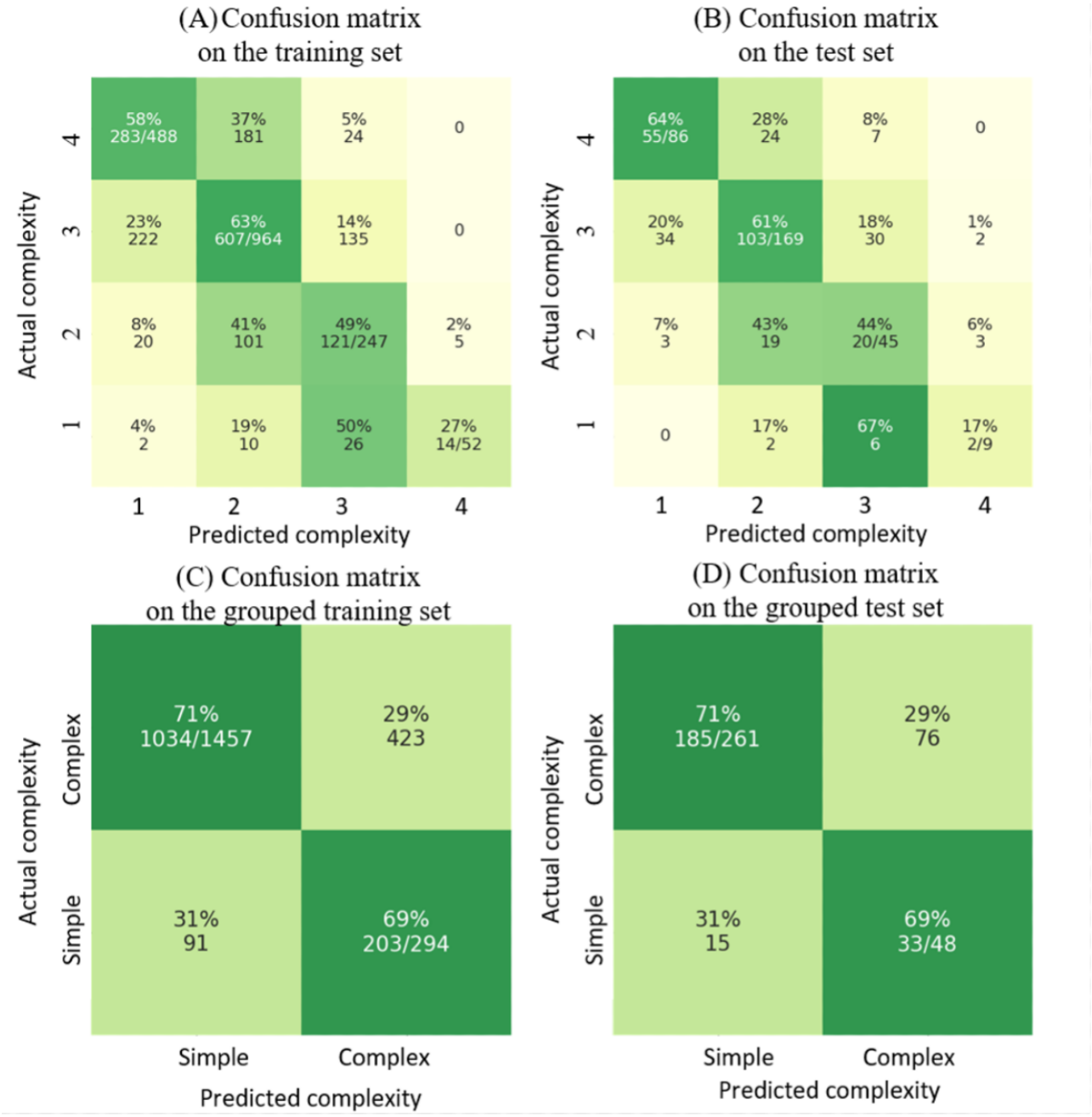


The y-axis labels in Figure 9 (A, B) were presented in reverse order. In the original published version, the labels appeared as 4, 3, 2, 1 from top to bottom, whereas they should read 1, 2, 3, 4 from top to bottom. The y-axis labels in Figure 9 (C, D) were presented in reverse order. In the original published version, the labels appeared as "Complex, Simple" from top to bottom, whereas they should read "Simple, Complex" from top to bottom. The underlying data and conclusions of the paper remained unchanged; however, the labeling error may cause confusion for readers.

The second figure originally appeared as (Figure 3):

XSL•FO
RenderX

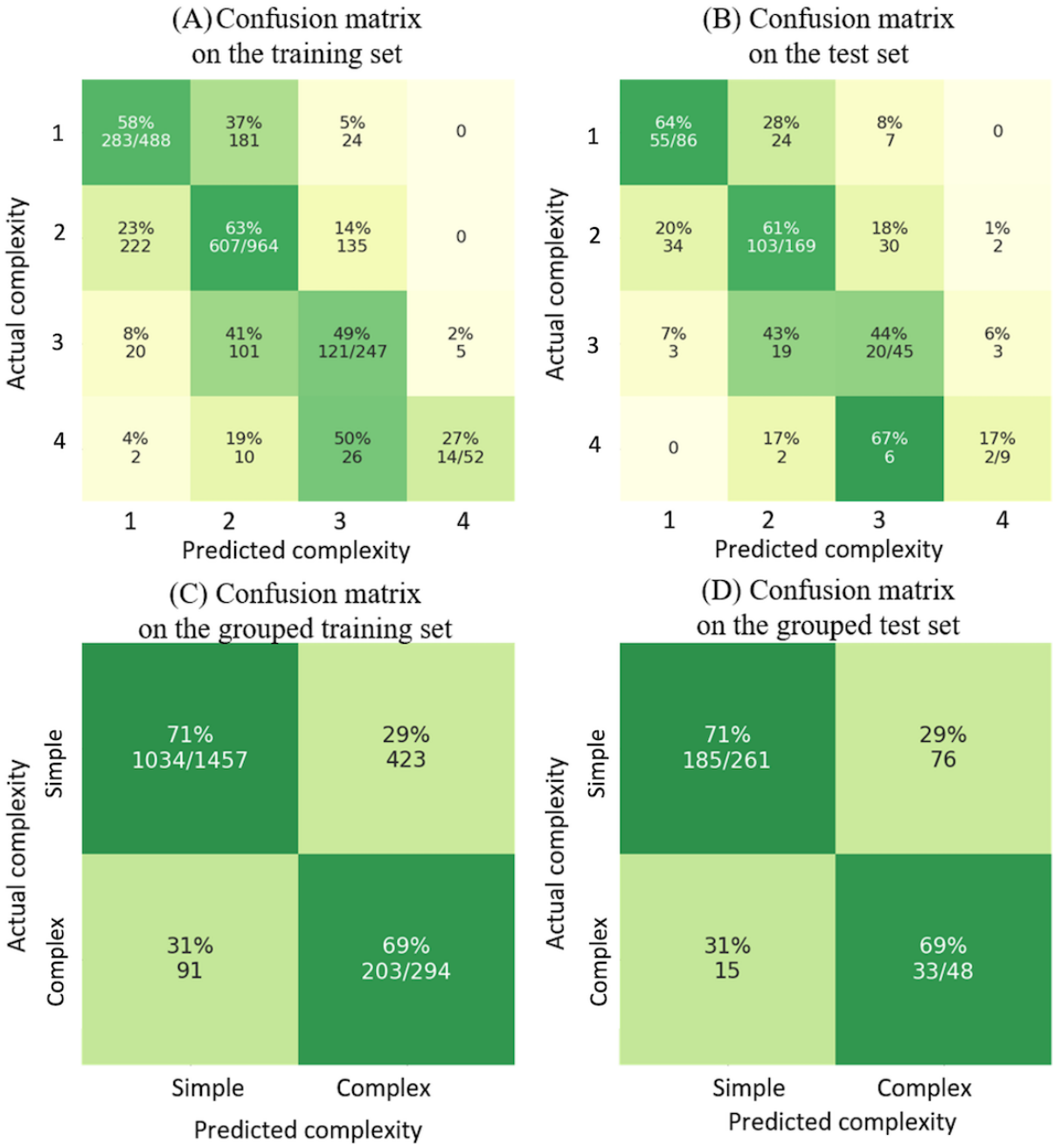
Figure 3. Original Figure 10, (A) and (B) The main model's performance on the training set (1751 cases) and the test set (309 cases). (C) and (D) The main model's performance on the grouped training set (1457 cases as simple and 294 cases as complex) and the test set (261 cases as simple and 48 cases as complex).



This figure has been revised to the following (Figure 4):

XSL•FO
RenderX

Figure 4. Revised Figure 10, (A) and (B) The main model's performance on the training set (1751 cases) and the test set (309 cases). (C) and (D) The main model's performance on the grouped training set (1457 cases as simple and 294 cases as complex) and the test set (261 cases as simple and 48 cases as complex).



The y-axis labels in Figure 10 (A, B) were presented in reverse order. In the published version, the labels appeared as 4, 3, 2, 1 from top to bottom, whereas they should read 1, 2, 3, 4 from top to bottom. The y-axis labels in Figure 10 (C, D) were presented in reverse order. In the published version, the labels appeared as "Complex, Simple" from top to bottom, whereas they should read "Simple, Complex" from top to bottom. The underlying data and conclusions of the paper remained unchanged; however, the labeling error may cause confusion for readers.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to thos

Reference

1. Xu HA, Maccari B, Guillain H, Herzen J, Agri F, Raisaro JL. An end-to-end natural language processing application for prediction of medical case coding complexity: algorithm development and validation. JMIR Med Inform 2023 Jan 19;11:e38150 [FREE Full text] [doi: 10.2196/38150] [Medline: 36656627]

XSL•FO
RenderX

Submitted 24.Feb.2026; this is a non–peer-reviewed article;accepted 27.Feb.2026; published 14.Apr.2026.

Please cite as:
Xu HA, Maccari B, Guillain H, Herzen J, Agri F, Raisaro JL
Correction: An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation
JMIR Med Inform 2026;14:e94047
URL: https://medinform.jmir.org/2026/1/e94047
doi:10.2196/94047
PMID:

©He Ayu Xu, Bernard Maccari, Hervé Guillain, Julien Herzen, Fabio Agri, Jean Louis Raisaro. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 14.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation

JMIR Publications Editorial Office

Related Article:

Correction of: https://medinform.jmir.org/2024/1/e65033

(JMIR Med Inform 2026;14:e88845) doi:10.2196/88845

Following the publication of "Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation" [1], concerns were raised regarding a potential conflict of interest between reviewer X Liu and members of the author list. After review, it was determined that the authors and the reviewer were both affiliated with Sichuan University West China Hospital. This conflict of interest was not disclosed to the editorial office as per journal policy.

This article has been subsequently reassessed by the *JMIR Medical Informatics* editor in chief, who determined that the content of this review did not unduly influence the decision to accept the article. The decision to accept the submission was based primarily on the remaining reviewers' comments and the previous editor in chief's contemporary appraisal of the manuscript.

As such, the JMIR Publications Editorial Office issues this corrigendum to update the Conflicts of Interest statement as follows:

> *Reviewer X Liu was found to have undisclosed conflicts of interest with members of the author list under the terms of JMIR Publications policy. The editor in chief determined that the content of the reviews did not unduly influence the decision to accept the article. The decision to accept the submission was based primarily on the remaining reviewers' comments and the editor in chief's contemporary appraisal of the manuscript.*

We regret that these issues were not identified before publication.

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Li X, Liu K, Lang Y, Chai Z, Liu F. Exploring the potential of claude 3 opus in renal pathological diagnosis: performance evaluation. JMIR Med Inform 2024 Nov 15;12:e65033 [FREE Full text] [doi: 10.2196/65033] [Medline: 39547661]

Submitted 02.Dec.2025; this is a non–peer-reviewed article;accepted 02.Dec.2025; published 23.Feb.2026.

Please cite as:
Editorial Office JMIRP
Correction: Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation
JMIR Med Inform 2026;14:e88845
URL: https://medinform.jmir.org/2026/1/e88845
doi:10.2196/88845
PMID:

©JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Experiences With Outpatient Pharmacy Services in Hospitals Using Automated Pharmacy Systems: Cross-Sectional Study

Ali Jasem Buabbas[1], MSc, PhD; Fatemah Mohammad Alsaleh[2], MSc, PhD; Mariam Alsanafi[2], MSc, PhD; Maryam Ali Safar[3], BMedSc, DMD; Ahmad T Al-Sultan[1], MSc, PhD; Talal Ahmad Zaman[4]

[1]Department of Community Medicine and Behavorial Sciences, College of Medicine, Kuwait University, Hawalli, Jabriya, Kuwait

[2]Department of Pharmacy Practice, College of Pharmacy, Kuwait University, Jabriya, Kuwait

[3]Ministry of Health, Kuwait, Kuwait

[4]College of Medicine, University of Alexandria, Alexandria, Egypt

Corresponding Author:
Ali Jasem Buabbas, MSc, PhD
Department of Community Medicine and Behavorial Sciences, College of Medicine, Kuwait University, Hawalli, Jabriya, Kuwait

Abstract

Background: Advancements in health technology and the adoption of electronic systems in hospital pharmacies have transformed pharmacy practice and service delivery, with patients and health care providers reporting perceived benefits related to patient care and safety. Therefore, it is of paramount importance to seek patients' opinions based on their experiences in receiving outpatient pharmacy services through automated pharmacy systems.

Objective: This study aimed to investigate patients' experiences and explore their satisfaction with outpatient pharmacy services in hospitals using automated pharmacy systems.

Methods: This cross-sectional study used an online questionnaire to assess patients' experiences of outpatient pharmacy services in governmental hospitals in Kuwait where automated pharmacy systems are implemented.

Results: Almost 95% (393/416) of the participants either agreed or strongly agreed that they were totally satisfied with the services delivered through the automated pharmacy systems. Participants reported the perceived benefits of these outpatient pharmacy services, including perceived prevention of medication errors (211/416, 50.7%), time savings (356/416, 85.6%), and easier access to prescription information at any time (249/416, 59.9%).

Conclusions: Patients' experiences with outpatient pharmacy services in automated pharmacy settings were generally positive, reflecting satisfaction with medication collection and service quality.

(JMIR Med Inform 2026;14:e80963) doi:10.2196/80963

KEYWORDS

automated pharmacy; pharmacy practice; patient experience; patient satisfaction; e-dispensing; e-prescribing

Introduction

Over the past years, pharmacy practice has undergone significant transformation with the adoption of electronic prescribing (e-prescribing) and electronic dispensing (e-dispensing; hospital information system [HIS]) systems, resulting in numerous advantages within professional practice settings. Notable examples include reductions in medication errors [1-4], time savings [5,6], and increased work efficiency [1,2,6]. However, such systems also present certain challenges. Reported drawbacks include risks to medication safety (eg, unclear or incorrect dosage instructions or strength) [4], technical issues [1,5], and resistance to adoption among staff [1]. Difficult adoption of automated systems has been reported in previous studies and has been attributed to their incapability to execute some functions that were previously completed manually [2,5,6]. The implementation of e-prescribing and/or e-dispensing systems has been extensively examined in the literature, with a focus on evaluating whether their intended purposes have been achieved and whether the anticipated benefits have been realized.

A systematic review of the literature on the barriers and facilitators to implementing an e-prescribing system highlighted the importance of identifying the perceptions of user groups, including pharmacists, managers, and patients [7]. Patients' involvement in the assessment process is essential because the quality of outpatient pharmacy service delivery is ultimately perceived by them; therefore, patient satisfaction can be considered an indicator of success [8,9]. Previous studies have supported the need for future research to evaluate the impact of

XSL•FO
RenderX

integrated e-prescribing and e-dispensing systems [8,10], as the combination of these 2 domains has the potential to improve various aspects of pharmacy schemes and practice [10].

In this study, the term "automated pharmacy" encompasses 2 key domains. The first is the receipt of prescriptions via an e-prescribing system, also referred to as a "pharmacy information system," which is a subset of a much greater HIS. Within such a system, patient data and medication lists are stored electronically. The second component of the "automated pharmacy" concept involves an e-dispensing system, in which electronic cabinets with the drugs on the shelves (most of which are prepackaged) are used to deliver the prescribed medications to the dispensing area. Briefly, the pharmacist receives the prescription from the patient, reviews it to ensure that the medications are accurately prescribed, and uses the e-prescribing system to send the order to the electronic or automated cabinet. The cabinet prepares the medications and delivers them to the pharmacist, who then dispenses them and provides patient counseling.

In Kuwait, there are 6 general governmental hospitals, all of which have transitioned their outpatient pharmacies from manual processes to automated systems. However, the extent of implementation varies across institutions. In some settings, automated dispensing systems are not fully connected to the e-prescribing platform, while in others, the e-prescribing system functions as a stand-alone tool that is not integrated with the HIS or electronic health records.

Following implementation, several challenges arise for staff operating integrated e-prescribing and automated dispensing systems, despite the technological advancements they offer. These challenges include technical difficulties, workflow and organizational issues, and factors related to users' attitudes, adaptation, and readiness to use technology. Such issues can affect the quality of the outpatient pharmacy services delivered to patients, particularly regarding waiting times, the adequacy of medication counseling, availability of medications, and the accuracy of medication dispensing. These challenges highlight the need to explore patients' experiences to understand their perceptions and satisfaction with outpatient pharmacy care in hospitals, which can help identify areas for improving and optimizing service quality.

Within this context, previous studies in Kuwait have examined the perspectives of health care providers regarding the use of HISs in primary care settings [11,12] or secondary care settings [13]. Patients' satisfaction with e-prescribing and electronic health records has also been explored [14,15]. In the existing literature, there is no research in Kuwait that examined patients' satisfaction based on their experiences with integrated automated outpatient pharmacy services. Thus, this study aimed to assess patients' satisfaction with their experiences in receiving outpatient pharmacy services through integrated e-prescribing and e-dispensing systems in the outpatient pharmacies of secondary care governmental hospitals. This enabled the development of recommendations to improve pharmacy practice in Kuwait.

Methods

Study Design, Sample, and Setting

The design of this study was a cross-sectional survey, using a questionnaire developed in Google Forms that was electronically administered to collect data from patients at their convenience between June and July 2024 at the outpatient pharmacies of 4 governmental hospitals in the state of Kuwait. The selected hospitals were Adan, Jahra, Farwaniya, and Jaber Al-Ahmad, which had all adopted and implemented the concept of automated pharmacy, which is using an integrated e-prescribing and e-dispensing system in the outpatient pharmacies.

A Survey Questionnaire

The items of the questionnaire were taken from previous validated studies [16]. Content validity was evaluated by the research team, which included 2 academic staff members from the Pharmacy Practice Department at the College of Pharmacy, Kuwait University, who are experts in their field, and 1 academic staff member from the College of Medicine with expertise in health informatics. This evaluation involved reviewing the questionnaire items to ensure their relevance to the study objectives and their clarity. Several modifications were made to ensure compatibility and appropriateness for the research aim, sample, and setting. The term "medication teaching" was replaced with "medication counseling," as this terminology is commonly used and understood by patients and pharmacy staff in Kuwait. The section "pharmacy place" was removed because it addressed the physical layout of the pharmacy, which was unrelated to the study's aim and objectives. Additionally, 1 item was excluded, which was about "the availability of a private area for counseling," because this service is not consistently provided across outpatient pharmacies in Kuwait and is not applicable to the local setting. A new section was added, which was about "dispensing medications," and was developed from previous studies [17,18] to explore aspects of communication between pharmacists and patients.

Questionnaire Structure

The questionnaire consisted of 40 items and seven sections: (1) time of prescription dispensing (3 items), (2) pharmacist attitudes (5 items), (3) medication supply (5 items), (4) medication counseling (5 items), (5) dispensing medications (5 items), (6) overall impact of automated pharmacy on outpatient pharmacy services (7 items), and (7) sociodemographic characteristics (10 items).

Inclusion and Exclusion Criteria

This study included secondary care hospitals that offer integrated automated outpatient pharmacy services. Patients and/or caregivers who provide close, hands-on care to the patient are familiar with the patient's experiences with outpatient pharmacy services and can offer valuable insights alongside the patients themselves. Eligible participants were those who were able to complete the questionnaire and were therefore invited to take part in the study. Private hospitals and private pharmacies were excluded from this study because they operate under different policies and procedures and do not use integrated e-prescribing

XSL•FO
RenderX

and e-dispensing systems within their outpatient pharmacy services.

Pilot Study

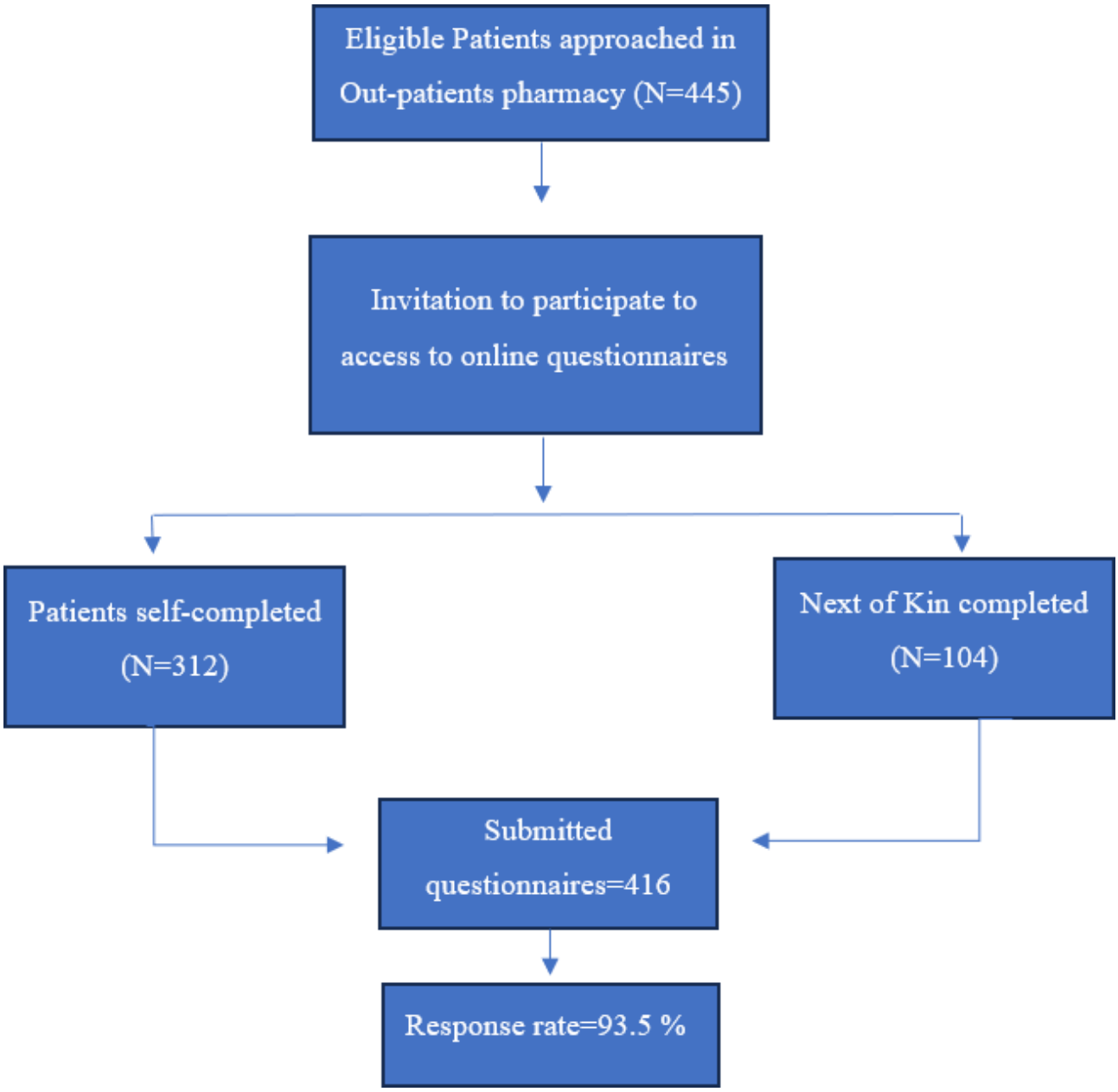
A pilot study was conducted to ensure clarity and readability with 25 patients who were waiting to receive their medications from outpatient pharmacies. The results demonstrated that the questionnaire was easy to understand and well received, confirming its readiness for full-scale distribution.

Recruitment and Data Collection

Patients waiting to receive their medications in the outpatient pharmacy areas were approached and invited to participate voluntarily. The questionnaire was administered to patients using a QR code or via a tablet device, providing direct access to the online questionnaire. No identifying patient information was collected, and responses could not be linked to individual patients. Pharmacy staff were not present while participants completed the survey, which helped maintain anonymity and reduce potential social desirability bias. For patients unable to complete the questionnaire themselves, a next of kin who was familiar with their experience (eg, a spouse, son, or sister) was permitted to respond on their behalf (Figure 1).

The questionnaire was offered in 2 languages, Arabic and English, to accommodate a diverse patient population. The original English version of the questionnaire was translated into Arabic, and the translation was validated using a back-translation method to ensure accuracy and clarity. All patients who met the inclusion criteria were invited to participate. Data collection was undertaken by a group of pharmacy students (n=6) from the College of Pharmacy, Kuwait University. They were available in the waiting areas of the outpatient pharmacies to share the online questionnaire with patients.

Figure 1. The recruitment protocol of the study participants.



XSL•FO
RenderX

Construct Validation Using Factor Analysis and Reliability Assessment

The Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity were used to evaluate sampling adequacy [19]. These tests help determine whether data are suitable for factor analysis. Data factorability was confirmed with a KMO value above 0.50 and a significant Bartlett's test value (P<.05) [19]. Principal component analysis with Promax rotation, an oblique method, was used. Promax was selected because it permits factors to correlate, which aligns with the expected relationships among the constructs [20]. The number of factors was decided based on eigenvalues, scree plots, and explained variance. Exploratory factor analysis was performed with a factor loading cutoff of 0.40 to extract scale dimensions [19,21].

Internal consistency reliability was assessed using Cronbach α coefficient based on a 4-point Likert scale (time of prescription dispensing, pharmacist attitudes, medication supply, medication counseling, and dispensing medications). A Cronbach α value of ≥0.70 was regarded as satisfactory evidence of internal consistency for the new instrument [19,20,22].

Ethical Considerations

Ethical approval for this study was obtained from the The Standing Committee for Coordinating Medical and Health Research prior to data collection (reference number 2352). Participation was voluntary, and informed consent was obtained from all participants before administering the survey. To protect privacy and confidentiality, no personally identifiable information was collected. Surveys were completed anonymously in the outpatient pharmacy waiting area, and participants were informed that they could decline to answer any question or withdraw at any time without any impact on their care or pharmacy services. All collected data were kept confidential, stored securely, and accessible only to the research team. Study findings are reported in aggregate form, ensuring that no individual participant can be identified.

Statistical Data Analysis

Statistical analyses were conducted using IBM SPSS Statistics (version 29; IBM Corp). Descriptive statistics were used to present the data. Categorical variables were presented as frequencies and percentages. To summarize the patients' perspective scores, responses for the questionnaire items were averaged across all items within the category. As an initial exploratory comparison, patient and caregiver perspective scores were compared using the Mann-Whitney U test due to nonnormal score distributions and unequal group sizes.

A normality test for the component scores was undertaken using the Shapiro-Wilk test, which indicated that the data were not normally distributed. Accordingly, results are presented as minimum, maximum, median, and IQR. Nonparametric statistical tests were used to assess associations between component scores and variables of interest, including the Mann-Whitney U test for binary comparisons, the Kruskal-Wallis test for comparisons across multiple groups, Spearman rank correlation for continuous variables, and the chi-square test for trend for ordinal categorical variables.

For analyses based on a priori hypotheses involving comparisons across multiple groups, post hoc pairwise comparisons were conducted following statistically significant Kruskal-Wallis tests, with P values adjusted for multiple comparisons using the Dunn-Bonferroni method. These analyses were planned and considered confirmatory. Effect sizes were reported as rank-biserial correlation coefficients (r_rb) for binary comparisons, epsilon squared (ε^2) for multigroup comparisons, and Spearman correlation coefficient (ρ) for continuous variables.

Multiple linear regression analyses were performed to identify independent factors associated with the overall score across domains of patients' perspectives on automated pharmacy services. Additional association analyses not specified a priori were treated as exploratory and were not adjusted for multiple comparisons; these results are presented for hypothesis generation. Ordinal survey responses were averaged across items within each component to generate composite scores for inferential analyses. Statistical significance was set at P value of <.05.

A sensitivity analysis was conducted to assess the results to potential selection bias arising from voluntary participation and convenience sampling. This analysis examined the possible impact of underrepresentation of less satisfied patients on the study outcomes. Specifically, the patients' perspective score was systematically reduced by 0.3, 0.5, and 1.0 SDs, representing mild, moderate, and severe dissatisfaction scenarios, respectively. These thresholds were selected based on established effect size conventions for continuous outcomes and their application in patient-reported outcome research [23]. For each scenario, the multiple linear regression analyses were repeated using the adjusted outcome variable, while all predictor variables and model specifications remained unchanged.

Results

Sociodemographic Characteristics of the Study Participants

A total of 445 patients were approached, of whom 416 completed the questionnaire, producing a response rate of 93.5%. The sociodemographic characteristics of the study participants are illustrated in Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics of participants (N=416).

Characteristics	Values, n (%)
Total	416 (100.0)
Hospital	
Adan	111 (26.7)
Jahra	95 (22.8)
Farwaniya	100 (24.0)
Jaber Al-Ahmad	110 (26.4)
Age (years)	
<29	142 (34.3)
30-39	108 (26.1)
40-49	105 (25.4)
≥50	59 (14.3)
Nationality	
Kuwaiti	351 (84.4)
Non-Kuwaiti	65 (15.6)
Sex	
Male	117 (28.1)
Female	299 (71.9)
Marital status	
Single	158 (38.0)
Married	258 (62.0)
Education	
High school or less	76 (18.3)
Diploma	108 (26.0)
Bachelor	208 (50.0)
Higher education	24 (5.8)
Estimated number of pharmacy visits in the last year	
Once to twice	92 (22.1)
3-4 times	155 (37.3)
Monthly	161 (38.7)
Weekly	8 (1.9)
Dispense medications	
Personally	312 (75.0)
Someone else	104 (25.0)
The time I have waited to receive my medication	
≤5 minutes	136 (35.5)
6-15 minutes	167 (43.6)
>15 minutes	80 (20.9)
The number of medications in my prescription	
1-5 items	331 (79.6)
6-10 items	62 (14.9)
>10 items	23 (5.5)

XSL•FO
RenderX

Most participants were female (299/416, 71.9%), Kuwaiti nationals (351/416, 84.4%), younger than 29 years (142/416, 34.3%), and married (258/416, 62.0%). Half of the participants (208/416, 50%) held a bachelor's degree. Regarding pharmacy visits, approximately one-third of the participants visited the pharmacy monthly (161/416, 38.7%), while another third reported visiting 3 or 4 times per year (155/416, 37.3%). Most participants were collecting medications for themselves (312/416, 75%) and had between 1 and 5 medications in their prescriptions (331/416, 79.6%). Dispensing waiting times varied, with the majority of participants waiting 6 - 15 minutes (167/416, 40.1%), followed by those waiting less than 5 minutes (136/416, 32.7%), and those waiting more than 15 minutes (80/416, 19.2%).

Based on the exploratory factor analysis results shown in Table 2, each item with a value greater than 0.40 was assigned to an extracted factor. The KMO value of 0.914 indicated that the data were appropriate for factor analysis, and the value of Bartlett's test of sphericity was significant (χ^2_{231}=5140.38; P<.001). The scree plot indicated a 5-component solution, with all items showing acceptable communalities, and eigenvalues greater than 1.0 were extracted, accounting for 67.27% of the total variance. The reliability of the 22-item factored scale was obtained. The overall Cronbach α for the scale showed excellent reliability (α=0.92).

Table . Results of the exploratory factor analysis.

Component	Item	Factor loadings[a]	Eigenvalue	VE[b] (%)	CVE[c] (%)
Time of prescription dispensing			1.79	8.12	8.12
	1.00	0.83			
	2.00	0.90			
Pharmacist attitude			4.03	18.32	26.44
	3.00	0.74			
	4.00	0.74			
	5.00	0.83			
	6.00	0.80			
	7.00	0.75			
Medication supply			3.56	16.18	42.62
	8.00	0.67			
	9.00	0.62			
	10.00	0.72			
	11.00	0.73			
	12.00	0.71			
Medication counseling			3.05	13.84	56.46
	13.00	0.52			
	14.00	0.40			
	15.00	0.84			
	16.00	0.85			
	17.00	0.68			
Dispensing medication			2.38	10.81	67.27
	18.00	0.43			
	19.00	0.79			
	20.00	0.69			
	21.00	0.79			
	22.00	0.54			

[a]Extraction method: principal component analysis. Rotation method: Promax. Factor loading cutoff: >0.40.

[b]VE: variance explained.

[c]CVE: cumulative variance explained.

XSL•FO
RenderX

Time of Prescription Dispensing

The results showed that most respondents agreed or strongly agreed that their prescriptions were dispensed within a reasonable time (364/416, 87.5%; Table 3).

XSL•FO
RenderX

Table . Perceptions of e-prescription and automated dispensing systems in outpatient pharmacies at secondary care hospitals (N=416).

Item	Values, n (%)								Range	Overall median score (IQR)[a]
	Strongly agree	Agree	Disagree	Strongly disagree	Never	Rarely	Sometimes	Many times		
Time of prescription dispensing									1.0-4.0	3.0 (0.5)
1. Receive medications within a reasonable time	136 (32.7)	228 (54.8)	38 (9.1)	14 (3.4)	—[b]	—	—	—		
2. Waiting time is acceptable considering the quantity of prescription medication	115 (27.6)	226 (54.3)	63 (15.1)	12 (2.9)	—	—	—	—		
Pharmacist attitude									4.0-1.2	3.4 (1.0)
3. Pharmacist helped me to get the medications	201 (48.3)	201 (48.3)	9 (2.2)	5 (1.2)	—	—	—	—		
4. Pharmacist helped to solve any problem while getting the medication	168 (40.4)	211 (50.7)	33 (7.9)	4 (1.0)						
5. Pharmacist answered my questions	197 (47.4)	200 (48.1)	18 (4.3)	1 (0.2)						
6. Pharmacist understood the medical case	172 (41.3)	222 (53.4)	21 (5.0)	1 (0.2)						
7. Pharmacist treated me with respect	254 (61.1)	153 (36.8)	5 (1.2)	4 (1.0)						
Medication supply									4.0-1.8	3.2 (0.8)
8. Medication quantity was sufficient	178 (42.8)	214 (51.4)	19 (4.6)	5 (1.2)						

XSL•FO
RenderX

Item	Values, n (%)								Range	Overall median score (IQR)[a]
	Strongly agree	Agree	Disagree	Strongly disagree	Never	Rarely	Sometimes	Many times		
9. All my medications were available in the pharmacy	131 (31.5)	180 (43.3)	87 (20.9)	18 (4.3)						
10. Medication name was clear and easy to read	197 (47.4)	200 (48.1)	14 (3.4)	5 (1.2)						
11. Medication label or sticker instructions were clear	192 (46.2)	202 (48.6)	18 (4.3)	4 (1.0)						
12. Medication appearance and quality was good	195 (46.9)	211 (50.7)	10 (2.4)	0 (0.0)						
Medication counseling									4.0-1.0	3.0 (0.8)
13. Pharmacist explained the reason of my medication	169 (40.6)	187 (45.0)	44 (10.6)	16 (3.8)						
14. Pharmacist told me how to take the correct medication dose	189 (45.4)	192 (46.2)	27 (6.5)	8 (1.9)						
15. Pharmacist explained my medication's possible side effects	84 (12.5)	159 (38.2)	159 (38.2)	52 (12.5)						
16. Pharmacist explained how to store my medication	82 (12.5)	128 (30.8)	154 (37.0)	52 (12.5)						
17. I had enough time with the pharmacist	123 (29.6)	194 (46.6)	77 (18.5)	22 (5.3)						
Dispensing medication									4.0-1.0	3.8 (0.4)

XSL•FO
RenderX

Item	Values, n (%)								Range	Overall median score (IQR)[a]
	Strongly agree	Agree	Disagree	Strongly disagree	Never	Rarely	Sometimes	Many times		
18. The communication with the pharmacist and/or the doctor was challenging	—	—	—	—	218 (52.4)	111 (26.7)	71 (17.1)	16 (3.8)		
19. The prescription was sent to the wrong pharmacy	—	—	—	—	349 (83.9)	53 (12.7)	12 (2.9)	2 (0.5)		
20. The wrong medication was dispensed to me	—	—	—	—	384 (92.3)	26 (6.3)	4 (1.0)	2 (0.5)		
21. I received wrong instructions regarding the drugs dispensed for me	—	—	—	—	371 (89.2)	31 (7.5)	12 (2.9)	2 (0.5)		
22. My medical prescription or information was not received by the pharmacy	—	—	—	—	294 (70.7)	79 (19.0)	32 (7.7)	11 (2.6)		

[a]For each perspective score, responses for the questionnaire items were averaged across all items within the category.

[b]Not applicable.

Most participants (341/416, 82.0%) also found the waiting time acceptable, considering the number of medications prescribed. A statistically significant association (P=.006; chi-square test) was found between the number of medications prescribed and the reported waiting time (Table 4). Patients prescribed 1 - 5 medications were more likely to experience shorter waiting times: 38.8% (118/304) received their medications within 5 minutes, 41.8% (127/304) waited 6 - 15 minutes, and 19.4% (59/304) waited more than 15 minutes.

Table . Association between the number of medications and the waiting time to receive the medications (P=.006[a]; chi-square for trend).

Number of medications	Waiting time		
	≤5 minutes	6-15 minutes	>15 minutes
1-5 items, n (raw %)	118 (38.8)	127 (41.8)	59 (19.4)
6-10 items, n (raw %)	14 (24)	31 (53)	13 (22)
>10 items, n (raw %)	4 (19)	9 (43)	8 (38)

[a]Chi-square test for trend.

XSL•FO
RenderX

In contrast, patients prescribed 6 - 10 medications experienced longer waiting times. Only 24% (14/58) received their medications within 5 minutes, while 53% (31/58) waited 6 - 15 minutes, and 22% (13/58) waited more than 15 minutes. Among those prescribed more than 10 medications, few (4/21, 19%) received their medications within 5 minutes or less, while many (9/21, 43%) waited 6 - 15 minutes, and a similar proportion (8/21, 38%) waited more than 15 minutes, as shown in Table 4.

Pharmacist Attitudes

As shown in Table 3, the results indicated that most participants had positive perceptions of pharmacist attitudes. The majority agreed or strongly agreed that the pharmacists helped obtain their medications (401/416, 96.5%) and assisted them in resolving issues related to medication access (379/416, 91.1%). In addition, a large proportion reported that the pharmacists answered their questions (397/416, 95.5%) and demonstrated knowledge and understanding toward their medical cases (394/416, 94.7%). Furthermore, most participants shared the perspective that the pharmacists treated them with respect and dignity (407/416, 97.9%).

Medication Supply

Most participants found the quantities of medications dispensed to be sufficient (392/416, 94.2%). Additionally, nearly three-quarters (311/416, 74.8%) reported that all their prescribed medications were available in the pharmacy (Table 3). However, a subset of participants (105/416, 25.2%) encountered instances where 1 or more prescribed medications were unavailable. Medication labels were clear and easy to read for most participants (397/416, 95.5%), and the usage instructions were clear and understandable (394/416, 94.8%). Furthermore, the majority expressed satisfaction with the appearance and quality of the medications provided (406/416, 97.6%).

Medication Counseling

The results revealed that the pharmacists performed well in several key areas of patient counseling (Table 3), which included explaining the purpose of medication use (356/416, 85.6%), providing instructions on correct medication dosage (381/416, 91.6%), and dedicating sufficient time to patients during consultations (317/416, 76.2%). However, notable gaps were identified in some specific aspects of counseling. Only half of the participants reported receiving information about the potential side effects of their medications (211/416, 50.7%). Similarly, counseling on proper medication storage was lacking, with approximately half of the participants not receiving such explanations (206/416, 49.5%).

Dispensing Medications

As shown in Table 3, most participants reported that communication challenges with pharmacists and/or physicians rarely or never occurred (329/416, 79.1%). Similarly, errors in the e-prescribing and e-dispensing process were uncommon. Specifically, the majority of participants agreed that rarely or never were prescriptions sent to the wrong pharmacy (402/416, 96.6%), the wrong medication was dispensed (410/416, 98.6%), incorrect instructions were received regarding the medications dispensed (402/416, 96.7%), or the prescription information was not received by the pharmacy (373/416, 89.7%).

Impact of Automated Pharmacy and Overall Satisfaction With Outpatient Pharmacy Services

The participants identified several perceived benefits of the integrated e-prescribing and e-dispensing systems, reflecting their positive impact on outpatient pharmacy services (Multimedia Appendix 1). These perceived benefits included time savings (356/416, 85.6%), perceived prevention of medication errors (211/416, 50.7%), and convenient access to prescription information at any time (249/416, 59.9%).

Also, the results showed high overall satisfaction with the automated outpatient pharmacy services, as 94.5% (393/416) of participants either agreed or strongly agreed that they were satisfied with the services delivered through the automated systems (Multimedia Appendix 2).

The overall median score for the time of prescription dispensing section was 3.0, while the median score for pharmacist attitudes was 3.4. The median scores for the medication supply, medication counseling, and dispensing medications domains were 3.2, 3.0, and 3.8, respectively (Table 3).

Tables 5 and 6 demonstrate the relationships between sociodemographic characteristics and the overall median scores for each questionnaire component. There were significant differences among hospitals, with small effect sizes for the time of prescription dispensing and the dispensing medications scores (ε^2=0.05 and ε^2=0.02, respectively). The median time of prescription dispensing score for Farwaniya hospital was significantly higher than for Adan hospital (3.2 vs 3.0; adjusted P=.03) and Jaber Al-Ahmad hospital (3.2 vs 3.0; adjusted P<.001). The median dispensing medications score for Jaber Al-Ahmad hospital was significantly higher than for Adan hospital (4.0 vs 3.8; adjusted P=.02) and Jahra hospital (4.0 vs 3.8; adjusted P=.04).

XSL•FO
RenderX

Table . Component scores for a survey assessing e-prescribing and e-dispensing systems in outpatient pharmacies at secondary care hospitals according to sociodemographic characteristics (N=416).

Characteristics	Median score (IQR)[a]				
	Time of prescription dispensing	Pharmacist attitude	Medication supply	Medication counseling	Dispensing medication
Hospital					
Adan	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.6)
Jahra	3.0 (1.0)	3.4 (1.0)	3.2 (1.0)	3.0 (1.0)	3.8 (0.6)
Farwaniya	3.2 (1.0)	3.4 (1.0)	3.3 (1.0)	3.0 (1.2)	3.8 (0.4)
Jaber Al-Ahmad	3.0 (1.5)	3.6 (1.0)	3.4 (0.8)	3.0 (0.9)	4.0 (0.4)
Age (years)					
<29	3.0 (0.6)	3.4 (1.0)	3.2 (1.0)	3.0 (1.0)	3.8 (0.6)
30-39	3.0 (1.0)	3.6 (1.0)	3.3 (0.8)	3.0 (0.9)	3.8 (0.4)
40-49	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.4)
≥50	3.0 (0.5)	3.4 (0.8)	3.2 (0.8)	2.6 (1.0)	4.0 (0.4)
Nationality					
Kuwaiti	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (1.0)	3.8 (0.6)
Non-Kuwaiti	3.0 (0.8)	3.6 (1.0)	3.2 (0.9)	3.0 (1.1)	3.8 (0.4)
Sex					
Male	3.0 (1.0)	3.6 (1.0)	3.4 (1.0)	3.0 (1.2)	3.8 (0.6)
Female	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.4)
Marital status					
Single	3.0 (0.5)	3.5 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.6)
Married	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (1.0)	3.8 (0.4)
Education					
High school or less	3.0 (1.0)	3.6 (1.0)	3.2 (1.0)	3.0 (1.0)	4.0 (0.4)
Diploma	3.0 (0.5)	3.2 (1.0)	3.4 (0.8)	3.0 (1.4)	3.8 (0.6)
Bachelor	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.6)
Higher education	3.0 (1.9)	3.7 (0.9)	3.6 (1.0)	3.1 (1.6)	3.9 (0.9)
Estimated number of pharmacy visits in the last year					
Once to twice	3.0 (1.4)	3.4 (1.0)	3.1 (1.0)	3.0 (1.0)	4.0 (0.6)
3-4 times	3.0 (0.5)	3.4 (1.0)	3.4 (0.8)	3.0 (0.8)	3.8 (0.4)
Monthly	3.0 (0.5)	3.6 (1.0)	3.2 (0.8)	3.0 (1.0)	3.8 (0.6)
Weekly	3.0 (1.0)	3.3 (1.4)	3.0 (1.1)	2.5 (1.8)	3.6 (0.9)
Dispense medications					
Personally	3.0 (0.5)	3.4 (1.0)	3.2 (0.8)	3.0 (0.8)	3.8 (0.4)
Someone else	3.0 (0.5)	3.3 (1.0)	3.2 (0.9)	2.9 (1.0)	3.8 (0.8)
The time I have waited to receive my medication					
≤5 minutes	3.5 (1.0)	3.6 (0.9)	3.6 (1.0)	3.0 (1.0)	4.0 (0.4)
6-15 minutes	3.0 (0.5)	3.4 (0.8)	3.2 (0.6)	2.8 (0.6)	3.8 (0.6)
>15 minutes	2.8 (1.0)	3.2 (1.0)	3.1 (1.0)	2.8 (1.0)	3.6 (0.6)
The number of medications in my prescription					
1-5 items	3.0 (0.5)	3.4 (1.0)	3.4 (0.8)	3.0 (0.8)	3.8 (0.4)
6-10 items	3.0 (0.5)	3.5 (1.0)	3.0 (1.0)	3.0 (0.9)	3.8 (0.6)

XSL•FO
RenderX

Characteristics	Median score (IQR)[a]				
	Time of prescription dispensing	Pharmacist attitude	Medication supply	Medication counseling	Dispensing medication
11-15 items	3.0 (1.0)	3.0 (0.6)	3.2 (0.8)	3.0 (1.0)	3.8 (0.6)

[a]For each perspective score, responses for the questionnaire items were averaged across all items within the category.

Table . Relationships between sociodemographic variables and component scores for a survey assessing e-prescribing and e-dispensing systems in outpatient pharmacies at secondary care hospitals (N=416).

Characteristics	Time of prescription dispensing		Pharmacist attitude		Medication supply		Medication counseling		Dispensing medication	
	Effect size[a]	*P* value[b]	Effect size[a]	*P* value[b]	Effect size[a]	*P* value[b]	Effect size[a]	*P* value[b]	Effect size[a]	*P* value[b]
Hospital	0.055	<.001	0.004	.71	0.004	.73	0.003	.60	0.018	.02
Age (years)	−0.050	.31	−0.050	.34	−0.003	.96	−0.130	.007	0.040	.46
Nationality	0.370	.72	−0.120	.90	0.150	.89	1.370	.17	0.450	.65
Gender	0.219	.13	0.225	.12	0.105	.09	0.053	.39	0.112	.06
Marital status	0.028	.61	0.008	.89	0.019	.74	0.057	.33	0.052	.35
Education	0.110	.03	0.010	.87	0.040	.43	−0.070	.15	0.090	.07
Number of visits (last year)	−0.010	.89	0.030	.54	0.001	.98	0.030	.63	−0.060	.22
Dispense medications	0.281	.12	0.242	.46	0.266	.22	0.272	.190	0.299	.06
Time waited	−0.460	<.001	−0.150	.003	−0.150	.003	−0.220	<.001	−0.170	.001
Number of medications	−0.110	.03	−0.030	.51	−0.090	.06	−0.020	.77	−0.050	.33

[a]Rank-biserial correlation (r_{rb}) for 2-group comparisons, epsilon squared (ε^2) for multigroup comparisons, and Spearman correlation coefficient for continuous variables. Interpretation thresholds: r_{rb}≈0.10=small, ≈0.30=moderate, and ≥0.50=large; ε^2≈0.01=small, ≈0.06=moderate, and ≥0.14=large; and value of P ranges from −1 to +1, P<.1=small, .1-.3=medium, and >.5=large.

[b]Mann-Whitney *U* tests (2 groups), Kruskal-Wallis tests (≥3 groups), or Spearman rank correlation test (continuous variables).

An indirect correlation coefficient of r=−0.13 was present between age and medication counseling score (P=.007). Likewise, there was an indirect correlation between dispensing medications score and education level (r=−0.11; P=.03). All 4 other domains of the questionnaire had scores that were significantly related to the amount of time it took to receive prescriptions (P<.05), with indirect Spearman rank correlation coefficients ranging from −0.15 to −0.46. Furthermore, a significant indirect correlation coefficient of r=−0.11 was found between number of medications and dispensing medications score (P=.03). For other patient characteristics, no statistically significant differences or associations were observed across any of the 5 perspective domains (all P>.05).

All 5 domains of the questionnaire were significantly related to each other (P≤.001), with Spearman rank correlation coefficients ranging from 0.30 to 0.62 (Table 7). The number of medications a patient was prescribed and the amount of time they waited to obtain those prescriptions had a significant association (P=.006), as seen in Table 4. In other words, the amount of waiting time rises as the number of medications increases (from 1 - 5 items to 6 - 10 items or 10 items or more).

XSL•FO
RenderX

Table . Spearman rank correlations between domains in a survey assessing e-prescribing and e-dispensing systems in outpatient pharmacies at secondary care hospitals (N=416).

Component	Time of prescription dispensing		Pharmacist attitude		Medication supply		Medication counseling		Dispensing medication	
	R	*P* value	*R*	*P* value	*R*	*P* value	*R*	*P* value	*R*	*P* value
Time of prescription dispensing	N/A[a]	N/A	0.48	<.001	0.47	<.001	0.44	<.001	0.3	<.001
Pharmacist attitude	0.48	<.001	N/A	N/A	0.62	<.001	0.62	<.001	0.39	<.001
Medication supply	0.47	<.001	0.62	<.001	N/A	N/A	0.62	<.001	0.48	<.001
Medication counseling	0.44	<.001	0.62	<.001	0.62	<.001	N/A	N/A	0.48	<.001
Dispensing medication	0.3	<.001	0.39	<.001	0.48	<.001	0.48	<.001	N/A	N/A

[a]Not applicable.

The results showed that there is an inverse relationship between medication dispensing waiting time and patient perceptions of automated outpatient pharmacy services, whereas increased waiting times were associated with less satisfaction of pharmacists' attitudes, medication counseling, and dispensing practices.

Table 8 shows the results of the multiple regression model examining patients' perceptions toward automated outpatient pharmacy services, with total perception scores modeled as a function of sociodemographic and other relevant variables. All variables were entered simultaneously to estimate their independent associations. The model was statistically significant (F_{231}=2.32; *P*=.001) and explained 11% of the variance in patients' perceptions. Compared with patients who waited 5 minutes or less, those who waited 6-15 minutes and more than 15 minutes reported significantly lower perception scores (β=−3.70, 95% CI −5.76 to −1.65, *P*<.001 and β=−6.19, 95% CI −8.82 to −3.56, *P*< .001, respectively).

XSL•FO
RenderX

Table . Multivariable linear regression predicting the total score across all key domains of patients' perspectives on automated pharmacy, adjusted for sociodemographic and other variables.

Characteristics	Regression coefficient (β)[a]	95% CI	SE	*t* statistic (*df*)	*P* value
Constant	78.69	74.42 to 82.96	2.17	36.22 (395)	<.001
Hospital					
Adan	N/A[b]	N/A	N/A	N/A	N/A
Jahraa	−1.57	−4.20 to 1.07	1.34	−1.17 (395)	.24
Farwaniya	−1.27	−3.95 to 1.41	1.36	−0.93 (395)	.35
Jaber AlAhmad	0.69	−1.88 to 3.25	1.30	0.53 (395)	.59
Age (years)					
<29	N/A	N/A	N/A	N/A	N/A
30 - 39	−0.08	−2.61 to 2.46	1.29	−0.06 (395)	.95
40 - 49	−1.40	−4.14 to 1.34	1.39	−1.01 (395)	.32
≥50	−1.83	−5.15 to 1.49	1.69	−1.08 (395)	.28
Nationality					
Kuwaiti	−0.95	−3.50 to 1.60	1.30	−0.73 (395)	.47
Non-Kuwaiti	N/A	N/A	N/A	N/A	N/A
Sex					
Male	N/A	N/A	N/A	N/A	N/A
Female	−0.68	−2.69 to 1.34	1.03	−0.66 (395)	.51
Marital status					
Single	N/A	N/A	N/A	N/A	N/A
Married	1.11	−1.14 to 3.36	1.15	0.97 (395)	.33
Education					
High school and less	N/A	N/A	N/A	N/A	N/A
Diploma	−1.23	−3.97 to 1.52	1.40	−0.88 (395)	.38
Bachelor	−2.16	−4.66 to 0.34	1.27	−1.70 (395)	.09
Higher education	−0.62	−4.93 to 3.69	2.19	−0.28 (395)	.78
Estimated number of pharmacy visits in the last year					
Once to twice	N/A	N/A	N/A	N/A	N/A
3-4 times	0.75	−1.64 to 3.13	1.21	0.62 (395)	.54
Monthly	1.74	−0.71 to 4.19	1.25	1.39 (395)	.16
Weekly	−2.96	−9.78 to 3.86	3.47	−0.85 (395)	.39
Dispense medications					
Personally	N/A	N/A	N/A	N/A	N/A
Someone else	−1.86	−3.94 to 0.21	1.06	−1.76 (395)	.08
The time I have waited to receive my medication					
≤5 minutes	N/A	N/A	N/A	N/A	N/A
6 - 15 minutes	−3.70	−5.76 to −1.65	1.04	−3.55 (395)	<.001
>15 minutes	−6.19	−8.82 to −3.56	1.34	−4.63 (395)	<.001
The number of medications in my prescription					
1 - 5 items	N/A	N/A	N/A	N/A	N/A
6 - 10 items	−0.23	−2.81 to 2.34	1.31	−0.18 (395)	.86

XSL•FO
RenderX

Characteristics	Regression coefficient (β)[a]	95% CI	SE	*t* statistic (*df*)	*P* value
>10 items	−2.58	−6.58 to 1.42	2.03	−1.27 (395)	.21

[a]Results from multivariable linear regression using the total score across all key components of patients' perspectives on automated pharmacy as a continuous outcome variable, along with hospital location, age, nationality, gender, marital status, education, number of pharmacy visits in the last year, dispensed medications, time waited to receive medication, and number of medications in the prescription. The regression coefficients (β) represent the mean difference in the overall patients' perspective score associated with each category increase, holding all other variables constant. The 95% CIs and *P* values are reported.

[b]Not applicable.

The results of the sensitivity analyses are shown in Table 9. Across all simulated dissatisfaction scenarios (−0.3 SD, −0.5 SD, and −1.0 SD), the direction, magnitude, and statistical significance of the regression coefficients remained consistent with those observed in the primary analysis (Table 8). In particular, the association between longer waiting times and lower patient perspective scores persisted across all scenarios, with nearly identical regression coefficients and 95% CIs. As expected, the model intercept decreased progressively with increasing downward shifts, reflecting lower assumed overall patient perspectives under more conservative assumptions. The results showed that there were no remarkable changes in the estimated effects of individual predictors, and no variables changed their significance status across the sensitivity analyses.

XSL•FO
RenderX

Table . Sensitivity analysis of multivariable linear regression outcomes across different dissatisfaction assumptions.

Characteristics	Mild (SD −0.3)			Moderate (SD −0.5)			Severe (SD −1.0)		
	B	95% CI	*P* value	*B*	95% CI	*P* value	*B*	95% CI	*P* value
Constant	75.9	71.6 to 80.1	<.001	74.0	69.7 to 78.3	<.001	69.3	65.0 to 73.6	<.001
Hospital									
Adan	N/A[a]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Jahraa	−1.6	−4.2 to 1.1	.24	−1.6	−4.2 to 1.1	.24	−1.6	−4.2 to 1.1	.24
Farwaniya	−1.3	−3.9 to 1.4	.35	−1.3	−3.9 to 1.4	.35	−1.3	−3.9 to 1.4	.35
Jaber AlAhmad	0.7	−1.9 to 3.3	.59	0.7	−1.9 to 3.3	.59	0.7	−1.9 to 3.3	.59
Age (years)									
<29	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30 - 39	−0.1	−2.6 to 2.5	.95	−0.1	−2.6 to 2.5	.95	−0.1	−2.6 to 2.5	.95
40 - 49	−1.4	−4.1 to 1.3	.32	−1.4	−4.1 to 1.3	.32	−1.4	−4.1 to 1.3	.32
≥50	−1.8	−5.1 to 1.5	.28	−1.8	−5.1 to 1.5	.28	−1.8	−5.1 to 1.5	.28
Nationality									
Kuwaiti	−0.9	−3.5 to 1.6	.47	−0.9	−3.5 to 1.6	.47	−0.9	−3.5 to 1.6	.47
Non-Kuwaiti	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Sex									
Male	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Female	−0.7	−2.7 to 1.3	.51	−0.7	−2.7 to 1.3	.51	−0.7	−2.7 to 1.3	.51
Marital status									
Single	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Married	1.1	−1.1 to 3.4	.33	1.1	−1.1 to 3.4	.33	1.1	−1.1 to 3.4	.33
Education									
High school and less	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Diploma	−1.2	−4.0 to 1.5	.38	−1.2	−4.0 to 1.5	.38	−1.2	−4.0 to 1.5	.38
Bachelor	−2.2	−4.7 to 0.3	.09	−2.2	−4.7 to 0.3	.09	−2.2	−4.7 to 0.3	.09
Higher education	−0.6	−4.9 to 3.7	.78	−0.6	−4.9 to 3.7	.78	−0.6	−4.9 to 3.7	.78
Estimated number of pharmacy visits in the last year									
Once to twice	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3-4 times	0.7	−1.6 to 3.1	.54	0.7	−1.6 to 3.1	.538	0.7	−1.6 to 3.1	.54
Monthly	1.7	−0.7 to 4.2	.16	1.7	−0.7 to 4.2	.164	1.7	−0.7 to 4.2	.16
Weekly	−3.0	−9.8 to 3.9	.39	−3.0	−9.8 to 3.9	.394	−3.0	−9.8 to 3.9	.39
Dispense medications									
Personally	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Someone else	−1.9	−3.9 to 0.2	.08	−1.9	−3.9 to 0.2	.08	−1.9	−3.9 to 0.2	.08
The time I have waited to receive my medication									
≤5 minutes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

XSL•FO
RenderX

Characteristics	Mild (SD −0.3)			Moderate (SD −0.5)			Severe (SD −1.0)		
	B	95% CI	*P* value	*B*	95% CI	*P* value	*B*	95% CI	*P* value
6 - 15 minutes	−3.7	−5.8 to −1.7	<.001	−3.7	−5.8 to −1.7	<.001	−3.7	−5.8 to −1.7]	<.001
>15 minutes	−6.2	−8.8 to −3.6	<.001	−6.2	−8.8 to −3.6	<.001	−6.2	−8.8 to −3.6]	<.001
The number of medications in my prescription									
1 - 5 items	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6 - 10 items	−0.2	−2.8 to 2.3	.86	−0.2	−2.8 to 2.3	.86	−0.2	−2.8 to 2.3	.86
>10 items	−2.6	−6.6 to 1.4	.21	−2.6	−6.6 to 1.4	.21	−2.6	−6.6 to 1.4	.21

[a]Not applicable.

Discussion

Principal Findings

Studying patients' opinions and satisfaction provides a clear picture of the perceived impact and success of implementing automated pharmacy systems in the real world. Patient feedback can help determine whether the system meets their needs, improves the quality of outpatient pharmacy services, and achieves its intended goals in practice.

The existing literature has shown numerous advantages of using such technology in pharmacy practice. In this context, this study revealed that patients perceived integrated e-prescribing and e-dispensing systems as beneficial to outpatient pharmacy services and expressed overall satisfaction with their experiences of care, as reported by 95% (393/416) of the participants. Participants cited several key perceived benefits, including time savings, perceived prevention of medication errors, convenient access to prescription information at any time, and effective communication between pharmacists and physicians. These findings are consistent with previous studies, which have similarly reported a perceived reduction in medication errors [1-3], time savings in the prescribing and dispensing process [5,6], and increased work efficiency [1,2,6] following the adoption of automated pharmacy systems.

Influence of Sociodemographic Data on Participant Satisfaction

The findings revealed that the satisfaction levels of the participants varied across several sociodemographic groups. This study found an indirect association between age and medication counseling and between dispensing medications score and educational level. This could be explained by the notion that people with higher education levels could have higher expectations of health care services, which could lead to dissatisfaction when these are unmet [24]. Similar findings were found in a previous study [25], which assessed patient satisfaction toward outpatient pharmacy services among outpatients attending public health clinics in Malaysia. The findings revealed that demographic factors (eg, age, education level, self-perceived health status, and general knowledge of pharmacists) significantly influenced patients' perceptions of health care services. Notably, older age and higher education levels were associated with a lower patient satisfaction mean score [25]. Similarly, a study conducted in Pakistan [26] reported that male patients were significantly more satisfied with community outpatient pharmacy services than females. Younger patients (aged 18 - 27 years) exhibited greater satisfaction than older age groups. Illiterate and less-educated individuals showed greater satisfaction levels, while those with higher education levels were less satisfied, likely due to higher expectations. Marital status and occupation also played a potential role, with married individuals showing greater satisfaction and government employees showing less satisfaction. No significant association was observed among students after statistical adjustment [26].

Patients' Satisfaction With Dispensing Time, Pharmacists' Attitudes, and Medication Supply

Although most patients in this study agreed that their prescriptions were dispensed in a reasonable time, the findings revealed that the patients with more medications experienced longer waiting times, suggesting challenges in processing large or complex orders using the automated pharmacy systems.

These findings aligned with those of a previous study into primary health care settings in Kuwait [27], which found that patients' satisfaction was strongly influenced by the timeliness of prescription fulfillment by physicians. Numerous studies have demonstrated that e-prescription systems reduce dispensing times and improve pharmacy workflow [28]. In this context, a study assessed the impact of e-prescription systems on pharmacy workflow and practice in Finland and found a reduction in both median and average delivery times across various prescription types [29]. In particular, the maximum dispensing time with the adoption of e-prescribing systems was shorter in 2012 (6 minutes and 22 seconds) than the time with the use of non–e-prescribing systems in 2006 (12 minutes and 40 seconds). The study concluded that e-prescribing systems significantly improved the efficiency of prescription processing in Finnish community pharmacies [29]. Additionally, a study conducted in Sweden into the impact of integrated e-prescribing systems on pharmacists' practice found that the pharmacists perceived the system as expediting prescription processing and reducing the risk of medication errors [1]. The results showed a positive

XSL•FO
RenderX

correlation between the perceived usefulness of the electronic system for improving work efficiency and medication safety.

This study revealed that the participants were highly satisfied with the pharmacists' attitudes. Nearly all the participants agreed that the pharmacists helped obtain their medications, effectively resolved issues related to medication access, answered their questions, demonstrated knowledge of their medical conditions, and treated them with dignity and respect. These findings were supported by previous studies showing that e-prescribing systems enhance communication between patients and pharmacists [7], patients and prescribers [5], and pharmacists and other health care providers [30,31]. Similarly, a study conducted in Qatar reported comparable levels of satisfaction among participants regarding pharmacists' empathy and communication skills [16]. Furthermore, effective communication, regardless of whether systems are manual or electronic, has been found to significantly contribute to overall patient satisfaction [7,32].

The findings of this study demonstrated statistically significant differences in patient satisfaction across outpatient pharmacies in the 4 general hospitals, particularly regarding medication dispensing speed and pharmacists' attitudes. These site-level variations may reflect differences in staffing levels, staff competencies, workload, maturity of automation, or local operational workflows. For instance, the higher satisfaction with medication dispensing speed observed at Farwaniya Hospital compared with Adan Hospital may be attributable to shorter waiting times, fewer medications per prescription, or more efficient dispensing practices. While patient satisfaction is an important indicator of service performance, it should be interpreted alongside technical and organizational factors to provide a comprehensive understanding of the effectiveness and efficiency of digital pharmacy systems.

The findings of this study indicated that approximately one-quarter of the surveyed participants experienced medication unavailability. Similar findings have been reported by previous studies [33-36], highlighting the need for improved inventory integration to overcome such issues and enhance patient satisfaction. In this context, a study showed that automated drug-dispensing systems carry advantages over traditional systems in improving inventory control and reducing storage errors, resulting in better patient satisfaction [37,38].

The participants in this study were satisfied with the perceived accuracy of the pharmacy's handling of medical prescription information. In support of the claim that prescription accuracy contributes to patient satisfaction, a systematic review study revealed that the integration of e-prescription and e-dispensing systems significantly reduced pharmacy errors [39]. This indicates that the adoption of automated systems may enhance perceived accuracy and is associated with higher levels of patient satisfaction and perceived safety.

Patients' Satisfaction With Medication Counseling and the Dispensing Process

Despite the growing use of e-prescription and e-dispensing systems, challenges remain in their utilization. The findings of this study indicated that the patients reported high levels of satisfaction with medication counseling overall. Specifically, pharmacists were commended for clearly explaining the purpose of medication use, providing accurate dosage instructions, and allocating sufficient time for patients during consultations. However, approximately half of the participants expressed dissatisfaction with the counseling received regarding potential drug side effects and proper medication storage.

A US study into outpatient use of electronic prescriptions found that the use of an e-prescribing system led to lower rates of adverse drug events among patients, suggesting that appropriate counseling on side effects and storage could further reduce such risks [40]. Therefore, in this study, the lack of counseling regarding the side effects of medications and storage conditions indicates a potential gap in quality of practice. This may pertain to numerous reasons, such as a long queue of patients waiting, leading to time pressures, or pharmacists' limited counseling skills, which would affect the quality of the counseling. This shortcoming may reflect the need for additional pharmacist training to improve their counseling skills and not neglect any important information from patient education. Previous research has shown that training physicians in e-prescribing significantly reduces errors and improves workflow, thus indicating that e-prescription systems alone do not resolve counseling-related challenges unless supplemented by targeted education [41]. Addressing these aspects could significantly enhance both patient satisfaction and medication safety.

In addition, despite the automation of the dispensing process, the findings of this study reported longer waiting times for patients with prescribed multiple medications. This association suggests that prolonged waiting times may negatively influence patients' overall perception of outpatient pharmacy domains, including pharmacists' attitudes, medication counseling, medication supply, and dispensing medications. The literature showed that previous studies are consistent with the findings of this study, indicating that waiting time is a critical determinant of patient satisfaction, irrespective of whether outpatient pharmacy services are delivered through conventional or automated systems [38,42,43]. Therefore, the findings revealed that the success of automated outpatient pharmacy services relies not only on technology use but also on efficient workflow that led to suitable time medication dispensing.

This finding highlights the need for process optimization within pharmacy workflows. As with counseling gaps, delays in dispensing may diminish the perceived benefits of automation. Similar studies conducted in Saudi Arabia [38,44] emphasized the value of integrated pharmacy systems while acknowledging the need for region-specific improvements. These studies support the idea that continuous pharmacist education and the implementation of structured feedback mechanisms are essential for identifying and addressing system inefficiencies. By incorporating such strategies, pharmacies can better meet patient expectations and improve the outpatient pharmacy services.

Impact of E-Prescribing and E-Dispensing Systems on Outpatient Pharmacy Services

This study found that most patients reported a positive impact based on their experiences with the medication dispensing process, with pharmacist communication emerging as a key

XSL•FO
RenderX

contributor to patient satisfaction. These findings are consistent with those of a study in Qatar in 2013, which emphasized the importance of pharmacist-patient communication [16]. While pharmacy staff provide irreplaceable expertise in prescription monitoring and dispensing, the integration of electronic systems may help alleviate workload and reduce cognitive burden. Participants perceived that reducing manual tasks enhanced efficiency and contributed to improved patient safety by minimizing the risk of human error [45]. Pharmacists and pharmacy technicians are susceptible to fatigue, stress, and other human limitations, which may compromise dispensing accuracy. In this context, e-prescription systems serve as a valuable support tool that may help mitigate such risks [18,45]. By providing an additional layer of safety and reliability, these systems can contribute to patients' perceptions of positive outcomes and higher satisfaction levels.

The absence of a comparison group involving nonautomated outpatient pharmacies is considered a key methodological limitation of this study. Consequently, the findings cannot isolate the specific effects of automation from other contextual or organizational factors and should be interpreted as reflecting patient experiences within automated pharmacy settings only, which limits their generalizability to other pharmacy models. Future studies using comparative or quasi-experimental designs are needed to evaluate the incremental impact of automation on patient experience and service performance.

Limitations

This study has several limitations. First, it focused exclusively on patients in Kuwait. While the findings provide valuable insights into local users of outpatient pharmacy services, the generalizability of the results to other populations and health care systems may be limited. In addition, the absence of a comparison group involving nonautomated outpatient pharmacies limits the ability to isolate the specific impact of automation.

Second, patient satisfaction data were self-reported, which may have introduced recall or self-reporting bias and may not accurately reflect the actual quality or safety of outpatient pharmacy services. Third, the presence of pharmacy students in the waiting area may have introduced courtesy or social desirability bias, which may have potentially inflated satisfaction scores (approximately 95% agreement), despite private survey completion. The consistently high levels of agreement observed may therefore partly reflect response tendencies rather than true satisfaction levels. Accordingly, future studies should consider alternative data collection approaches or comparative analyses across sites or time periods to better assess this potential bias. Fourth, despite the sensitivity analyses conducted, voluntary participation and convenience sampling may have introduced selection bias, as patients who were more satisfied or more engaged with automated outpatient pharmacy services may have been more likely to participate. Although the sensitivity analyses suggest that potential underrepresentation of less satisfied patients is unlikely to have materially altered the study conclusions, the findings may not be fully representative of all users of automated outpatient pharmacy systems in Kuwait, particularly individuals who actively avoid or disengage from these systems. Fifth, the study was conducted in 4 governmental hospitals and did not include private hospitals or pharmacies, which may differ in service quality, system implementation, and operational policies. Sixth, the cross-sectional design captured perceptions at a single point in time, limiting the ability to assess changes in patient-reported satisfaction over time.

Finally, although responses were obtained from both patients and caregivers, caregivers represented a smaller proportion of the sample (104/416, 25%). An exploratory comparison using the Mann-Whitney *U* test did not detect statistically significant differences between the 2 groups; therefore, further subgroup analyses were not pursued.

Future Research

Future research should link patient-reported perceptions with objective safety and quality indicators, such as medication dispensing error rates, adverse drug events, dispensing accuracy, workflow measures, and prescription turnaround time, to determine whether perceived benefits correspond to measurable improvements in patient outcomes. Longitudinal and mixed methods designs, including qualitative interviews, may provide deeper insight into how and why perceived benefits evolve over time and whether anticipated benefits are realized in routine practice. Methodologically, probability-based sampling approaches should be considered to improve representativeness and reduce selection bias. In addition, remote or independent survey administration and comparative designs across data collection settings may help assess and minimize social desirability bias. Comparative studies involving governmental and private sector hospitals or pharmacies are also warranted, given potential differences in implementation maturity, staffing, workflows, and service policies. Finally, future studies specifically powered to compare patient and caregiver perspectives, alongside technical and organizational assessments, would provide a more comprehensive evaluation of system effectiveness and help identify actionable improvement strategies.

Conclusions

This study demonstrated that integrated e-prescribing and e-dispensing systems offer clear perceived benefits in outpatient pharmacies within governmental hospitals in Kuwait. This is reflected in the positive experiences reported by patients, reflecting overall perceived satisfaction and perceived service performance within the operating systems.

However, the findings also highlight areas requiring improvement to maximize patients' satisfaction toward pharmacy automation systems. The following clinical implications are made:

1. Workflow design optimization: operational strategies should be implemented to improve pharmacy workflow design within automated pharmacy settings to enhance work efficiency, especially for processing lengthy prescriptions and managing drug inventories to reduce dispensing delays, as well as making sure that a counseling area is available for patient education.

XSL•FO
RenderX

2. E-prescribing system enhancement: counseling prompts should be integrated to ensure that important information is delivered completely and properly to patients.
3. Automated inventory management: enhanced inventory control should be implemented, including proactive stock monitoring to clarify what is available and what is near expiry to improve the supply chain management. This is essential because the results showed that approximately 25% (105/416) of patients reported medication unavailability.
4. Staff training: pharmacy staff should receive comprehensive training on best practices for using integrated e-prescription and e-dispensing systems. This would improve patient counseling, interprofessional communication, patient safety, and overall patient satisfaction.
5. Feedback mechanisms: regular collection of feedback from both patients and pharmacy staff should be instituted to identify and address issues related to waiting times, medication availability, and the quality of counseling services.

By addressing these identified gaps, pharmacists can enhance patient satisfaction and improve the overall performance of outpatient pharmacy services supported by integrated digital systems. Strengthening the digital infrastructure and workflow integration would enable hospitals to optimize outpatient pharmacy services and move toward patient-centered digital pharmacy models that support efficient, reliable, and safe medication services. These findings have relevance beyond the study setting and may inform the development of digital outpatient pharmacy services across hospitals in the Arabian Gulf region.

Acknowledgments

The authors would like to thank the students from the College of Pharmacy at Kuwait University who undertook the data collection: Yousef Bahman, Ritaj Alawadhi, Sadeem Almubarak, Nourah Alshammeri, Zainab Moradi, and Sara Alorf.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The impact of integrated e-prescribing and e-dispensing systems on outpatient pharmacy services.
[PNG File, 49 KB - medinform_v14i1e80963_app1.png]

Multimedia Appendix 2
Overall satisfaction with automated outpatient pharmacy services.
[PNG File, 5 KB - medinform_v14i1e80963_app2.png]

References

1. Rahimi B, Timpka T. Pharmacists' views on integrated electronic prescribing systems: associations between usefulness, pharmacological safety, and barriers to technology use. Eur J Clin Pharmacol 2011 Feb;67(2):179-184. [doi: 10.1007/s00228-010-0936-9] [Medline: 21069521]
2. Abdel-Qader DH, Cantrill JA, Tully MP. Satisfaction predictors and attitudes towards electronic prescribing systems in three UK hospitals. Pharm World Sci 2010 Oct;32(5):581-593. [doi: 10.1007/s11096-010-9411-z] [Medline: 20593237]
3. Bright HR, Peter JV, Chandy SJ. Electronic prescribing system in a Teaching Hospital-user satisfaction and factors affecting successful implementation. Der Pharmacia Letter 2019;11(2):10-24 [FREE Full text]
4. Kauppinen H, Ahonen R, Timonen J. The impact of electronic prescriptions on medication safety in Finnish community pharmacies: a survey of pharmacists. Int J Med Inform 2017 Apr;100:56-62. [doi: 10.1016/j.ijmedinf.2017.01.014] [Medline: 28241938]
5. Hammar T, Nyström S, Petersson G, Rydberg T, Åstrand B. Swedish pharmacists value ePrescribing: a survey of a nationwide implementation. J Phara Health Ser Res 2010 Mar;1(1):23-32. [doi: 10.1211/jphsr.01.01.0012]
6. McLeod M, Karampatakis GD, Heyligen L, McGinley A, Franklin BD. The impact of implementing a hospital electronic prescribing and administration system on clinical pharmacists' activities—a mixed methods study. BMC Health Serv Res 2019 Mar 12;19(1):156. [doi: 10.1186/s12913-019-3986-4] [Medline: 30866925]
7. Gagnon MP, Nsangou É, Payne-Gagnon J, Grenier S, Sicotte C. Barriers and facilitators to implementing electronic prescription: a systematic review of user groups' perceptions. J Am Med Inform Assoc 2014;21(3):535-541. [doi: 10.1136/amiajnl-2013-002203] [Medline: 24130232]

XSL•FO
RenderX

8. Beard RJ. Electronic prescribing and robotic dispensing: the impact of integrating together on practice and professionalism. In: Dekoulis G, editor. Robotics—Legal, Ethical and Socioeconomic Impacts: Intech Open; 2017. [doi: 10.5772/intechopen.69769]
9. Tsyben A, Gooding N, Kelsall W. Assessing the impact of a newly introduced electronic prescribing system across a paediatric department— lessons learned. Arch Dis Child 2016 Sep;101(9):e2. [doi: 10.1136/archdischild-2016-311535.20] [Medline: 27540199]
10. Marotti R. Can we translate the UK automation experience? Pharm Pract Res 2016 Sep;46(3):253-255. [doi: 10.1002/jppr.1184]
11. Almutairi BA, Potts HWW, Al-Azmi SF. Physicians' perceptions of electronic prescribing with electronic medical records in Kuwaiti primary healthcare centres. Sultan Qaboos Univ Med J 2019;18(4):e476-e482. [doi: 10.18295/squmj.2018.18.04.008]
12. Buabbas A, Al-shawaf H, editors. Pharmacists perspectives on the effectiveness of primary care information system on pharmacy practice in Kuwait health care centres. Presented at: Proceedings of the 6th International Symposium on Health Informatics and Bioinformatics; May 2-5, 2011. [doi: 10.1109/HIBIT.2011.6450800]
13. Al-Otaibi J, Tolma E, Alali W, Alhuwail D, Aljunid SM. The factors contributing to physicians' current use of and satisfaction with electronic health records in Kuwait's public health care: cross-sectional questionnaire study. JMIR Med Inform 2022 Oct 7;10(10):e36313. [doi: 10.2196/36313] [Medline: 36206039]
14. Al-Jafar E. Exploring patient satisfaction before and after electronic health record (EHR) implementation: the Kuwait experience. Perspect Health Inf Manag 2013;10(Spring):1c. [Medline: 23805063]
15. Al-Azmi SF, Mohammed AM, Hanafi MI. Patients' satisfaction with primary heath care in Kuwait after electronic medical record implementation. J Egypt Public Health Assoc 2006;81(5-6):277-300. [Medline: 18706302]
16. Khudair IF, Raza SA. Measuring patients' satisfaction with pharmaceutical services at a public hospital in Qatar. Int J Health Care Qual Assur 2013;26(5):398-419. [doi: 10.1108/IJHCQA-03-2011-0025] [Medline: 23905301]
17. Frail CK, Kline M, Snyder ME. Patient perceptions of e-prescribing and its impact on their relationships with providers: a qualitative analysis. J Am Pharm Assoc (2003) 2014 Nov;54(6):630-633. [doi: 10.1331/JAPhA.2014.13176]
18. Odukoya OK, Stone JA, Chui MA. E-prescribing errors in community pharmacies: exploring consequences and contributing factors. Int J Med Inform 2014 Jun;83(6):427-437. [doi: 10.1016/j.ijmedinf.2014.02.004] [Medline: 24657055]
19. Polit DF, Beck CT. Nursing Research Generating and Assessing Evidence for Nursing Practice, 10th edition: Wolters Kluwer; 2017.
20. Moridi M, Pazandeh F, Hajian S, Potrata B. Development and psychometric properties of Midwives' Knowledge and Practice Scale on Respectful Maternity Care (MKP-RMC). PLoS One 2020;15(11):e0241219. [doi: 10.1371/journal.pone.0241219] [Medline: 33141835]
21. Fu E, White MA, Hughto JMW, Steiner B, Willis EA. Development of the physical activity tracking preference questionnaire. Int J Exerc Sci 2019;12(5):297-309. [doi: 10.70252/WEJR4154] [Medline: 31156755]
22. Ekolu SO, Quainoo H. Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods. Global J Eng Educ 2019;21(1):24-29 [FREE Full text]
23. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 1988.
24. Jackson JL, Chamberlin J, Kroenke K. Predictors of patient satisfaction. Soc Sci Med 2001 Feb;52(4):609-620. [doi: 10.1016/s0277-9536(00)00164-7] [Medline: 11206657]
25. Ismail A, Gan YN, Ahmad N. Factors associated with patient satisfaction towards pharmacy services among out-patients attending public health clinics: questionnaire development and its application. PLoS One 2020;15(11):e0241082. [doi: 10.1371/journal.pone.0241082] [Medline: 33170854]
26. Alanazi AS, Shah S, Abbas G, et al. Assessing patient satisfaction with community pharmacy services: a large regional study at Punjab, Pakistan. Patient Prefer Adherence 2023;17:13-22. [doi: 10.2147/PPA.S389053] [Medline: 36636289]
27. Alhashem AM, Alquraini H, Chowdhury RI. Factors influencing patient satisfaction in primary healthcare clinics in Kuwait. Int J Health Care Qual Assur 2011;24(3):249-262. [doi: 10.1108/09526861111116688] [Medline: 21938970]
28. Porterfield A, Engelbert K, Coustasse A. Electronic prescribing: improving the efficiency and accuracy of prescribing in the ambulatory care setting. Perspect Health Inf Manag 2014;11(Spring):1g. [Medline: 24808808]
29. Peltoniemi T, Suomi R, Peura S, Lähteenoja MNY. Electronic prescription as a driver for digitalization in Finnish pharmacies. BMC Health Serv Res 2021 Sep 27;21(1):1017. [doi: 10.1186/s12913-021-07003-0] [Medline: 34565354]
30. Campbell C, Morris C, McBain L. Electronic transmission of prescriptions in primary care: transformation, timing and teamwork. J Prim Health Care 2021 Dec;13(4):340-350. [doi: 10.1071/HC21050] [Medline: 34937647]
31. Darwesh BM, Machudo SY, John S. The experience of using an automated dispensing system to improve medication safety and management at King Abdul Aziz University Hospital. J Pharm Pract Community Med 2017 Jun 15;3(3):114-119. [doi: 10.5530/jppcm.2017.3.26]
32. Callaway Kim K, Rothenberger SD, Tadrous M, et al. Drug shortages prior to and during the COVID-19 pandemic. JAMA Netw Open 2024 Apr 1;7(4):e244246. [doi: 10.1001/jamanetworkopen.2024.4246] [Medline: 38578641]
33. Paulo Mário Elobo M. Factors influencing drug stock-outs at health facilities—Mocuba. MOJ Biol Med 2024;9(1):8-10. [doi: 10.15406/mojbm.2024.09.00209]

XSL•FO
RenderX

34. Phuong JM, Penm J, Chaar B, Oldfield LD, Moles R. The impacts of medication shortages on patient outcomes: a scoping review. PLoS One 2019;14(5):e0215837. [doi: 10.1371/journal.pone.0215837] [Medline: 31050671]
35. Atif M, Malik I, Mushtaq I, Asghar S. Medicines shortages in Pakistan: a qualitative study to explore current situation, reasons and possible solutions to overcome the barriers. BMJ Open 2019 Sep 4;9(9):e027028. [doi: 10.1136/bmjopen-2018-027028] [Medline: 31488466]
36. Amirthalingam P, Alruwaili AS, Albalawi OA, et al. Comparing patient satisfaction with automated drug dispensing system and traditional drug dispensing system: a cross-sectional study. Patient Prefer Adherence 2024;18:2337-2345. [doi: 10.2147/PPA.S492802] [Medline: 39583135]
37. Oldland AR, Golightly LK, May SK, Barber GR, Stolpman NM. Electronic inventory systems and barcode technology: impact on pharmacy technical accuracy and error liability. Hosp Pharm 2015 Jan;50(1):34-41. [doi: 10.1310/hpj5001-34] [Medline: 25684799]
38. Momattin H, Arafa S, Momattin S, Rahal R, Waterson J. Robotic pharmacy implementation and outcomes in Saudi Arabia: a 21-month usability study. JMIR Hum Factors 2021 Sep 1;8(3):e28381. [doi: 10.2196/28381] [Medline: 34304149]
39. Osmani F, Arab-Zozani M, Shahali Z, Lotfi F. Evaluation of the effectiveness of electronic prescription in reducing medical and medical errors (systematic review study). Ann Pharm Fr 2023 May;81(3):433-445. [doi: 10.1016/j.pharma.2022.12.002] [Medline: 36513154]
40. Gabriel MH, Powers C, Encinosa W, Bynum JPW. E-prescribing and adverse drug events: an observational study of the Medicare part D population with diabetes. Med Care 2017 May;55(5):456-462. [doi: 10.1097/MLR.0000000000000684] [Medline: 28060051]
41. Craxford S, Taylor L, Duguid A, Shivji F, Pickering S. The learning curve of electronic prescribing. Br J Hosp Med 2015 Sep 2;76(9):538-540. [doi: 10.12968/hmed.2015.76.9.538]
42. Leemanza Y, Kristin E. Effectiveness of wait time targets and patient satisfaction feedback in decreasing wait times for prescription services in an outpatient pharmacy. Can J Hosp Pharm 2024;77(1):e3438. [doi: 10.42112/cjhp.3438] [Medline: 38204510]
43. Yulia R, Hartono R, Indrayanti M, Ayumuyas NP, Herawati F. Studying waiting time in pharmacy: a strategy for improving patient satisfaction. MethodsX 2025 Jun;14:103282. [doi: 10.1016/j.mex.2025.103282] [Medline: 40230556]
44. Almaghaslah D, Alsayari A, Almaghaslah S, Alsanna H. Patients' satisfaction with e-prescribing (Wasfaty) in Saudi Arabia: a survey of country-level implementation. Healthcare (Basel) 2022 Apr 27;10(5):806. [doi: 10.3390/healthcare10050806] [Medline: 35627943]
45. Whitaker M, Lester C, Rowell B. Handing off electronic prescription data from prescribers to community pharmacies: a qualitative analysis of pharmacy staff perspectives. J Patient Saf 2024 Sep 1;20(6):397-403. [doi: 10.1097/PTS.0000000000001244] [Medline: 38742931]

Abbreviations

e-dispensing: electronic dispensing
e-prescribing: electronic prescribing
HIS: hospital information system
KMO: Kaiser-Meyer-Olkin

Edited by A Sen; submitted 19.Jul.2025; peer-reviewed by C Duclos, H Rajesh, M Hany; revised version received 22.Jan.2026; accepted 23.Jan.2026; published 06.Mar.2026.

Please cite as:
Buabbas AJ, Alsaleh FM, Alsanafi M, Safar MA, Al-Sultan AT, Zaman TA
Patient Experiences With Outpatient Pharmacy Services in Hospitals Using Automated Pharmacy Systems: Cross-Sectional Study
JMIR Med Inform 2026;14:e80963
URL: https://medinform.jmir.org/2026/1/e80963
doi:10.2196/80963

© Ali Jasem Buabbas, Fatemah Mohammad Alsaleh, Mariam Alsanafi, Maryam Ali Safar, Ahmad T Alsultan, Talal Ahmad Zaman. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Scribes: Are We Measuring What Matters?

Enrico Coiera, PhD; David Fraile-Navarro, PhD

Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, Sydney, NSW, Australia

Corresponding Author:
Enrico Coiera, PhD
Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, Sydney, NSW, Australia

Abstract

Artificial intelligence (AI) scribes, software that can convert speech into concise clinical documents, have achieved remarkable clinical adoption at a pace rarely seen for digital technologies in health care. The reasons for this are understandable: the technology works well enough, it addresses a genuine pain point for clinicians, and it has largely sidestepped regulatory requirements. In many ways, clinical adoption of AI scribes has also occurred well ahead of robust evidence of their safety and efficacy. The papers in this theme issue demonstrate real progress in the technology and evidence of its benefit: documentation times are reported to decrease when using scribes, clinicians report feeling less burdened, and the notes produced are often of reasonable quality. Yet as we survey the emerging evidence base, there remains one outstanding and urgent unanswered question: Are AI scribes safe? We need to know the clinical outcomes achievable when scribes are used compared to other forms of note taking.

(JMIR Med Inform 2026;14:e89337) doi:10.2196/89337

KEYWORDS

artificial intelligence; AI; digital scribes; medical scribes; electronic documentation; patient safety; error prevention

Beyond Efficiency

Early evaluation of any technology naturally focuses on its primary promise, in the case of scribes [1], reducing documentation burden. The studies in this collection confirm that artificial intelligence (AI) scribes deliver on this front. Kanaparthy and colleagues' [2] rapid review found a general trend toward reduced self-reported documentation time and improvements in clinician satisfaction. The comparative analysis by Ha and colleagues [3] demonstrates that current commercial systems can generate reasonable-quality SOAP (subjective, objective, assessment, and plan) notes in about a minute after a 15-minute encounter. These are meaningful findings. Having shown that scribes likely save time in the settings they have so far been evaluated in (eg, primary care and outpatient settings), we can turn to harder questions of safety, clinical reasoning, and wider system-level effects.

Toward Systematic Harm Measurement

How safe are clinical scribes? Do they make errors, what type of errors might see when using digital scribes, and are these errors clinically consequential? What are the causes of errors, and what then are the harm mitigations we need to put into place?

Several papers in this collection begin this work. Ha and colleagues [3] highlight that none of the systems they evaluated are error free. Biro and colleagues [4] have developed and validated an instrument specifically designed to assess the accuracy and safety of AI scribe outputs, an essential foundation since, without standardized measurement, we cannot compare across systems or track changes over time. Their early work confirms that AI scribes do make errors, and some have patient safety implications.

Digital technologies have the capacity to both reduce human error and generate new error classes [5]. This raises another question worthy of further investigation: Are digital scribe errors equivalent to human errors, or do they have different risk profiles? A human might lose attention momentarily; a digital scribe will not fatigue but may misrecognize words through speech recognition errors. An AI might confidently fabricate a medication or symptom that was never mentioned or omit clinically significant details in the pursuit of conciseness, each reshaping the clinical narrative in different ways.

Equally, we should recognize that the status quo carries its own safety risks: clinician burnout and cognitive overload contribute to errors that scribes may help reduce. A complete safety evaluation must weigh new risks introduced by AI against existing risks that the technology may mitigate and consider what analogous error-detection and correction mechanisms we need to build for AI-generated documentation.

Clinical Reasoning When Using a Digital Scribe: An Underexplored Frontier

Perhaps the richest opportunity for future research relates to the quality of clinical care when documentation is outsourced to AI. Note-taking is not merely administrative work. When a clinician summarizes their thoughts into a document, they are actively processing information, prioritizing what matters, and forming and testing hypotheses in real time. The clinical note

XSL•FO
RenderX

is not just a record of the consultation; it is a cognitive artifact that supports clinical reasoning [6].

What happens to this human sense-making process when documentation is delegated to an AI? We may be freeing the attention of clinicians to be more present with patients, or we may be altering how they would normally think. The Y-KNOT (Your-Knowledgeable Navigator of Treatment) implementation study offers an intriguing signal: expert ratings suggested that while most AI-generated drafts were rated positively, around 1 in 6 preanesthetic assessments were judged to have a negative impact on clinical decision-making [7]. As we document faster, are we also documenting differently? Does that difference matter for patient care? If so, can we mitigate the potential for harm, for example, through clinical training or redesigning the user interaction with a scribe to bring clinicians back into the document-and-reason loop?

Ecosystem Perspectives

AI scribes sit at a critical position within a clinical workflow. They determine what gets recorded and how it is structured, which in turn shapes what downstream systems, including other AI tools, will see and act on. In this sense, scribes are a gateway technology, creating data layers that propagate through the health system. As we have argued elsewhere in the context of generative AI more broadly, it is helpful to view these technologies through an ecosystem lens that emphasizes system-level properties over isolated components [8]. With this perspective, we can evaluate the broader scribe ecosystem against system-level dimensions: resilience (how does care adapt when the scribe fails?), sustainability (what happens when cloud-based systems change or disappear?), and service interactions (does optimizing documentation affect other aspects of care?).

The patients in Leiserowitz and colleagues' [9] survey were generally open to an AI scribe when it was framed as supporting clinician focus. Their study also showed one interaction worth monitoring: a meaningful proportion of patients indicated they might withhold sensitive information if an always-listening device was present. Understanding these broader system effects will require looking beyond the scribe itself to the clinical environment it inhabits.

Regulatory Evolution

There are several reasons why we still lack safety evidence. AI scribes have proliferated [10] in part because they sit outside traditional medical device classification [11]. Many commercial scribes skirt the software as a medical device definition of a decision support system and so have evaded regulation. With no regulatory demands for robust safety evaluation, there is little commercial value, it seems, in publishing commercial safety data. The evidence in this collection suggests that we may need new regulatory thinking, not because current systems are demonstrably unsafe, but because they are different. These systems do influence clinical decisions, and unlike traditional medical devices, they are not static; the large language model underlying a scribe may be updated or retrained over time. Regulatory frameworks designed for deterministic, frozen technologies may need to evolve alongside the technology itself. The question is not whether to regulate but how to do so in ways that preserve innovation while ensuring ongoing safety. Potential mechanisms might include postmarket surveillance requirements, mandatory incident reporting for generative medical AI, or periodic re-evaluation as underlying models are updated.

Conclusion

The papers in this theme issue represent important progress. AI scribes appear to deliver on their core promise of reducing documentation burden, and the field is developing increasingly rigorous evaluation approaches. Building on this foundation, we see an opportunity to expand what we measure: systematic assessment of errors and harms, investigation of effects on clinical reasoning, and attention to ecosystem-level dynamics. Documentation burden is real, and technologies that address it are welcome. The next chapter of research can help ensure that the time saved translates into better care.

Funding

This work was supported by the National Health and Medical Research Council (NHMRC) Centre for Research Excellence in Digital Health. EC and DF-N are supported by an NHMRC investigator grant (GNT2008645). The funders had no role in this manuscript. A large language model (Claude, Anthropic) was used to assist with editing this manuscript.

Authors' Contributions

Conceptualization: EC
Writing – original draft: EC, DF-N
Writing – review & editing: EC, DF-N

Conflicts of Interest

None declared.

References

1. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. NPJ Digit Med 2018;1(1):58. [doi: 10.1038/s41746-018-0066-9] [Medline: 31304337]

XSL•FO
RenderX

2. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes: rapid review. JMIR AI 2025 Oct 10;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
3. Ha E, Choon-Kon-Yune I, Murray L, et al. Evaluating the usability, technical performance, and accuracy of artificial intelligence (AI) scribes for primary care: a competitive analysis. JMIR Hum Factors 2025 Jul 23;12:e71434. [doi: 10.2196/71434] [Medline: 40700466]
4. Biro J, Handley JL, Cobb NK, et al. Accuracy and safety of AI-enabled scribe technology: instrument validation study. J Med Internet Res 2025 Jan 27;27:e64993. [doi: 10.2196/64993] [Medline: 39869899]
5. Magrabi F, Ong MS, Runciman W, Coiera E. An analysis of computer-related patient safety incidents to inform the development of a classification. J Am Med Inform Assoc 2010;17(6):663-670. [doi: 10.1136/jamia.2009.002444] [Medline: 20962128]
6. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform 2025 Aug 1;13:e80898. [doi: 10.2196/80898] [Medline: 40749188]
7. Kim J, Lee SY, You SC, et al. A bilingual on-premises AI agent for clinical drafting: implementation report (Y-KNOT project). JMIR Med Inform 2025 Nov 24;13:e76848. [Medline: 41284981]
8. Coiera E, Fraile-Navarro D. AI as an ecosystem — ensuring generative AI is safe and effective. NEJM AI 2024 Aug 22;1(9). [doi: 10.1056/AIp2400611]
9. Leiserowitz G, Mansfield J, MacDonald S, Jost M. Patient attitudes toward ambient voice technology to support an AI scribe program. JMIR Med Inform 2025 Nov 27;13:e77901. [doi: 10.2196/77901] [Medline: 41308194]
10. Henry TA. 2 in 3 physicians are using health AI—up 78% from 2023. American Medical Association. 2025 Feb 26. URL: https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023 [accessed 2026-01-28]
11. Digital scribes. Australian Therapeutic Goods Administration. 2025. URL: https://www.tga.gov.au/products/medical-devices/software-and-artificial-intelligence/manufacturing/artificial-intelligence-ai-and-medical-device-software/digital-scribes [accessed 2025-12-08]

Abbreviations

AI: artificial intelligence
SOAP: subjective, objective, assessment, and plan
Y-KNOT: Your-Knowledgeable Navigator of Treatment

Edited by A Coristine; submitted 10.Dec.2025; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 06.Feb.2026.

Please cite as:
Coiera E, Fraile-Navarro D
AI Scribes: Are We Measuring What Matters?
JMIR Med Inform 2026;14:e89337
URL: https://medinform.jmir.org/2026/1/e89337
doi:10.2196/89337

© Enrico Coiera, David Fraile-Navarro. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 6.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identification and Localization of Breast Tumor Components via a Convolutional Neural Network Based on High-Frequency Ultrasound Combined With Histopathologic Registration: Prospective Study

Jia-Qian Yao[1*], MD; Wen-Wen Zhou[1,2*], MD; Zhi-Fei Chai[3], ME; Fei Ren[3], PhD; Tong-Yi Huang[1], MD; Tian-Tian Zhen[4], MD; Hui-Juan Shi[4], MD, PhD; Xiao-Yan Xie[1], MD, PhD; Ze Zhao[3], PhD; Ming Xu[1], MD, PhD

[1]Department of Medical Ultrasonics, The First Affiliated Hospital, Sun Yat-sen University, 58 Zhongshan 2nd Road, Guangzhou, China

[2]Department of Medical Ultrasonics, Suzhou Municipal Hospital Affiliated with Nanjing Medical University, Suzhou, China

[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[4]Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Ming Xu, MD, PhD
Department of Medical Ultrasonics, The First Affiliated Hospital, Sun Yat-sen University, 58 Zhongshan 2nd Road, Guangzhou, China

Abstract

Background: Given the highly heterogeneous biology of breast cancer, a more effective noninvasive diagnostic tool that unravels microscopic histopathology patterns is urgently needed.

Objective: This study aims to identify cancerous regions in ultrasound images of breast cancer via convolutional neural network based on registered grayscale ultrasound images and readily accessible biopsy whole slide images (WSIs).

Methods: This single-center study prospectively included participants undergoing ultrasound-guided core needle biopsy procedures for Breast Imaging Reporting and Data System category 4 or 5 breast lesions for whom breast cancer was pathologically confirmed from July 2022 to February 2023 consecutively. The basic information, ultrasound image data, biopsy tissue specimens, and corresponding WSIs were collected. After core needle biopsy procedures, the stained breast tissue specimens were sliced and coregistered with an ultrasound image of a needle tract. Convolutional neural network models for identifying breast cancer cells in ultrasound images were developed using FCN-101 and DeepLabV3 networks. The image-level predictive performance was evaluated and compared quantitatively by pixel accuracy, Dice similarity coefficient, and recall. Pixel-level classification was illustrated through confusion matrices. The cancerous region in the testing dataset was further visualized in ultrasound images. Potential clinical applications were qualitatively assessed by comparing the automatic segmentation results and the actual pathological tissue distributions.

Results: A total of 105 participants with 386 ultrasound images of breast cancer were included, with 270 (70%), 78 (20.2%), and 38 (9.8%) images in the training, validation, and test datasets, respectively. Both models performed well in predicting the cancerous regions in the biopsy area, whereas the FCN-101 model was superior to the DeepLabV3 model in terms of pixel accuracy (86.91% vs 69.55%; P=.002) and Dice similarity coefficient (77.47% vs 69.90%; P<.001). The two models yielded recall values of 54.64% and 58.46%, with no significant difference between them (P=.80). Furthermore, the FCN-101 model had an advantage in predicting cancerous regions, while the DeepLabV3 model achieved more accurate predictive pixels in normal tissue (both P<.05). Visualization of cancerous regions on grayscale ultrasound images demonstrated high consistency with those identified on WSIs.

Conclusions: The technique for spatial registration of breast WSIs and ultrasound images of a needle tract was established. Breast cancer regions were accurately identified and localized on a pixel level in high-frequency ultrasound images via an advanced convolutional neural network with histopathologic WSI as the reference standard.

(JMIR Med Inform 2026;14:e81181) doi:10.2196/81181

KEYWORDS

breast; neoplasms; biopsy; ultrasonography; registration

XSL•FO
RenderX

Introduction

Breast cancer heterogeneity has induced challenges in treatment planning and follow-up management, which leads to unfavorable outcomes [1,2]. Currently, ultrasound is a widely used diagnostic tool for breast cancer management, particularly valuable in screening, positive diagnosis, and treatment response assessment [3]. However, the biological heterogeneity of breast cancers leads to varied morphological features on ultrasound [4], often resulting in malignancy underestimation and overestimation. Furthermore, there is a heterogeneous response to treatment among patients with breast cancer [5-7]. Curative effect assessment secondary to preoperative neoadjuvant treatment is largely based on cancer volume changes [8], as well as biopsy for further validation [9]. A more accurate noninvasive diagnostic tool that indicates living cancer cells in breast cancer is urgently needed [10-12].

Hematoxylin and eosin (H&E) staining of breast tissue captured via core needle biopsy (CNB) has been introduced to reflect the underlying cellular and molecular information [13-15]. Preoperative diagnosis and curative effect assessment of breast cancer can be undermined by insufficient and nonrepresentative tissue owing to the heterogeneous distribution of breast cancer [16]. Likewise, the partial samples obtained by CNB may not represent the entire lesion [17]. There remains a need for standardized methods or imaging biomarkers available for accurately localizing histopathological cancerous subregions.

The convolutional neural network (CNN), a developed type of deep learning algorithm, has shown remarkable performance in correlating macroscopic imaging and microscopic histopathologic microstructure. A previous study showed that a multimodal radiomics model combining ultrasound and whole slide image (WSI) can effectively distinguish between luminal and nonluminal breast cancers [18]. Other studies have explored using the deep learning algorithm for correlation between magnetic resonance imaging (MRI) and whole-mount specimen images to localize prostate cancer [19]. Theoretically, these approaches may also be applicable in the ultrasound identification and localization of cancerous regions in breast cancer. Nonetheless, few studies have focused on this aspect. The correlation of ultrasound modality and readily accessible biopsy WSI remains open to question.

Therefore, this study aims to identify and localize cancerous regions in breast cancer based on a CNN algorithm that integrates high-frequency ultrasound (HFUS) images with WSI histopathology. The predictive performance of the model will be assessed.

Methods

Study Population

Consenting participants were recruited between July 2022 and February 2023 and were included as the training, validation, and test population. The eligibility criteria included the following: (1) index lesion was defined as category 4 or 5 according to the fifth edition of the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) of ultrasound [20]; (2) index lesion was visible on HFUS, and complete imaging data were stored; (3) underwent ultrasound-guided breast lesion biopsy and the histopathology indicated breast cancer. Participants were excluded if (1) the biopsy specimen was incomplete or inaccessible; (2) the breast ultrasound images were incomplete; (3) there was a history of treatment for breast cancer (surgery, antihormonal therapy, immunotherapy, and radiation therapy); (4) the pathologic diagnosis was incomplete. The study sample included 163 consecutive participants undergoing ultrasound-guided breast biopsy for suspicion of cancer. Of these, three participants whose ultrasound images had poor quality and one who did not provide research consent were excluded. Participants with incomplete, fractional tissue specimens (n=2) and benign histology reports (n=52) were also excluded.

Ultrasound and CNB Examination

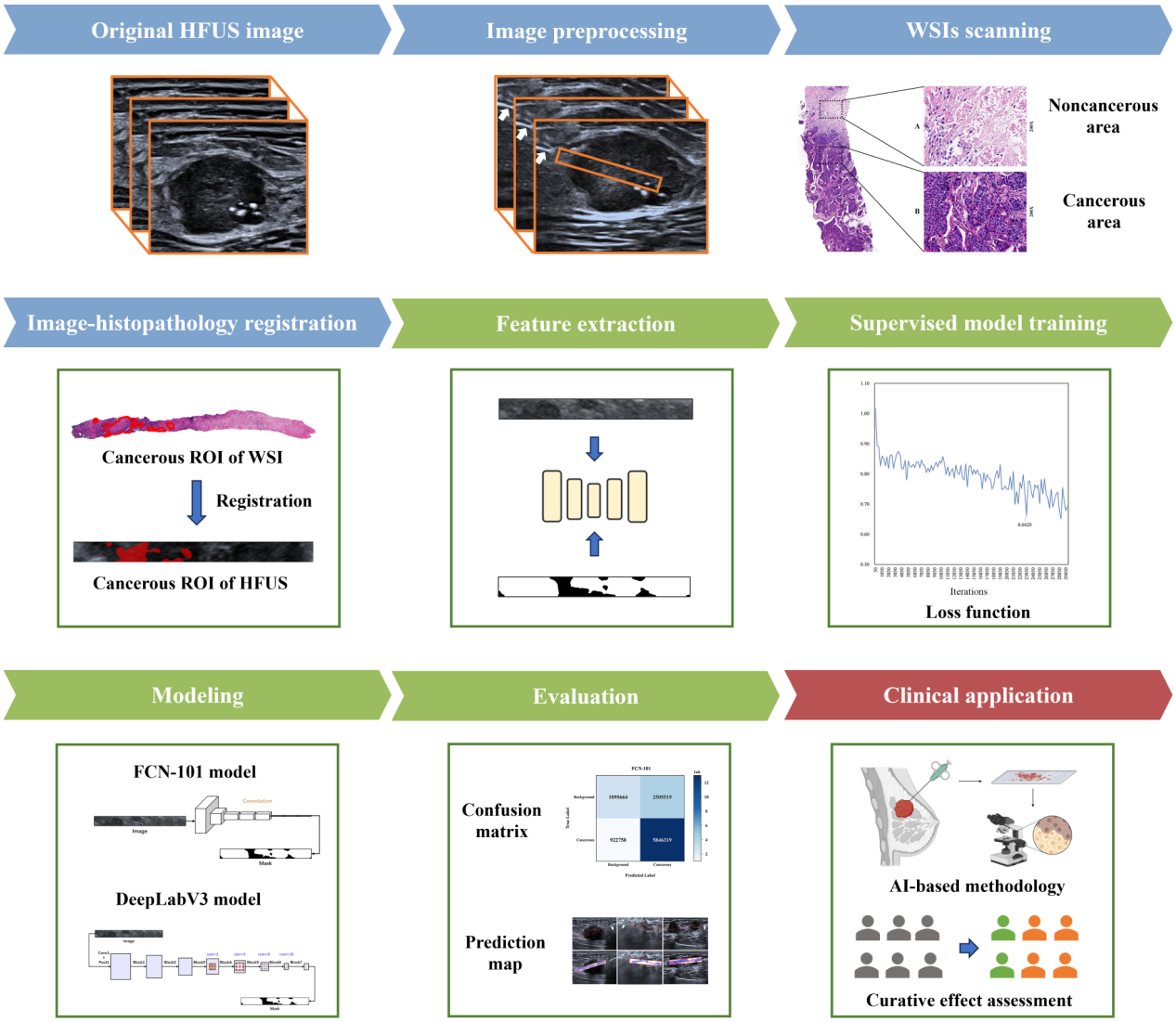
The overall design of this study is shown in Figure 1. Expert radiologists with at least 10 years of experience performed breast ultrasound examinations and ultrasound-guided CNB procedures following the standard practice protocol. Location, number, and morphologic characteristics (size, shape, orientation, margin, echo pattern, vascularity, and calcifications) of lesions were identified and categorizations were assigned by the expert radiologists according to the fifth edition of ultrasound BI-RADS [20]. An Aplio i900 Ultrasound System (Canon) with an i24LX8 high-frequency linear probe (frequency range: 8.0 - 18.0 MHz) was used to generate breast ultrasound images. All images were stored in DICOM format for subsequent analysis.

WSI Acquisition

When breast cancer was suspected, an ultrasound-guided CNB procedure was performed by expert radiologists to determine the diagnosis. MAGNUM biopsy instruments (BARD) with disposable core tissue biopsy needles (16G, 22 mm, MN1620, BARD) were adopted. During the procedure, the needle tip was positioned approximately 0.5 cm from the target biopsy region. The radiologists aligned the ultrasound probe and needle to visualize the entire needle tract. Subtle needle deflections were occasionally observed. These deflections were dynamically corrected in real time by adjusting the needle trajectory. The radiologists captured two B-mode HFUS images per biopsy for subsequent registration, including one pre-fire and one post-fire. To ensure the specimens were representative, biopsies were taken by the radiologist from different regions of the lesion, typically 4 - 6 samples, focusing on solid areas on B-mode ultrasound or areas with abundant vascularity on color Doppler flow imaging. When multiple lesions were encountered, the most suspicious lesion for malignancy was chosen for analysis. To facilitate the follow-up spatial registration of ultrasound and histopathologic images, the needle tip side of the biopsy specimen from the index lesion was stained with biological tissue dye (BIOGNOST), depicted in Figure S1 in Multimedia Appendix 1. After 2 to 5 minutes of coloration, the biopsy specimens were placed in 10% neutral buffered formalin for fixation and sent for histopathologic examination.

XSL•FO
RenderX

Figure 1. Workflow of the proposed cancerous regions identification protocol in this study. First, we got cropped HFUS images and corresponding spatially aligned biopsy WSIs. Second, a registration process was applied to achieve anatomic correlation. Third, the segmentation model was constructed using the FCN-101 and DeepLabV3 architectures. AI: artificial intelligence; HFUS: high-frequency ultrasound; ROI: region of interest; WSI: whole slide image.



The tissue samples were fixed and oriented during the embedding process to preserve the longitudinal axis. Each 3 µm histological slide was then sectioned parallel to the initial needle trajectory, thereby ensuring that the analyzed WSI represented the same gross anatomical plane captured by HFUS. The tissue strips were very small, thus the dimensional changes and distortion introduced by histological processing were assumed to be limited. Each section was stained with H&E. For each participant, all slides stained with H&E were reviewed by two experienced breast pathologists (with 5 y and 10 y of experience), and the histopathologic type was reported. Scanning of the H&E slides was performed using a KFBIO Digital Pathology Slide Scanner (KF-PRO-020) with a 200X objective lens. Representative heterogeneous cancer cell distribution of breast cancer in biopsy WSI is depicted in Figure S2 in Multimedia Appendix 1.

Imaging Registration

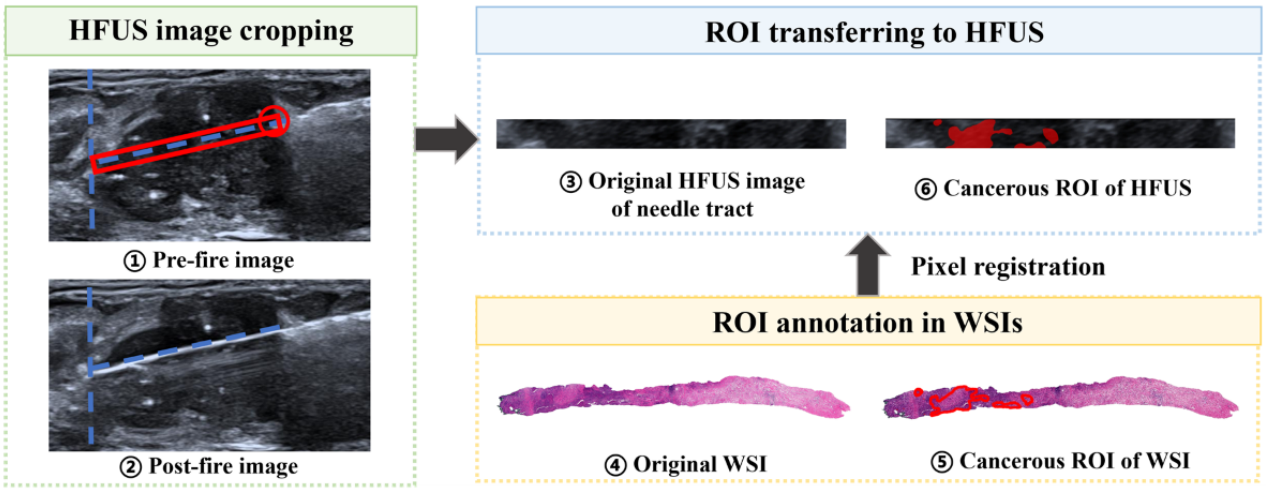
The expert pathologist with 10 years of experience used the open-source software QuPath (version 0.4.0) for digital histopathology analysis. The maximal cross-section of the extracted core was used for analysis. Cancer cells were identified based on nuclear atypia and mitotic figures. The regions of interest (ROIs) of cancerous regions were manually outlined on all high-resolution WSI slices, generating a per-pixel cancer cell labeling (Figure 2).

For 4 - 6 samples from one lesion, the expert pathologist selected 1 - 4 specimens for further registration processing. The method for selecting was based on three criteria: (1) intact and well-formed; (2) length matches the needle notch (22 mm); (3) clear and distinguishable staining. For each selected tissue sample, one radiologist with at least 5 years of experience correlated the WSI to two captured B-mode HFUS images. To ensure the spatial registration from WSI to the HFUS images, the radiologist compares the pre-fire and post-fire HFUS images using Photoshop software (version CC 2019; Adobe Inc). In the software, the needle tip position in the pre-fire image was taken as the starting point; actual length of the biopsy specimen was determined by the needle projection distance in the post-fire

XSL•FO
RenderX

image. Then, the radiologist cropped the biopsy area in the pre-fire image for subsequent annotation and analysis. The detailed image cropping process is demonstrated in Figure 2. The HFUS image of the needle tract was then saved in JPG format. The ROIs outlined in WSIs were converted to correlated HFUS images via Labelme software, an open-source image annotation tool (version 5.1.0).

Figure 2. Spatial registration and mapping protocol between HFUS and WSI. We compared the pre-fire image and post-fire image to crop the needle tract in HFUS images. An experienced pathologist manually outlined the cancerous cells on WSI images using QuPath software (version 0.4.0 [21]), which were transferred to the cropped HFUS images, creating the labeled dataset for training. HFUS: high-frequency ultrasound; ROI: region of interest; WSI: whole slide image.



Data Preprocessing

The HFUS images and their ROIs annotation were obtained to establish the CNN prediction model. Open-source libraries including Python (version 3.8.11; Python Software Foundation), Imgviz (version 1.5.1), and Numpy (version 1.21.6) were used to convert the dataset into the standard VOC format. To enhance the robustness and generalization capability of the model, data augmentation strategies, such as random horizontal flipping, rotation, and pixel transformations, were used to increase the generalizability of the model. The dataset was randomly divided into training (n=270, 70%), validation (n=78, 20.2%), and test (n=38, 9.8%) subsets using a 7:2:1 ratio with no overlap between the subsets. Image pixel values were normalized using the *z*-score method to reduce computation burden and accelerate model convergence.

Model Development

The fully automated segmentation CNN model was designed to identify cancerous regions in breast cancer based on HFUS images, with the ROIs in WSI as the reference standard. The independent CNN model was trained with advanced FCN-101 [22] and DeepLabV3 [23] networks as the backbone, separately. During training, the models were adapted using the AdamW optimized algorithm and CosineAnnealing learning rate adjustment curve to improve efficiency. The weighted Dice loss function was used to mitigate the data imbalance during the training process, as plotted in Figure S3 in Multimedia Appendix 1. The iteration parameters with the lowest loss values were chosen, which were 0.6626 for the FCN-101 model and 0.7285 for the DeepLabV3 model. In the fine-tuned process, parameters were selected to construct the segmentation model when the best performance on the test dataset was achieved. Finally, cancerous regions in breast cancer were localized. More details on the CNN model are provided in Multimedia Appendix 1. The source code for our CNN models is available on GitHub [24].

Statistical Analysis

All statistical analysis was conducted with IBM SPSS Statistics 25.0 (version 25.0; IBM Corp), Python (version 3.8.11; Python Software Foundation), and R software (version 4.2.1, R Foundation for Statistical Computing). Continuous variables were exhibited as means (SDs) or medians and IQRs and compared using a 1-sample *t* test or Mann-Whitney *U* test where appropriate. Categorical variables were expressed as counts and percentages and compared through χ^2 tests. To assess and compare the segmentation performance of CNN models based on independent networks, pixel accuracy (PA), Dice similarity coefficient (DSC), mean Intersection over Union, precision, and recall were calculated. These metrics were computed separately for each test image, and the final performance values were obtained by averaging the image-level results across the entire dataset. To avoid inflated degrees of freedom arising from multiple images per participant, a paired-sample *t* test was performed at the participant level to compare the performance of FCN-101 and DeepLabV3 models. Normality of the paired differences was confirmed using the Shapiro-Wilk test. Detailed evaluation metrics were demonstrated in the Supplementary Methods in Multimedia Appendix 1. The confusion matrices were adopted to show the pixel-level classification of cancerous regions and normal tissue based on aggregated raw pixel counts. Two-tailed $P<.05$ was considered statistically significant.

Ethical Considerations

This single-center prospective study involving human participants was approved by the local institutional review board of the First Affiliated Hospital of Sun Yat-sen University (Ethics [2023]842). Written informed consent was obtained from all

XSL•FO
RenderX

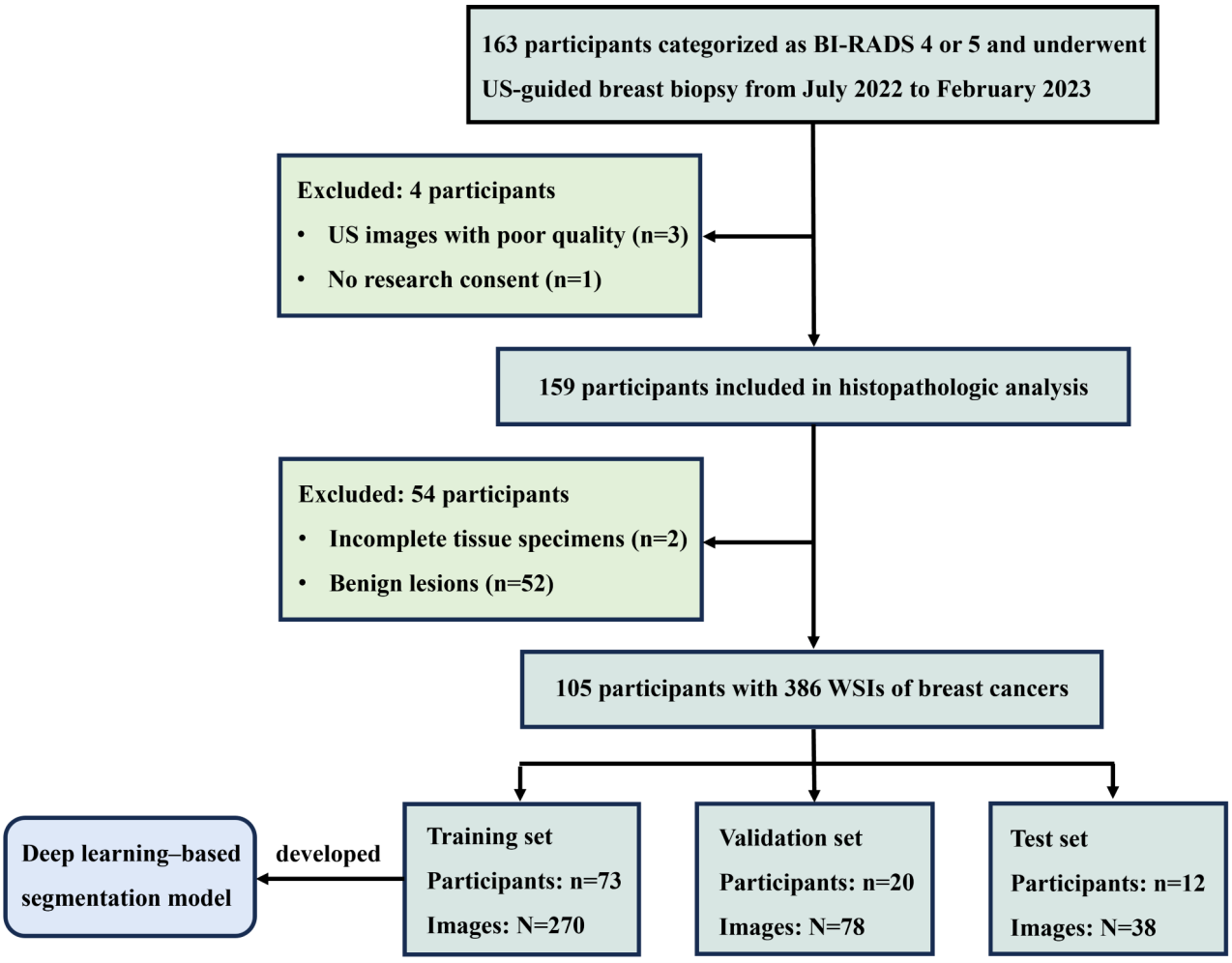
participants prior to their inclusion in the study. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. All data were deidentified before analysis to ensure participant privacy. Participants did not receive any financial or material compensation, as all procedures were part of routine clinical care, and the analytical use of the resulting data was clearly explained to participants during enrollment. No identifiable images or personal information are included.

Results

Participant Characteristics

A total of 163 individual participants received ultrasound-guided breast biopsy, and 58 participants were excluded (Figure 3). The final prospective dataset contained 105 participants (mean age, 53.7, SD 11.3 y old; all female) diagnosed as breast cancer with 386 HFUS images of the needle tract. All biopsies ultimately achieved successful targeting and adequate sampling in the 105 participants. Among them, 64 (87.7%), 19 (95%), and 9 (75%) invasive breast cancer and 9 (12.3%), 1 (5%), and 3 (25%) ductal carcinomas in situ were identified in the training, validation, and test subsets, respectively. Characteristics were compared between three subsets in Table 1 and no evidence of a statistical difference was observed, except for vascularity, where internal vascularity of breast lesions was observed in the training dataset (P=.03).

Figure 3. Flowchart showing inclusion and exclusion criteria of participants. BI-RADS: Breast Imaging Reporting and Data System; US: ultrasound; WSI: whole slide image.



XSL•FO
RenderX

Table . Participants and breast cancer characteristics. Unless otherwise indicated, data are numbers of lesions, with the percentages in parentheses.

Characteristic	Training set (n=73)	Validation set (n=20)	Testing set (n=12)	*P* value
Age (y), mean (SD)	52.7 (SD 11.2)	57.8 (SD 13.2)	52.8 (SD 7.4)	.21
Lesion number, n (%)				.08
Solitary	64 (87.7)	19 (95)	8 (66.7)	
Numerous	9 (12.3)	1 (5)	4 (33.3)	
BI-RADS[a] categorization, n (%)				.96
BI-RADS 4a	5 (6.8)	1 (5)	1 (8.3)	
BI-RADS 4b	9 (12.3)	3 (15)	1 (8.3)	
BI-RADS 4c	18 (24.7)	3 (15)	2 (16.7)	
BI-RADS 5	40 (54.8)	13 (65)	8 (66.7)	
Lesion location, n (%)				.68
Left breast	32 (43.8)	11 (55)	6 (50)	
Right breast	41 (56.2)	9 (45)	6 (50)	
Quadrant, n (%)				.30
Upper outer quadrant	45 (61.6)	9 (45)	5 (41.7)	
Lower outer quadrant	10 (13.7)	1 (5)	2 (16.7)	
Upper inner quadrant	12 (16.4)	6 (30)	3 (25)	
Lower inner quadrant	6 (8.2)	4 (20)	2 (16.7)	
Volume (mm^3), mean (SD)	4.1 (SD 5.1)	3.4 (SD 2.8)	4.3 (SD 7.0)	.82
Orientation, n (%)				.22
Parallel	82 (89.1)	15 (75)	11 (91.7)	
Not parallel	10 (10.9)	5 (25)	1 (8.3)	
Echo pattern, n (%)				.78
Hypoechoic	86 (93.5)	20 (100)	12 (100)	
Heterogeneous	6 (6.5)	0 (0)	0 (0)	
Margin, n (%)				.52
Circumscribed	6 (8.2)	2 (10)	2 (16.7)	
Not circumscribed	67 (91.8)	18 (90)	10 (83.3)	
Shape, n (%)				.47
Oval/Round	3 (4.1)	0 (0.0)	1 (8.3)	
Irregular	70 (95.9)	20 (100.0)	11 (91.7)	
Calcification, n (%)				.40
Absence	36 (49.3)	25 (59.5)	8 (66.7)	
Calcification in a mass	37 (50.7)	17 (40.5)	4 (33.3)	
Vascularity, n (%)				.03
Absent	11 (15.1)	7 (35)	2 (16.7)	
Internal vascularity	56 (76.7)	11 (55)	7 (58.3)	
Vessels in rim	6 (18.2)	2 (10)	4 (33.3)	
Histopathologic pattern, n (%)				.13
Invasive breast cancer	64 (87.7)	19 (95)	9 (75)	
Ductal carcinoma in situ	9 (12.3)	1 (5)	3 (25)	

[a]BI-RADS: Breast Imaging Reporting and Data System.

XSL•FO
RenderX

Model Evaluation

The segmentation capacity of the image-based CNN model based on FCN-101 and DeepLabV3 in the test dataset is provided in Table 2. The table presents image-level descriptive statistics, while all inferential comparisons were conducted at the participant level (n=12). In the test dataset, FCN-101 showed higher accuracy (PA: 86.91% vs 69.55%, *P*=.002), similarity (DSC: 77.47% vs 69.90%, *P*<.001), mean Intersection over Union (67.47% vs 60.29%, *P*<.001), and precision (66.01% vs 56.15%, *P*<.001) compared to DeepLabV3. There was no evidence of a difference in recall (54.64% vs 58.46%, *P*=.80) between the two algorithms.

Table . Comparison of prediction performance of FCN-101 and DeepLabV3 models in predicting cancerous regions in breast cancer in the test dataset. All metrics reflect image-level performance and are expressed as percentages. Statistical comparisons (*P* values) were performed at the participant level using paired analyses.

Evaluation metric	FCN-101, % (95% CI)	DeepLabV3, % (95% CI)	*P* value
PA[a]	86.91 (80.30-94.77)	69.55 (65.82-73.99)	.002
DSC[b]	77.47 (70.74-85.88)	69.90 (63.43-75.02)	<.001
mIoU[c]	67.47 (59.70-75.54)	60.29 (54.19-66.57)	<.001
Precision	66.01 (55.69-73.45)	56.15 (48.12-64.17)	<.001
Recall	54.64 (45.57-63.75)	58.46 (49.03-65.66)	.80

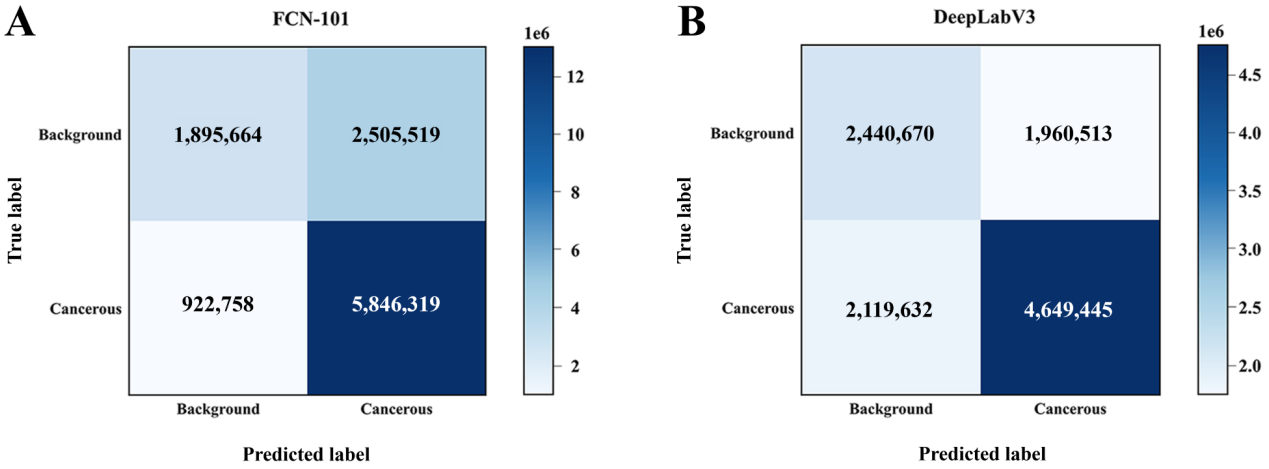
[a]PA: pixel accuracy.

[b]DSC: Dice similarity coefficient.

[c]mIoU: mean Intersection over Union.

The confusion matrix presented in Figure 4 evaluates the pixel categorization capacity, providing further insight into the per-pixel predictions of each model. The results indicated that the FCN-101 model successfully predicted the majority of cancerous pixels (5,846,319 vs 4,649,445 pixels; *P*<.05). However, the DeepLabV3 model demonstrated more accurate predictive pixels in background components (2,440,670 vs 1,895,664 pixels; *P*<.05). Based on the pixel-wise confusion matrices, the recall and specificity of FCN-101 were 86.37% and 43.07%, while those of DeepLabV3 reached 68.69% and 55.46%.

Figure 4. Confusion matrices for predicting cancerous regions. Confusion matrices were applied to summarize pixel-level classification outcomes aggregated across the entire test dataset for (A) FCN-101 and (B) DeepLabV3, indicating discordance or concordance with the ground truth (cancerous cell/background) from WSI results. WSI: whole slide image.

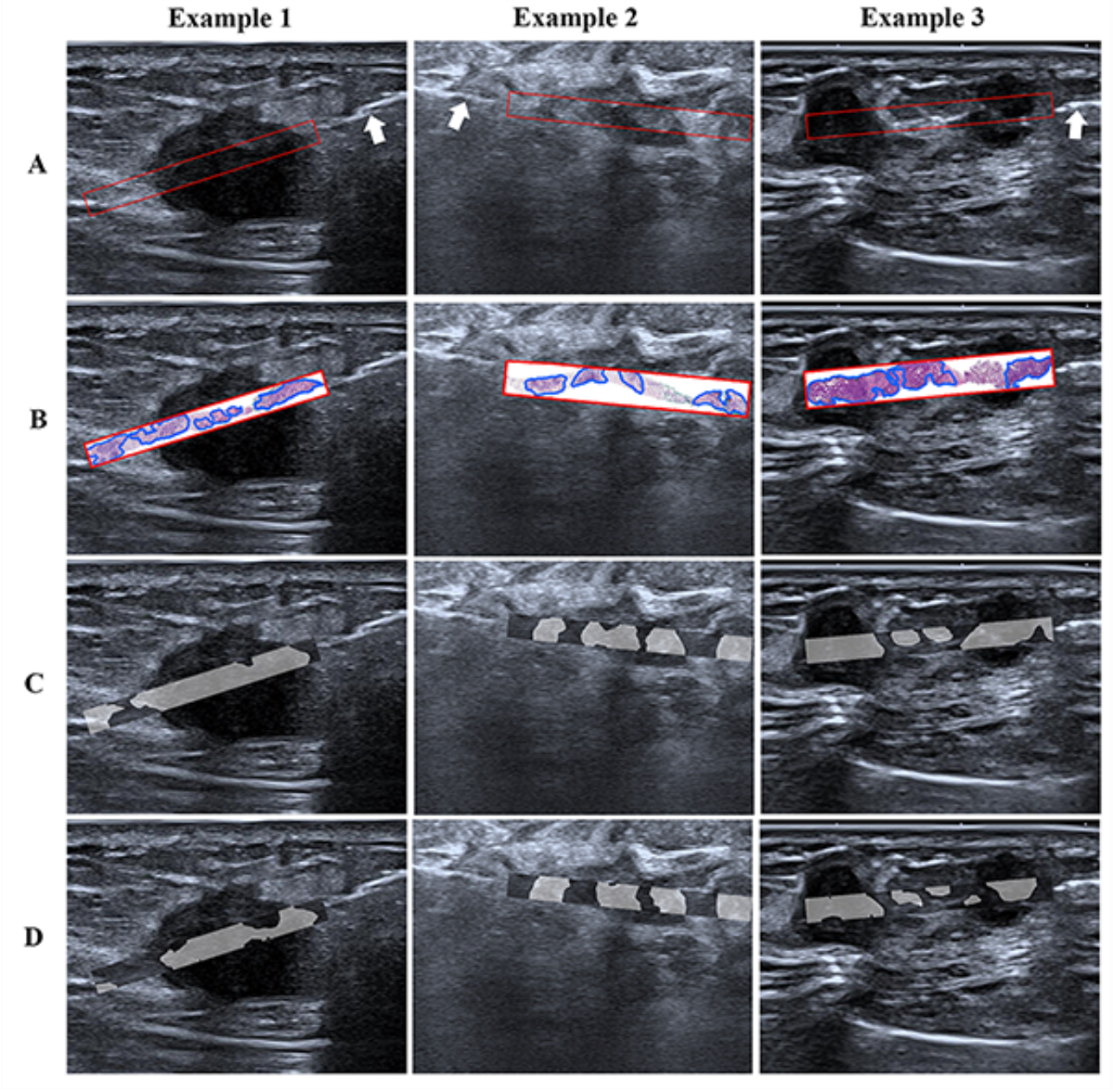


Model Visualization

In the test dataset, the automatic segmentation results of cancerous regions based on the two networks were visually displayed and qualitatively compared with the gold standard to evaluate the predictive performance of the models. Cancerous regions in the CNB biopsy area of three different participants' breast HFUS images were identified and localized in Figure 5, as generated by the FCN-101 and DeepLabV3 models. Cancerous region predictions by the two CNN models align closely with the actual histopathology. Additionally, even in challenging cases with unclear boundaries and mixed internal echoes, as represented in Example 2, both models were able to accurately predict the specific distribution of cancerous regions within the lesions.

XSL•FO
RenderX

Figure 5. Examples of identifying cancerous regions in three participants with invasive breast cancer. (**A**) Original HFUS images, with biopsy needle (white arrow) pointing to the needle tract area (red frame). (**B**) Annotation of ground truth labels by an expert pathologist. (**C**) Prediction results of the FCN-101 model (gray area). (**D**) Prediction results of the DeepLabV3 model (gray area). HFUS: high-frequency ultrasound.



Discussion

Principal Results

Breast cancer is a group of highly heterogeneous diseases with varying imaging features. Thus, differential diagnosis through imaging is limited [25,26]. Development and validation of a fast and noninvasive method equal to histologic results is urgently needed. Here, we developed a cancerous region classifier using a deep learning network with true labels from radiology-pathology registration. The FCN-101 model was superior to the DeepLabV3 model in terms of PA (86.91% vs 69.55%; *P*=.002) and DSC (77.47% vs 69.90%; *P*<.001). Recall values were 54.64% and 58.46%, with no significant difference observed between them (*P*=.80). The FCN-101 model excelled in identifying cancerous regions (5,846,319 vs 4,649,445 pixels; *P*<.05), whereas DeepLabV3 was more accurate for normal tissue (2,440,670 vs 1,895,664 pixels; *P*<.05) in pixel-level predictions. In the clinical workflow, the model could segment cancerous regions in grayscale ultrasound images of the breast. The results highlight the model's potential for advancing breast cancer assessment at the microscopic level via ultrasound imaging.

Efforts have been made to establish a methodology pertaining to image-histopathology registration [27-32]. For example, Ward et al [27] and Kwak et al [28] demonstrated accurate alignment between MRIs and digital histopathologic analyses in patients with prostate cancer. In addition, Wildeboer et al [29] established a multiparametric machine learning on ultrasound for histopathology localization of prostate cancer. However, similar research in breast imaging remains limited. In our study, WSIs from CNB provided accessible clinical data suitable for state-of-the-art deep learning algorithms.

XSL•FO
RenderX

Thus, a strength of this study is that we combined accessible CNB tissue samples with real-time ultrasound imaging to facilitate the registration process. The biopsy needle targeted the desired section plane while the tissue sample was obtained at the same level, which was crucial for registration. Moreover, HFUS offers spatial tissue distribution data, enhancing dynamic biopsy procedures and clinical utility. However, subtle registration errors existed; for example, the biopsy needle was deflected when encountering small breast tumors or rigid glandular tissue [33]. To address this, our skilled operators actively mitigated needle deflections. Pre-fire and post-fire ultrasound images were compared to ensure that the extracted tissue precisely corresponded to the visualized ultrasound plane.

Studies have investigated the biopsy efficacy between existing methods. Currently, three image-guided breast biopsy techniques are used: stereotactic-, ultrasound-, and MRI-guided biopsies. Stereotactic biopsy is indicated for calcifications, and MRI-guided biopsy is indicated for lesions visible only on MRI; both primarily assess ductal proliferative lesions. Ultrasound-guided biopsy, however, applies to a broader range of breast lesions [34]. Yashima et al [35] retrospectively compared the positive biopsy rate in 453 patients with 500 lesions that underwent ultrasound-guided core needle biopsy or vacuum-assisted biopsy and reported positive biopsy rates of 61.9% and 72.4% (P=.032), respectively. Unrepresentative CNB specimens might not fully reflect the overall characteristics of the tumor [17]. Although multipoint sampling and repetitive biopsy increased the detection rate, they also increased complications like bleeding, infection, and tumor spreading [36]. Based on this, our model may generate real-time cancer prediction heatmaps, where suspicious regions are highlighted to assist biopsy site selection during CNB sampling. However, prior to clinical use, the proposed approach should undergo rigorous supervision and ethical evaluation to ensure its safety and reliability in guiding biopsy decisions.

The visualized cancerous map was used to enhance the interpretability of the model. Here, we found that distribution maps facilitated the assessment of cancerous regions by highlighting the hypoechoic area in HFUS images (illustrated in Figure 5), which is in accordance with clinical routine. Furthermore, areas of abnormal echogenicity or edges of lesions often corresponded to cancerous regions identified in histopathology, which was also correctly predicted by the deep learning models. Guided by the ultrasound-based CNN algorithm, doctors could identify suspicious regions based on predicted cancerous regions, even in small lesions, helping to ensure sufficient and representative CNB samples for accurate histopathological assessment. However, the model is intended to assist clinicians and cannot replace their judgment in biopsy decision-making.

Notably, there is a discrepancy between the per-image recall and the pixel-level aggregated recall, which can be explained by the scale-dependent nature of these two evaluation strategies. This scale-dependent discrepancy indicates that the current model achieves higher sensitivity for larger tumors, while the detection of small lesions remains more challenging. It highlights an important direction for future optimization aimed at improving detection stability across different tumor sizes.

Although the two models had advantages in predicting cancerous regions, the confusion matrix analysis revealed a relatively low specificity, particularly for the FCN-101 model (43.07%). This indicates the presence of false-positive segmentation, with a proportion of normal background pixels being incorrectly segmented as cancerous. Such oversegmentation may be acceptable in identifying suspicious regions when emphasizing sensitivity. However, in the context of biopsy guidance, such behavior highlights the need for further optimization toward more precise boundary discrimination. Future work will focus on improving specificity through better loss function design, boundary-aware learning, and postprocessing strategies.

Limitations

This study combined registered WSIs and HFUS images to enhance cancerous region recognition in breast cancer, which has not been well-established in the literature. Still, we acknowledge the limitations of this study. First, it is a single-center prospective study with a small and potentially homogenous dataset, which undermines the model's generalizability and heightens the risk of overfitting. Besides, the small sample size in the test set limits statistical power and the generalizability of the prospective findings. To address this, we plan to incorporate data from multiple centers for robust external validation and conduct prospective studies to explore the models' role in assisting breast biopsy and postoperative follow-up after neoadjuvant therapy. Second, the relatively low recall indicates that some malignant regions may be missed; future studies will focus on improving model sensitivity through ensemble and data balancing approaches. Besides, benchmarking against widely adopted baseline models (eg, U-Net) could be performed in future research. Third, there is an absence of molecular subtype or pathological classification analysis and an imbalance of specific subtypes with small sample sizes (ductal carcinoma in situ in the test set). Given that different subtypes and cancer types exhibit distinct morphological features, a subtype-specific analysis could reveal performance differences and lead to refined models optimized for specific subtypes. Fourth, other imaging modalities such as color Doppler flow imaging, elastography, and contrast-enhanced ultrasound also play paramount roles in breast cancer diagnosis. For example, the lower prevalence of internal vascularity in the test set compared with the training set may have influenced model performance. Combining the information from multiple modalities could potentially further improve the performance of the CNN model. Fifth, the ultrasound-histopathology registration was based on biopsy WSIs and ultrasound images of the needle tract. Spatial correspondence should be regarded as approximate, and pixel-level metrics may overestimate the physical precision, which necessitates further refinement and validation via the whole-mount specimen.

Conclusion

In conclusion, we have proposed and evaluated deep learning models to identify cancerous regions in breast cancer in HFUS images through spatial registration of breast biopsy WSIs and HFUS images. This technique is potentially useful in conventional ultrasound examinations and ultrasound-guided breast biopsy procedures.

XSL•FO
RenderX

Acknowledgments

During drafting and revision of the manuscript, we used ChatGPT 5.0 (OpenAI) to polish the work to refine the language and improve readability. The authors reviewed and edited the content to ensure accuracy and take full responsibility for the content of the publication.

Funding

This study was supported by the National Scientific Foundation Committee of China (grants 82071951, 82102057, 82402297) and the Major Research Plan of the National Natural Science Foundation of China (grant 92059201). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

All data supporting the findings are available from the corresponding author on reasonable request. Source code for our convolutional neural network models is deposited on GitHub [37].

Authors' Contributions

JQY contributed to conceptualization, formal analysis, investigation, methodology, and writing – original draft. WWZ contributed to data curation, formal analysis, investigation, methodology, and writing – original draft. ZFC contributed to formal analysis, software, and validation. FR contributed to software, methodology, and supervision. TYH contributed to data curation, investigation, and methodology. TTZ contributed to data curation and investigation. HJS contributed to data curation and supervision. XYX contributed to conceptualization, funding acquisition, and supervision. ZZ contributed to conceptualization, methodology, software, supervision, and visualization. MX contributed to conceptualization, funding acquisition, project administration, and writing – review and editing. JQY and WWZ are cofirst authors. ZZ and MX are cocorresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures and notes illustrating the modeling framework and statistical evaluation methodologies used in this study.
[DOCX File, 3635 KB - medinform_v14i1e81181_app1.docx]

References

1. Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. Nat Rev Clin Oncol 2015 Jul;12(7):381-394. [doi: 10.1038/nrclinonc.2015.73] [Medline: 25895611]
2. Lüönd F, Tiede S, Christofori G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. Br J Cancer 2021 Jul;125(2):164-175. [doi: 10.1038/s41416-021-01328-7] [Medline: 33824479]
3. Berg WA, Bandos AI, Mendelson EB, Lehrer D, Jong RA, Pisano ED. Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666. J Natl Cancer Inst 2016 Apr;108(4):djv367. [doi: 10.1093/jnci/djv367] [Medline: 26712110]
4. Kutasovic JR, McCart Reed AE, Sokolova A, Lakhani SR, Simpson PT. Morphologic and genomic heterogeneity in the evolution and progression of breast cancer. Cancers (Basel) 2020 Mar 31;12(4):848. [doi: 10.3390/cancers12040848] [Medline: 32244556]
5. Haque W, Verma V, Hatch S, Suzanne Klimberg V, Brian Butler E, Teh BS. Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy. Breast Cancer Res Treat 2018 Aug;170(3):559-567. [doi: 10.1007/s10549-018-4801-3] [Medline: 29693228]
6. Fayanju OM, Ren Y, Thomas SM, et al. The clinical significance of breast-only and node-only pathologic complete response (pCR) after neoadjuvant chemotherapy (NACT): a review of 20,000 breast cancer patients in the National Cancer Data Base (NCDB). Ann Surg 2018 Oct;268(4):591-601. [doi: 10.1097/SLA.0000000000002953] [Medline: 30048319]
7. Romeo V, Accardo G, Perillo T, et al. Assessment and prediction of response to neoadjuvant chemotherapy in breast cancer: a comparison of imaging modalities and future perspectives. Cancers (Basel) 2021 Jul 14;13(14):3521. [doi: 10.3390/cancers13143521] [Medline: 34298733]
8. Fowler AM, Mankoff DA, Joe BN. Imaging neoadjuvant therapy response in breast cancer. Radiology 2017 Nov;285(2):358-375. [doi: 10.1148/radiol.2017170180] [Medline: 29045232]
9. Kuerer HM, Smith BD, Krishnamurthy S, et al. Eliminating breast surgery for invasive breast cancer in exceptional responders to neoadjuvant systemic therapy: a multicentre, single-arm, phase 2 trial. Lancet Oncol 2022 Dec;23(12):1517-1524. [doi: 10.1016/S1470-2045(22)00613-1] [Medline: 36306810]

XSL•FO
RenderX

10. Shi Z, Huang X, Cheng Z, et al. MRI-based quantification of intratumoral heterogeneity for predicting treatment response to neoadjuvant chemotherapy in breast cancer. Radiology 2023 Jul;308(1):e222830. [doi: 10.1148/radiol.222830] [Medline: 37432083]
11. Fatayer H, Sharma N, Manuel D, et al. Serial MRI scans help in assessing early response to neoadjuvant chemotherapy and tailoring breast cancer treatment. Eur J Surg Oncol 2016 Jul;42(7):965-972. [doi: 10.1016/j.ejso.2016.03.019] [Medline: 27260848]
12. Ali HR, Jackson HW, Zanotelli VRT, et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. Nat Cancer 2020 Feb;1(2):163-175. [doi: 10.1038/s43018-020-0026-6] [Medline: 35122013]
13. Shamai G, Binenbaum Y, Slossberg R, Duek I, Gil Z, Kimmel R. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. JAMA Netw Open 2019 Jul 3;2(7):e197700. [doi: 10.1001/jamanetworkopen.2019.7700] [Medline: 31348505]
14. Liu H, Xu WD, Shang ZH, et al. Breast cancer molecular subtype prediction on pathological images with discriminative patch selection and multi-instance learning. Front Oncol 2022;12:858453. [doi: 10.3389/fonc.2022.858453]
15. Park KH, Loibl S, Sohn J, et al. Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis, treatment and follow-up of patients with early breast cancer. ESMO Open 2024 May;9(5):102974. [doi: 10.1016/j.esmoop.2024.102974] [Medline: 38796284]
16. Tasoulis MK, Lee HB, Yang W, et al. Accuracy of post-neoadjuvant chemotherapy image-guided breast biopsy to predict residual cancer. JAMA Surg 2020 Dec 1;155(12):e204103. [doi: 10.1001/jamasurg.2020.4103] [Medline: 33026457]
17. Brogi E, Krystel-Whittemore M. Papillary neoplasms of the breast including upgrade rates and management of intraductal papilloma without atypia diagnosed at core needle biopsy. Mod Pathol 2021 Jan;34(Suppl 1):78-93. [doi: 10.1038/s41379-020-00706-5] [Medline: 33106592]
18. Huang Y, Yao Z, Li L, et al. Deep learning radiopathomics based on preoperative US images and biopsy whole slide images can distinguish between luminal and non-luminal tumors in early-stage breast cancers. EBioMedicine 2023 Aug;94:104706. [doi: 10.1016/j.ebiom.2023.104706] [Medline: 37478528]
19. Bhattacharya I, Seetharaman A, Kunder C, et al. Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework. Med Image Anal 2022 Jan;75:102288. [doi: 10.1016/j.media.2021.102288] [Medline: 34784540]
20. Morris E, Comstock C, Lee C, et al. ACR BI-RADS magnetic resonance imaging ACR BI-RADS atlas, breast imaging reporting and data system. 2013 URL: https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/BI-RADS [accessed 2026-01-08]
21. QuPath. URL: https://qupath.github.io/ [accessed 2026-01-21]
22. Dung CV, Anh LD. Autonomous concrete crack detection using deep fully convolutional neural network. Automation in Construction 2019 Mar;99:52-58. [doi: 10.1016/j.autcon.2018.11.028]
23. Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. arXiv. Preprint posted online on Jun 17, 2017. [doi: 10.1007/978-3-030-01234-2_49]
24. Chai ZF. Breast cancer component segmentation. GitHub repository. 2024. URL: https://github.com/ChaiZhifeiF/Breast-Cancer-Component-Segmentation [accessed 2024-07-23]
25. Lee HJ, Kim EK, Kim MJ, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. Eur J Radiol 2008 Feb;65(2):293-298. [doi: 10.1016/j.ejrad.2007.04.008] [Medline: 17531417]
26. Abdullah N, Mesurolle B, El-Khoury M, Kao E. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. Radiology 2009 Sep;252(3):665-672. [doi: 10.1148/radiol.2523080670] [Medline: 19567644]
27. Ward AD, Crukley C, McKenzie CA, et al. Prostate: registration of digital histopathologic images to in vivo MR images acquired by using endorectal receive coil. Radiology 2012 Jun;263(3):856-864. [doi: 10.1148/radiol.12102294] [Medline: 22474671]
28. Kwak JT, Sankineni S, Xu S, et al. Prostate cancer: a correlative study of multiparametric MR imaging and digital histopathology. Radiology 2017 Oct;285(1):147-156. [doi: 10.1148/radiol.2017160906] [Medline: 28582632]
29. Wildeboer RR, Mannaerts CK, van Sloun RJG, et al. Automated multiparametric localization of prostate cancer based on B-mode, shear-wave elastography, and contrast-enhanced ultrasound radiomics. Eur Radiol 2020 Feb;30(2):806-815. [doi: 10.1007/s00330-019-06436-w] [Medline: 31602512]
30. Zhang Z, Wu HH, Priester A, et al. Prostate microstructure in prostate cancer using 3-T MRI with diffusion-relaxation correlation spectrum imaging: validation with whole-mount digital histopathology. Radiology 2020 Aug;296(2):348-355. [doi: 10.1148/radiol.2020192330] [Medline: 32515678]
31. Schoop RAL, de Roode LM, de Boer LL, Dashtbozorg B. Framework for deep learning based multi-modality image registration of snapshot and pathology images. IEEE J Biomed Health Inform 2024 Nov;28(11):6699-6711. [doi: 10.1109/JBHI.2024.3444908] [Medline: 39150810]
32. Schalk SG, Postema A, Saidov TA, et al. 3D surface-based registration of ultrasound and histology in prostate cancer imaging. Comput Med Imaging Graph 2016 Jan;47:29-39. [doi: 10.1016/j.compmedimag.2015.11.001] [Medline: 26647110]

XSL•FO
RenderX

33. Ma K, Kornecki A, Bax J, Mundt Y, Fenster A. Development and validation of a new guidance device for lateral approach stereotactic breast biopsy. Med Phys 2009 Jun;36(6):2118-2129. [doi: 10.1118/1.3130017] [Medline: 19610301]
34. Nakano S, Imawari Y, Mibu A, Otsuka M, Oinuma T. Differentiating vacuum-assisted breast biopsy from core needle biopsy: is it necessary? Br J Radiol 2018 Dec;91(1092):20180250. [doi: 10.1259/bjr.20180250] [Medline: 29975150]
35. Yashima Y, Fujioka T, Kubota K, et al. Comparison of the clinical and pathological characteristics of ultrasound-guided biopsy for breast masses and non-mass lesions between 16-gauge spring-loaded core needle biopsy and 12-gauge spring-loaded vacuum-assisted biopsy. J Med Ultrasonics 2023 Apr;50(2):205-212. [doi: 10.1007/s10396-022-01279-3]
36. Loving VA, Johnston BS, Reddy DH, et al. Antithrombotic therapy and hematoma risk during image-guided core-needle breast biopsy. Radiology 2023 Jan;306(1):79-86. [doi: 10.1148/radiol.220548] [Medline: 35997610]
37. ChaiZhifeiF/breast-cancer-component-segmentation. GitHub. URL: https://github.com/ChaiZhifeiF/Breast-Cancer-Component-Segmentation [accessed 2026-01-21]

Abbreviations

BI-RADS: Breast Imaging Reporting and Data System
CNB: core needle biopsy
CNN: convolutional neural network
DSC: Dice similarity coefficient
H&E: hematoxylin and eosin
HFUS: high-frequency ultrasound
MRI: magnetic resonance image/imaging
PA: pixel accuracy
ROI: region of interest
WSI: whole slide image

Edited by A Coristine; submitted 23.Jul.2025; peer-reviewed by L Yongping, P Bhatnagar; accepted 17.Dec.2025; published 23.Jan.2026.

Please cite as:
Yao JQ, Zhou WW, Chai ZF, Ren F, Huang TY, Zhen TT, Shi HJ, Xie XY, Zhao Z, Xu M
Identification and Localization of Breast Tumor Components via a Convolutional Neural Network Based on High-Frequency Ultrasound Combined With Histopathologic Registration: Prospective Study
JMIR Med Inform 2026;14:e81181
URL: https://medinform.jmir.org/2026/1/e81181
doi:10.2196/81181

© Jia-Qian Yao, Wen-Wen Zhou, Zhi-Fei Chai, Fei Ren, Tong-Yi Huang, Tian-Tian Zhen, Hui-Juan Shi, Xiao-Yan Xie, Ze Zhao, Ming Xu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automatic Recognition and Prognostic Prediction of Colorectal Liver Metastases Using a Multi-Scale Deep Learning Framework: Model Development and Validation Study

Houwen Long, BSc; Feng Ying, BSc; Shi Wu, MSc; Long Li, MSc

Department of Anus and Intestine Surgery, Yongkang First People's Hospital Affiliated to Hangzhou Medical College, 599 Jinshan West Road, Dongcheng Street, Yongkang, China

Corresponding Author:
Long Li, MSc
Department of Anus and Intestine Surgery, Yongkang First People's Hospital Affiliated to Hangzhou Medical College, 599 Jinshan West Road, Dongcheng Street, Yongkang, China

Abstract

Background: Colorectal cancer liver metastasis (CRLM) presents considerable challenges in both diagnosis and prognosis, as conventional approaches often are limited by subjectivity, variability, and limited efficiency. Recent advances in deep learning have shown great potential for automated extraction of pathological features, offering improved diagnostic accuracy and more reliable prognostic predictions.

Objective: This study aimed to develop and validate a multi-model ensemble deep learning framework (colon cancer liver metastasis network [CLM-Net]) for the automatic recognition and prognostic prediction of CRLM from pathological images, thereby enhancing diagnostic accuracy and clinical applicability.

Methods: A total of 197 pathologically annotated CRLM cases were collected and integrated from publicly available datasets (Kaggle and The Cancer Imaging Archive) to construct high-quality training and independent test sets. CLM-Net was built upon base convolutional neural network architectures including VGG16, DeepLab-v3, and U-Net, incorporating multi-scale atrous convolutions, a squeeze-and-excitation attention mechanism, a conditional random field refinement module, and transfer learning strategies. The framework was comprehensively evaluated using 5-fold cross-validation and independent testing for classification, segmentation, and prognostic prediction tasks. For survival prediction, 1024-dimensional image feature vectors extracted by CLM-Net were analyzed using logistic regression and random forest classifiers, with Kaplan-Meier curves and log-rank tests used for survival analysis.

Results: CLM-Net demonstrated superior performance in pathological image recognition, achieving 94% accuracy, 92% recall, 93% F_1-score, and an area under the receiver operating characteristic curve of 0.96 on the independent test set, outperforming single models. For survival prediction, CLM-Net with multi-scale attention achieved an area under the receiver operating characteristic curve of 0.864. Kaplan-Meier analysis revealed its significantly stronger risk stratification ability compared with VGG16, U-Net, and DeepLab-v3. Precision-recall curves and heatmaps further confirmed the model's high robustness and generalizability in unseen samples. Clinical evaluation by pathologists indicated strong interpretability and diagnostic utility, with a concordance rate of 90%.

Conclusions: The proposed CLM-Net framework, through the integration of diverse deep learning architectures and mechanisms, markedly improved the recognition and prognostic prediction of CRLM, demonstrating excellent generalization and clinical translation potential. Future studies will focus on multi-center validation and integration of multimodal features to further optimize its role in precision medicine.

(JMIR Med Inform 2026;14:e73311) doi:10.2196/73311

KEYWORDS

colorectal liver metastasis; deep learning; pathological images; prognostic prediction; model evaluation; integrated learning

Introduction

Colorectal cancer (CRC) is among the most prevalent malignant tumors worldwide, with liver metastasis being a major cause of mortality in affected patients [1-3]. In high-income countries in particular, both the incidence and mortality rates of CRC remain high. According to global cancer statistics, CRC consistently ranks among the top cancers in terms of incidence, and approximately 25% to 50% of patients with CRC will eventually develop liver metastases. Hepatic metastasis not only represents the primary driver of CRC-related death but also markedly affects patient quality of life and treatment efficacy.

XSL•FO
RenderX

Early and accurate diagnosis, as well as reliable prediction of colorectal cancer liver metastasis (CRLM), is therefore essential for improving patient outcomes and guiding personalized therapeutic strategies [4-6]. Deep learning, a leading technology in the field of artificial intelligence, has demonstrated great potential in this area. By leveraging large-scale pathological image datasets, deep learning algorithms can autonomously learn and extract clinically relevant features, thereby enhancing the diagnostic precision and prognostic assessment of CRLM [7,8].

In recent years, the rapid advancement of deep learning technologies has revolutionized medical image analysis [9,10], enabling the efficient interpretation of complex pathological data through automated feature extraction [11]. Despite these advances, substantial gaps remain in the application of deep learning to the diagnosis and prognostic prediction of CRLM. Most existing studies rely on single deep learning models, which often struggle to accurately capture complex pathological features such as tumor boundaries and microarchitectural details. This limitation compromises the precision and robustness of predictive outcomes. Furthermore, how to effectively integrate the strengths of different deep learning algorithms to construct more comprehensive and reliable models remains an urgent challenge [8,12].

In recent years, deep learning techniques have been extensively applied to the diagnosis and prognosis prediction of CRLM. Most existing studies have relied on single-model frameworks—such as U-Net, ResNet, or temporal long short-term memory networks—for lesion segmentation, image classification, or dynamic disease prediction, respectively. However, these approaches often experience limited feature extraction capabilities, an inability to simultaneously address both diagnostic and prognostic tasks, and poor adaptability to the inherent heterogeneity of pathological images. To provide a comprehensive overview of the current landscape, we systematically reviewed representative deep learning-based studies in CRLM analysis, summarizing their specific tasks, data sources, advantages, and limitations (Table S1 in Multimedia Appendix 1). In contrast, the colon cancer liver metastasis network (CLM-Net) model proposed in this study integrates the architectural strengths of VGG16, DeepLab-v3, and U-Net, and incorporates multi-scale feature fusion, squeeze-and-excitation (SE) attention modules, and a conditional random field postprocessing component. These enhancements significantly improve image recognition performance while enabling accurate prognostic prediction and robust clinical applicability, effectively addressing the limitations of existing single-task models.

To overcome these challenges, we designed and developed a novel multi-model ensemble deep learning framework, CLM-Net, designed to enhance both the efficiency and accuracy of pathological image analysis in CRLM. CLM-Net harnesses the strengths of multiple deep learning architectures, including convolutional neural networks (CNNs) (eg, VGG16), segmentation models (eg, DeepLab-v3), and multi-scale processing modules. The model integrates multi-resolution pathological features to enhance representational power and generalizability. In the feature extraction stage, an SE attention mechanism is introduced to strengthen the model's focus on key lesion areas, thereby improving analysis precision. Additionally, the adoption of transfer learning not only accelerates model convergence but also enhances performance on limited-sample datasets.

CLM-Net is designed to address critical challenges in the diagnosis and prognostic evaluation of CRLM by automatically extracting and integrating microscopic features from pathological images, in combination with relevant clinical indicators. This integrated approach provides a highly efficient and accurate prediction tool. The primary objective of this study is to enhance diagnostic precision and prognostic assessment for CRLM, while overcoming limitations of current models in feature representation and overall performance.

This work not only bridges a significant technological gap in the field of CRLM diagnosis and prognosis but also opens new avenues for the development of personalized medicine. By coupling pathological feature extraction with predictive analytics, CLM-Net holds strong potential as a valuable tool in precision oncology for CRLM, contributing to both technological innovation and improved clinical decision-making.

In summary, we propose CLM-Net, a novel deep learning framework that incorporates multi-model ensemble learning, attention mechanisms, and transfer learning to enhance automated recognition and prognostic prediction of CRLM pathology images. We hypothesize that CLM-Net will demonstrate superior performance in diagnostic accuracy, prognostic capability, and generalizability compared with conventional single-model approaches and will be highly endorsed by clinical professionals for its practical utility.

Methods

Ethical Considerations

This study was conducted using publicly available datasets, including the Liver Tumor Segmentation dataset from Kaggle and The Cancer Imaging Archive (TCIA). Because the data are publicly accessible and fully deidentified, institutional review board approval was not required for this retrospective study in accordance with institutional policies of Yongkang First People's Hospital Affiliated to Hangzhou Medical College (policy document available upon request). The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

All datasets used in this study contain anonymized patient information, and no identifiable personal data were accessed. Therefore, informed consent from individual participants was not required.

The study adhered to principles of data privacy and confidentiality, as only deidentified imaging and clinical metadata were used for research purposes.

No financial or other compensation was provided to participants, as this study did not involve direct patient recruitment or intervention.

XSL•FO
RenderX

Public Data Download and Integration

The pathological image data used in this study were primarily derived from 2 public databases: the Liver Tumor Segmentation Challenge dataset on Kaggle and TCIA. The Kaggle dataset contained 129 CRLM-related histopathological slides with corresponding clinical information, including age, sex, and tumor stage. To further enhance the sample size and diversity, an additional 68 CRLM cases were obtained from TCIA. In total, image data from 197 patients were included. After quality control and image augmentation, a final set of 1000 high-quality images was generated for model training and evaluation.

During preprocessing, low-resolution or unevenly stained images were excluded. All images were resized to 256×256 pixels to meet the input requirements of the model. Annotations were performed by experienced pathologists to ensure the accuracy of metastatic lesions and surrounding regions. Each image was linked to a unique patient identifier to guarantee patient-level independence across training, validation, testing, and prognostic modeling tasks.

Detailed Annotation of Pathological Images

Pathological image annotation was conducted by board-certified pathologists with over 10 years of professional experience. Using Aperio ImageScope software (version 12.3.3; Leica Biosystems, Germany), the experts meticulously annotated tumor regions and liver metastasis features within each slide. In addition, they recorded relevant clinical parameters, including tumor size and morphology, clarity of tumor boundaries with adjacent normal liver tissue, the spatial distribution of tumor-associated macrophages, and the density and localization of immune cells (eg, T cells and B cells) within the tumor microenvironment. All annotation data were saved in XML format to ensure compatibility and efficient parsing by the subsequent deep learning models.

Independent Verification and Annotation Error Control

To ensure annotation accuracy, all pathological images in this study were independently annotated by 2 experienced pathologists. Interrater reliability was assessed using the Cohen κ coefficient. In cases of disagreement, a third senior pathologist reviewed and confirmed the final annotation, ensuring both precision and consistency. To minimize manual annotation errors, the following quality control strategies were implemented:

- Dual annotation and cross-validation: Each image was independently annotated by 2 pathologists, and results were cross-verified to reduce single-observer bias.
- Standardized training protocols: All participating pathologists underwent unified training and followed standardized annotation guidelines to ensure interobserver consistency.
- Tool-assisted annotation: Semiautomated annotation tools were used to assist in lesion identification, thereby enhancing both efficiency and accuracy.

These measures collectively ensured high-quality annotations, establishing a robust and reliable dataset for subsequent deep learning model training and analysis.

Application of Data Augmentation Techniques

To improve model generalization across diverse pathological image features, various data augmentation strategies were applied, tailored to the specific requirements of each deep learning architecture. For VGG16, images were first resized to 224×224 pixels using bilinear interpolation to match the model's input specification. Standard normalization was then performed by subtracting the dataset's mean pixel value and dividing by the SD, thereby transforming the pixel distribution to approximate a standard normal distribution. In the case of DeepLab-v3, numerical data were similarly normalized to a mean of 0 and an SD of 1. For textual data inputs, word embedding techniques were used to convert words into dense vector representations using pretrained models such as Word2Vec or GloVe for initialization. For U-Net, input images were also resized and normalized as described above. In addition, random cropping and horizontal flipping were implemented to increase data variability and reduce overfitting. All preprocessing steps were conducted using standard APIs provided by mainstream deep learning frameworks such as TensorFlow and PyTorch. Specifically, the torchvision.transforms module in PyTorch was used for image transformations.

Development of Basic Convolutional Neural Network Models

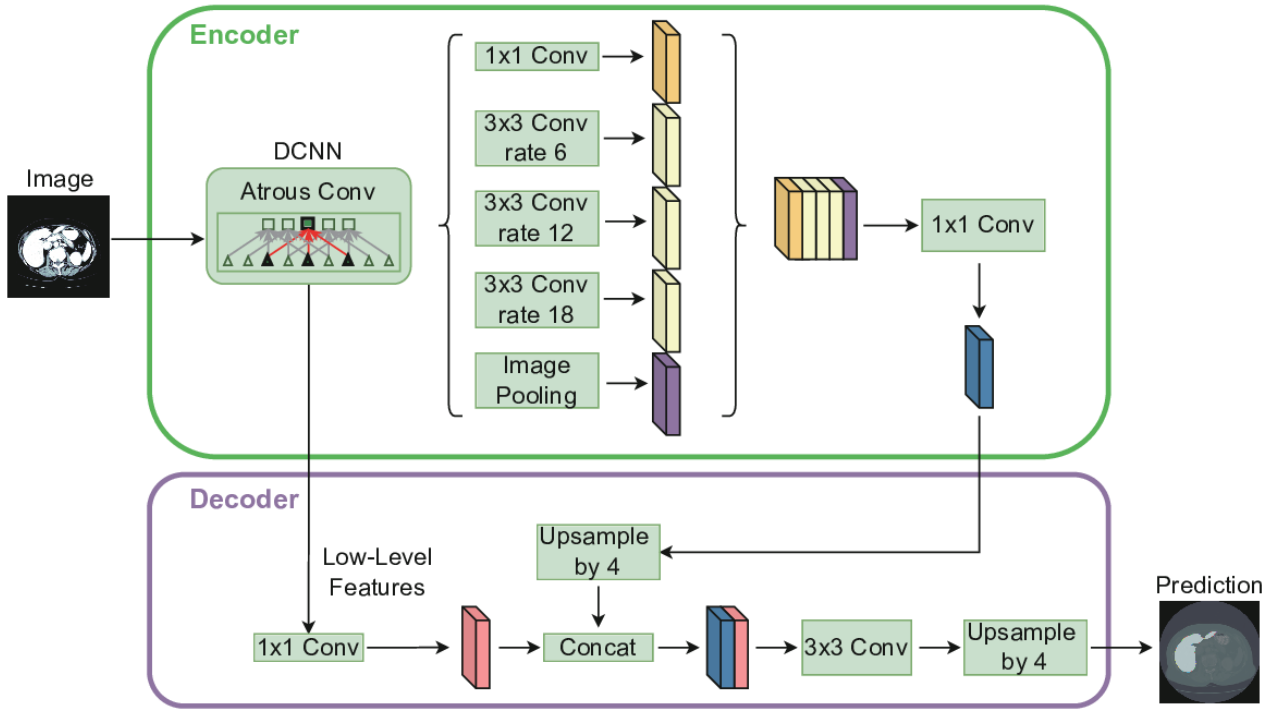
Several classical CNN models, including DeepLab-v3, VGG16, and U-Net, were used for the classification and segmentation of pathological images. All models were implemented using the PyTorch framework (version 1.7.1; Meta AI, Meta Platforms, Inc). To accelerate convergence and improve generalization, model parameters were initialized with pretrained weights from the ImageNet dataset. Training was conducted using the Adam optimizer, with a learning rate of 1×10^{-4}, a batch size of 32, and 50 training epochs.

DeepLab-V3

DeepLab-v3 is a state-of-the-art semantic segmentation model that incorporates dilated (atrous) convolutions to capture multi-scale contextual information without reducing spatial resolution. Its core component, atrous spatial pyramid pooling (ASPP), applies parallel atrous convolutions with varying dilation rates to extract rich semantic features at multiple scales, thereby enhancing model adaptability to complex structures in histopathological images. The use of dilated convolutions expands the receptive field without increasing the number of parameters or computational cost, enabling more comprehensive feature extraction. DeepLab-v3 follows an encoder-decoder architecture, where the encoder captures high-level semantic features, and the decoder restores spatial details for precise segmentation. Commonly, pretrained backbone networks such as Xception or MobileNet are adopted to serve as feature extractors, followed by the ASPP module and decoder to complete the segmentation task. The model structure is illustrated in Figure 1.

XSL•FO
RenderX

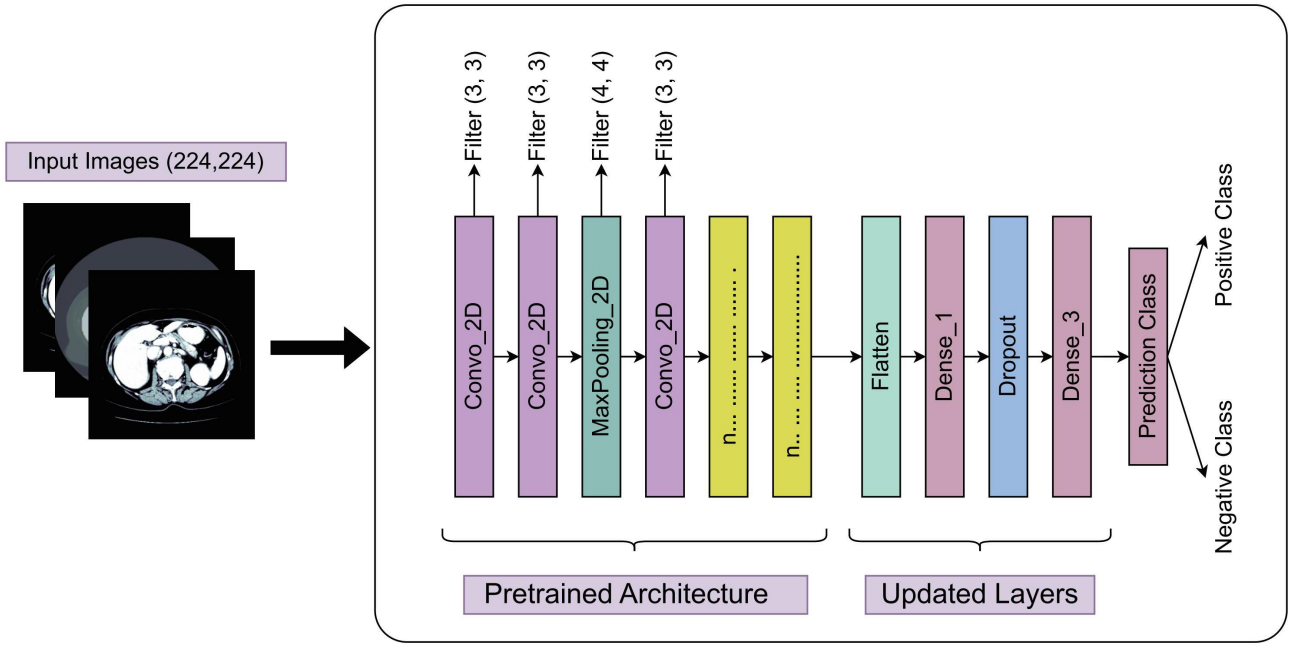
Figure 1. Schematic diagram of the DeepLab-v3 algorithm. DCNN: deep convolutional neural network.



VGG16

VGG16 is a classic deep CNN composed of 16 layers, including 13 convolutional layers and 3 fully connected layers. It captures image features by repetitively stacking 3×3 convolutional kernels and 2×2 max-pooling layers. By using multiple small kernels, VGG16 achieves a larger receptive field while maintaining a relatively low number of parameters and a deeper network structure. Its architecture is straightforward and consistent, with fixed strides in all convolutional and pooling layers, which simplifies network design and optimization. In our study, VGG16 was initialized with pretrained weights from the ImageNet dataset. During training, images were preprocessed by mean subtraction and optimized using mini-batch stochastic gradient descent. The algorithmic framework is illustrated in Figure 2.

Figure 2. Schematic diagram of the VGG16 algorithm.



U-Net

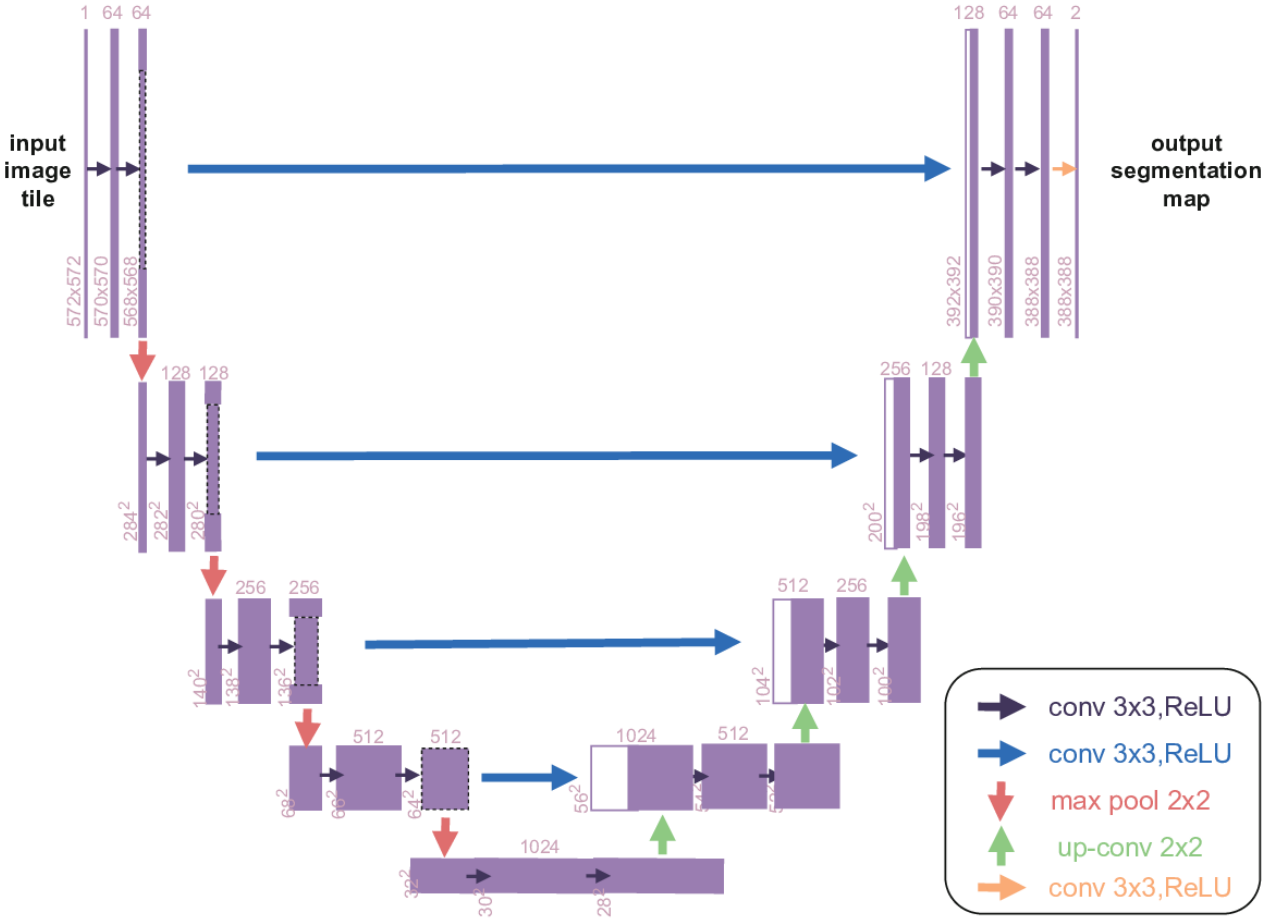
U-Net is a neural network model specifically designed for medical image segmentation, characterized by its U-shaped architecture consisting of a contracting path (encoder) and an expansive path (decoder). A key feature of U-Net is the use of skip connections that merge encoder feature maps with corresponding decoder outputs, thereby preserving spatial

XSL•FO
RenderX

context and enhancing the reconstruction of fine-grained details. U-Net relies heavily on data augmentation techniques during training, such as rotation, scaling, and elastic deformation, to improve generalizability. In our implementation, the model was trained using the Adam optimizer, with a learning rate of 1×10^{-4}, a batch size of 8, and a cross-entropy loss function. The algorithm structure is shown in Figure 3.

Each of the 3 models demonstrates unique strengths for different image analysis tasks. DeepLab-v3 is suited for fine-grained segmentation and multi-scale feature extraction. VGG16, with its efficient and compact design, is often used as a backbone for various visual recognition tasks. U-Net, in contrast, excels in medical image segmentation, particularly in scenarios with limited datasets. Based on the complementary strengths of these architectures, we developed a customized CNN framework, CLM-Net, specifically tailored for the automatic recognition of pathological features in CRLM.

Figure 3. Schematic diagram of the U-Net algorithm. ReLU: rectified linear unit.



Design of the CLM-Net Model

CLM-Net is a CNN framework specifically developed for the automatic recognition of pathological features in CRLM histopathological images. The architecture adopts a parallel fusion structure that integrates the strengths of 3 widely used CNN models: VGG16, DeepLab-v3, and U-Net. Initially, input images are processed through a VGG16 backbone pretrained on ImageNet to extract low-level texture features. These features are then passed into 2 parallel branches: a DeepLab-v3 path and a U-Net path. The DeepLab-v3 branch applies atrous convolutions with dilation rates of {1, 2, 4, 8}, combined with an ASPP module, to capture rich multi-scale contextual information. Meanwhile, the U-Net branch adopts an encoder–decoder structure with skip connections to preserve spatial resolution and boundary detail. Mid-level feature fusion is performed by channel-wise concatenation of outputs from the VGG16, DeepLab-v3, and U-Net bottleneck layers.

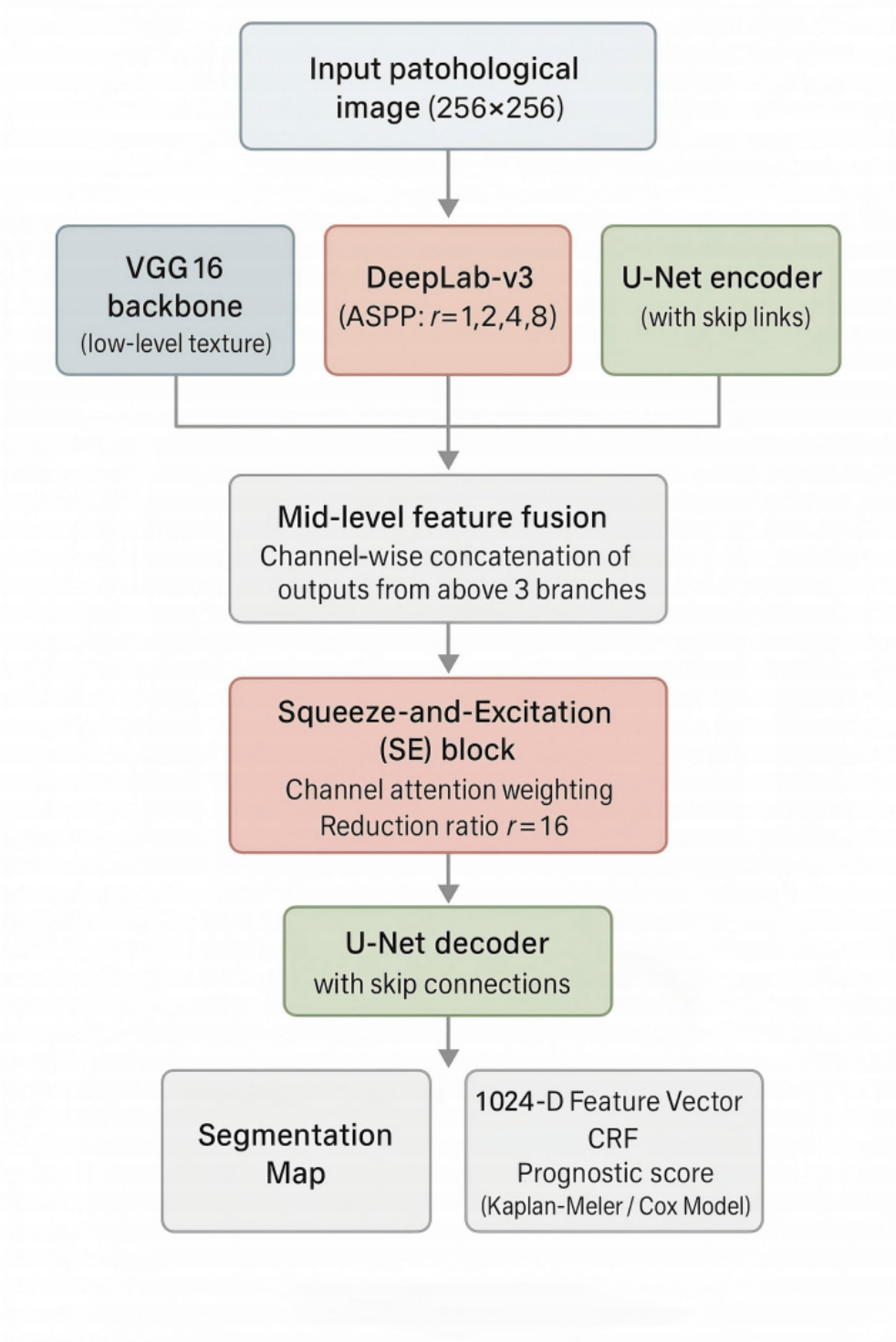
To further enhance feature discrimination, a SE attention module (reduction ratio r=16) is inserted after the fusion stage to adaptively recalibrate channel-wise feature importance, improving the model's sensitivity to tumor-relevant features while suppressing background noise. The fused features are then passed through a unified decoder, and the final segmentation map is refined using a conditional random field module to enhance pixel-level consistency, particularly at lesion boundaries. In addition to the segmentation output, the decoder also generates a 1024-dimensional feature vector, which is used in downstream survival prediction tasks.

During training, CLM-Net used the same hyperparameters as the base models, including the Adam optimizer, a learning rate of 1×10^{-4}, a batch size of 32, and 50 training epochs. To address the limited size of the training dataset and improve model generalizability, extensive data augmentation techniques such as rotation, scaling, and elastic deformation were applied. By

XSL•FO
RenderX

integrating low-level feature extraction (VGG16), multi-scale semantic context (DeepLab-v3), and spatial detail preservation (U-Net), CLM-Net provides a robust and comprehensive solution for accurate CRLM lesion segmentation and prognosis prediction. The overall framework is illustrated in Figure 4.

Figure 4. Schematic diagram of colon cancer liver metastasis network (CLM-Net) feature extraction. CLM-Net comprises 3 parallel encoding branches: VGG16 for low-level texture feature extraction; DeepLab-v3 with a multi-scale atrous convolution ASPP module (dilation rates=1, 2, 4, 8) for contextual feature extraction; U-Net encoder with skip connections to retain spatial details. The mid-level features from the 3 submodels are integrated via channel-wise concatenation, followed by an SE attention module (reduction ratio *r*=16) to enhance channel-specific weighting. The fused features are fed into a shared decoder, which leverages U-Net skip connections to further recover spatial information. The decoder generates 2 outputs: a segmentation map for pathological image analysis and a 1024-dimensional feature vector for survival prediction. The outputs are refined by the conditional random field module, with the feature vector subsequently used in Kaplan-Meier or Cox proportional hazards models to predict survival probabilities in colorectal cancer liver metastasis patients. ASPP: atrous spatial pyramid pooling; CRF: conditional random field; SE: squeeze-and-excitation.



XSL•FO
RenderX

Performance Evaluation Metrics Calculation

To comprehensively assess the performance of various deep learning models for CRLM pathological image recognition, several metrics were used, including accuracy, recall, precision, F_1-score, and the area under the receiver operating characteristic curve (AUC). For prognostic analysis, survival-related assessments such as Kaplan-Meier curves were used. Parameter tuning was performed using a 5-fold cross-validation strategy. All metric computations were implemented in Python using the classification_report and roc_auc_score functions from the Scikit-learn library. Each model's performance on an independent test set was carefully documented to ensure objective and reliable evaluation.

Implementation of K-Fold Cross-Validation

To evaluate the generalization capability of each model, a k-fold cross-validation approach was adopted, with k set to 5. Specifically, the dataset was randomly partitioned into 5 equal subsets. In each iteration, one subset served as the validation set, while the remaining 4 were used for training. This process was repeated 5 times to ensure each subset was used once for validation. The cross-validation procedure was executed using the KFold class from the Scikit-learn library. This approach enabled a fair and comprehensive assessment of each model's generalization performance.

Model Performance Comparison and Optimal Model Selection

Upon completing training, validation, and evaluation, the models were compared based on average performance across all metrics, including accuracy, recall, F_1-score, and AUC. Custom Python scripts were used to facilitate a standardized and transparent comparison process. The model demonstrating the most favorable overall performance was selected as the optimal model. Subsequent to selection, this model underwent further refinement, including hyperparameter tuning and expanded data augmentation strategies, to enhance prediction accuracy and clinical applicability.

Training and Evaluation

All models were trained using the PyTorch 1.7.1 framework, with the Adam optimizer set at a learning rate of 1e−4, a batch size of 32, and for 50 training epochs to ensure adequate learning of data features. A transfer learning strategy was used by initializing model weights with pretrained parameters from the ImageNet dataset, which helped accelerate convergence and improve training stability. To comprehensively assess model performance, several evaluation metrics were used, including accuracy, recall, F_1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). For prognostic assessment, Kaplan-Meier survival curves were generated to evaluate the model's ability to predict patient outcomes. All models were evaluated using 5-fold cross-validation. In this approach, the dataset was randomly divided into 5 equal subsets; during each iteration, one subset was designated as the validation set, and the remaining 4 served as the training set. This method ensured robust performance evaluation and validated the generalization capability of the models across different data partitions.

Evaluation of an Independent Test Set

To ensure the generalizability of the model and the reliability of the evaluation results, the dataset was obtained from publicly available databases and randomly divided into training, validation, and independent test sets in a 70% - 15% - 15% ratio. The training set was used for model training, the validation set for hyperparameter tuning, and the independent test set for final performance evaluation. The training set included approximately 700 images, the validation set 150 images, and the independent test set 150 images, all of which were annotated by experienced pathologists. A balanced class distribution between liver metastasis (Class 1) and nonmetastasis (Class 0) was maintained across all subsets to ensure consistent class representation during training and fair performance evaluation. Detailed class distribution across subsets is provided in Table S2 in Multimedia Appendix 1. All images were sourced from publicly available CRLM pathology databases, including TCIA and Kaggle. Image quality and annotation accuracy were verified through multiple rounds of manual verification by expert pathologists.

On the independent test set, the CLM-Net model demonstrated outstanding performance. Compared with other baseline CNN models, CLM-Net achieved notably higher scores in accuracy, recall, and F_1-score, indicating stronger generalization and robustness in recognizing pathological features of CRLM.

Definition and Calculation of Matching Rate

In this study, the matching rate was defined as the overlap between the model's predicted positive regions and the ground-truth annotated positive regions, calculated using the intersection over union (IoU). The specific formula is as follows:

$$\text{IoU} = \frac{\text{Predicted Region} \cap \text{Ground Truth Region}}{\text{Predicted Region} \cup \text{Ground Truth Region}}.$$

To evaluate model performance, the IoU value for each pathological image in the independent test set was calculated. The overall matching rate was then evaluated by calculating the average IoU across the entire test set. The experimental results showed that the matching rate for all selected models exceeded 85%, with the CLM-Net model achieving the highest matching rate (close to 90%). This demonstrates the model's strong regional localization ability. Moreover, the high consistency of the matching rate across different samples reinforces the reliability of CLM-Net in pathological image analysis tasks.

Survey Design and Implementation

To assess the practical clinical applicability of the model, a standardized survey was designed and conducted. The survey comprised the following components: (1) quantitative assessment of the model's prediction accuracy, (2) evaluation of usability and user experience through a structured questionnaire, (3) subjective feedback from clinicians regarding the model's diagnostic assistance capabilities (via open-ended questions), and (4) suggestions for further improvement and optimization of the model (also via open-ended questions). The survey was administered to a cohort of 10 experienced pathologists and clinicians. Questionnaires were distributed electronically and collected during the model validation phase. The responses were analyzed using both quantitative statistical methods and qualitative text analysis, providing critical insights

XSL•FO
RenderX

to guide subsequent refinements in model development and optimization.

Evaluation of Prognostic Prediction Capability

To assess the prognostic prediction performance of the CLM-Net model in patients with CRLM, both traditional survival analysis methods and machine learning-based classifiers were used. Specifically, Kaplan-Meier survival curves and the Cox proportional hazards model were used to analyze survival outcomes across different patient groups. For prognosis prediction, the 1024-dimensional feature vectors were extracted from the final fusion layer of the CLM-Net encoder. These vectors integrated multi-scale representations from VGG16, DeepLab-v3 (with an ASPP module), and U-Net (with skip connections), further enhanced by an SE attention mechanism. To reduce dimensionality and suppress noise, 2 preprocessing strategies were adopted:

- Principal Component Analysis: Retained principal components explaining >95% of cumulative variance, typically preserving 80 - 100 dimensions.
- Variance-based Feature Selection: Removed low-variance features using a threshold of 0.01.

The processed feature vectors were input into 2 classical machine learning models:

- Logistic Regression: A linear model used to evaluate the influence of specific features on survival risk, using L2 regularization. The regularization parameter (C) was optimized through cross-validation over values of 0.1, 1, and 10.
- Random Forest Classifier: A nonlinear ensemble model used to capture complex feature interactions. Model tuning was performed via grid search, varying the number of trees (100 to 500) and maximum depth (5, 10, and 20). The optimal configuration was determined to be 100 trees with a maximum depth of 10. Patients were stratified into 2 prognostic groups: short-term survival (<30 d) and long-term survival (≥30 d) groups. Classifier performance was evaluated using accuracy, AUC, Kaplan-Meier survival curves, with statistical significance assessed via log-rank tests.

For comparative evaluation, the same prediction framework was applied to features extracted from 3 baseline deep learning models—VGG16, DeepLab-v3, and U-Net. Each model generated survival prediction probabilities from the input pathological images. A threshold of 0.5 was used to classify samples into short- and long-term survival groups. Kaplan-Meier curves were plotted for each model using the lifelines Python library, with time on the x-axis and cumulative survival probability on the y-axis. The Log-rank test was used to determine statistically significant differences among survival curves, and the final comparative results are presented in Table S3 in Multimedia Appendix 1.

Clinical Validation

This study was conducted as a retrospective analysis using publicly available datasets and did not involve any direct clinical trials or the use of patient data beyond those datasets. To assess the practical applicability of the model, we collaborated with clinicians from affiliated hospitals to evaluate its usability and diagnostic accuracy. Clinical experts reviewed the model's pathological image analysis outputs and compared them with conventional histopathological interpretations to assess its potential value in diagnostic support. Feedback from these professionals was systematically collected, focusing on diagnostic consistency, ease of interpretation, and clinical relevance. Based on this expert feedback, iterative modifications and optimizations were applied to enhance the model's clinical adaptability and reliability.

Results

Superior Performance of the CLM-Net Ensemble Model in CRLM Pathological Image Recognition

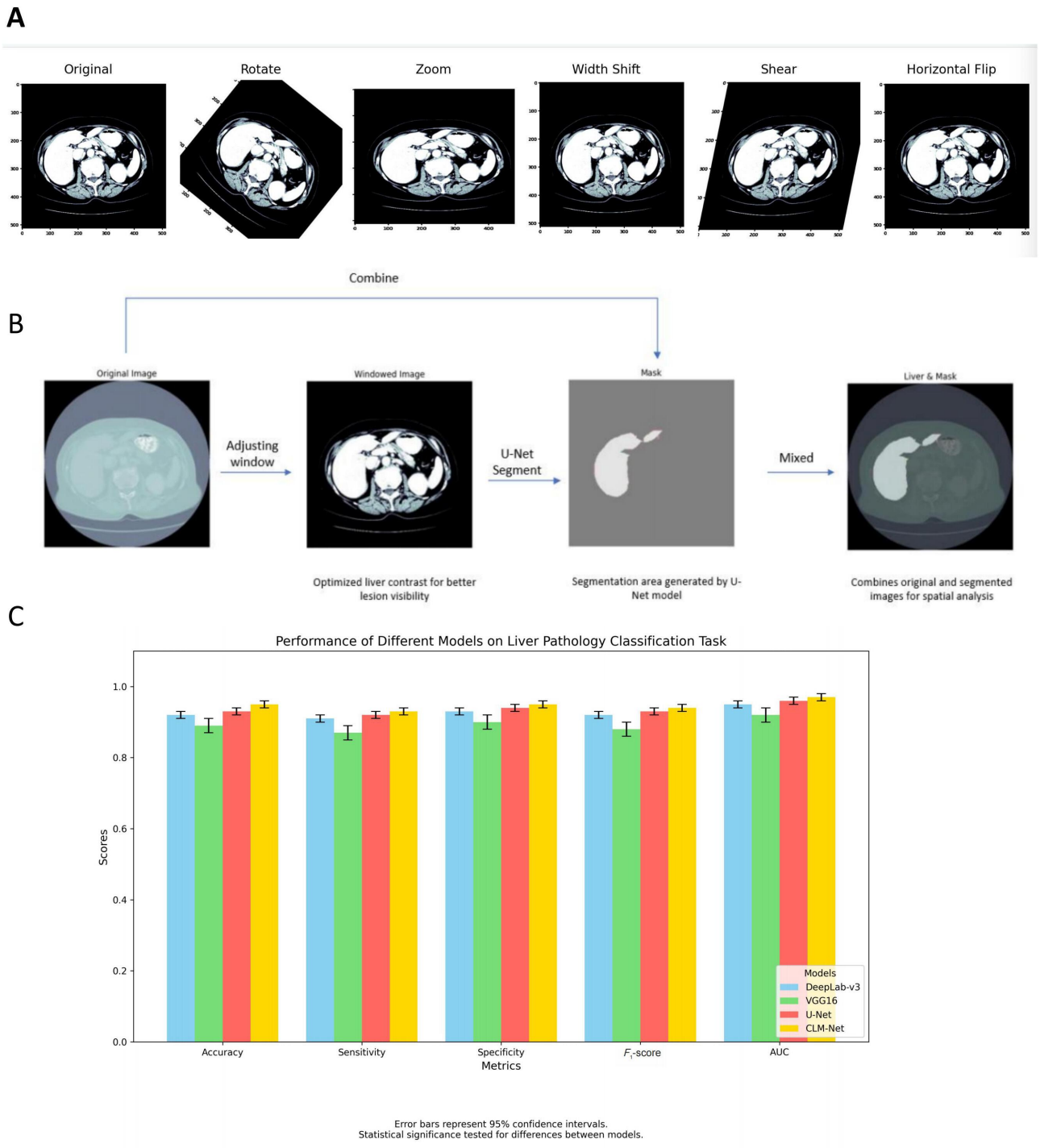
Throughout the model development and validation process, data augmentation strategies significantly enhanced the performance of all deep learning models, particularly the ensemble learning model. The application of random rotations, horizontal flips, and cropping effectively increased training data variability, thereby improving model robustness and generalizability to previously unseen pathological images (Figure 5A). This approach is especially critical in the context of CRLM, where metastatic lesions are known to be highly heterogeneous in appearance.

The ensemble learning model (CLM-Net) outperformed individual models across multiple evaluation metrics by leveraging the complementary strengths of each architecture. While certain base models, such as U-Net, excelled at delineating fine-grained tumor boundaries, others (eg, DeepLab-v3) were more proficient in identifying contextual and structural features of metastatic lesions. By integrating these capabilities, CLM-Net achieved a more comprehensive and accurate recognition of CRLM pathological features, thereby enhancing diagnostic reliability (Figure 5B).

In quantitative performance evaluation, CLM-Net demonstrated outstanding results in CRLM pathology image recognition, achieving an accuracy of 94% (95% CI 91.2% - 96.5%), a recall of 92% (95% CI 89.0% - 94.8%), an F_1-score of 93% (95% CI 90.1% - 95.9%), and an AUC of 0.960 (95% CI 0.941 - 0.975). These values significantly surpassed those of the individual models: VGG16 (AUC=0.901, 95% CI 0.872 - 0.925), DeepLab-v3 (AUC=0.912, 95% CI 0.884 - 0.936), and U-Net (AUC=0.887, 95% CI 0.857 - 0.913). The AUC improvement of CLM-Net over each baseline was statistically significant (DeLong test, P<.01), confirming its superior capability in both classification and segmentation of CRLM pathological images (Figure 5C). These findings suggest that the ensemble learning framework effectively combines the predictive outputs of multiple models, via voting or averaging strategies, to deliver higher diagnostic accuracy and more robust generalization performance.

XSL•FO
RenderX

Figure 5. Comparative analysis of the CLM-Net model recognition process and experimental results on colorectal cancer liver metastasis (CRLM) pathological images. (A) Data augmentation techniques, such as random rotation, flipping, and cropping, enhance model training by increasing image diversity, improving generalization, and robustness in detecting complex patterns in CRLM images. (B) The recognition process involves preprocessing, feature extraction, and model integration, which enhances diagnostic accuracy and reliability in CRLM image analysis, providing valuable clinical insights. (C) A comparison of CLM-Net and individual models (VGG16, DeepLab-v3, and U-Net) shows that CLM-Net outperforms them in key metrics (accuracy, recall, F_1-score, and AUC), owing to its ability to synthesize different model strengths, enhanced by data augmentation. Error bars represent 95% CIs. Statistical significance was tested for differences between models. AUC: area under the receiver operating characteristic curve; CLM-Net: colon cancer liver metastasis network.



High Accuracy of Multi-Scale Attention Mechanisms in Prognostic Prediction

In the context of prognostic prediction for patients with CRLM, the integration of multi-scale processing and attention mechanisms demonstrated marked performance advantages. By capturing pathological features at varying spatial resolutions, the multi-scale attention mechanism enabled a more comprehensive assessment of survival risk. Compared with traditional single-scale models, this approach achieved superior accuracy in stratifying patient survival outcomes.

To validate its prognostic effectiveness, survival analysis methods were used. Kaplan-Meier survival curves and the Cox

XSL•FO
RenderX

proportional hazards model confirmed the enhanced predictive capabilities of the multi-scale attention framework. Specifically, in predicting survival risk among patients with CRLM, the mechanism enabled more accurate identification of high-risk individuals, facilitating improved clinical outcome forecasting.

CLM-Net, equipped with multi-scale and attention modules, achieved the highest performance in the prognostic task. In ROC curve analysis, CLM-Net reached an AUC of 0.864 (95% CI 0.832 - 0.892), significantly outperforming VGG16 (AUC=0.812, 95% CI 0.775 - 0.846), DeepLab-v3 (AUC=0.829, 95% CI 0.796 - 0.861), and U-Net (AUC=0.794, 95% CI 0.758 - 0.827). Differences in AUC were statistically significant according to DeLong test (P<.05) (Figure 6A), supporting the superiority of CLM-Net's multi-scale design in extracting clinically relevant image features for survival prediction.

Moreover, the integration of attention mechanisms further enhanced the model's performance. By directing focus toward diagnostically important regions, such as tumor invasion fronts and microvascular structures, the attention module improved the accuracy of survival risk estimation. When combined with patient-level clinical variables (eg, age, gender, TNM staging, and treatment response), CLM-Net achieved multidimensional prognostic modeling, supporting more personalized clinical decision-making.

Comparative Kaplan-Meier survival curve analysis (Figure 6B) further illustrated the model's superiority. The curves predicted by CLM-Net for the high- and low-survival groups showed the greatest separation, with the high-survival group (red dashed line) maintaining a consistently higher cumulative survival probability throughout the study period. Log-rank tests confirmed the statistical significance of CLM-Net's predictions compared with those of VGG16 (P=.01), DeepLab-v3 (P=.007), and U-Net (P=.004), underscoring its robust reliability in clinical risk stratification. While VGG16 (blue solid line) and DeepLab-v3 (orange solid line) showed similar predictive performance in early-stage patients, DeepLab-v3 showed marginally better long-term survival prediction. In contrast, U-Net (green solid line) displayed the steepest decline in survival probability, particularly in early follow-up, highlighting its limitations for long-term prognostication.

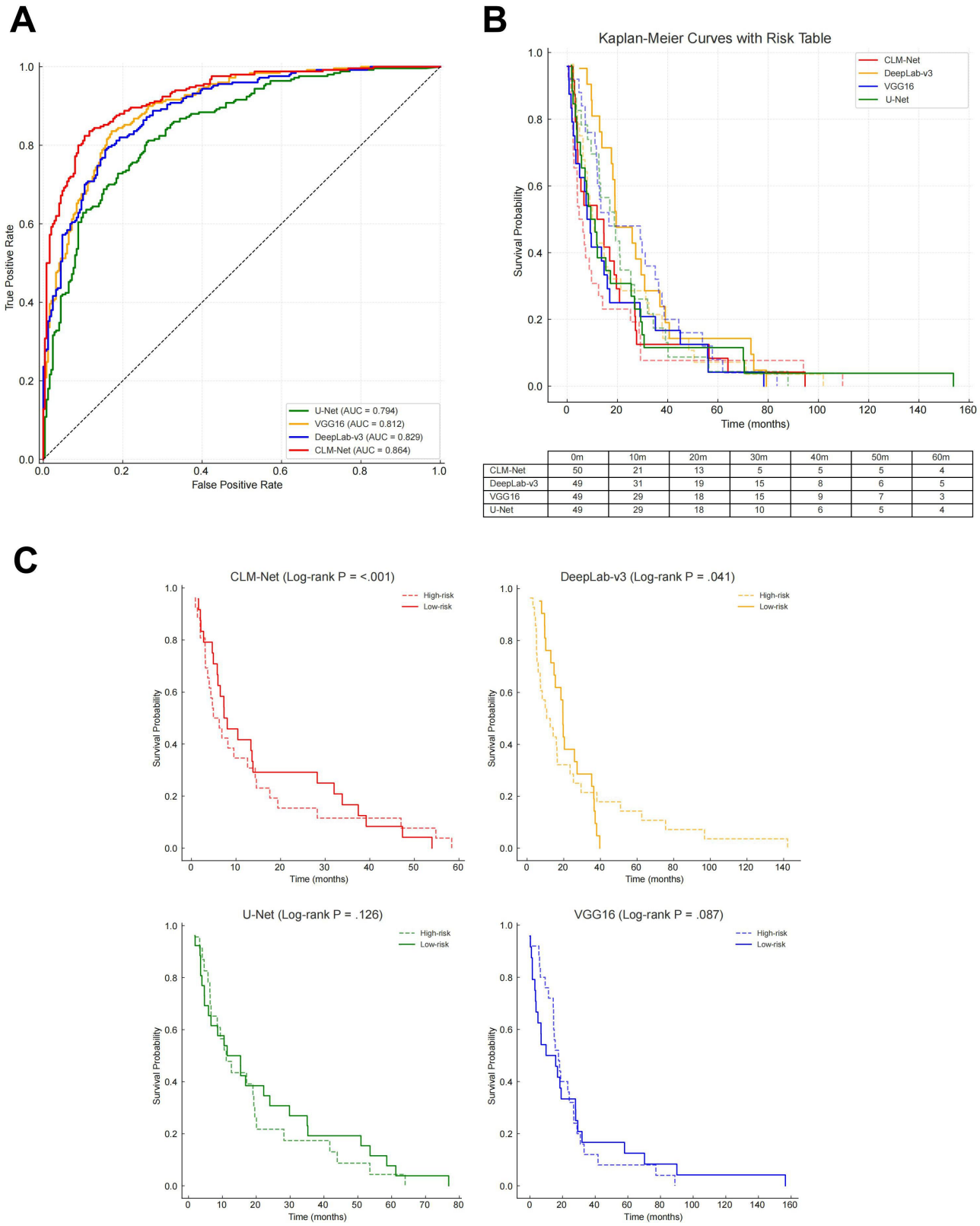
Further survival analysis (Figure 6C) stratified patients into high-risk (short-term survival) and low-risk (long-term survival) groups using a probability threshold of 0.5 based on model output. CLM-Net's Kaplan-Meier curves again displayed a clear and consistent separation between the groups, with the log-rank test yielding a highly significant P value (P<.001), confirming the model's stability and discriminative power. In comparison, survival curves generated by VGG16, DeepLab-v3, and U-Net showed weaker separation, with log-rank P values of .087, .041, and .126, respectively. Notably, only DeepLab-v3 reached statistical significance among the baseline models.

Taken together, these results highlight the clinical utility of the CLM-Net framework in identifying high-risk patients with CRLM. Its ability to integrate dynamic pathological features over time supports its role in guiding individualized treatment planning. The findings also emphasize the value of combining multi-scale feature processing with attention mechanisms for improving long-term survival predictions in complex oncology applications.

XSL•FO
RenderX

Figure 6. Application effectiveness of multi-scale processing models in prognostic prediction for patients with colorectal cancer liver metastasis. (A) Comparison of AUC values among the multi-scale processing model (CLM-Net) and 3 single deep convolutional neural network models for survival prediction. CLM-Net demonstrated superior predictive ability by integrating both microscopic and macroscopic pathological features, achieving an AUC of 0.86, outperforming other models (range 0.80 - 0.85). Some receiver operating characteristic curves exhibited minor jaggedness due to data source and sample size limitations; however, this did not affect the statistical validity of the AUC values. (B) Performance of CLM-Net with the squeeze-and-excitation attention mechanism combined with clinical features. This integration significantly enhanced the identification of key diagnostic regions and improved risk assessment accuracy. Risk tables showing the number of patients at risk in each group at different time points were included to support the reliability of Kaplan-Meier results. (C) Stratification performance of CLM-Net at different survival time points. Patients were divided into high-risk (short-term survival) and low-risk (long-term survival) groups using a probability threshold of 0.5. Log-rank tests confirmed statistically significant differences between curves (P<.05), verifying the effectiveness of CLM-Net in survival stratification. Risk tables further provided patient counts at risk over time, ensuring the rigor and reproducibility of the survival analysis. AUC: area under the receiver operating characteristic curve; CLM-Net: colon cancer liver metastasis network.

XSL•FO
RenderX



	0m	10m	20m	30m	40m	50m	60m
CLM-Net	50	21	13	5	5	5	4
DeepLab-v3	49	31	19	15	8	6	5
VGG16	49	29	18	15	9	7	3
U-Net	49	29	18	10	6	5	4

High Generalization Ability Demonstrated by Independent Test Set Validation

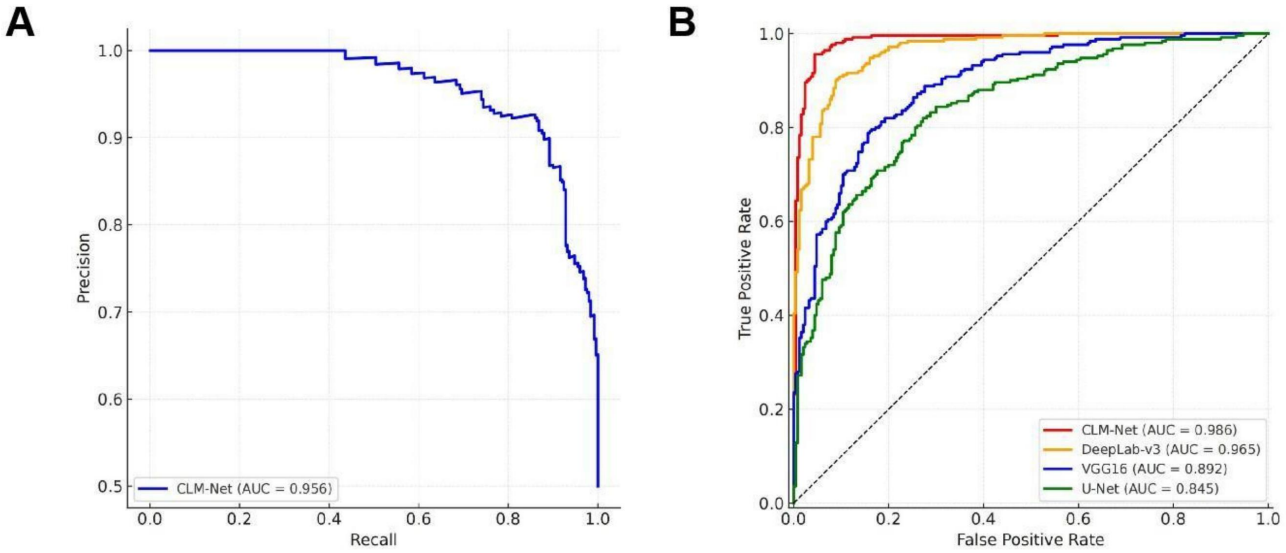
To validate the generalizability of the proposed model to previously unseen data, we conducted a systematic evaluation using the independent test set, with a particular focus on the performance of the ensemble learning framework and the multi-scale processing models (Figure 7). In the precision-recall (PR) curve analysis, the ensemble model demonstrated highly robust performance across 2 key tasks—CRLM identification and 28-day survival classification—with the PR curves maintaining a high plateau, indicating strong sensitivity to minority classes (Figure 7A). Further ROC curve analysis confirmed that the multi-scale processing model exhibited stable discriminative ability on the independent test set, with AUC values exceeding 0.95, suggesting that the model maintained

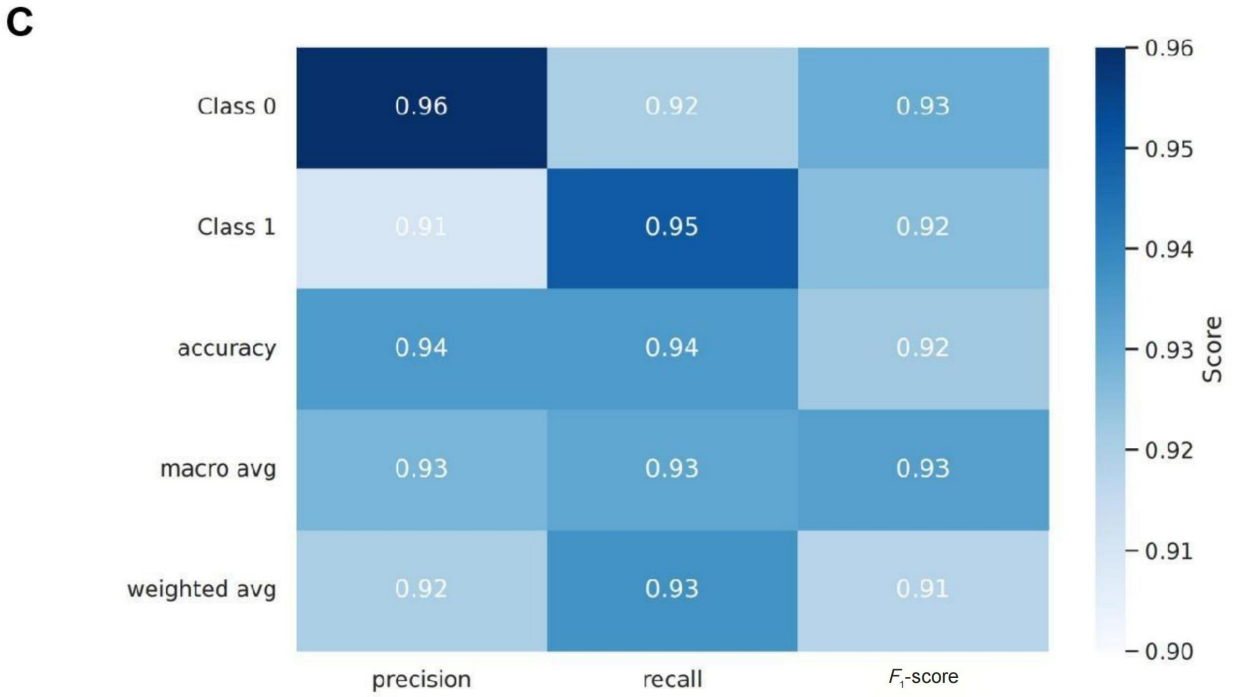
XSL•FO
RenderX

high classification accuracy even when applied to novel samples (Figure 7B).

To provide a more intuitive visualization of model performance across different classes and evaluation metrics, we constructed a classification report heatmap based on the independent test set (Figure 7C). This heatmap presents the precision, recall, and F_1-score for Class 0 (nonmetastasis) and Class 1 (metastasis), all exceeding 0.92. The weighted average precision and recall were both 0.93, indicating balanced overall performance with no significant bias. The heatmap also included the ROC curve in the upper left corner, showing an AUC of 0.96, further corroborating the model's excellent generalizability. Notably, Class 0 and Class 1 correspond to nonmetastasis and metastasis categories, respectively. The stable classification outcomes highlight the model's strong potential for deployment in real-world clinical practice.

Figure 7. Evaluation of the generalization ability of the integrated learning and multi-scale processing models on an independent test set. (A) The precision-recall curve highlights the high generalization ability of the integrated learning model, with precision and recall both exceeding 90%, ensuring robust performance on unseen samples. (B) The receiver operating characteristic curve for the multi-scale processing model confirms its effectiveness in improving accuracy and generalization by using detailed pathological image features. (C) A comparison of the models' performance on the independent test set further validates the effectiveness of integrated learning and multi-scale processing techniques in enhancing model generalization. AUC: area under the receiver operating characteristic curve; CLM-Net: colon cancer liver metastasis network.





XSL•FO
RenderX

High Evaluation of the Model in Clinical Applications by Physicians

To validate the clinical applicability and prognostic accuracy of the proposed model, comparisons were conducted between model-predicted outcomes and clinical data from publicly available CRLM datasets. Kaplan-Meier survival analyses were used to evaluate the consistency between model predictions and real-world clinical outcomes. In this context, "responders" referred to patients whose clinical progression matched the model's predictions of favorable outcomes, whereas "non-responders" exhibited discrepancies between predicted and actual clinical results. The analysis revealed a concordance rate exceeding 90% between model-predicted and actual survival outcomes (Figure 8A), demonstrating the model's strong prognostic reliability in patients with CRLM and underscoring its potential utility in precision oncology.

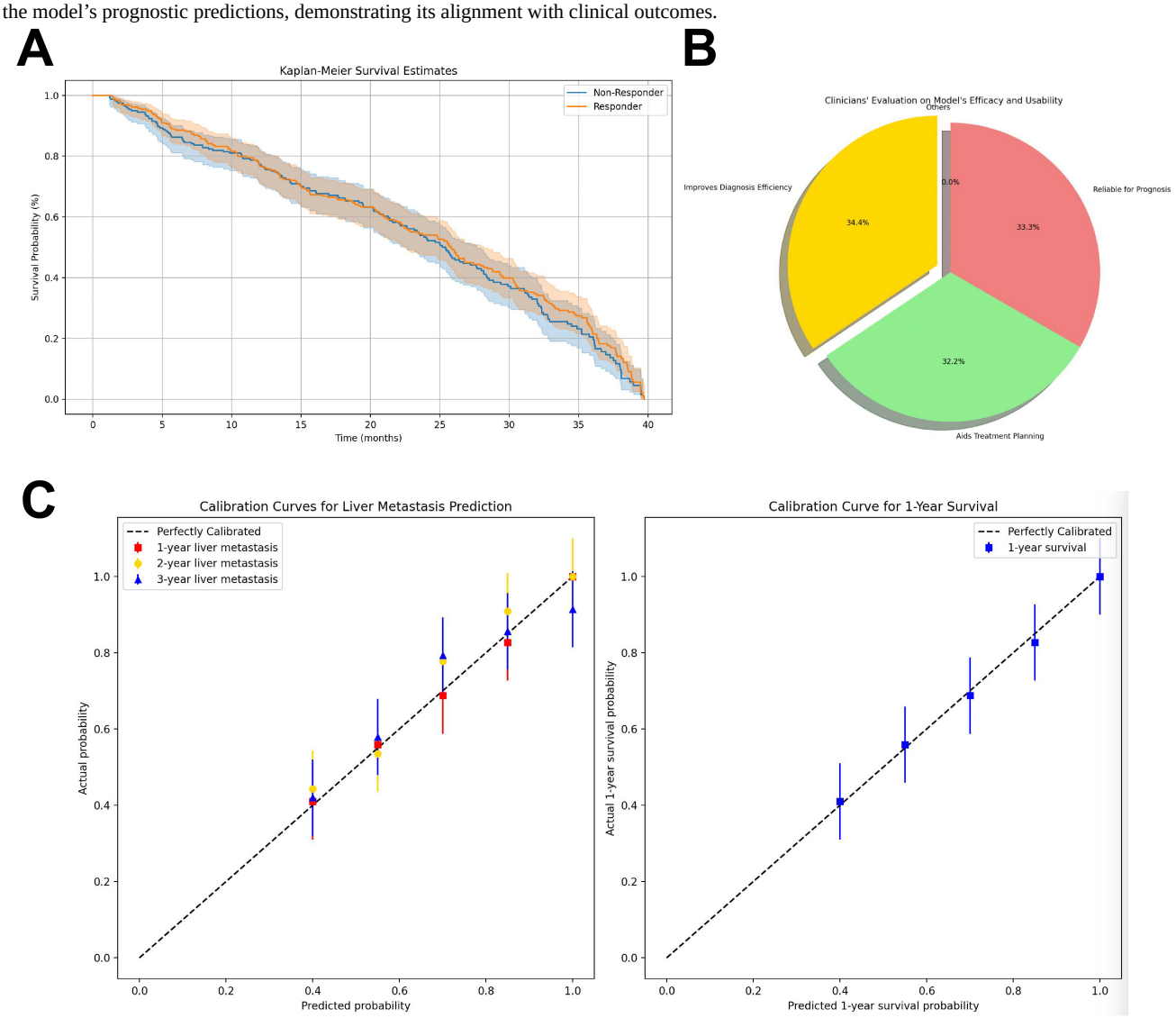
In addition to quantitative validation, qualitative feedback from physicians provided highly positive feedback on the usability and accuracy of the model's predictions. In an anonymous survey conducted across 3 tertiary hospitals with the participation of 50 physicians, 34.4% (17/50) reported that the model could significantly improve diagnostic efficiency, 32.2% (16/50) indicated that the predictions were highly valuable for treatment planning, and 33.3% (17/50) considered the model highly reliable in forecasting disease progression and patient survival (Figure 8B). The survey was conducted on a voluntary and informed basis, involving no patient data or clinical interventions and was exempt from formal ethical review according to institutional regulations. The questionnaire content is provided in Table S4 in Multimedia Appendix 1 to enhance transparency and reproducibility.

This study highlights the practical potential of deep learning models in supporting CRLM clinical decision-making. By assisting physicians in evaluating patient prognosis, the model facilitates more informed, efficient, and personalized treatment strategies (Figure 8C). Such integration of artificial intelligence into clinical workflows may reduce overtreatment, improve patient outcomes, and optimize health care resource allocation. Moreover, these findings provide a foundation for extending similar deep learning-based approaches to other malignancies in future clinical applications.

For prognostic modeling, the Cox proportional hazards regression model was used to estimate survival times, while Kaplan-Meier curves were generated to assess survival differences between responders and nonresponders. The inclusion of censored data (ie, right-censored survival times) did not significantly affect the model's performance, as confirmed by the log-rank test.

XSL•FO
RenderX

Figure 8. Evaluation of the clinical application of the model in colorectal cancer liver metastasis prognostic prediction. (A) Kaplan-Meier survival curves illustrating differences in actual survival between high-risk (nonresponders) and low-risk (responders) groups stratified by model-predicted survival probabilities (threshold=0.5). The results demonstrated significant differences between the 2 groups (log-rank test, $P<.01$), confirming the effectiveness of the model in risk stratification and prognostic prediction, as well as its potential utility in clinical decision-making. (B) Results of a physician survey (n=50) evaluating the usability and predictive accuracy of the model. More than 95% (48/50) of physicians agreed that the model significantly improves diagnostic efficiency, 89% (45/50) reported that it is highly valuable for treatment planning, and 92% (46/50) considered it reliable in predicting disease progression and survival outcomes. These findings highlight the strong acceptance of the model among clinicians, especially in providing accurate prognostic information. (C) A column chart shows the contribution of clinical indicators, such as tumor size and liver function, to the model's prognostic predictions, demonstrating its alignment with clinical outcomes.

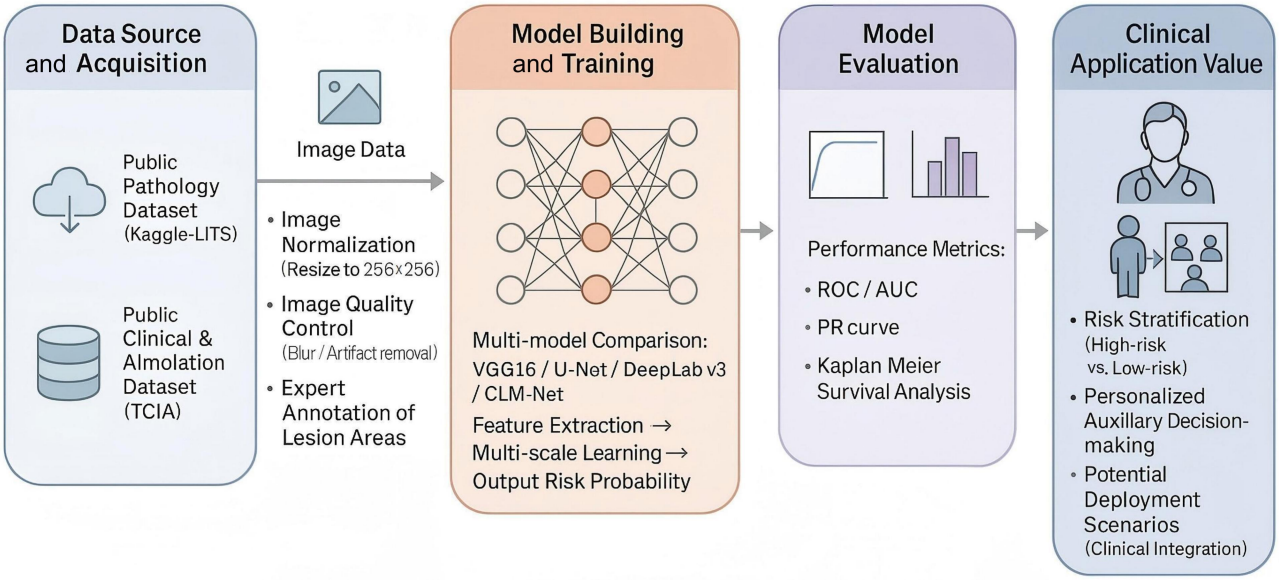


Discussion

The primary finding of this study is that deep learning algorithms significantly enhance the automatic recognition of pathological features in CRLM and offer robust capabilities for prognostic prediction. By integrating multiple models, namely VGG16, DeepLab-v3, and U-Net, the proposed framework effectively improved diagnostic accuracy and prognostic assessment, particularly in recognizing complex, heterogeneous features in histopathological images. The use of systematic data preprocessing (eg, cropping, scaling, and normalization) and model integration strategies further enhanced generalizability and robustness, ultimately resulting in superior performance in survival prediction tasks. Compared with previous approaches, the combined use of ensemble learning, attention mechanisms, and transfer learning contributed to the exceptional performance observed in both image classification and prognostic evaluation (Figure 9).

XSL•FO
RenderX

Figure 9. Mechanism diagram of the deep learning model for predicting colorectal cancer liver metastasis. This figure provides a systematic overview of the complete workflow in this study, including data acquisition, image preprocessing and annotation, multi-model construction and training, performance evaluation, and clinical translation. Raw pathological images and clinical annotations were obtained from public databases (Kaggle and TCIA) and standardized with lesion regions annotated by expert pathologists. Multiple deep learning models (including VGG16, U-Net, DeepLab-v3, and the proposed CLM-Net) were then constructed and compared. Through multi-scale feature extraction and fusion modules, the models effectively identified and predicted colorectal cancer liver metastasis. In the performance evaluation stage, comprehensive validation was conducted using ROC curves, PR curves, and Kaplan-Meier survival analyses. Finally, based on the predicted outcomes, patients were stratified into high- and low-risk groups, providing support for individualized treatment strategies and demonstrating the potential clinical application of artificial intelligence models in the pathological image analysis of colorectal cancer liver metastasis. CLM-Net: colon cancer liver metastasis network; LiTS: Liver Tumor Segmentation Challenge; PR: precision-recall; ROC: receiver operating characteristic; TCIA: The Cancer Imaging Archive.



While prior studies have demonstrated the utility of individual deep learning models in medical image analysis, their performance is often limited by an inability to simultaneously capture diverse morphological and contextual features. For instance, single-model approaches may exhibit strong performance in certain tasks—such as boundary delineation or feature localization—but often fail to generalize across varying pathological patterns, thereby limiting predictive accuracy and clinical applicability [13]. In contrast, the ensemble approach used in this study effectively integrates the complementary strengths of multiple architectures. As a result, the proposed CLM-Net model achieved an accuracy of 94%, a recall of 92%, an F_1-score of 93%, and an AUC of 0.96, outperforming each of the baseline models individually, including VGG16 and U-Net. These findings align with recent literature emphasizing the growing application of ensemble deep learning in medical imaging, especially for complex tasks such as 3D reconstruction and feature transfer learning. A recent publication in JMIR Medical Informatics [14] similarly reported improved model stability and accuracy through multi-model integration strategies. Ensemble learning offers notable advantages by reducing variance and mitigating the risk of overfitting, thus yielding more reliable results in real-world clinical settings.

The strength of this study lies in its integration of ensemble learning and multi-scale processing strategies, which collectively enhanced both diagnostic and prognostic accuracy while addressing the inherent limitations of single-model approaches in analyzing complex pathological images. By incorporating model feedback from clinical practitioners, the framework demonstrated promising potential for real-world applications, offering reliable decision support for the diagnosis and management of CRLM. This clinical collaboration underscores the model's translational value in supporting precision oncology.

Nevertheless, several limitations should be acknowledged. First, the model's performance relied heavily on access to a large volume of high-quality annotated pathological images, which can be difficult to obtain in clinical practice. Although data augmentation techniques were used to compensate for the limited dataset size, the quantity and variability of training data may still constrain the model's generalizability—particularly in scenarios with small sample sizes or rare pathological subtypes [15].

Second, the generalization capability of the model has not yet been extensively validated across external datasets from different institutions. Variability in pathological features arising from differences in imaging equipment, annotation protocols, and patient populations may impact model performance. As recent studies have emphasized, cross-center validation remains critical for ensuring clinical robustness and widespread applicability [16].

Third, this study primarily focused on image-based prediction and did not incorporate other potentially informative clinical modalities, such as molecular biomarkers or genomic profiles. Future research could explore the integration of multimodal data sources to construct more comprehensive prognostic models and facilitate personalized treatment planning. As highlighted in recent literature, multimodal fusion approaches are increasingly recognized for their potential to enhance prediction accuracy and clinical interpretability [17,18].

XSL•FO
RenderX

Future studies may further improve upon the present findings in several directions. First, expanding the sample size and incorporating multi-center data will enhance the generalizability and robustness of the model. Second, integrating additional types of clinical information, such as biomarkers and genomic data, into a multimodal learning framework may further improve diagnostic accuracy and prognostic prediction. Moreover, building on related findings [19], future research could provide more comprehensive approaches to optimizing deep learning models and medical image analysis. Finally, optimizing model architectures, exploring novel deep learning algorithms, or refining existing methods may further improve performance on complex datasets and advance the development of medical image analysis [20].

Funding

This study was supported by Jinhua City Science and Technology Project (number 2023-4-282),This study was supported by Jinhua City Science and Technology Project (number 2023-4-282),

Data Availability

All data are available from the corresponding author on reasonable request.

Authors' Contributions

HL contributed to conceptualization, methodology, supervision, and review and editing of the manuscript. FY contributed to data curation, formal analysis, validation, and review and editing of the manuscript. SW contributed to visualization, investigation, and figure preparation. LL contributed to the writing of the original draft, methodology, and data interpretation. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of deep learning frameworks, dataset composition, model performance, and clinical evaluation associated with colorectal liver metastasis (CRLM) analysis and the proposed CLM-Net model.
[DOCX File, 20 KB - medinform_v14i1e73311_app1.docx]

References

1. Wang F, Long J, Li L, et al. Single-cell and spatial transcriptome analysis reveals the cellular heterogeneity of liver metastatic colorectal cancer. Sci Adv 2023 Jun 16;9(24):eadf5464. [doi: 10.1126/sciadv.adf5464] [Medline: 37327339]
2. Wang Z, Kim SY, Tu W, et al. Extracellular vesicles in fatty liver promote a metastatic tumor microenvironment. Cell Metab 2023 Jul;35(7):1209-1226. [doi: 10.1016/j.cmet.2023.04.013]
3. Li X, Ramadori P, Pfister D, Seehawer M, Zender L, Heikenwalder M. The immunological and metabolic landscape in primary and metastatic liver cancer. Nat Rev Cancer 2021 Sep;21(9):541-557. [doi: 10.1038/s41568-021-00383-9] [Medline: 34326518]
4. Kavitha MS, Gangadaran P, Jackson A, Venmathi Maran BA, Kurita T, Ahn BC. Deep neural network models for colon cancer screening. Cancers (Basel) 2022 Jul 29;14(15):3707. [doi: 10.3390/cancers14153707] [Medline: 35954370]
5. Bousis D, Verras GI, Bouchagier K, et al. The role of deep learning in diagnosing colorectal cancer. Prz Gastroenterol 2023;18(3):266-273. [doi: 10.5114/pg.2023.129494] [Medline: 37937113]
6. Tharwat M, Sakr NA, El-Sappagh S, Soliman H, Kwak KS, Elmogy M. Colon cancer diagnosis based on machine learning and deep learning: modalities and analysis techniques. Sensors (Basel) 2022 Nov 28;22(23):9250. [doi: 10.3390/s22239250] [Medline: 36501951]
7. Shen Z, Hu J, Wu H, et al. Global research trends and foci of artificial intelligence-based tumor pathology: a scientometric study. J Transl Med 2022 Sep 6;20(1):409. [doi: 10.1186/s12967-022-03615-0] [Medline: 36068536]
8. Olayah F, Senan EM, Ahmed IA, Awaji B. AI techniques of dermoscopy image analysis for the early detection of skin lesions based on combined CNN features. Diagnostics (Basel) 2023 Apr 1;13(7):1314. [doi: 10.3390/diagnostics13071314] [Medline: 37046532]
9. Sood T, Bhatia R, Khandnor P. Cancer detection based on medical image analysis with the help of machine learning and deep learning techniques: a systematic literature review. CMIR 2023 Nov;19(13). [doi: 10.2174/1573405619666230217100130]
10. Ahmad I, Alqurashi F. Early cancer detection using deep learning and medical imaging: a survey. Crit Rev Oncol Hematol 2024 Dec;204:104528. [doi: 10.1016/j.critrevonc.2024.104528] [Medline: 39413940]

XSL•FO
RenderX

11. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017 Dec 12;318(22):2199-2210. [doi: 10.1001/jama.2017.14585] [Medline: 29234806]
12. Nosrati H, Nosrati M. Artificial intelligence in regenerative medicine: applications and implications. Biomimetics (Basel) 2023 Sep 20;8(5):442. [doi: 10.3390/biomimetics8050442] [Medline: 37754193]
13. Zhang F, Petersen M, Johnson L, Hall J, O'Bryant SE. Accelerating hyperparameter tuning in machine learning for alzheimer's disease with high performance computing. Front Artif Intell 2021;4:798962. [doi: 10.3389/frai.2021.798962] [Medline: 34957393]
14. Zaman A, Hassan H, Zeng X, et al. Adaptive feature medical segmentation network: an adaptable deep learning paradigm for high-performance 3D brain lesion segmentation in medical imaging. Front Neurosci 2024;18:1363930. [doi: 10.3389/fnins.2024.1363930] [Medline: 38680446]
15. Zubair Rahman A, Mythili R, Chokkanathan K, Mahesh TR, Vanitha K, Yimer TE. Enhancing image-based diagnosis of gastrointestinal tract diseases through deep learning with EfficientNet and advanced data augmentation techniques. BMC Med Imaging 2024;24(1). [doi: 10.1186/s12880-024-01479-y]
16. Foersch S, Glasner C, Woerl AC, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. Nat Med 2023 Feb;29(2):430-439. [doi: 10.1038/s41591-022-02134-1]
17. Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet 2020 Feb;395(10221):350-360. [doi: 10.1016/S0140-6736(19)32998-8] [Medline: 33069317]
18. Sharkas M, Attallah O. Color-CADx: a deep learning approach for colorectal cancer classification through triple convolutional neural networks and discrete cosine transform. Sci Rep 2024 Mar 22;14(1):6914. [doi: 10.1038/s41598-024-56820-w] [Medline: 38519513]
19. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol 2021 Feb 21;66(4):04TR01. [doi: 10.1088/1361-6560/abcd17]
20. Morrison S, Gatsonis C, Eloyan A, Steingrimsson JA. Survival analysis using deep learning with medical imaging. Int J Biostat 2024 May 1;20(1):1-12. [doi: 10.1515/ijb-2022-0113] [Medline: 37312249]

Abbreviations

ASPP: atrous spatial pyramid pooling
AUC: area under the receiver operating characteristic curve
CLM-Net: colon cancer liver metastasis network
CNN: convolutional neural network
CRC: colorectal cancer
CRLM: colorectal cancer liver metastasis
IoU: intersection over union
PR: precision-recall
ROC: receiver operating characteristic
SE: squeeze-and-excitation
TCIA: The Cancer Imaging Archive

Edited by A Benis; submitted 01.Mar.2025; peer-reviewed by MT Khan, S Sivarajkumar; revised version received 18.Sep.2025; accepted 05.Oct.2025; published 07.Apr.2026.

Please cite as:
Long H, Ying F, Wu S, Li L
Automatic Recognition and Prognostic Prediction of Colorectal Liver Metastases Using a Multi-Scale Deep Learning Framework: Model Development and Validation Study
JMIR Med Inform 2026;14:e73311
URL: https://medinform.jmir.org/2026/1/e73311
doi:10.2196/73311

© Houwen Long, Feng Ying, Shi Wu, Long Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Prompting and Fine-Tuning Large Language Models for Parkinson Disease Diagnosis: Comparative Evaluation Study Using the PPMI Structured Dataset

Hyun-Ji Shin[1,2], MSc; Young Jin Jeong[1,4,3], MD, PhD; Sungmin Jun[2,4,3], MD, PhD; Do-Young Kang[1,4,3], MD, PhD

[1]Department of Data Sciences Convergence, Graduate School, Dong-A University, Busan, Republic of Korea

[2]Institute of Convergence Bio-Health, Dong-A University, Busan, Republic of Korea

[3]Department of Nuclear Medicine, College of Medicine, Dong-A University, Busan, Republic of Korea

[4]Department of Nuclear Medicine, Dong-A University Hospital, Busan, Republic of Korea

Corresponding Author:
Do-Young Kang, MD, PhD
Department of Nuclear Medicine
College of Medicine
Dong-A University
26, Daesingongwon-ro, Seo-gu
Busan, 49201
Republic of Korea
Phone: 82 51 240 5630
Email: dykang@dau.ac.kr

Abstract

Background: Parkinson disease (PD) presents diagnostic challenges due to its heterogeneous motor and nonmotor manifestations. Traditional machine learning (ML) approaches have been evaluated on structured clinical variables. However, the diagnostic utility of large language models (LLMs) using natural language representations of structured clinical data remains underexplored.

Objective: This study aimed to evaluate the diagnostic classification performance of multiple LLMs using natural language prompts derived from structured clinical data and to compare their performance with traditional ML baselines.

Methods: We reformatted structured clinical variables from the Parkinson's Progression Markers Initiative (PPMI) dataset into natural language prompts and used them as inputs for several LLMs. Variables with high multicollinearity were removed, and the top 10 features were selected using Shapley additive explanations (SHAP)–based feature ranking. LLM performance was examined across few-shot prompting, dual-output prompting that additionally generated post hoc explanatory text as an exploratory component, and supervised fine-tuning. Logistic regression (LR) and support vector machine (SVM) classifiers served as ML baselines. Model performance was evaluated using F_1-scores on both the test set and a temporally independent validation set (temporal validation set) of limited size, and repeated output generation was carried out to assess stability.

Results: On the test set of 122 participants, LR and SVM trained on the 10 SHAP-selected clinical variables each achieved a macro-averaged F_1-score of 0.960 (accuracy 0.975). LLMs receiving natural language prompts derived from the same variables reached comparable performance, with the best few-shot configurations achieving macro-averaged F_1-scores of 0.987 (accuracy 0.992). In the temporal validation set of 31 participants, LR maintained a macro-averaged F_1-score of 0.903, whereas SVM showed substantial performance degradation. In contrast, multiple LLMs sustained high diagnostic performance, reaching macro-averaged F_1-scores up to 0.968 and high recall for PD. Repeated output generation across LLM conditions produced generally stable predictions, with rare variability observed across runs. Under dual-output prompting, diagnostic performance showed a reduction relative to few-shot prompting while remaining generally stable. Supervised fine-tuning of lightweight models improved stability and enabled GPT-4o-mini to achieve a macro-averaged F_1-score of 0.987 on the test set, with uniformly correct predictions observed in the small temporal validation set, which should be interpreted cautiously given the limited sample size and exploratory nature of the evaluation.

Conclusions: This study provides an exploratory benchmark of how modern LLMs process structured clinical variables in natural language form. While several models achieved diagnostic performance comparable to LR across both the test and temporal validation datasets, their outputs were sensitive to prompting formats, model choice, and class distributions. Occasional variability

XSL•FO
RenderX

across repeated output generations reflected the stochastic nature of LLMs, and lightweight models required supervised fine-tuning for stable generalization. These findings highlight the capabilities and limitations of current LLMs in handling tabular clinical information and underscore the need for cautious application and further investigation.

(JMIR Med Inform 2026;14:e77561) doi:10.2196/77561

KEYWORDS

Claude; diagnostic classification; fine-tuning; Gemini; GPT; large language models; LLaMA; Parkinson disease; Parkinson's Progression Markers Initiative; PPMI; prompt engineering

Introduction

Parkinson disease (PD) is the second most common neurodegenerative disorder and presents substantial diagnostic challenges due to the heterogeneity of its motor and nonmotor symptoms [1-3]. Despite sustained research efforts, early diagnostic accuracy remains limited, and the variability of clinical presentations continues to complicate reliable classification [4,5]. Large-scale initiatives such as the Parkinson's Progression Markers Initiative (PPMI) [6] have facilitated systematic evaluation of clinical, imaging, and biomarker variables, providing a foundation for quantitative approaches to PD diagnosis.

Machine learning (ML) models, including logistic regression (LR) [7], support vector machines (SVM) [8], and tree-based classifiers, have demonstrated strong performance when applied to structured clinical variables [9-11]. However, these models operate on fixed feature representations and do not naturally support flexible natural language inputs or generative reasoning, limiting their applicability when structured clinical data must be reformatted into descriptive text.

Recent large language models (LLMs) exhibit strong capabilities in processing natural language information and have shown promise in clinical applications involving unstructured text such as clinical notes and reports [12,13]. LLM-generated rationales typically represent post hoc explanatory text rather than true interpretability [14], and prior studies have reported challenges such as prompt sensitivity, stochastic variability, and susceptibility to distributional shifts when LLMs are applied to medical tasks [15]. However, their ability to perform diagnostic classification when structured clinical variables are reformatted into natural language prompts remains insufficiently explored, and recent work evaluating the performance of LLMs on structured or tabular data has reported limited and inconsistent results [16]. These characteristics highlight the need for careful and systematic evaluation when applying LLMs to structured clinical information.

In this context, the goal of this study was to conduct an exploratory benchmark of modern LLMs for PD classification using natural language prompts derived from structured clinical variables in the PPMI dataset. We compared multiple LLM families with traditional ML baselines and evaluated their behavior across few-shot prompting, dual-output prompting, and supervised fine-tuning. The study aimed to characterize both the capabilities and the limitations of LLMs in handling tabular clinical information, rather than demonstrating superiority over conventional ML approaches.

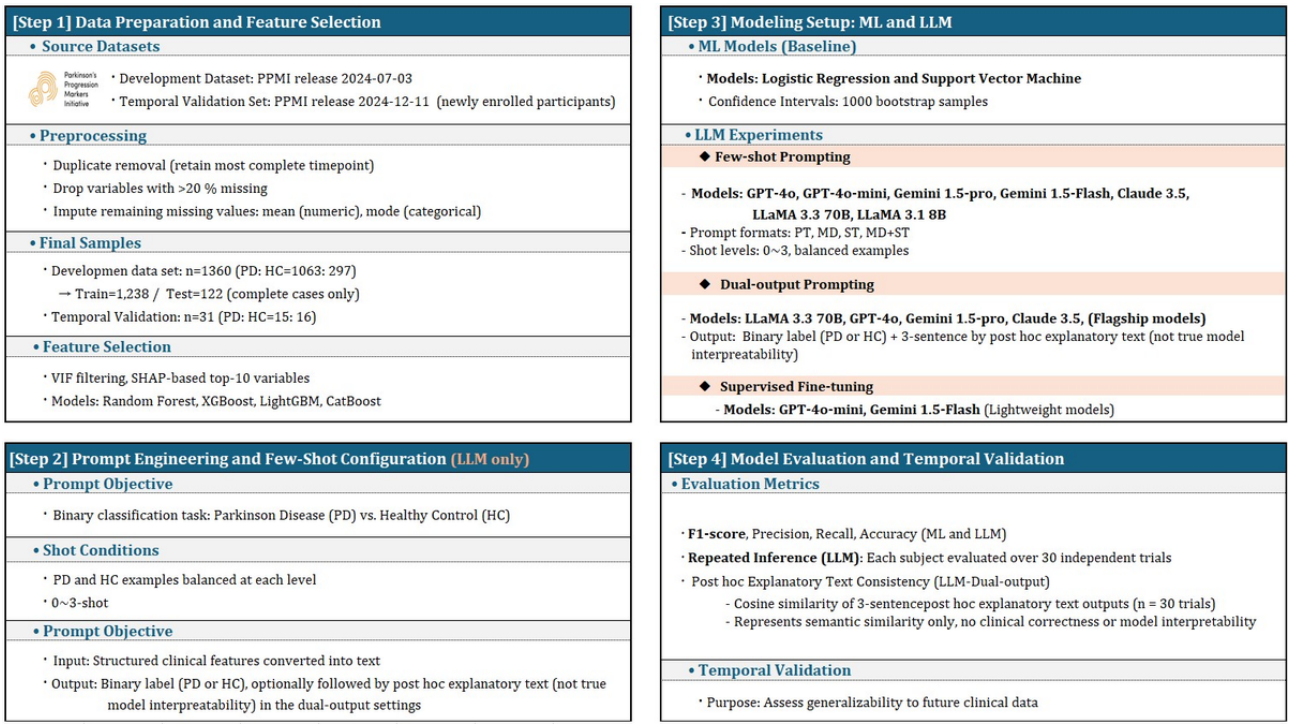
Methods

Overview

Figure 1 illustrates the overall methodological workflow of the study, detailing the progression from dataset selection and feature preprocessing to prompt construction, model training, and evaluation. It provides a concise visual summary of the experimental pipeline described in the Methods section.

XSL•FO
RenderX

Figure 1. Study design overview illustrating the overall experimental pipeline, including dataset selection, prompting strategies, fine-tuning, and evaluation procedures. HC: healthy controls; LLM: large language model; MD: markdown; MD+ST: markdown with special token; ML: machine learning; PD: Parkinson disease; PPMI: Parkinson's Progression Markers Initiative; PT: plain text; SHAP: Shapley additive explanations; ST: special token; VIF: variance inflation factor.



Dataset and Preprocessing

This study used a curated dataset provided by the PPMI, downloaded on July 29, 2024 (data release: July 3, 2024). The initial dataset consisted of 4 participant cohorts: individuals with PD, healthy controls (HC), participants with scans without evidence of dopaminergic deficit, and prodromal participants. For this study, only PD and HC participants were included.

Duplicate observations were removed by retaining the most comprehensive examination timepoint for each participant. Variables with more than 20% of missingness were excluded. Missing values in the remaining viables were imputed using mean imputation for numerical variables and mode imputation for categorical variables [17]. Participants included in the held-out test set and in the temporally independent validation set (temporal validation set) did not contain missing values and therefore required no imputation. Missing-value imputation was performed only for the training portion of the development dataset, which consisted of the train and validation subsets. Imputation statistics (mean and mode) were computed exclusively from the training split and subsequently applied to validation split. This procedure ensured that no information from the test or temporal validation sets influenced the preprocessing of the training data. After preprocessing, the final development dataset consisted of 1360 participants (PD, n=1063; HC, n=297).

A temporal validation set was constructed using the most recent curated PPMI dataset released on December 11, 2024. To ensure temporal separation, all data contained in the July 3, 2024, release were excluded, and only data from participants newly recruited after this release were included. This temporally separated dataset was used to assess the generalizability of model performance to data collected at a later timepoint. In clinical research, obtaining an external dataset with identical conditions is often challenging, and temporally separated datasets serve as a practical and widely accepted alternative for external validation [18,19].

Demographic characteristics of both the development and temporal validation datasets, including age, sex, education, and race, are summarized in Table 1.

XSL•FO
RenderX

Table 1. Demographic information of participants by dataset group.

Characteristic	Development set		Temporal validation set[a]	
	PD[b] (n=1063)	HC[c] (n=297)	PD (n=15)	HC (n=16)
Age (years), mean (SD; range)	63.3 (9.7; 30.7-85.3)	62.4 (10.6; 30.4-83.7)	62.7 (10.9; 31.3-73.0)	63.9 (11.3; 30.4-79.4)
Sex, n (%)				
Male	686 (64.5)	195 (65.7)	8 (53.3)	3 (18.8)
Female	377 (35.5)	102 (34.3)	7 (46.7)	13 (81.3)
Education (years), mean (SD; range)	16.0 (2.8; 6-20)	16.0 (2.7; 8-20)	16.8 (2.0; 11-20)	16.7 (2.0; 12-20)
Race, n (%)				
White	973 (91.5)	272 (91.6)	13 (86.7)	16 (100)
Black	30 (2.8)	10 (3.4)	0 (0)	0 (0)
Asian	21 (2)	2 (0.7)	1 (6.7)	0 (0)
Other	39 (3.7)	13 (4.4)	1 (6.7)	0 (0)

[a]Temporal validation set includes participants newly enrolled after the July 3, 2024, data release (excluded from the development set) and serves as a temporally separated dataset for external validation.

[b]PD: Parkinson disease.

[c]HC: healthy controls.

The development dataset was randomly divided into training, validation, and test subsets following an approximately 7:2:1 ratio. Only complete cases were included in the test set to ensure reliable evaluation. Feature selection and fine-tuning were performed exclusively using the training set. The test set was used solely for final model evaluation and was not involved in any feature selection or model fitting procedures. Class distributions for each subset are summarized in Table 2.

Table 2. Dataset splits and class distribution for diagnostic modeling.

Datasets and subsets	Total, n	PD[a], n (%)	HC[b], n (%)
Development set			
Training set	990	771 (77.9)	219 (22.1)
Validation set	248	193 (77.8)	55 (22.2)
Test set	122	99 (81.1)	23 (18.9)
Temporal validation set	31	15 (48.4)	16 (51.6)

[a]PD: Parkinson disease.

[b]HC: healthy controls.

Feature Selection

Multicollinearity Evaluation

To address multicollinearity, we generated a correlation matrix and computed the variance inflation factor for all candidate variables [20]. Variables with variance inflation factor values exceeding 10, a commonly accepted threshold for severe multicollinearity, were removed to prevent unstable coefficients and improve model reliability [21,22]. A total of 70 variables remained after this screening step and were used for Shapley additive explanations (SHAP)–based feature selection.

SHAP-Based Feature Importance

Feature importance was evaluated using a weighted average of SHAP values obtained from 4 tree-based models, including random forest [23], XGBoost [24], LightGBM [25], and CatBoost [26]. These models were selected because they are widely used for their tabular data and are known to mitigate issues related to multicollinearity while accommodating mixtures of numerical and categorical features effectively [27].

For each model, SHAP values were computed for all remaining variables, normalized, and then combined using a weighted averaging approach based on the diagnostic performance of each model. The performance metrics and assigned weights are summarized in Multimedia Appendix 1.

The weighted average SHAP value for each feature was calculated as [28]:

where:

- φ_{ij} represents the SHAP value of feature j computed by model i
- w_i represents the performance-based weight assigned to model i

XSL•FO
RenderX

- the denominator ensures that weights are normalized so they sum to 1.

The top 10 variables were selected based on these weighted average SHAP values, as shown in Multimedia Appendix 2. The final features set consisted of clinically meaningful motor, olfactory, and imaging features. Specifically, the selected variables were updrs3_score, con_putamen, updrs_totscore, updrs2_score, lowput_expected, upsit_pctl, DATSCAN_PUTAMEN_L, con_striatum, mean_putamen, and DATSCAN_PUTAMEN_R. These features were used consistently across all subsequent ML and LLM experiments.

Prompt Construction

To evaluate the performance of LLMs, we developed 4 prompting formats that express input variables in plain text (PT), markdown (MD), special token (ST) annotations, and a combined markdown with special token (MD+ST) structure [29]. The PT format presents the input features as natural language sentences, whereas the MD format uses a structured table to delineate characteristics and enhance feature recognition, and the combined MD+ST format integrates both structured layout and explicit ST annotation.

Each prompting format was applied under 0-shot conditions to 3-shot conditions. All few-shot settings incorporated balanced PD and HC examples, with 1 pair for 1-shot, 2 pairs for 2-shot, and 3 pairs for 3-shot prompting. For dual-output prompting, the diagnostic label was followed by 3 sentences of post hoc explanatory text, which served as an exploratory output and did not represent true model interpretability [30].

For dual-output prompting, each LLM was instructed to generate a diagnostic label (PD or HC) followed by 3 sentences of post hoc explanatory text with a maximum length of 180 tokens to ensure comparability across models. Examples of all prompting formats, few-shot configurations, and sample outputs are provided in Multimedia Appendix 3.

Modeling Approaches

This study compared traditional ML classifiers with multiple LLMs in order to evaluate diagnostic performance under consistent experimental conditions. As deterministic baselines, we trained LR with L2 regularization and a SVM with a radial basis function (RBF) kernel, using the top 10 SHAP-selected features. Both models were implemented using standard scikit-learn procedures without extensive hyperparameter tuning, since the goal was to establish transparent and reproducible baselines rather than to maximize predictive performance. The complete hyperparameter configurations for the ML classifiers are summarized in Multimedia Appendix 4. These models were evaluated following the procedures described in the "Training and Evaluation Settings" section.

For LLM-based classification, we selected model families that collectively capture diversity in architecture, accessibility, and usage constraints. The included families were LLaMA, GPT, Gemini, and Claude. Each family provided one or more model variants with different parameter sizes or intended applications. The LLaMA models consisted of LLaMA 3.1 8B Instruct and LLaMA 3.3 70B Instruct, both of which were run locally using open-access model weights. The GPT models consisted of GPT-4o-mini and GPT-4o, accessed through the OpenAI application programming interface (API). The Gemini models included Gemini 1.5 Flash and Gemini 1.5 Pro, accessed through the Google Gemini API. For the Claude family, the Claude 3.5 Sonnet model was used. Although lightweight variants such as Claude Instant were available, the Sonnet model was selected due to budget considerations and its expected performance.

All LLMs were accessed through their officially supported APIs or locally hosted implementations using Python-based programmatic interfaces. The same prompting formats, shot settings, and evaluation procedures were applied across all models, except where input-format restrictions required specific handling. Additional implementation details, including software environments and API configurations, are provided in Multimedia Appendix 5.

Training and Evaluation Settings

Overview

The evaluation schedule and prompting phases for each model are summarized in Multimedia Appendix 6, and the hardware and software environments used in all experiments are described in Multimedia Appendix 7. To address the class imbalance between PD and HC, model performance was primarily assessed using the F_1-score, with precision and recall reported as supporting metrics. Accuracy was treated as a secondary measure, given that F_1-score provides a more balanced assessment under imbalanced classification tasks [31].

For all binary classification tasks, PD was assigned as the positive class (label=1) and HC as the negative class (label=0) across all ML and LLM experiments. Numerical features were standardized using *z* score scaling, where each variable was transformed by subtracting the mean and dividing by the SD computed from the training split. This ensured a consistent feature scale for both LR and SVM while preventing data leakage. Standardization using *z* score was applied exclusively to the ML baselines. For LLM-based experiments, models were provided with the original clinical values embedded in natural language prompts, rather than normalized numerical vectors. For ML models, performance uncertainty was quantified using 95% bootstrap CIs based on 1000 resampling iterations [32].

For LLMs, diagnostic classification was evaluated under 0-shot settings to 3-shot settings using 4 prompting formats including PT, MD, ST, and MD+ST. All few-shot experiments incorporated balanced PD and HC examples at each shot level. Across all LLM conditions, each participant was evaluated through 30 repeated model executions under identical prompting conditions to account for the inherent nondeterminism of LLM outputs [33]. Any change in diagnostic labels across these executions was recorded as a label inconsistency event. For dual-output prompting experiments, semantic consistency of the generated post hoc explanatory texts was also evaluated.

For each LLM and prompting configuration, macro-averaged performance metrics (F_1-score, precision, recall, and accuracy) were computed for each of the 30 model executions. Nonparametric 95% CIs for LLM performance were estimated

XSL•FO
RenderX

only for the temporal validation set using a hierarchical bootstrap procedure. In each iteration, participants were sampled with replacement and, for each selected participant, one prediction was randomly drawn from the 30 repeated trials. The 2.5th and 97.5th percentiles of the resulting performance distributions were reported in the temporal validation tables.

All combinations of prompting format and shot number were evaluated for each LLM on the development test set. The prompting configuration reported as the best for each model corresponds to the combination that achieved the highest macro-averaged F_1-score on this test set. During temporal validation, each LLM was evaluated using only under its best-performing configuration. This ensured that temporal validation assessed the generalizability of each model's optimal setting rather than retesting all prompt-shot combinations on temporally independent data.

ML Baseline Training Protocols

Two conventional ML classifiers were trained as deterministic baselines for comparison with LLM-based diagnostic approaches. The models were LR with L2 regularization and a SVM with an RBF kernel, both of which were trained using the same feature subset identified in the SHAP-based feature selection step. Class weights were applied to address the imbalance between PD and HC samples in the development set.

Both models were implemented using standard scikit-learn procedures without extensive hyperparameters optimization, as the primary goal was to establish transparent and reproducible baselines rather than to maximize predictive performance. LR was trained using the liblinear solver with balanced class weights, while SVM used an RBF kernel with probability estimation enabled. Performance for both models was evaluated for direct comparability with all LLM-based experiments.

Few-Shot Prompting Evaluation of LLMs

The LLMs were evaluated using the 4 prompting formats described earlier, applied across 0-, 1-, 2-, and 3-shot settings. Examples of PD and HC were balanced in all few-shot conditions to ensure consistent contextual exposure. All 4 prompting formats were applied to the LLaMA models. In contrast, only the PT and MD formats were applied to the GPT, Gemini, and Claude models because these API-based models do not support ST annotations in their input structure. Importantly, the inclusion of STs substantially influenced both diagnostic performance and output stability. Accordingly, ST-based prompting (including MD+ST) was applied exclusively to the LLaMA models as a model-specific input design choice, and the resulting differences in prompt structuring capability were explicitly considered as a limitation in cross-model comparisons.

Dual-Output Prompting of LLMs for Diagnosis and Post Hoc Explanatory Text

This experiment examined whether requiring LLMs to generate post hoc explanatory text could influence their diagnostic classification performance. Recent studies have shown that step-by-step prompting can improve LLM accuracy in complex decision tasks [34], yet fully stepwise reasoning is often impractical in clinical contexts due to token limitations. Rather than instructing the models to articulate their reasoning step by step, the models were prompted to produce concise post hoc explanatory text. This approach enabled assessment of whether a reduced explanatory demand could affect the diagnostic output.

Each model generated a binary diagnosis of PD or HC followed by exactly 3 sentences of post hoc explanatory text. To ensure consistency across models and to manage token usage, the length of these explanatory outputs was fixed. This dual-output prompting was applied to 4 models, including GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, and LLaMA 3.3 70B. Among these, the LLaMA model was prompted using the ST format, whereas the other models were prompted using the PT format because these models do not support ST inputs.

Although the generated explanations were not reviewed by clinical experts, their semantic consistency was evaluated to assess the stability of the post hoc explanatory text under repeated prompting. High semantic consistency reflects only the reproducibility of the generated text across repeated runs and does not imply that the explanations are clinically accurate or factually grounded. For each participant, their explanatory outputs per model were collected, and pairwise cosine similarity between sentence embeddings was computed using the all-mpnet-base-v2 model from the Sentence-Transformers library [35,36]. This semantic consistency analysis was performed only on the development set, while the temporal validation set was used exclusively for the primary diagnostic evaluation.

LLM Supervised Fine-Tuning

Supervised fine-tuning was conducted to evaluate whether labeled training data could improve the diagnostic performance and output stability of lightweight LLMs. Unlike prompt-based learning, which is constrained by the number of examples that can be included within a single prompt, supervised fine-tuning allows the model to learn diagnostic patterns directly from the full training dataset.

Among the models tested in this study, GPT-4o-mini and Gemini 1.5 Flash were selected as compact and computationally efficient alternatives to assess how smaller LLMs perform relative to larger models. Both models were fine-tuned using the same dataset that was used in the prompting experiments, and all evaluations were performed under 0-shot conditions on the identical test set. The generation temperature was fixed at 0.1 to ensure consistent and deterministic behavior. Further details regarding fine-tuning platforms, sampling strategies, and training parameters are provided in Multimedia Appendix 8.

The GPT-4o-mini model was fine-tuned through the OpenAI API. Training jobs were submitted from a local workstation after converting the dataset into JSONL format. A total of 1052 samples were used for training and 186 samples for validation. The fine-tuning process followed OpenAI's job-based workflow, in which each training job is versioned and automatically made accessible through the OpenAI API upon completion.

The Gemini 1.5 Flash model was fine-tuned using the Google AI Studio platform. Because the platform restricts training to a maximum of 500 samples, stratified sampling was applied to

XSL•FO
RenderX

preserve the original class ratio, resulting in 383 PD samples and 117 HC samples. The model was trained for 1 epoch with a batch size of 1 and a learning rate of 1, after which the fine-tuned model was deployed via the Gemini API using the assigned model identifier.

All fine-tuned models were trained and evaluated using the PT prompt format to maintain consistency between training and testing conditions.

Ethical Considerations

This study used publicly available, deidentified data provided by PPMI. According to institutional policies and the PPMI Data Use Agreement, additional institutional review board approval was not required for secondary analyses of this publicly accessible dataset. The original PPMI study was conducted under institutional review board approval at all participating institutions, and all participants provided written informed consent permitting data sharing and secondary analyses. All analyses in this study were conducted on anonymized records with no personally identifiable information, and no participant compensation was applicable.

Results

ML Baseline Performance

To establish a deterministic baseline for comparison with LLM-based diagnostic approaches, we evaluated 2 conventional ML classifiers trained on the top 10 SHAP-selected features. LR and a SVM with an RBF kernel were implemented using standard scikit-learn procedures.

As summarized in Table 3, both models achieved identical performance on the test subset of the development set, yielding a macro-averaged F_1-score of 0.960 and an accuracy of 0.975. Precision and recall were balanced across PD and HC, and the 95% CIs ranged from approximately 0.94 to 1.00, reflecting the modest sample size of the test subset (n=122).

Table 3. Performance of machine learning models on the test subset of the development set (n=122; CIs were estimated via 1000 bootstrap iterations).

Model	F_1-score[a] (95% CI)	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy
LR_L2[e]	0.960 (0.943-1.000)	0.953/0.990/0.917	0.968/0.980/0.957	0.975
SVM_RBF[f]	0.960 (0.944-1.000)	0.953/0.990/0.917	0.968/0.980/0.957	0.975

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]LR_L2: logistic regression (L2 regularization).

[f]SVM_RBF: support vector machine (RBF kernel).

Table 4 presents the results from the temporal validation set (n=31). LR demonstrated moderately high generalizability, achieving a macro-averaged F_1-score of 0.903 and an accuracy of 0.903. In contrast, SVM showed a substantial decline in generalizability, producing a macro-averaged F_1-score of 0.484 with markedly reduced recall for PD. The wide CIs observed for both models reflect the limited size of the temporal validation set and the inherent uncertainty associated with evaluating performance on small external samples.

Table 4. Performance of machine learning models on the temporal validation set (n=31; CIs were estimated via 1000 bootstrap iterations).

Model	$F1$-score[a] (95% CI)	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy
LR_L2[e]	0.903 (0.791-1.000)	0.917/0.833/1.0	0.906/1.0/0.813	0.903
SVM_RBF[f]	0.484 (0.157-0.506)	0.241/0.483/0	0.5/0/1.0	0.484

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]LR_L2: logistic regression (L2 regularization).

[f]SVM_RBF: support vector machine (RBF kernel).

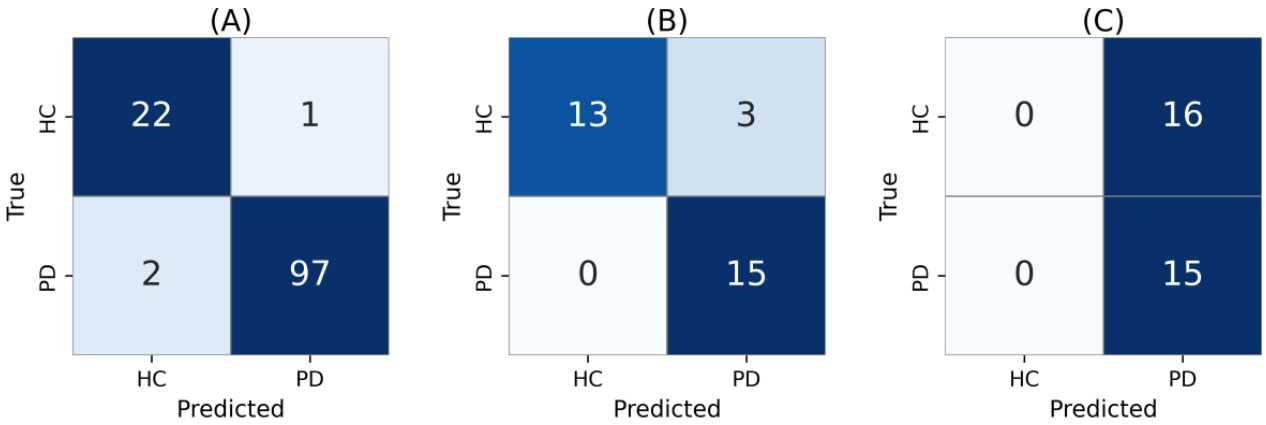
Given the small size of the temporal validation set (n=31), results obtained on this dataset should be interpreted with caution, as performance estimates may be sensitive to individual misclassifications.

Figure 2 illustrates the prediction patterns for the 2 classifiers across both datasets. On the development test subset (Figure 2A), both models produced identical predictions, misclassifying 1 HC participant as PD and misclassifying 2 PD participants as HC. In the temporal validation set, LR misclassified only 3 HC cases as PD (Figure 2B), whereas SVM misclassified nearly all HC cases (Figure 2C). Although the two models performed similarly on the development test subset, their stability diverged

XSL•FO
RenderX

substantially when evaluated on data collected at a later timepoint.

Figure 2. Confusion matrix for machine learning classifiers. (A) Test subset of the development set. (B) Logistic regression of temporal validation set. (C) Support vector machine of temporal validation set. HC: healthy controls; PD: Parkinson disease.



Few-Shot Prompting Performance of LLMs

A total of 4 prompting formats, including PT, MD, ST, and MD+ST, were evaluated under 0-shot conditions to 3-shot conditions to assess the diagnostic performance of LLMs. ST-based prompts were applied exclusively to the LLaMA models because the inclusion of STs substantially affected both classification accuracy and output stability. Non-LLaMA models such as GPT, Gemini, and Claude did not support ST formatting and were therefore evaluated using PT and MD formats.

Table 5 presents the best-performing results for each model on the test dataset (n=122). The best-performing configuration for each model corresponds to the prompt type and shot number that achieved the highest macro-averaged F_1-score on the development test set.

LLaMA 3.3 70B and Gemini 1.5 Pro achieved the highest macro-averaged F_1-score of 0.987 with an accuracy of 0.992 and stable predictions across 30 repeated trials. LLaMA 3.1 8B reached the same F_1-score using the MD+ST prompt at the 3-shot setting but showed one instance of prediction inconsistency. Claude 3.5 Sonnet demonstrated strong performance with an F_1-score of 0.972 and an accuracy of 0.984 under 1-shot MD setting, maintaining complete consistency across repeated runs. GPT-4o and GPT-4o-mini achieved slightly lower but stable performance, with macro-averaged F_1-scores of 0.961 and 0.910, respectively. Gemini 1.5 Flash produced the lowest F_1-score (0.894) despite consistent predictions, reflecting reduced sensitivity for HC cases. Detailed per-participant consistency analyses across 30 trials are provided in Multimedia Appendix 9.

Figure 3 displays the confusion matrices for the best-performing configuration of each model. Most classification errors occurred in HC samples, indicating high sensitivity for PD across the LLMs.

XSL•FO
RenderX

Table 5. Best performance of large language models on the test dataset (n=122).

Model	Prompt	Shot	F_1-score[a]	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e], n	Inconsistent cases[f]
LLaMA 3.1 8B	MD+ST[g]	3	0.987	0.995/0.990/1	0.978/1/0.957	0.992	1	PD #73 (13:17)
LLaMA 3.3 70B	ST[h]	2	0.987	0.995/0.990/1	0.978/1/0.957	0.992	0	—[i]
GPT-4o-mini	MD[j]	2	0.910	0.972/0.943/1	0.869/1/0.739	0.951	1	HC #103 (19:11)
GPT-4o	PT[k]	0	0.961	0.954/0.990/0.917	0.969/0.980/0.957	0.975	2	HC #26 (28:2); HC #68 (17:13)
Gemini 1.5 Flash	PT, MD	2	0.894	0.967/0.934/1	0.848/1/0.696	0.938	0	—
Gemini 1.5 Pro	PT, MD	2	0.987	0.995/0.990/1	0.978/1/0.957	0.992	0	—
Claude 3.5 Sonnet	MD	1	0.972	0.990/0.980/1	0.957/1/0.913	0.984	0	—

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]Number of participants (out of 122) whose predictions were inconsistent at least once across 30 repeated trials.

[f]Example of inconsistent participants showing the final label (eg, PD #73) and the number of predicted labels across 30 runs (eg, 13:17 indicates 13 HC and 17 PD predictions).

[g]MD+ST: markdown with special token.

[h]ST: special token.

[i]Not applicable.

[j]MD: markdown.

[k]PT: plain text.

Figure 3. Confusion matrices of large language models under the best-performing configurations on the test dataset (n=122). HC: healthy controls; PD: Parkinson disease.

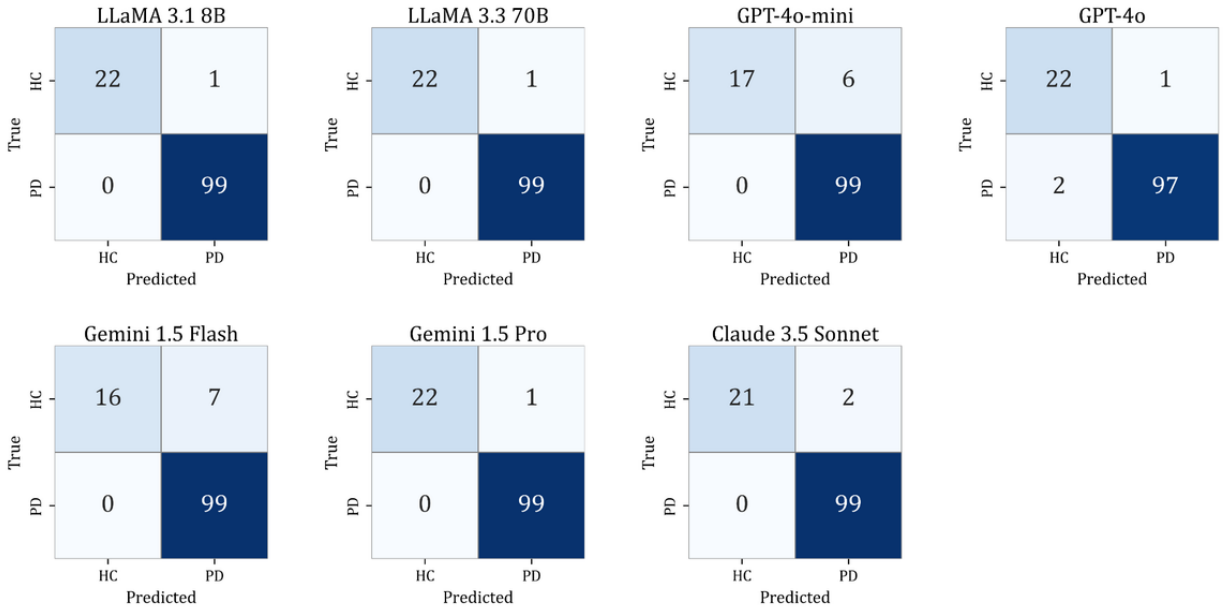


Table 6 and Figure 4 summarize the best-performing configuration of each LLM on the temporal validation set (n=31). LLaMA 3.3 70B achieved the highest macro-averaged F_1-score of 0.968 using the ST format at the 2-shot setting, with recall of 1.00 and consistent predictions across all participants. LLaMA 3.1 8B, Gemini 1.5 Pro, and Claude 3.5 Sonnet showed

XSL•FO
RenderX

comparable performance, each achieving a macro-averaged F_1-score of 0.936 with stable predictions. GPT-4o and Gemini 1.5 Flash achieved moderately high performance (F_1-score=0.903), with consistent predictions across all participants. In contrast, GPT-4o-mini achieved the lowest F_1-score (0.836), driven primarily by reduced recall for HC cases. Across all models, recall for PD remained consistently high (recall≥0.938), demonstrating strong sensitivity even when evaluated on data collected at a later timepoint. Figure 4 shows the confusion matrices for each model under the prompt and shot condition listed in Table 6.

These findings indicate that LLMs generally maintained high diagnostic accuracy and prediction stability under temporally separated data, supporting their potential generalizability beyond the development dataset.

Table 6. Best performance of large language models on the temporal validation set (n=31).

Model	Prompt	Shot	F_1-score[a] (95% CI)	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e], n	Inconsistent cases[f]
LLaMA 3.1 8B	MD+ST[g]	3	0.936 (0.854-1.000)	0.941/0.882/1	0.938/1/0.875	0.935	2	HC #13 (9:21); HC #23 (7:23)
LLaMA 3.3 70B	ST[h]	2	0.968 (0.899-1.000)	0.969/0.938/1	0.969/1/0.938	0.968	0	—[i]
GPT-4o-mini	MD[j]	2	0.836 (0.708-0.966)	0.875/0.750/1	0.844/1/0.688	0.839	2	HC #10 (4:26); HC #28 (8:22)
GPT-4o	PT[k]	0	0.903 (0.773-1.000)	0.917/0.833/1	0.906/1/0.813	0.903	2	PD #1 (1:29); HC #10 (28:2)
Gemini 1.5 Flash	PT	2	0.903 (0.773-1.000)	0.917/0.833/1	0.906/1/0.813	0.903	0	—
Gemini 1.5 Pro	PT	2	0.936 (0.832-1.000)	0.941/0.882/1	0.938/1/0.875	0.935	0	—
Claude 3.5 Sonnet	MD	1	0.936 (0.832-1.000)	0.941/0.882/1	0.938/1/0.875	0.935	31	—

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]Number of participants (out of 31) whose predictions were inconsistent at least once across 30 repeated trials.

[f]Example of inconsistent participants showing the final label (eg, HC #13) and the number of predicted labels across 30 runs (eg, 9:21 indicates 9 HC and 21 PD predictions).

[g]MD+ST: markdown with special token.

[h]ST: special token.

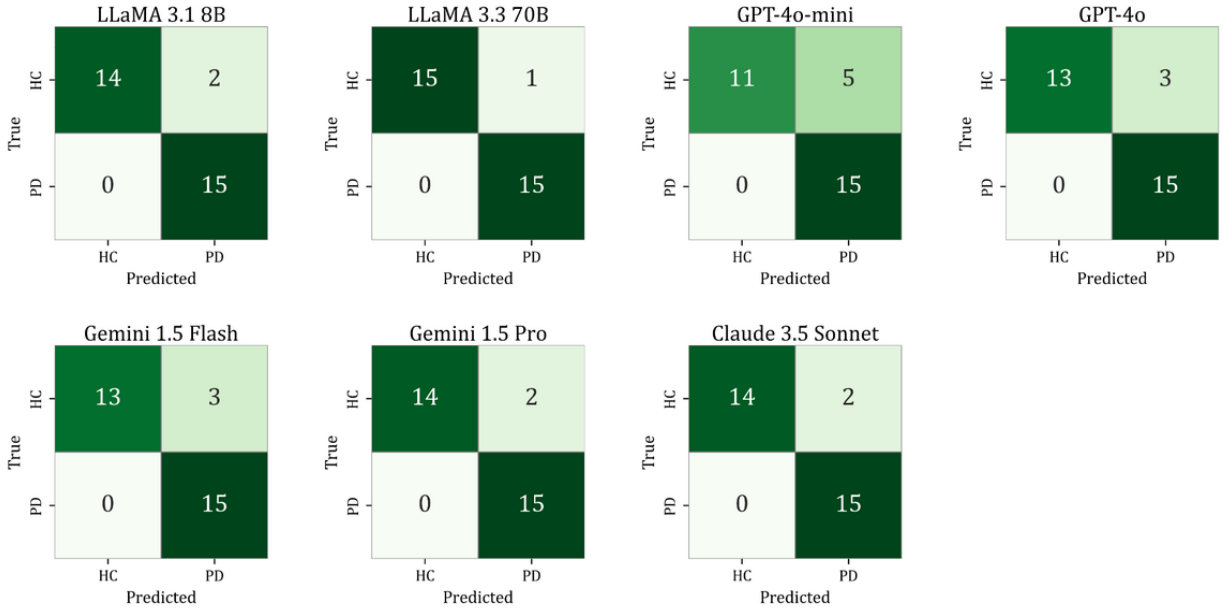
[i]Not applicable.

[j]MD: markdown.

[k]PT: plain text.

XSL•FO
RenderX

Figure 4. Confusion matrices of large language models under their best-performing configurations on the temporal validation set (n=31). HC: healthy controls; PD: Parkinson disease.



Diagnostic Performance Under Dual-Output Prompting

Dual-output prompting was used to evaluate whether requiring LLMs to generate post hoc explanatory text influenced their diagnostic reliability. This experimental setting included 4 representative models, namely LLaMA 3.3 70B, GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Unlike diagnostic-only prompting, this setup instructed each model to output a binary diagnostic label followed by 3 sentences of post hoc explanatory text. Because the LLaMA models are highly sensitive to input formatting, dual-output prompts for LLaMA 3.3 70B were constructed using the ST format. GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet do not support ST inputs, and therefore PT prompts were used for these models to maintain compatibility.

As summarized in Table 7, all models achieved high diagnostic performance under dual-output prompting, although overall F_1-scores were slightly lower than those obtained under few-shot prompting. LLaMA 3.3 70B and Claude 3.5 Sonnet each achieved a macro-averaged F_1-score of 0.972 and correctly identified all PD cases. GPT-4o and Gemini 1.5 Pro showed similar results and achieved F_1-scores of 0.958 with an accuracy of 0.975. Most prediction errors occurred in HC cases, whereas all PD samples were correctly classified. Instances of inconsistency across 30 repeated trials were rare, typically affecting no more than 2 participants for each model. The full repeated-trial results are presented in Multimedia Appendix 10.

Figure 5 displays the confusion matrices corresponding to each model's best-performing dual-output configuration.

XSL•FO
RenderX

Table 7. Best performance of large language models on the test dataset under dual-output prompting (n=122).

Model	Prompt	Shot	F_1-score[a]	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e], n	Inconsistent cases[f]
LLaMA 3.3 70B	ST[g]	3	0.972	0.990/0.980/1	0.957/1/0.913	0.984	122	HC #23 (29:1); HC #68 (28:2)
GPT-4o	PT[h]	1	0.958	0.985/0.970/1	0.935/1/0.870	0.975	122	HC #26 (26:4); HC #77 (6:24)
Gemini 1.5 Pro	PT	3	0.958	0.985/0.970/1	0.935/1/0.870	0.975	122	HC #97 (28:2)
Claude 3.5 Sonnet	PT	0	0.972	0.990/0.980/1	0.957/1/0.913	0.984	122	PD #23 (1:29)

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

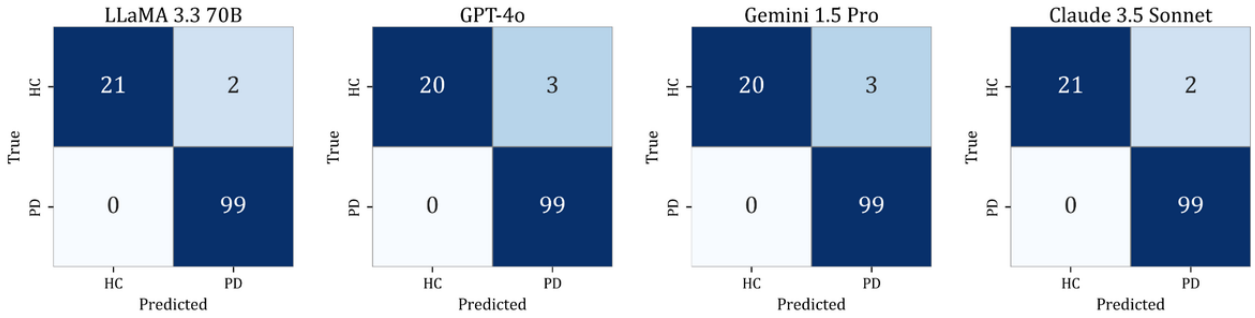
[e]Number of participants (out of 122) whose predictions were inconsistent at least once across 30 repeated trials.

[f]Example of inconsistent participants showing the final label (eg, HC #23) and the number of predicted labels across 30 runs (eg, 29:1 indicates 29 HC and 1 PD predictions).

[g]ST: special token.

[h]PT: plain text.

Figure 5. Confusion matrices of large language models under best-performing dual-output prompting conditions on the test subset of the development set (n=122). HC: healthy controls; PD: Parkinson disease.



Generalizability was further assessed using the temporal validation set (n=31). As shown in Table 8, all models maintained strong diagnostic sensitivity, and recall for PD remained equal to 1 across all participants. GPT-4o achieved the highest macro-averaged F_1-score of 0.968 with an accuracy of 0.968. LLaMA 3.3 70B followed with a macro-averaged F_1-score of 0.935. Gemini 1.5 Pro and Claude 3.5 Sonnet showed modest decreases in performance with F_1-scores of 0.903 and 0.869, respectively, primarily due to HC misclassifications.

Figure 6 presents the confusion matrices for the temporal validation set under the best-performing configuration for each model.

XSL•FO
RenderX

Table 8. Best performance of large language models under dual-output prompting on the temporal validation set (n=31).

Model	Prompt	Shot	F_1-score[a] (95% CI)	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e]	Inconsistent cases[f]
LLaMA 3.3 70B	ST[g]	3	0.935 (0.832-1.000)	0.941/0.882/1	0.938/1/0.875	0.935	2	HC #11 (23:7); HC #21 (8:22)
GPT-4o	PT[h]	1	0.968 (0.896-1.000)	0.969/0.938/1	0.969/1/0.938	0.968	1	HC #21 (29:1)
Gemini 1.5 Pro	PT	3	0.903 (0.774-1.000)	0.917/0.833/1	0.906/1/0.813	0.903	0	—[i]
Claude 3.5 Sonnet	PT	0	0.869 (0.735-0.968)	0.895/0.789/1	0.875/1/0.750	0.871	1	HC #10 (1:29)

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]Number of participants (out of 31) whose predictions were inconsistent at least once across 30 runs.

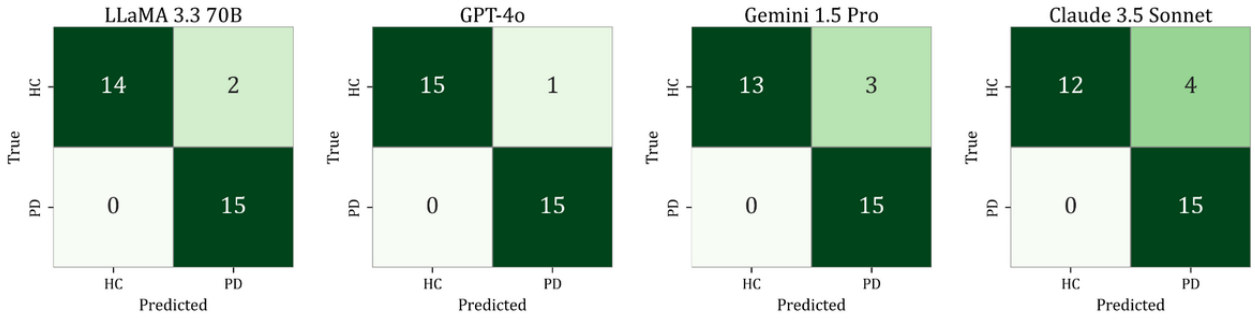
[f]Example of inconsistent participants showing the final label (eg, HC #11) and the number of predicted labels across 30 runs (eg, 23:7 indicates 23 HC and 7 PD predictions).

[g]ST: special token.

[h]PT: plain text.

[i]Not applicable.

Figure 6. Confusion matrices of large language models under best-performing dual-output prompting conditions on the temporal validation set (n=31). HC: healthy controls; PD: Parkinson disease.



Following the diagnostic classification results presented in Table 7, the semantic consistency of the post hoc explanatory texts was evaluated to assess the stability of generated explanations under dual-output prompting. Pairwise cosine similarity was computed among 30 post hoc explanatory texts generated for each participant, and the resulting mean and SD were averaged across all 122 test participants. As shown in Table 9, all models maintained high semantic consistency, with mean cosine similarity values exceeding 0.95. LLaMA 3.3 70B achieved the highest value under 0-shot prompting (0.997 ± 0.005), and the variation across models and shot settings was small (≤ 0.03). This evaluation was limited to the development set, and additional exploratory metrics for the temporal validation set are provided in Multimedia Appendix 11.

XSL•FO
RenderX

Table 9. Semantic consistency of reasoning outputs on the test dataset (n=122). Semantic consistency was assessed based on pairwise cosine similarity among 30 post hoc explanatory texts generated per participant.

Model	Cosine similarity, mean (SD)
LLaMA 3.3 70B	
0-shot	0.997 (0.005)
1-shot	0.995 (0.009)
2-shot	0.966 (0.016)
3-shot	0.980 (0.011)
GPT-4o	
0-shot	0.973 (0.016)
1-shot	0.980 (0.011)
2-shot	0.981 (0.011)
3-shot	0.975 (0.016)
Gemini 1.5 Pro	
0-shot	0.969 (0.019)
1-shot	0.968 (0.025)
2-shot	0.956 (0.029)
3-shot	0.951 (0.026)
Claude 3.5 Sonnet	
0-shot	0.985 (0.017)
1-shot	0.986 (0.016)
2-shot	0.987 (0.016)
3-shot	0.979 (0.015)

Fine-Tuned Prompting Performance of LLMs

Supervised fine-tuning substantially improved the diagnostic performance of lightweight LLMs and resulted in more consistent classification across the evaluated datasets. Table 10 summarizes the results of GPT-4o-mini and Gemini 1.5 Flash on the development test set (n=122).

Table 10. Fine-tuning performance of lightweight large language models on the development test set (n=122).

Model	Prompt	F_1-score[a]	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e]	Inconsistent cases[f]
GPT-4o-mini	PT[g]	0.987	0.995 / 0.990 / 1	0.978 / 1 / 0.957	0.992	0	—[h]
Gemini 1.5 Flash	PT	0.973	0.990 / 0.980 / 1	0.957 / 1 / 0.913	0.984	1	HC #23 (9:21)

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

[e]Number of participants (out of 122) whose predictions were inconsistent at least once across 30 repeated trials.

[f]Example of inconsistent participants showing the final label (eg, HC #23) and the number of predicted labels across 30 runs (eg, 9:21 indicates 9 HC and 21 PD predictions).

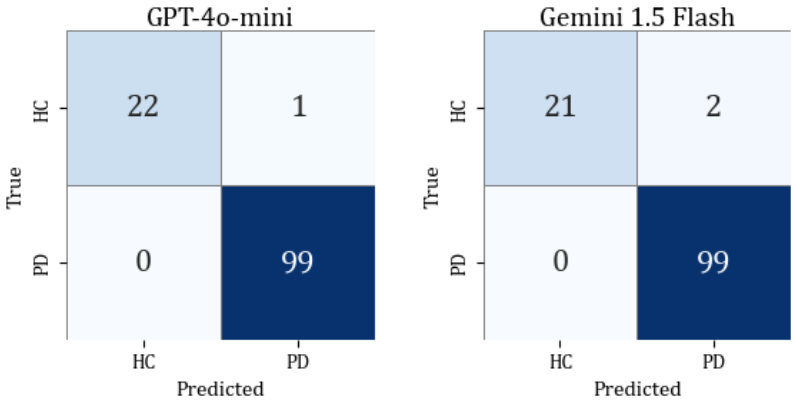
[g]PT: plain text.

[h]Not applicable.

The fine-tuned GPT-4o-mini achieved the highest macro-averaged F_1-score of 0.987, with recall of 1.00 and stable predictions across 30 repeated trials. Gemini 1.5 Flash followed closely with an F_1-score of 0.973, maintaining an accuracy of 0.984 and showing only 1 inconsistent prediction (participant HC #23). Both models were trained and evaluated using PT prompts. The corresponding confusion matrices are shown in Figure 7. Figure 7 displays the confusion matrices of the fine-tuned LLMs on the development test set, showing that nearly all misclassifications occurred among HC participants, whereas both models correctly identified all PD cases in the sample.

XSL•FO
RenderX

Figure 7. Confusion matrices of lightweight large language models after fine-tuning on the development test set (n=122). HC: healthy controls; PD: Parkinson disease.



To further assess model generalizability, both fine-tuned LLMs were evaluated on the temporal validation set (n=31). As summarized in Table 11, GPT-4o-mini achieved F_1-score, precision, recall, and accuracy of 1.000 on the temporal validation set, correctly classifying all participants in this sample. In contrast, Gemini 1.5 Flash demonstrated slightly lower but still strong performance with an F_1-score of 0.903 and an accuracy of 0.903. All PD participants were correctly identified, and the few errors were limited to HC participants. Figure 8 presents the confusion matrices under the same fine-tuned PT prompting configuration. These findings confirm that fine-tuning notably improved classification stability while preserving sensitivity to PD in temporally independent data.

Table 11. Fine-tuning performance of lightweight large language models on the temporal validation set (n=31).

Model	Prompt	F_1-score[a] (95% CI)	Precision (macro avg[b]/PD[c]/HC[d])	Recall (macro avg/PD/HC)	Accuracy	Inconsistency[e]	Inconsistent cases[f]
GPT-4o-mini	PT[g]	1 (1.000-1.000)	1/1/1	1/1/1	1	0	—[h]
Gemini 1.5 Flash	PT	0.903 (0.770-1.000)	0.917/0.833/1	0.906/1/0.813	0.903	1	HC #21 (25:5)

[a]F_1-scores represent macro avg values across PD and HC classes.

[b]macro avg: macro-averaged.

[c]PD: Parkinson disease.

[d]HC: healthy controls.

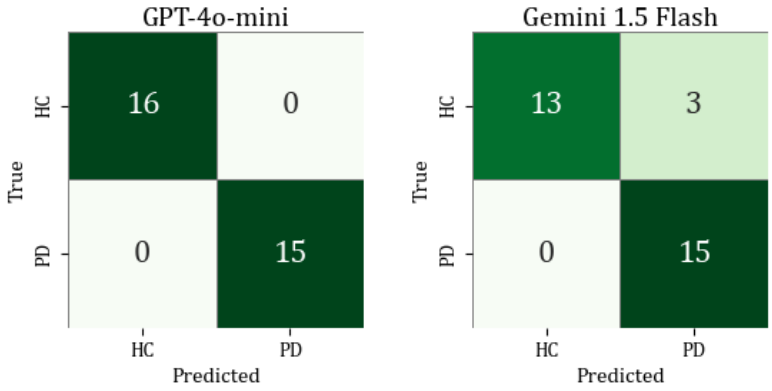
[e]Number of participants (out of 31) whose predictions were inconsistent at least once across 30 runs.

[f]Example of inconsistent participants showing the final label (eg, HC #21) and the number of predicted labels across 30 runs (eg, 25:5 indicates 25 HC and 5 PD predictions).

[g]PT: plain text.

[h]Not applicable.

Figure 8. Confusion matrices of lightweight large language models after fine-tuning on the temporal validation set (n=31). HC: healthy controls; PD: Parkinson disease.

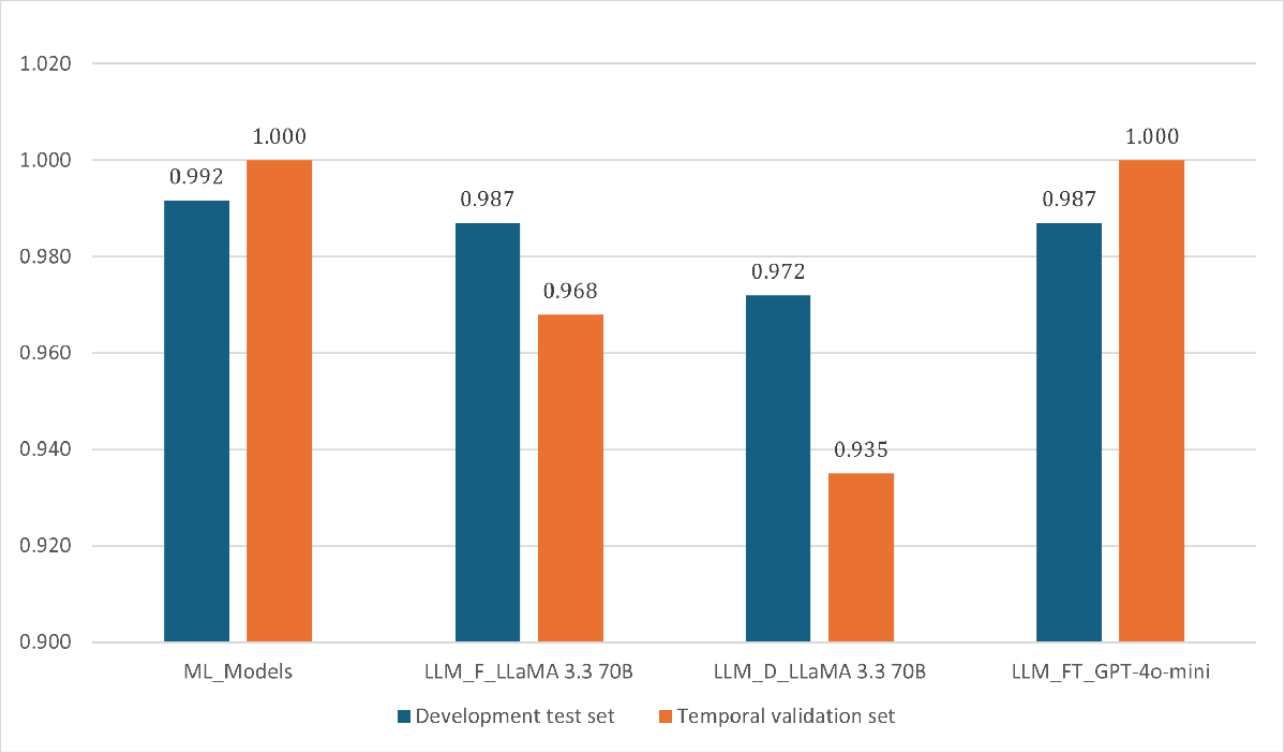


These results confirm that with sufficient training data and consistent prompting formats, even small-scale LLMs can achieve classification accuracy comparable to larger models while maintaining stable performance.

XSL•FO
RenderX

Figure 9 provides an integrated comparison of the best-performing models across all experimental settings, including traditional ML baselines, few-shot prompting with LLM (LLM_F), dual-output prompting with LLM (LLM-D), and fine-tuned prompting with LLM (LLM_FT). Across the development dataset, multiple LLM configurations achieved macro-averaged F_1-scores comparable to LR, and fine-tuned lightweight models reached the highest overall performance.

Figure 9. Comparison of the best macro-averaged F1-scores from the top-performing models in machine learning (ML) and large language model experiments across few-shot, dual-output, and fine-tuned prompting on the development and temporal validation sets.



On the temporal validation dataset, LR maintained moderate generalizability, whereas SVM showed substantial degradation when applied to temporally separated data. In contrast, several LLM configurations preserved high recall for PD and sustained overall performance, particularly under fine-tuned and few-shot prompting conditions. These results demonstrate that while ML baselines provide deterministic reference points, LLMs exhibit greater flexibility across prompting strategies and maintain stable sensitivity to PD in both datasets.

Discussion

Overview

This study examined how modern LLMs process structured clinical variables when these variables are reformatted into natural language prompts for the diagnostic classification of PD. Using SHAP-selected features derived from the PPMI dataset, we compared multiple LLM families and prompting strategies with conventional ML baselines. Three main findings emerged. First, several LLMs achieved diagnostic performance comparable to LR while maintaining high sensitivity for PD across both the development test set and the temporal validation set. Second, the diagnostic behavior of LLMs varied depending on prompt format, model family, and shot configuration, whereas the ML baselines produced deterministic and highly stable predictions. Third, supervised fine-tuning markedly improved both accuracy and output stability in lightweight LLMs, allowing a compact model such as GPT-4o-mini to correctly classify all participants in the temporal validation set. In addition, because the ML baselines in this study were minimally tuned, part of any performance gap between ML models and LLMs may reflect limited optimization of the ML baselines rather than true methodological differences. Accordingly, the present comparisons should be interpreted as exploratory rather than definitive.

The performance of the LLaMA family was strongly influenced by input formatting. In particular, the inclusion or removal of STs resulted in notable differences in accuracy and sensitivity, although their classification results varied more widely across different shot settings [29]. Dual-output prompting, which required models to generate diagnostic labels along with post hoc explanatory text, resulted in slightly lower F_1-scores compared with diagnostic-only prompting but did not substantially destabilize predictions. The generated text exhibited high semantic consistency across repeated trials. These explanations should be regarded as post hoc natural language outputs rather than indicators of true model interpretability, since they are produced after the model's primary diagnostic prediction step [14].

To evaluate whether inconsistent predictions reflected clinically ambiguous participants rather than model-level variability, we conducted a qualitative review of cases that exhibited the highest numbers of label inconsistencies. A total of 2 HC participants and 2 PD participants were selected for detailed examination. All 10 SHAP-selected variables, including Unified Parkinson

XSL•FO
RenderX

Disease Rating Scale (UPDRS) motor scores, University of Pennsylvania Smell Identification Test (UPSIT) percentiles, and dopamine transporter single-photon emission computed tomography putaminal uptake metrics, were compared with the overall distributions of the PD and HC groups. None of the reviewed cases demonstrated borderline or contradictory clinical profiles. The HC cases showed preserved dopaminergic uptake and normal motor assessments, with only mild olfactory reductions typical of healthy older adults. The PD cases exhibited reduced dopaminergic activity, clear asymmetry, and motor impairment consistent with established PD patterns. These observations suggest that label inconsistencies are unlikely to arise from underlying clinical ambiguity. Instead, they appear to reflect model stochasticity and prompt-dependent variability [33].

Supervised fine-tuning clarified the role of training data in stabilizing LLM predictions. When provided with labeled examples, both GPT-4o-mini and Gemini 1.5 Flash demonstrated substantial improvements in diagnostic accuracy and showed consistently high sensitivity to PD on the temporal validation set. GPT-4o-mini classified all 31 participants correctly after fine-tuning. This result suggests that compact models can approximate or exceed the performance of larger models when trained on appropriately structured datasets. It also indicates that fine-tuning can reduce susceptibility to prompt-level variability and may support more reliable behavior in clinical decision-support environments. However, this comparison should be interpreted with caution. Although architectural differences may also contribute to the observed performance gap, GPT-4o-mini was fine-tuned on more than twice as many labeled samples as Gemini 1.5 Flash (1052 vs 500) due to platform constraints. Part of GPT-4o-mini's superior performance may therefore reflect the larger amount of training data rather than inherent model advantages, which limits how directly the two models can be compared.

Overall, this study illustrates both the potential and the limitations of modern LLMs for processing structured clinical variables that are presented in natural language form. While several models achieved strong diagnostic performance and generalized well to temporally separated data, their outputs remained sensitive to prompt structures, model architectures, and few-shot configurations. Occasional inconsistencies across repeated runs further highlight the stochastic nature of LLM output generation [33]. These characteristics reinforce the importance of careful interpretation and the need for rigorous evaluation frameworks before LLMs can be integrated safely into real-world diagnostic workflows.

Limitations

Several limitations should be considered when interpreting these findings. First, although temporal validation provided an important assessment of model generalizability, the temporal validation set was relatively small (n=31), which resulted in wide CIs for the reported performance metrics. In addition, models were trained on datasets that included imputed values, whereas evaluation was conducted on datasets restricted to complete cases without missing data, including both the development test set and the temporal validation set. This mismatch may introduce a distributional shift and result in performance estimates that reflect a best-case evaluation scenario rather than real-world clinical conditions where missing data are common.

Second, the 10 features used for model input were selected using SHAP values from tree-based models. Feature sets obtained using alternative selection strategies may differ, so the current feature subset may not fully represent model-agnostic feature selection. Furthermore, the explanatory text generated by LLMs under the dual-output prompting framework was not reviewed by clinical experts. Accordingly, the semantic consistency metric reflects internal textual stability rather than clinically accurate or factually grounded post hoc explanatory text, and the clinical validity of the generated explanations remains unverified.

Finally, several methodological constraints limit direct model-to-model comparisons. Prompt format and shot configuration were selected based on performance observed on the development test dataset, rather than using a separate validation set for configuration selection. This design choice reflects the study's aim to broadly compare model behaviors rather than to establish definitive optimal configurations. In addition, platform-specific constraints limited the extent of fine-tuning that could be performed across LLMs. While GPT-4o-mini was fine-tuned using 1052 training samples with an additional held-out validation set of 186 samples, the Gemini 1.5 Flash fine-tuning interface restricts supervised training to a maximum of 500 samples, and the machine learning baselines were trained on 990 samples. As a result, the observed performance differences across models may reflect differences in training data availability rather than inherent architectural superiority and should be interpreted as exploratory.

Prompt structuring flexibility also differed across platforms, as the ST format was applied only to the LLaMA models due to platform-specific input constraints, further limiting the degree to which direct model-to-model comparisons can be made in this study.

Future Work

Future research should expand these results in several directions. Larger temporal or external datasets, including real-world clinical settings, are needed to strengthen generalizability assessments. Additional work is warranted to examine optimization strategies for prompt design, temperature settings, and calibration methods that may reduce stochastic variability. Expert-based evaluation of generated post hoc explanatory text may clarify how these outputs can be used to support clinical decision-making. Further exploration of supervised fine-tuning for additional lightweight LLMs could help identify resource-efficient models suitable for deployment in constrained clinical environments. Finally, integrating imaging, sensor-derived digital biomarkers, and longitudinal clinical trajectories may clarify how LLMs can combine multimodal biomedical data for diagnostic tasks.

Conclusions

This study provides an exploratory benchmark of how LLMs process structured clinical variables when presented in natural language form. Multiple LLMs achieved diagnostic performance

XSL•FO
RenderX

comparable to conventional ML baselines and maintained high sensitivity for PD under temporal validation. However, their predictions were influenced by prompt format, shot configuration, and model architecture, and occasional inconsistencies reflected inherent stochasticity rather than clinical ambiguity. Supervised fine-tuning substantially improved reliability in lightweight models, demonstrating that compact architectures can achieve stable and high-performing classification when trained on sufficient labeled examples. These findings highlight both the opportunities and the challenges associated with applying LLMs to structured clinical data and emphasize the need for rigorous evaluation before clinical implementation.

Acknowledgments

The authors would like to thank Editage for English language editing. We also acknowledge the use of generative artificial intelligence (AI) tools (ChatGPT by OpenAI and Grok by xAI) under author supervision for language editing and phrasing refinement during manuscript preparation. All analyses, experimental designs, and interpretations related to LLM diagnostics were fully conceived, conducted, and verified solely by the authors.

Funding

This work was supported by 2 funding sources. First, the Regional Innovation System & Education (RISE) program through the Institute of Regional Innovation System & Education in Busan Metropolitan City, funded by the Ministry of Education (MOE) and the Busan Metropolitan City, Republic of Korea (2025-RISE-02-003). Second, the Global-Learning & Academic research institution for Master's·PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF), grant funded by the Ministry of Education (RS-2025-2540216). The funders had no role in the study design, data collection, analysis, interpretation of results, manuscript writing, or the decision to submit the work for publication.

Data Availability

The datasets analyzed during this study are available from the Parkinson's Progression Markers Initiative (PPMI) repository, subject to PPMI data access policies. Selected code and processed data supporting the findings of this study are included in Multimedia Appendix 12. Additional materials can be shared upon reasonable request.

Authors' Contributions

HJS was responsible for the conceptualization, methodology, data curation, formal analysis, investigation, visualization, and writing–original draft. DYK, as the corresponding author, provided overall supervision, project administration, funding acquisition, and contributed to conceptualization and writing–review and editing. YJJ and SMJ contributed to validation and provided clinical insights and feedback during manuscript reviews. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performance and weights of tree-based models.
[DOC File , 59 KB - medinform_v14i1e77561_app1.doc]

Multimedia Appendix 2
Weighted mean SHAP values for the top 10 selected variables.
[DOC File , 59 KB - medinform_v14i1e77561_app2.doc]

Multimedia Appendix 3
Prompt formats by structure type and output design.
[DOC File , 51 KB - medinform_v14i1e77561_app3.doc]

Multimedia Appendix 4
Hyperparameter configurations for logistic regression and support vector machine classifiers.
[DOC File , 58 KB - medinform_v14i1e77561_app4.doc]

Multimedia Appendix 5
Description of model families and API.

XSL•FO
RenderX

[DOC File , 63 KB - medinform_v14i1e77561_app5.doc]

Multimedia Appendix 6
Experimental timeline of model evaluations.
[DOC File , 60 KB - medinform_v14i1e77561_app6.doc]

Multimedia Appendix 7
Hardware and software configuration.
[DOC File , 57 KB - medinform_v14i1e77561_app7.doc]

Multimedia Appendix 8
Summary of fine-tuning configurations for lightweight language models.
[DOC File , 60 KB - medinform_v14i1e77561_app8.doc]

Multimedia Appendix 9
Diagnostic performance of large language models in Parkinson disease classification using prompt engineering and few-shot learning.
[DOC File , 169 KB - medinform_v14i1e77561_app9.doc]

Multimedia Appendix 10
Diagnostic generalization of large language models under dual-output prompting on the test dataset.
[DOC File , 85 KB - medinform_v14i1e77561_app10.doc]

Multimedia Appendix 11
Semantic consistency of reasoning outputs on the temporal validation set under dual-output prompting.
[DOC File , 33 KB - medinform_v14i1e77561_app11.doc]

Multimedia Appendix 12
Complete code and dataset package used for all experiments, including preprocessing, few-shot prompting, reasoning consistency analysis, and supervised fine-tuning.
[ZIP File (Zip Archive), 306 KB - medinform_v14i1e77561_app12.zip]

References

1. Shakya S, Prevett J, Hu X, Xiao R. Characterization of Parkinson's disease subtypes and related attributes. Front Neurol 2022;13:810038 [FREE Full text] [doi: 10.3389/fneur.2022.810038] [Medline: 35677337]
2. Tolosa E, Garrido A, Scholz SW, Poewe W. Challenges in the diagnosis of Parkinson's disease. Lancet Neurol 2021;20(5):385-397 [FREE Full text] [doi: 10.1016/S1474-4422(21)00030-2] [Medline: 33894193]
3. Virameteekul S, Revesz T, Jaunmuktane Z, Warner TT, De Pablo-Fernández. Clinical diagnostic accuracy of Parkinson's Disease: Where do we stand? Mov Disord 2023;38(4):558-566. [doi: 10.1002/mds.29317] [Medline: 36602274]
4. Jankovic J. Parkinson's disease: Clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79(4):368-376. [doi: 10.1136/jnnp.2007.131045] [Medline: 18344392]
5. Statistics. Parkinson’s Foundation. URL: https://www.parkinson.org/understanding-parkinsons/statistics [accessed 2025-02-02]
6. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol 2011;95(4):629-635 [FREE Full text] [doi: 10.1016/j.pneurobio.2011.09.005] [Medline: 21930184]
7. Hosmer DJ, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd Edition. Hoboken: John Wiley & Sons; 2013.
8. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273-297. [doi: 10.1007/bf00994018]
9. Gerraty RT, Provost A, Li L, Wagner E, Haas M, Lancashire L. Machine learning within the Parkinson's progression markers initiative: review of the current state of affairs. Front Aging Neurosci 2023;15:1076657 [FREE Full text] [doi: 10.3389/fnagi.2023.1076657] [Medline: 36861121]
10. Alex JSR. Early detection of Parkinson’s disease using motor symptoms and machine learning. ArXiv Preprint posted online on April 18, 2023 [FREE Full text]
11. Pradeep P, Kamalakannan J. Comprehensive review of literature on Parkinson's disease diagnosis. Comput Biol Chem 2024;113:108228. [doi: 10.1016/j.compbiolchem.2024.108228] [Medline: 39413446]
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
13. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics 2024;11(3):57. [doi: 10.3390/informatics11030057]

XSL•FO
RenderX

14. Wiegreffe S, Marasović A. Teach me to explain: a review of datasets for explainable natural language processing. ArXiv Preprint posted online on February 24, 2021 [FREE Full text]
15. Wang Z, Zhang G, Yang K, Shi N, Zhou W, Hao S, et al. Interactive natural language processing. ArXiv Preprint posted online on May 22, 2023 [FREE Full text]
16. Sui Y, Zhou M, Zhou M, Han S, Zhang D. Table meets LLM: can large language models understand structured table data? Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24) 2024:645-654. [doi: 10.1145/3616855.3635752]
17. Afkanpour M, Hosseinzadeh E, Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Res Methodol 2024;24(1):188 [FREE Full text] [doi: 10.1186/s12874-024-02310-6] [Medline: 39198744]
18. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. NPJ Digit Med 2023;6(1):86 [FREE Full text] [doi: 10.1038/s41746-023-00832-9] [Medline: 37149704]
19. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245-247 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
20. Thompson CG, Kim RS, Aloe AM, Becker BJ. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. Basic and Applied Social Psychology 2017;39(2):81-90. [doi: 10.1080/01973533.2016.1277529]
21. Neter J, Nachtsheim CJ, Kutner MH. Applied Linear Regression Models. 4th Edition. New York: McGraw-Hill/Irwin; 2004.
22. O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007;41(5):673-690. [doi: 10.1007/s11135-006-9018-6]
23. Breiman L. Random forests. Mach Learn 2011;45(1):5-32 [FREE Full text] [doi: 10.1186/1478-7954-9-29]
24. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: 22nd ACM SIGKDD International Conference. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16); August 13-17, 2016; San Francisco p. 785-794.
25. Ke G, Meng Q, Finley T. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). New York: Curran Associates, Inc; 2017:3146-3154.
26. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018). New York: Curran Associates, Inc; 2018:6638-6648.
27. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022;81:84-90. [doi: 10.1016/j.inffus.2021.11.011]
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017. URL: https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [accessed 2025-06-21]
29. He J, Rungta M, Koleczek D, Sekhon A, Wang F, Hasan S. Does prompt formatting have any impact on LLM performance? ArXiv Preprint posted online on November 15, 2024. [doi: 10.48550/arXiv.2411.10541]
30. Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. ArXiv Preprint posted online on May 19, 2022 [FREE Full text]
31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
32. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist. Med 2000;19(9):1141-1164. [doi: 10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f]
33. Song Y, Gao Q, Bai Y. The good, the bad, and the greedyvaluation of LLMs should not ignore non-determinism. ArXiv Preprint posted online on July 15, 2024 [FREE Full text] [doi: 10.18653/v1/2025.naacl-long.211]
34. Kojima T, Gu SS, Reid M. Large language models are zero-shot reasoners. ArXiv Preprint posted online on May 24, 2022. [doi: 10.48550/arXiv.2205.11916]
35. Reimers N, Gurevych I. Sentence-BERTntence embeddings using Siamese BERT-networks. ArXiv Preprint posted online on August 27, 2019 [FREE Full text]
36. Song K, Tan X, Qin T, Lu J, Liu TY. MPNet: masked and permuted pre-training for language understanding. ArXiv. Preprint posted online on April 20, 2020 [FREE Full text]

Abbreviations

API: application programming interface
HC: healthy controls
LLM: large language model

XSL•FO
RenderX

LLM_D: dual-output prompting with LLM
LLM_F: few-shot prompting with LLM
LLM_FT: fine-tuned prompting with LLM
LR: logistic regression
MD: markdown
MD+ST: markdown with special token
ML: machine learning
PD: Parkinson disease
PPMI: Parkinson's Progression Markers Initiative
PT: plain text
RBF: radial basis function
SHAP: Shapley additive explanations
ST: special token
SVM: support vector machine
UPDRS: Unified Parkinson Disease Rating Scale
UPSIT: University of Pennsylvania Smell Identification Test

Edited by A Coristine; submitted 15.May.2025; peer-reviewed by A Hungbo, N Acharya; comments to author 22.May.2025; revised version received 24.Dec.2025; accepted 24.Dec.2025; published 15.Jan.2026.

Please cite as:
Shin HJ, Jeong YJ, Jun S, Kang DY
Prompting and Fine-Tuning Large Language Models for Parkinson Disease Diagnosis: Comparative Evaluation Study Using the PPMI Structured Dataset
JMIR Med Inform 2026;14:e77561
URL: https://medinform.jmir.org/2026/1/e77561
doi:10.2196/77561
PMID:

©Hyun-Ji Shin, Young Jin Jeong, Sungmin Jun, Do-Young Kang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

CO2 Laser Popularity in Germany: A Five-Year Google Trends Analysis (2020 2025)

Michael Constantin Kirchberger[1], Priv.-Doz. Dr. med. habil; Andreas Eisenried[2], Priv.-Doz. Dr. med. habil

[1]Lehrstuhl Dermatologie Friedrich-Alexander-Universität Erlangen-Nürnberg, Ulmenweg 18, Erlangen, Germany

[2]Lehrstuhl Anästhesiologie Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Michael Constantin Kirchberger, Priv.-Doz. Dr. med. habil
Lehrstuhl Dermatologie Friedrich-Alexander-Universität Erlangen-Nürnberg, Ulmenweg 18, Erlangen, Germany

Abstract

Background: Fractional carbon dioxide (CO_2) laser resurfacing is widely used for the treatment of scars and photoaging. In recent years, public interest in minimally invasive esthetic procedures has grown, influenced by social media exposure and changing beauty norms. However, data quantifying population-level attention to CO_2 laser treatments in Germany are limited.

Objective: This study aimed to assess the long-term trajectory and seasonal patterns of public information-seeking behavior regarding fractional CO_2 laser treatments in Germany from January 2020 to December 2025 using Google Trends data.

Methods: Monthly normalized search volume (NSV) for the category "Health" and the term "CO2 laser" was retrieved for the period January 2020 to December 2025. Seasonal-Trend decomposition (STL) using LOESS (locally estimated scatterplot smoothing) was applied to isolate the long-term trend from seasonal fluctuations. The significance of the upward trajectory was assessed using linear regression on the extracted trend component, and seasonal differences were evaluated via the seasonal component amplitude.

Results: Public interest in CO_2 lasers increased significantly, with the annual mean NSV rising from 15.0 in 2020 to 68.1 in 2025. Regression analysis of the STL trend component revealed a steady, statistically significant monthly increase (slope=0.87 NSV/month; 95% CI 0.83 - 0.92; P<.001). Furthermore, a robust seasonal pattern was identified (P<.001), with search interest consistently peaking in winter (January mean SD 13.8) and reaching a nadir during the summer months (August SD=–14.4).

Conclusions: Digital information-seeking behavior regarding fractional CO_2 laser treatments in Germany increased by 354% over the past six years, accompanied by consistent, clinically relevant seasonal peaks in winter. These findings reflect broader shifts in esthetic awareness. The identified temporal patterns provide valuable insights for timing educational messaging, managing patient inquiries, and addressing safety considerations in esthetic medicine.

(JMIR Med Inform 2026;14:e77651) doi:10.2196/77651

KEYWORDS

fractional CO_2 laser; infodemiology; google trends; aesthetic dermatology; Germany

Introduction

Fractional carbon dioxide (CO_2) laser therapy has become a cornerstone in dermatological and esthetic medicine, particularly for the treatment of scars and skin aging. This modality combines the efficacy of traditional ablative laser techniques with the benefits of fractional technology, resulting in significant skin remodeling with reduced downtime. The fractional approach delivers precise microthermal zones, promoting collagen production and skin rejuvenation while preserving surrounding tissue, thereby minimizing recovery time and adverse effects [1]. Clinical studies have demonstrated the effectiveness of fractional CO_2 lasers in improving the appearance of surgical scars, acne scars, and photoaged skin, with high patient satisfaction rates [1-3].

The increasing demand for minimally invasive cosmetic procedures has been fueled by advancements in laser technology and a growing emphasis on esthetic self-care. Social media platforms have played a significant role in popularizing these treatments, often showcasing dramatic before-and-after results and promoting the benefits of fractional CO_2 laser therapy. This widespread exposure has led to heightened public interest and a surge in the number of individuals seeking information about such treatments [4,5].

Understanding public interest and behavior regarding fractional CO_2 laser treatments is essential for health care providers and policymakers. Google Trends, a tool that analyzes the popularity of search queries over time, has emerged as a valuable resource in health research for gauging public interest and information-seeking behavior. This methodology is rooted in the conceptual framework of infodemiology, defined by

XSL•FO
RenderX

Eysenbach [6] as the science of distribution and determinants of information in an electronic medium, specifically with the aim to inform public health and policy. Studies have used Google Trends data to monitor public awareness and interest in various health topics, including mental health, infectious diseases, and cosmetic procedures. The tool's ability to provide real-time, population-level data makes it particularly useful for tracking trends and informing public health strategies [7-10].

In this study, we aim to analyze public information-seeking behavior regarding fractional CO_2 laser treatments in Germany from 2020 to 2025. Using Google Trends data filtered by the "Health" category and Seasonal-Trend decomposition, we isolate long-term growth from recurring seasonal fluctuations. This analysis is highly timely, as digital curiosity for minimally invasive esthetic procedures surges, regulatory frameworks for medical laser operators face intense professional scrutiny. Understanding these search dynamics provides clinicians and policymakers with critical insights to anticipate patient inquiries, optimize public health education, and mitigate safety risks associated with improper device use.

Methods

Data Retrieval and Ethics

Public interest data for the search term "CO2 laser" in Germany were retrieved using Google Trends. To ensure dermatological relevance and exclude industrial applications, the query was restricted to the "Health" category (Category ID: 45). The study period covered six full calendar years, from January 1, 2020, to December 31, 2025. As the dataset consists of aggregated and anonymized search volumes, the study did not involve human subjects and was exempt from institutional review board approval.

Statistical Analysis

To address the autocorrelation inherent in time-series data and isolate underlying patterns, a Seasonal-Trend decomposition (STL) using LOESS (locally estimated scatterplot smoothing) was applied. This method decomposed the monthly normalized search volume (NSV) into three distinct components: the long-term trend, seasonal variation, and residual noise. The long-term trajectory was evaluated using ordinary least squares linear regression performed on the extracted trend component to calculate the monthly rate of change and its 95% CI. Seasonal variation was quantified by averaging the seasonal component across the six-year period to identify recurring monthly peaks and troughs. All statistical tests were two-tailed, with a significance threshold of α=.05. For regional analysis, the "Interest by Region" output was exported as a single dataset for the entire period (2020 - 2025) to ensure consistent normalization across all 16 federal states.

All analyses and visualizations were conducted using Python (version 3.11). Data processing was performed with Pandas and NumPy. Time-series decomposition was executed using the Statsmodels library. Statistical regressions were calculated via SciPy. For geospatial visualization, a high-resolution GeoJSON file of German state borders was merged with the search data using GeoPandas. Regional interest was represented as a choropleth map using the *Matplotlib* library, with federal states labeled using official two-letter abbreviations.

Ethical Considerations

This study used publicly accessible, fully anonymized, and aggregated search volume data provided by Google Trends. As the research did not involve human subjects, clinical interventions, or the collection of any personally identifiable information, institutional review board approval and informed consent were not required in accordance with local and institutional guidelines.

Results

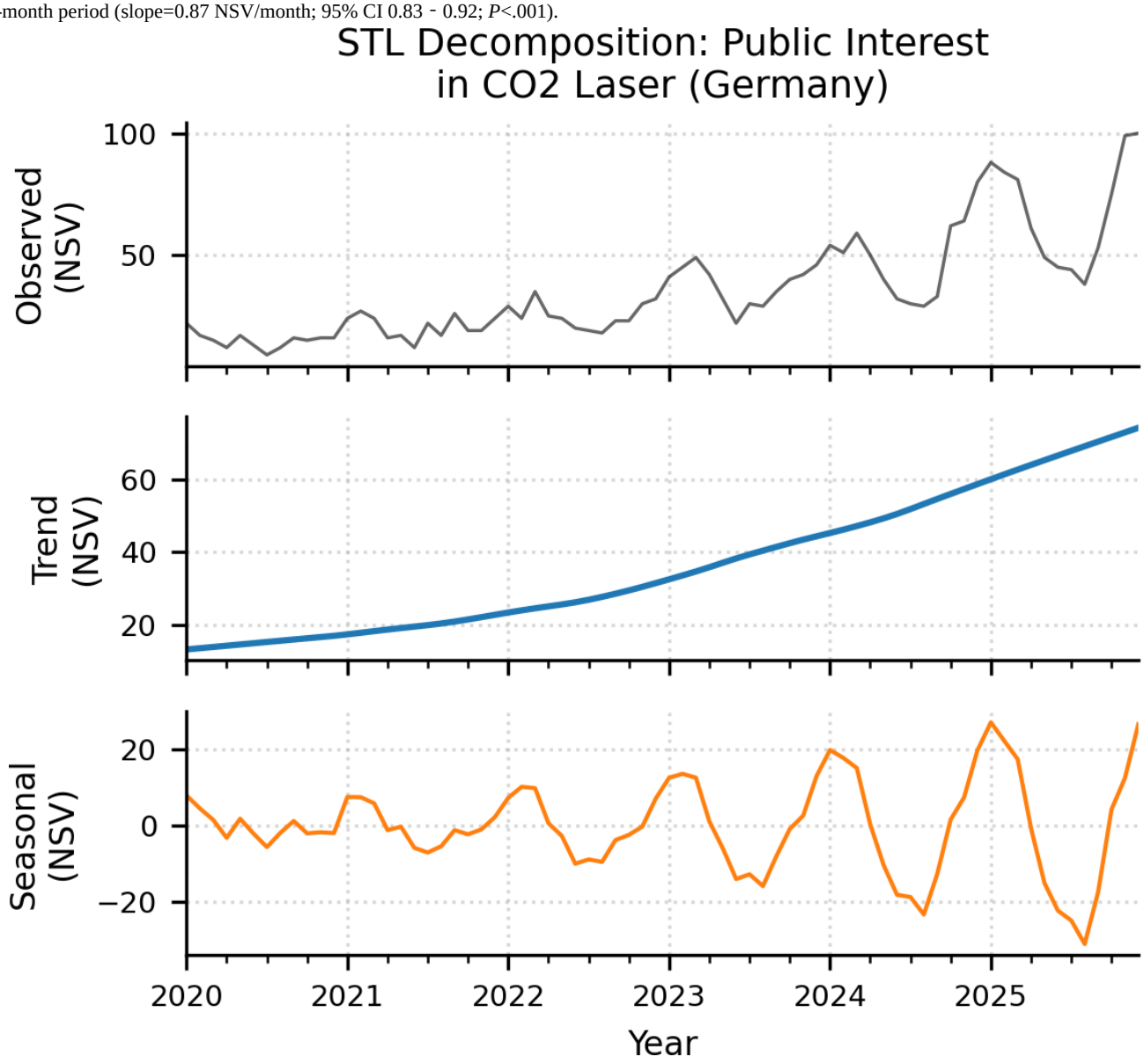
Trends and Seasonal Variation

The longitudinal analysis of search interest for the search term "CO2 laser" in Germany—filtered by the "Health" category to ensure medical relevance reveals a significant and sustained upward trajectory (Figure 1). By applying STL LOESS, the underlying long-term trend was isolated from seasonal noise and residuals, confirming that the growth is not merely a byproduct of transient spikes.

During the initial phase of the study (2020 - 2021), the trend component remained at a stable baseline, with annual mean NSV values increasing from 15.0 to 20.6. A more pronounced transition occurred in mid-2022, marking the beginning of a continuous growth phase that reached its peak in 2025 with an annual mean of 68.1 NSV. This represents a total increase in sustained public interest of 354% compared to the 2020 baseline. Linear regression analysis of the isolated trend component confirmed the high statistical significance of this growth (Slope=0.87 NSV/month; 95% CI 0.83 - 0.92; R^2=.96; P<.001).

XSL•FO
RenderX

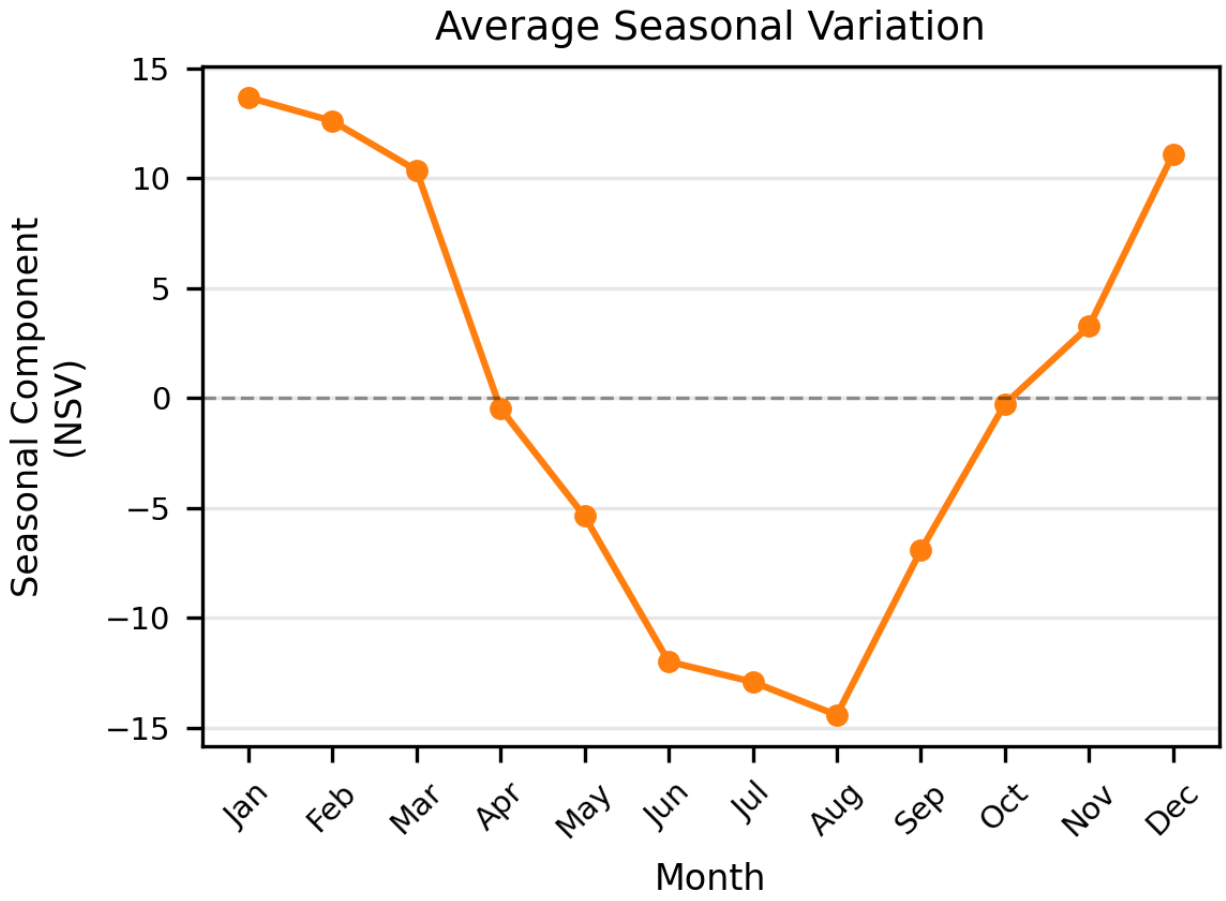
Figure 1. Long-term trend of public interest in CO_2 laser treatments in Germany (2020 - 2025). The graph illustrates the isolated trend component extracted via Seasonal-Trend decomposition using locally estimated scatterplot smoothing from monthly Google Trends data (filtered by the "health" category). The analysis reveals a sustained 354% increase in digital information-seeking behavior, with the annual mean normalized search volume (NSV) rising from 15.0 in 2020 to 68.1 in 2025. Linear regression of this isolated component confirms a highly significant upward trajectory over the 72-month period (slope=0.87 NSV/month; 95% CI 0.83 - 0.92; *P*<.001).



Complementing this long-term growth, a robust and recurring seasonal rhythm was identified. While Figure 1 (bottom panel) shows the consistency of these waves over time, Figure 2 illustrates the averaged seasonal deviations to highlight the typical annual cycle. Public interest follows a predictable pattern characterized by a "winter peak" and a "summer nadir." Information-seeking behavior typically reaches its zenith in January (+13.8 NSV units) and February (+12.6 NSV units).

XSL•FO
RenderX

Figure 2. Seasonal variation in search interest for CO_2 laser treatments. The chart displays the isolated seasonal component derived from the STL decomposition, representing the average monthly deviation from the underlying long-term trend. A robust, recurring annual cycle is evident (*P*<.001), with public interest consistently peaking during the winter months, reaching its zenith in January (+13.8 NSV units). Conversely, search activity drops to a predictable nadir during the summer, reaching its lowest point in August (–14.4 NSV units). This cyclical rhythm closely aligns with clinical recommendations to avoid ablative laser procedures during periods of high environmental UV radiation.



In contrast, interest consistently declines as annual UV radiation increases, reaching its lowest levels in August (–14.4 NSV units). This seasonal component remained highly consistent across the six-year period, with statistical analysis confirming that these monthly variations are non-random (*P*<.001), likely reflecting clinical recommendations to avoid laser treatments during periods of high solar exposure.

Yearly Growth

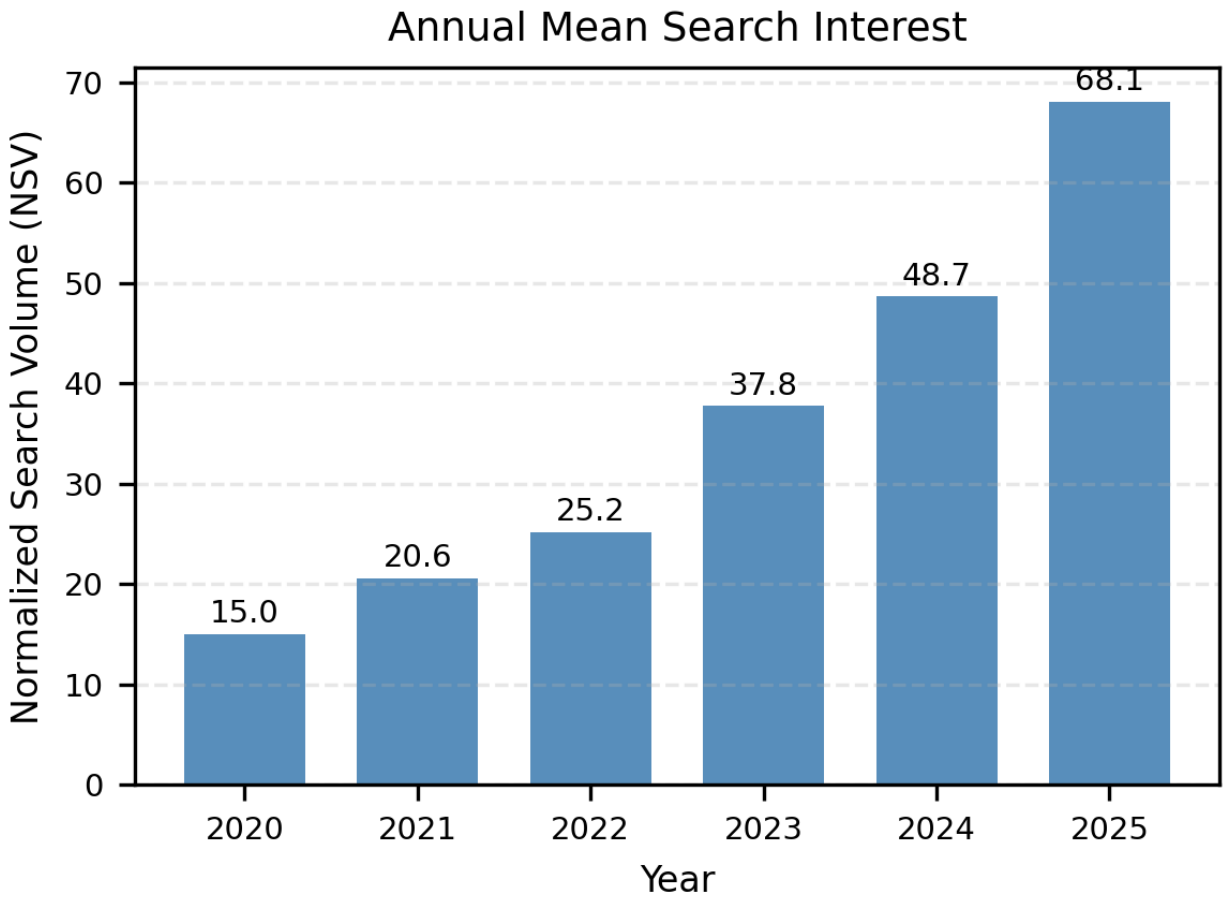
Annual mean normalized search interest is summarized in Figure 3. The data demonstrate a continuous increase in public attention toward “CO2 laser” in Germany throughout the six-year observation period.

The annual average was lowest in 2020 (15.0) and increased incrementally in each subsequent year: 20.6 in 2021, 25.2 in 2022, 37.8 in 2023, 48.7 in 2024, and 68.1 in 2025. These absolute values correspond to year-over-year increases of 37% (2020 - 2021), 22% (2021 - 2022), 50% (2022 - 2023), and 29% (2023 - 2024). By 2025, search interest reached a total increase of 354% compared to the 2020 baseline.

To quantify the overall trajectory while accounting for seasonality, a linear regression was performed on the monthly trend component. The model estimated a monthly increase of 0.87 units (95% CI 0.83 - 0.92; R^2=.96; *P*<.001), confirming a robust and well-fitting linear upward trend over the multi-year period. These findings indicate that the rise in interest reflects a sustained long-term development in population-level information-seeking behavior rather than transient search spikes.

XSL•FO
RenderX

Figure 3. Annual growth of public interest in CO_2 laser treatments in Germany (2020 - 2025). The chart illustrates the yearly mean normalized search volume (NSV) derived from the isolated STL trend component. Public information-seeking behavior demonstrated a continuous and statistically significant upward trajectory, growing from a baseline of 15.0 in 2020 to 68.1in 2025, representing a total sustained increase of 354%. Linear regression analysis confirmed a robust linear upward trend (slope=0.87 NSV/month; 95% CI 0.83 - 0.92; R^2=.96; *P*<.001), indicating a long-term shift in population-level interest.

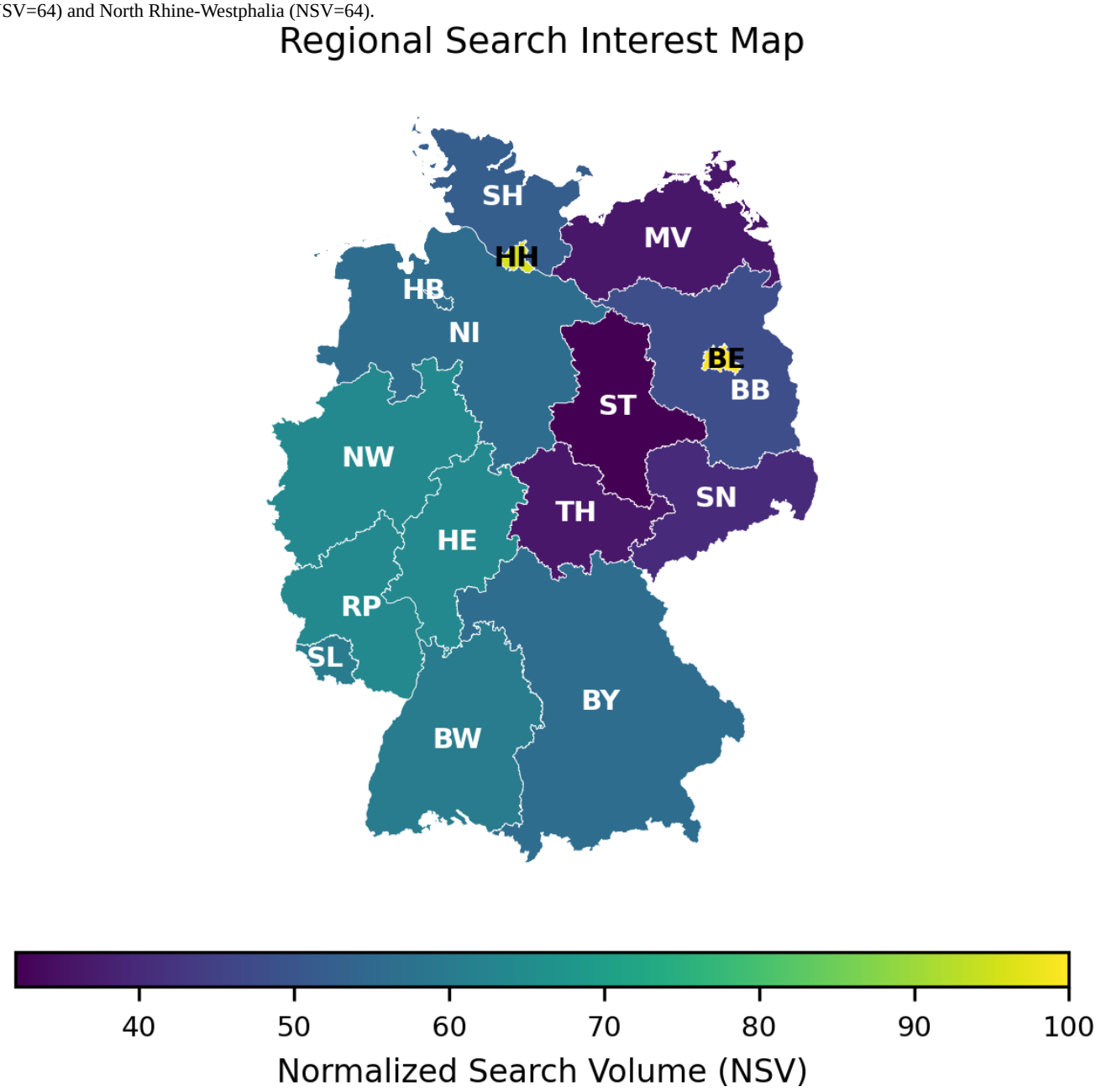


Regional Trends

Analysis of state-level search interest for the 2020 - 2025 period, revealed distinct regional concentrations of public attention toward CO2 laser treatments (Figure 4). The results indicate that search activity is most pronounced in major metropolitan regions. The highest search volumes were recorded in the city-states of Berlin (NSV 100) and Hamburg (NSV 96), followed by the populous states of Hesse (NSV 64) and North Rhine-Westphalia (NSV 64).

XSL•FO
RenderX

Figure 4. Geographic distribution of search interest for CO_2 laser treatments across German federal states. This choropleth map visualizes the average normalized search volume (NSV) 2020 - 2025. The spatial data demonstrate a pronounced concentration of public interest in major metropolitan hubs. The highest search activity was observed in the city-states of Berlin (NSV=100) and Hamburg (NSV=96), followed by the populous states of Hesse (NSV=64) and North Rhine-Westphalia (NSV=64).



Discussion

Principal Findings

This study used Google Trends data for "CO_2 laser" in Germany (2020 – 2025) in the category „Health" as a real-time proxy for public interest in this dermatologic procedure. Two robust patterns emerged: (1) a 354% sustained rise in the baseline trend component, increasing from a mean of 15.0 in 2020 to 68.1 in 2025 (Slope=0.87; 95% CI 0.83 - 0.92; P<.001) and (2) a predictable seasonal rhythm peaking in winter and reaching its nadir in summer.

Search-based infodemiology has repeatedly been shown to mirror real-world utilization of esthetic interventions [10-12]. Our upward trend corroborates earlier global observations by Tijerina et al [12] who reported growing interest in CO_2 -laser resurfacing over time [12]. The summer trough aligns with routine clinical practice, where CO_2 -laser treatments are avoided during periods of high UV exposure because of pigmentary risk; this concordance further validates Google Trends as a behavioral signal. From an operational perspective, the yearly rebound that starts each September implies that public-facing education or advertising campaigns could be most effective when launched in late summer.

While an immediate surge in search interest might have been expected during the initial COVID-19 lockdowns in 2020, our data indicate a more sustained upward trajectory that intensified later in the study period. This temporal alignment provides a compelling contextual hypothesis for the „Zoom effect". the

XSL•FO
RenderX

theory that heightened self-scrutiny during frequent videoconferencing acts as a catalyst for esthetic interest.

Survey-based research supports this possibility, with over half of respondents attributing increased interest in cosmetic procedures to videoconferencing, and two-thirds reporting habitual self-observation on screen [10,11]. However, it is important to note that without a formal interrupted time series analysis to identify a specific step change, this link remains speculative. The observed growth likely reflects a complex interplay between the „Zoom effect“ and a broader, pre-existing trend toward minimally invasive esthetic treatments.

Social-media ecosystems likely amplify these dynamics. Platforms such as TikTok and Instagram serve simultaneously as marketing channels and informal information sources; their algorithmic curation immerses users in procedure-related content, reinforcing norms and sustaining interest [5]. Such amplification may intensify both the secular growth and the seasonal peaks identified here.

Further social media platforms increasingly function as both promotional channels and informal information sources for not only esthetic procedures. Algorithm-driven content delivery creates personalized environments where users are repeatedly exposed to procedure-related imagery, reinforcing interest and perceived norms. This algorithmic amplification may contribute to sustained search activity and intensify seasonal peaks observed in our data.

Limitations

This study has several limitations inherent to the use of Google Trends data. First, the platform provides normalized search volumes without absolute user counts, limiting the ability to quantify population-level interest. Second, demographic details (eg, age, gender) are unavailable, restricting subgroup analysis. Most importantly, it is essential to distinguish between online search interest and actual clinical treatment uptake. The observed increase in search volume reflects heightened public curiosity and information-seeking behavior, rather than a direct measure of performed procedures.

Conclusions

Google search behavior offers a reliable lens on public interest in CO_2 -laser resurfacing, capturing both long-term growth and cyclical demand. This increasing digital curiosity heightens the need for rigorous practitioner training, regulatory oversight, and evidence-based patient education to ensure that online interest translates into safe, informed clinical decisions and mitigates potential complications.

Acknowledgments

We acknowledge the use of the Google Trends platform, which provided the anonymized search volume data necessary for this analysis. Google LLC had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Google Gemini was used for grammar and spelling.

Funding

The authors received no specific funding for this work.

Data Availability

The datasets generated and analyzed during the current study are publicly available via the Google Trends platform [13].

Conflicts of Interest

None declared.

References

1. Tierney EP, Hanke CW, Petersen J. Ablative fractionated CO2 laser treatment of photoaging: a clinical and histologic study. Dermatol Surg 2012 Nov;38(11):1777-1789. [doi: 10.1111/j.1524-4725.2012.02572.x] [Medline: 23110481]
2. Zhang DD, Zhao WY, Fang QQ, et al. The efficacy of fractional CO2 laser in acne scar treatment: a meta-analysis. Dermatol Ther 2021 Jan;34(1):e14539. [doi: 10.1111/dth.14539] [Medline: 33190373]
3. Chen L, Xu H, Wang QY, Chen P, Wang LQ, Qin XM. Treatment of surgical scars with fractional carbon dioxide (CO 2) laser: a randomized controlled trial. Adv Wound Care . [doi: 10.1089/wound.2024.0213]
4. Taylor MA, Ituarte BE, Wysong A, Sulewski R, Wei EX. Evaluating public interest in CO2 laser treatment: an analysis of Google Trends. J Cosmet Dermatol 2024 Apr;23(4):1498-1500. [doi: 10.1111/jocd.16121] [Medline: 38088501]
5. Wojtara MS. Use of social media for patient education in dermatology: narrative review. JMIR Dermatol 2023 Apr 14;6:e42609. [doi: 10.2196/42609] [Medline: 37632938]
6. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11. [doi: 10.2196/jmir.1157] [Medline: 19329408]
7. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. J Med Internet Res 2018 Nov 6;20(11):e270. [doi: 10.2196/jmir.9366] [Medline: 30401664]

XSL•FO
RenderX

8. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. JMIR Public Health Surveill 2019 May 29;5(2):e13439. [doi: 10.2196/13439] [Medline: 31144671]
9. Sivesind TE, Szeto MD, Kim W, Dellavalle RP. Google Trends in dermatology: scoping review of the literature. JMIR Dermatol 2021 May 25;4(1):e27712. [doi: 10.2196/27712] [Medline: 37632813]
10. Shen CZ, Zhao AT. The Influence of popular media on public interest in red-light therapy: longitudinal trend analysis. JMIR Dermatol 2025 Jun 12;8:e69796. [doi: 10.2196/69796] [Medline: 40504616]
11. Motosko C, Zakhem G, Ho R, Saadeh P, Hazen A. Using Google to trend patient interest in botulinum toxin and hyaluronic acid fillers. J Drugs Dermatol 2018 Nov 1;17(11):1245-1246. [Medline: 30500150]
12. Tijerina JD, Morrison SD, Nolan IT, Parham MJ, Nazerali R. Predicting public interest in nonsurgical cosmetic procedures using Google Trends. Aesthet Surg J 2020 Oct 24;40(11):1253-1262. [doi: 10.1093/asj/sjz264] [Medline: 31574152]
13. Interesse im zeitlichen verlauf. Google Trends. URL: https://trends.google.com/trends/explore?cat=45&date=today%205-y&geo=DE&q=%2Fm%2F047pqn&hl=de [accessed 2026-04-06]

Abbreviations

CO_2: carbon dioxide
LOESS: locally estimated scatterplot smoothing
NSV: normalized search volume
STL: seasonal-trend decomposition

Edited by A Benis; submitted 17.May.2025; peer-reviewed by A Kalluchi, J Raubenheimer, JL Bron, M Heppt; revised version received 21.Feb.2026; accepted 06.Mar.2026; published 21.Apr.2026.

Please cite as:
Kirchberger MC, Eisenried A
CO2 Laser Popularity in Germany: A Five-Year Google Trends Analysis (2020 - 2025)
JMIR Med Inform 2026;14:e77651
URL: https://medinform.jmir.org/2026/1/e77651
doi:10.2196/77651

© Michael Constantin Kirchberger, Andreas Eisenried. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applicability of Existing Gender Scores for German Clinical Research Data: Scoping Review and Data Mapping

Lea Schindler[1]; Hilke Beelich[1]; Elpiniki Katsari[2], Dr med; Daniele Liprandi[3], Dr ing; Sylvia Stracke[4], Dr med; Dagmar Waltemath[1,5], Dr ing

[1]Medical Informatics Laboratory, University Medicine Greifswald, Walther-Rathenau-Straße 48, Greifswald, Germany

[2]Heart Surgery, University Medicine Greifswald, Greifswald, Germany

[3]Laboratory of Evolutionary Biomechanics, Zoological Institute and Museum, University Greifswald, Greifswald, Germany

[4]Internal Medicine A, Nephrology, University Medicine Greifswald, Greifswald, Germany

[5]Core Unit Data Integration Center, University Medicine Greifswald, Greifswald, Germany

Corresponding Author:
Dagmar Waltemath, Dr ing
Medical Informatics Laboratory, University Medicine Greifswald, Walther-Rathenau-Straße 48, Greifswald, Germany

Abstract

Background: Considering sex and gender improves research quality, innovation, and social equity, while ignoring them leads to inaccuracies and inefficiency in study results. Despite increasing attention on sex- and gender-sensitive medicine, challenges remain with accurately representing gender due to its dynamic and context-specific nature.

Objective: This work aims to contribute to the implementation of a standard for collecting and assessing gender-specific data in German university hospitals and associated research facilities.

Methods: We carried out a review to identify and categorize state-of-the-art gender scores. We systematically assessed 22 publications regarding the applicability and practicability of their proposed gender scores. Specifically, we evaluated the use of these gender scores on German research data from routine clinical practice, using the Medical Informatics Initiative core dataset (MII CDS).

Results: Different methods for assessing gender have been proposed, but no standardized and validated gender score is available for health research. Most gender scores target epidemiological or public health research where questions about social aspects and life habits are already part of the questionnaires. However, it is challenging to apply concepts for gender scoring on clinical data. The MII CDS, for example, lacks all variables currently being recorded in gender scores. Although some of the required variables are indeed present in routine clinical data, they need to become part of the MII CDS.

Conclusions: To enable gender-specific retrospective analysis of routine clinical data, we recommend updating and expanding the MII CDS by including more gender-relevant information. For this purpose, we provide concrete action steps on how gender-related variables can be captured in routine clinical practice and represented in a machine-readable way.

International Registered Report Identifier (IRRID): RR2-10.2196/57669

(JMIR Med Inform 2026;14:e74162) doi:10.2196/74162

KEYWORDS

gender score; gender research; review; clinical routine; clinical research data; German Medical Informatics Initiative

Introduction

Background

Taking sex and gender (see Textbox 1 and Textbox 2) into account improves the quality of research and care, and it supports social equity [1]. Considering sex and gender from the outset can lead to new discovery and foster innovation. Ignoring sex and gender, however, leads to inaccuracies, inefficiencies, and difficulties generalizing results. This is, for example, highlighted by a recent study showing that men with low femininity reported a significant decrease in anxiety during the COVID-19 pandemic, meanwhile women with low femininity reported a significant increase [2].

XSL•FO
RenderX

Textbox 1. Definition of biological sex [3,4].

Biological sex is defined through biological attributes, such as:

- Genetics (ie, chromosomes, gene expression)
- Hormone levels
- Anatomy (ie, internal and external sex organs/reproductive organs)

Deviations regarding the hormonal or chromosomal attributes are possible, where the external appearance can be intersex and it is not possible to decide.

Textbox 2. Definition of (social) gender [3,4].

Gender refers to socially constructed roles, behaviors, expectations, and norms associated with being a woman, man, girl, or boy. It is formed through biological and psychosexual factors and an individual's social biography, and it is shaped by one's role in society. Gender influences how people perceive themselves and others and how they act and interact.

Many health care decisions are influenced by the patients' sex and gender, through social norms and experiences. Sex- and gender-sensitive medicine is receiving increasing attention in international research [5]. However, several publications have shown that women are still not considered equally to men and are poorly represented due to biased study designs [6-8]. It is well known today that this practice leads to an increase in inequality in medical treatment due to inappropriate therapies. Moreover, sex- and gender-specific reporting is often deficient, which leads to a lack of reproducibility and reduced effectiveness of research studies [1].

Even though the situation has improved, it can still be challenging to represent psychosocial gender aspects, as gender is a context-specific construct that is dynamic and multidimensional and can differ across time, geographical regions, and societies [9].

Gender scores are developed and used to assess gender roles and gender identity. They can be created and applied retrospectively to extract frequently missing gender information from already existing data. For this purpose, gender-sensitive variables in the existing data are determined by expert knowledge or statistical algorithms. Gender scores can then be calculated to predict social gender roles. For example, Lacasse et al [10] and Nauman et al [11] each developed a retrospective gender score and applied it to specific datasets from population studies. Examples for the corresponding gender-sensitive variables are specific *professions*, like in health care or construction, and typical *personality trait*s, like agreeableness or neuroticism; *loneliness* or *stress* was considered. In general, retrospective scores cannot capture participants' current gender identity (see Textbox 3).

Textbox 3. Definition of gender identity [3,4].

Gender identity is determined through individual self-perception and the sexual identity of a person, which may differ from someone's physical appearance or biological sex.

Prospective assessments can enhance the collection of gender-specific data from the outset of a study by incorporating further social data or information regarding the identity. Fraser et al [12] and Lagos and Compton [13], for example, evaluated a prospective 2-step gender identity measure on data from huge studies. The 2-step measure differentiates between a person's current gender identity and birth-assigned sex (see Textbox 3 and Textbox 4). The given examples show that gender scores are mostly used in surveys and studies and hardly used on routine clinical data.

Textbox 4. Definition of birth-assigned sex (BAS) [3,4].

BAS refers to the information that can be found on official documents, like a birth certificate, and is assigned based on the child's external anatomy or sex organs. BAS is often measured by a binary choice of response (ie, male or female). Biological sex is not a dichotomy, nor is BAS, where diversity is given.

All 38 university hospitals in Germany are part of the Medical Informatics Initiative [14] (MII), which serves to close the gap between research and routine care. A core dataset (CDS) is being developed by the MII consortia and the participating university hospitals to foster interoperability and allow a shared use of routine clinical data across Germany. Data Integration Centers collect and process routine clinical data and make them accessible to researchers for secondary use. If and how this established dataset considers gender and gender score applications are open questions.

It is important to note that the discussion about the applicability and state of gender scores does not only affect Germany. Critical assessments of sex and gender scores reveal that the problem affects different western cultures [15-18]. Ignoring gender differences in clinical research is a widespread global issue that undermines the validity and applicability of scientific findings across diverse populations [16]. A study conducted in Australia using the Bem Sex Role Inventory (BSRI) [19] found that the majority of participants received different gender scores in at least 1 of the 3 consecutive years, highlighting the complexity and fluid nature of gender identity [18]. This is particularly

XSL•FO
RenderX

concerning because the BSRI yielded inconsistent results even among a relatively stable population of patients older than 75 years — where minimal shifts in habits or identity would be expected over time. Despite being based on outdated social stereotypes, the BSRI is one of the most used gender scores, and other scores are partly based on it, as we later describe.

In this context, where classic gender scores are struggling, new tools are being considered to annotate medical data according to the gender of patients, thanks to the rise of machine learning methods [20]. The results of such methods can then be systematically used to mitigate biases in research, as done by Neufang et al [21], who developed an artificial intelligence model to improve the fairness of attention deficit hyperactivity disorder diagnosis with respect to gender. It is important to understand whether these algorithms are also applicable to general clinical data datasets, such as the MII CDS, to allow large-scale integration of gender debiasing in future research.

The issues outlined in the previous paragraphs indicate that more work needs to be done to update gender scores, taking into account the changes in commonly defined social rules that have occurred in the last century. In practice, outdated scores, such as the BSRI, are often used without considering proper validation and potential biases [22]. Moreover, a gold standard is still missing for how sex and gender should be considered in research with clinical research data. A practicable score should be balanced, comprehensive, and easily applicable. Interestingly, no gender score has so far been established as an international standard.

We therefore conducted a systematic evaluation of gender scores regarding their applicability in health research using German efforts on Data Integration in Medicine as a benchmark. As the world's third-largest economy and one of the World Health Organization's top two donors, we expected the German MII CDS to be a good candidate for our evaluation. In this article, the term gender score refers to the assessment of gender in general (social gender and gender identity), and the term gender identity measure refers exclusively to gender identity.

Related Work

The focus of our scoping review protocol published in JMIR Review Protocols [23] was the applicability and practicability of gender scores in health research. Related to the work described in this paper, Horstmann et al [24] published a review on recently used instruments for the operationalization of sex and gender in health research, and the work by Miani et al [25] summarizes epidemiological aspects in gender scores. The reviews included publications until 2020 and 2021, respectively. For this work, we identified the latest scores and therefore deliberately limited our search to publications between 2019 and 2024. Of the 18 articles from our primary result set, 13 were published after 2020.

In parallel to our work on the systematic review of recent gender scores, another systematic review about the operationalization of gender via composite gender scores in epidemiological studies was published in 2024 by Ballering et al [26]. The authors searched 3 databases (PubMed, Web of Science, and CINAHL) and identified 24 articles with a total of 26 gender scores developed in Europe or North America. They aggregated information on gender scores regarding author, year of publication, methodology, cohort information, and the included variables of the respective gender score and found that many variables overlapped across multiple studies.

Ballering et al [26] criticized theory-driven approaches — developed solely based on expert knowledge — as experts are not free from bias, potentially enhancing sexism or other biases. The authors further claim that data-driven approaches — using statistical models to develop gender scores — only predict female or male sex through identified psychosocial variables and do not consider gender itself. Psychosocial variables that do not associate with sex are excluded, and the quality of the gender scores is strongly dependent on the quality of the dataset. Overall, Ballering et al [26] concluded that data-driven approaches might be good to personalize health care, as gendered variables point out differences between women and men that could lead to improved treatment decisions and patient care.

The work by Ballering et al [26] is a great foundation for our work, and we are able to contribute further insights on the applicability of gender scores on routine clinical data. We had an overlap of 13 articles on both sides, with 13 of 22 articles in our review and 13 of 24 articles in the review by Ballering et al [26] (ie, there was an overlap of more than 50%: 59% ours and 54% theirs). We found further articles regarding gender identity that can be applied from the outset of a study or data collection, which could be of great use to implement gender information in routine clinical practice.

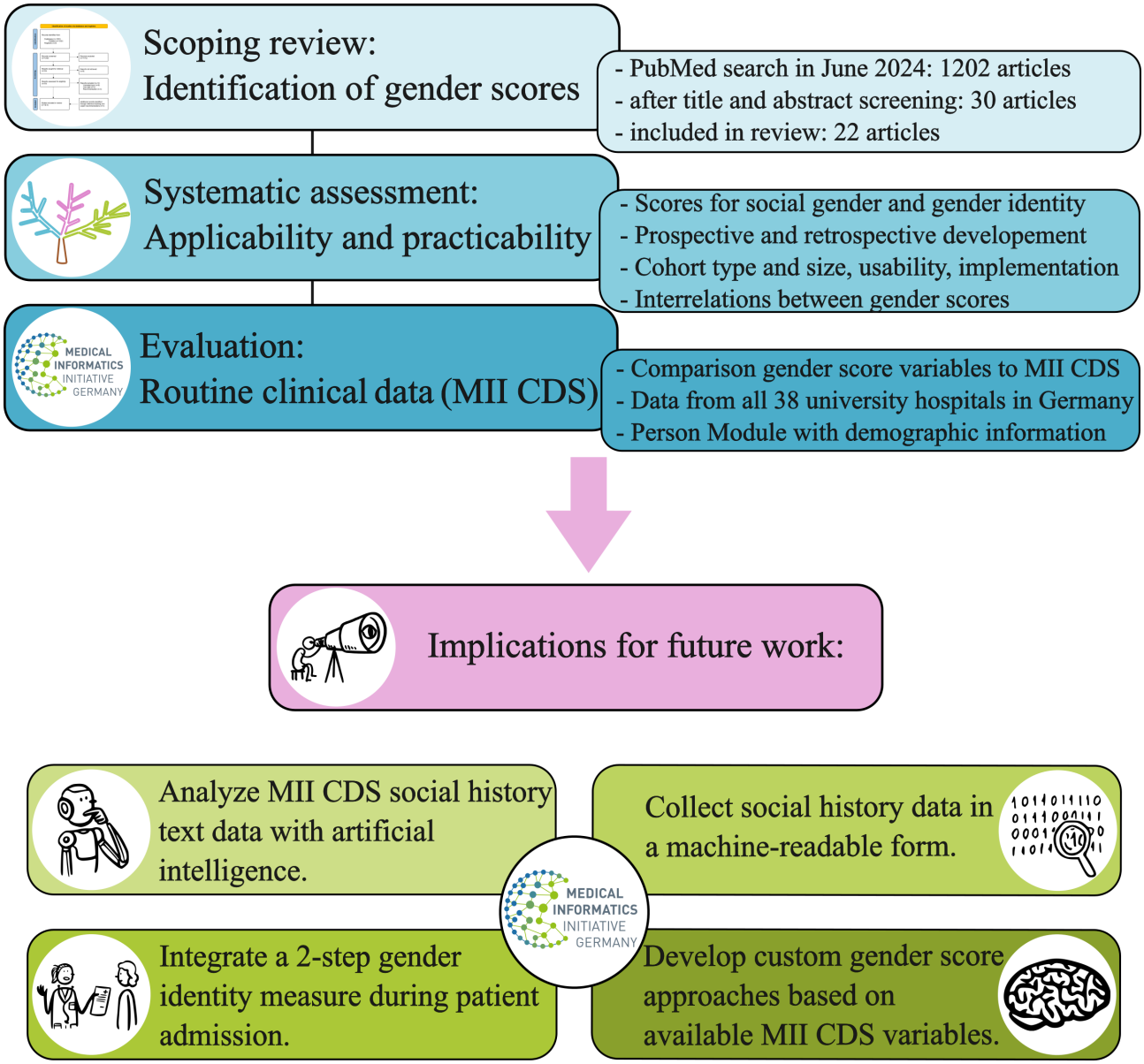
Aim of This Work

In the scope of this work, we identified and categorized state-of-the-art gender scores, systematically assessed their applicability and practicability, and evaluated their applicability on German research data from routine clinical practice (see the Applicability of Gender Scores on the MII CDS section).

The process of our work is illustrated in Figure 1. This work aimed to contribute to the necessary implementation of a standard for collecting and assessing gender-specific data in German university hospitals and the respective research facilities; therefore, we formulated 4 action steps for an extension in health research to enable gender-specific analyses (see Figure 1), which are shown at the bottom of the figure and will be described in detail in the Discussion section.

XSL•FO
RenderX

Figure 1. Overview of the implemented methods. MII CDS: Medical Informatics Initiative core dataset.



Methods

Scoping Review

We conducted a scoping review to identify articles that reported on the development and implementation of gender scores in health-related studies (see Figure 1). It resulted in 22 articles, which were identified through a title and abstract screening followed by a full-text evaluation. Moreover, we developed a data charting form to extract information on article type and applicability of the gender score in health research. The review protocol was published in JMIR Research Protocols [23].

Assessment of Applicability and Practicability

To assess the applicability and practicability of gender scores, we followed a systematic approach (see Figure 1). In the scope of this work, we extended the data charting process to fully assess the applicability of gender scores on routine clinical data.

Social Gender and Gender Identity

First, we distinguished between scores for gender identity (see Table 1) versus social gender (see Tables 2 and 3). Gender identity (see Textbox 3) is usually measured in a prospective way because it is a purely subjective perception and it is not possible to extract this information when it was not collected. However, social gender includes norms that are expected in society, stereotypes, and certain behaviors (see Textbox 2), and it can be measured prospectively and retrospectively. Typical behavior (eg, specific hobbies, occupational segregation, doing housework or odd jobs) can be analyzed and interpreted after data collection. A typical female or male lifestyle or behavior will lead to different exposure to risk factors for accidents or disease. Social expectations will also influence the process of recovery.

XSL•FO
RenderX

Table . Prospective gender identity measures [12,13,27,28].

Author(s) (year)	Description	Questions and answer option
Lagos and Compton (2021) [13]	2-step gender identity measure	1. What sex were you assigned at birth? (male, female, intersex) 2. What is your current gender? (woman, man, transgender, a gender not listed here [space for answer])
Fraser et al (2020) [12]	Open-ended gender measure	What is your gender? (open-ended item)
McGuire et al (2019) [27]	Genderqueer Identity Scale	1. I am non binary, genderqueer, or an identity other than male or female. 2. I don't want to be seen in the gender binary (as either male or female) 3. I try to deliberately confuse people about whether I am male or female. 4. I try to do things that are masculine and feminine at the same time. 5. I enjoy it when people are not sure if I am male or female 6. The way I think about my gender has always been the same. 7. My gender comes naturally from within me. 8. My gender is something I have spent a lot of time figuring out. 9. The way I show my gender changes depending on who I am with. 10. The way I think about my gender has been influenced by experiences in my life. 11. The way I think about my gender will probably continue to change further as I age 12. I have done research about gender theory and gender roles. 13. I try to convince others that society should not insist on a gender binary. 14. I try to convince others that society expects people to be too gender conforming. 15. Around me, I make sure people are free to express whatever gender roles they want. 16. The way I show my gender is important because I push society to question traditional gender roles. 17. I encourage others to be more open minded about gender and gender roles. 18. In the future, my gender expression will be traditional. 19. In the future, it will upset me if people misgender me. 20. The way I show my gender will probably be mostly the same from day to day. 21. In the future, I expect that people will rarely question my gender. 22. In the future, I think my gender will be fluid or change over time. 23. I will have a non-traditional gender role (be gender non-conforming).
Lombardi and Banik (2016) [28]	2-step gender identity measure	1. What is your sex or gender? (Check ALL that apply; male, female, other: please specify [space for answer]) 2. What sex were you assigned at birth? (Check one; male, female, unknown or question not asked, decline to state)

XSL•FO
RenderX

Table . Prospective gender scores regarding social gender [17,19,20,29-32].

Author(s) (year)	Gender score variables
Pohrt et al (2022)[a] [29]	1. Level of education[b] 2. Civil status[b] 3. Personal net income 4. Households primary earner status 5. Responsibility for caring for other people[b] 6. Responsibility educating and upbringing people[b] 7. Responsibility for picking up people when they become ill 8. Hours per week spent on housework 9. Being mainly responsible for housework, 10. Social support[b] 11. PAQ[c] 12. BSRI[d]: masculinity score 13. BSRI: femininity score 14. Perceived social standing
Tibubos et al (2022) [30]	
PAQ-8	1. Not at all emotional - very emotional 2. Not at all aware of feelings of others -very aware of feelings of others 3. Never gives up easily - gives up very easily 4. Not at all self-confident - very self-confident 5. Feels superior - feels very inferior 6. Not at all understanding of others - very understanding of others 7. Very warm in relations with others - very cold in relations with others 8. Goes to pieces under pressure - stands up well under pressure
Demuth et al (2021) [31]	The questionnaire is not available.
Nielsen et al (2021) [17]	1. Caregiver strain[b] 2. Work strain[b] 3. Independence 4. Risk-taking 5. Emotional intelligence 6. Social support[b] 7. Discrimination
Pelletier et al (2015) [20]	1. Primary earner status 2. Personal income 3. Number of hours per week doing housework 4. Primary responsibility for doing housework 5. Level of stress at home[b] 6. BSRI: masculinity score 7. BSRI: femininity score
Spence et al (1975) [32]	

XSL•FO
RenderX

Author(s) (year)	Gender score variables
PAQ: male score	1. Independent 2. Not easily influenced 3. Good at sports 4. Not excitable, minor crisis 5. Active 6. Competitive 7. Skilled in business 8. Knows ways of world 9. Adventurous 10. Outspoken 11. Interested in sex 12. Makes decisions easily 13. Not give up easily 14. Outgoing 15. Acts as leader 16. Intellectual 17. Self confident 18. Feels superior 19. Takes a stand 20. Ambitious 21. Stands up under pressure 22. Forward 23. Not timid
PAQ: female score	1. Emotional 2. Not hide emotions 3. Considerate 4. Grateful 5. Devotes self to others 6. Tactful 7. Strong conscience 8. Gentle 9. Helpful to others 10. Kind 11. Aware, other feelings 12. Neat 13. Creative 14. Understanding 15. Warm to others 16. Likes children 17. Enjoys art and music 18. Expresses tender feelings
PAQ: sex-specific score	1. Aggressive (M) 2. Dominant (M) 3. Likes math and science (M) 4. Excitable, major crisis (F) 5. Home-oriented (F) 6. Mechanical aptitude (M) 7. Needs approval (F) 8. Feelings hurt (F) 9. Cries easily (F) 10. Loud (M) 11. Religious (F) 12. Sees self running show (M) 13. Needs for security (F)
Bem (1974) [19]	

XSL•FO
RenderX

Author(s) (year)	Gender score variables
BSRI: masculinity score	1. Acts as a leader 2. Aggressive 3. Ambitious 4. Analytical 5. Assertive 6. Athletic 7. Competitive 8. Defends own beliefs 9. Dominant 10. Forceful 11. Has leadership abilities 12. Independent 13. Individualistic 14. Makes decisions easily 15. Masculine 16. Self-reliant 17. Self-sufficient 18. Strong personality 19. Willing to take a stand 20. Willing to take risks
BRSI: femininity score	1. Affectionate 2. Cheerful 3. Childlike 4. Compassionate 5. Does not use harsh language 6. Eager to soothe hurt feelings 7. Feminine 8. Flatterable 9. Gentle 10. Gullible 11. Loves children 12. Loyal 13. Sensitive to the needs of others 14. Shy 15. Soft spoken 16. Sympathetic 17. Tender 18. Understanding 19. Warm 20. Yielding
BRSI: social desirability score (neutral)	1. Adaptable 2. Conceited 3. Conscientious 4. Conventional 5. Friendly 6. Happy 7. Helpful 8. Inefficient 9. Jealous 10. Likable 11. Moody 12. Reliable 13. Secretive 14. Sincere 15. Solemn 16. Tactful 17. Theatrical 18. Truthful 19. Unpredictable 20. Unsystematic

[a]Overlaps with the already published review by Ballering et al [26].

[b]Variables that are potentially covered in routine clinical care in Germany.

[c]PAQ: Personal Attributes Questionnaire.

XSL•FO
RenderX

[d]BSRI: Bem Sex Role Inventory.

XSL•FO
RenderX

Table . Retrospective gender scores regarding social gender [9-11,33-40].

Author(s) (year)	Gender score variables
Cipriani et al (2024)[a] [36]	1. Hostile behavior during childhood 2. Hostile behavior during adulthood 3. Having at least a secondary school diploma[b] 4. Having at least a bachelor's degree[b] 5. Sleep satisfaction[b] 6. Sleep efficacy[b] 7. Having private housing 8. Sexual violence during childhood
Gisinger et al (2023)[a] [35]	1. Level of education[b] 2. Household income 3. Household size[b] 4. Civil status[b] 5. Perceived life stress 6. Sense of belonging to community
Teterina et al (2023)[a] [37]	
ICD-10-CA[c] diagnostic codes with highest effects for predicting male/female gender: male score	1. Fall on and from scaffolding[b] 2. Occupant of heavy transport vehicle injured in noncollision transport accident, driver, traffic accident[b] 3. Driver of other all-terrain or other off road motor vehicle injured in traffic accident[b] 4. Motorcycle rider injured in collision with fixed or stationary object, driver, traffic accident[b] 5. Asymptomatic human immunodeficiency virus (HIV) infection status[b] 6. Motorcycle rider injured in collision with car, pick-up truck or van, driver, traffic accident[b] 7. Motorcycle rider injured in noncollision transport accident, driver, nontraffic accident[b] 8. Motorcycle rider injured in unspecified nontraffic accident[b] 9. Fracture of malar and maxillary bones, LeFort 2, closed[b] 10. Contact with nonpowered hand tool[b]
ICD-10-CA diagnostic codes with highest effects for predicting male/female gender: female score	1. Assault by spouse or partner 2. Esthetic sports[b] 3. Physical abuse[b] 4. Horse riding sports[b] 5. Animal-rider or occupant of animal-drawn vehicle injured by fall from or being thrown from animal or animal-drawn vehicle in noncollision accident[b] 6. Fall while being carried or supported by other persons[b] 7. Other specified gymnastic and esthetic sports and recreational activity[b] 8. Bitten or struck by dog[b] 9. Animal-rider or occupant of animal-drawn vehicle injured in other and unspecified transport accidents[b] 10. Problems in relationship with spouse or partner

XSL•FO
RenderX

Author(s) (year)	Gender score variables
Vader et al (2023)[a] [38]	

XSL•FO
RenderX

Author(s) (year)	Gender score variables
Masculine gender score	1. Work and education (compared with partner)[b] 2. Informal care: household chores, odd jobs, taking care of sick people[b] 3. Lifestyle: physical intensity, type of sport, smoking, alcohol[b] 4. Emotions: emotional problems, nervousness, energetic and vibrant, exhausted and tired
de Breij et al (2022)[a] [33]	1. Working hours[b] 2. Income 3. Occupation segregation[b] 4. Education[b] 5. Informal caregiving[b] 6. Time spent on household chores
Wandschneider et al (2022)[a] [9]	1. A person who is living with their partner for the long term should get married. 2. Children below the age of 6 suffer if their mother works. 3. A same-sex couple can raise a child just as well as a man and woman. 4. It would be good for society if transgender people were recognized as normal. 5. Working experience part-time employment 6. Hours/weekday housework 7. Hours/weekday repairs 8. Hours/weekday leisure, hobbies 9. Worried about global terrorism 10. Worried about crime in Germany 11. Satisfaction with housework 12. Willingness to take risks 13. Worried about own retirement pension
Nauman et al (2021)[a] [11]	1. Chronic stress 2. Civil status[b] 3. Risk-taking behavior 4. Agreeableness 5. Neuroticism 6. Extraversion 7. Conscientiousness 8. Loneliness 9. Level of education
Yuan et al (2021)[a] [39]	1. Risk willingness 2. Loneliness[b] 3. Less participation in household tasks 4. Regular drinking[b] 5. Depression[b]
Ballering et al (2020)[a] [40]	1. Leisure activities (eg, type of sport)[b] 2. Occupation-related components (eg, profession, housewife/-husband)[b] 3. Time spend on household tasks 4. Time spend on odd jobs 5. Lifestyle (eg, dieting, preparing dinner, alcohol, smoking) 6. Experiencing long-term difficulties or negative life events 7. Personality traits and emotions (eg, discipline, impulsiveness, self-consciousness, vulnerability, competence

XSL•FO
RenderX

Author(s) (year)	Gender score variables
Lacasse et al (2020)[a] [10]	1. Civil status[b] 2. Racial/ cultural group 3. Level of education[b] 4. Household income, size, and composition[b] 5. Household food insecurity 6. Ownership of the household 7. Sense of belonging to the local community 8. Occupation: type, working hours, self-employment[b] 9. Amount of stress
Smith and Koehoorn (2016)[a] [34]	1. Responsibility for caring for children[b] 2. Occupational segregation[b] 3. Hours of work relative to partner/spouse 4. Education relative to partner/spouse[b]

[a]Overlaps with the already published review by Ballering et al [26].

[b]Variables that are potentially covered in routine clinical care in Germany.

[c]ICD-10-CA: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada.

Prospective and Retrospective Scores

Second, we distinguished between prospective (see Tables 1 and 2) and retrospective (see Table 3) scores. It is advisable to include gender variables from the outset of a study; this results in a prospective gender score and is easy to apply to the study dataset. However, the huge amount of existing data drives the development of retrospective gender scores, where gender-specific information can be extracted in hindsight.

Cohort Type and Size

Third, it needs to be considered whether the gender scores themselves are representative. In practice, some gender scores are developed on restricted datasets, for instance, for specific patient groups or other samples that are not representative (eg, older age groups, workers, or specific diagnoses). This is underlined by the finding of Wandschneider et al [9] who showed that gendered practices vary between eastern and western Germany, reflecting a different social development due to the historical political context. As gender norms, stereotypes, or behaviors vary between societies, the relevant gender variables can differ across social groups, countries, and regions or even between different societies within a country.

Furthermore, since retrospective scores depend on statistical modeling, we also considered the sample size on which the models were based. Another analysis we performed was to further classify the background under which the score was developed. We took into account whether the research group used patient or population data to categorize which gender score might be more appropriate for the MII CDS. We also considered the continent where the data collection took place, which is important for consistency due to varying gender-sensitive variables depending on the society or geographical location.

Usability

Fourth, we examined the validity, usability, and practicability of the gender scores. Validity is the main issue because it is of high importance that the score is representative regarding gender. Usability refers to how well the score is established. If a new standard is to be introduced, it is of high importance to use an accepted model to ensure that the scores are usable, comparable, and applicable to large-scale studies and assessments. Practicability refers to how difficult a score would be to integrate into health research.

Implementation

The fifth point we included in our assessment was the level of implementation of scores. To apply scores systematically on existing data, a theoretical model was not sufficient. Instead, it was necessary to have the executable code of a model to integrate it directly. Therefore, we further investigated the applicability of gender scores from the technical viewpoint to evaluate whether models can, realistically, be implemented in a clinical setting. We used the following categories to describe this aspect: code published — refers to a gender score for which an executable model was published; data published — refers to a gender score for which the underlying data are available and a score could be reproduced; statistical parameters published — refers to a model for which all statistical parameters were published and the model could be constructed from them; not published — a model for which not all relevant parameters are available. We considered “code published” the gold standard regarding applicability. “Data published” and “parameters published,” on the other hand, might contain sufficient information to construct an executable model. In the scope of this study, we considered code or data available “on request” as unavailable. However, in practice, there might still be limitations if the description of the article is not concise enough to reproduce all steps.

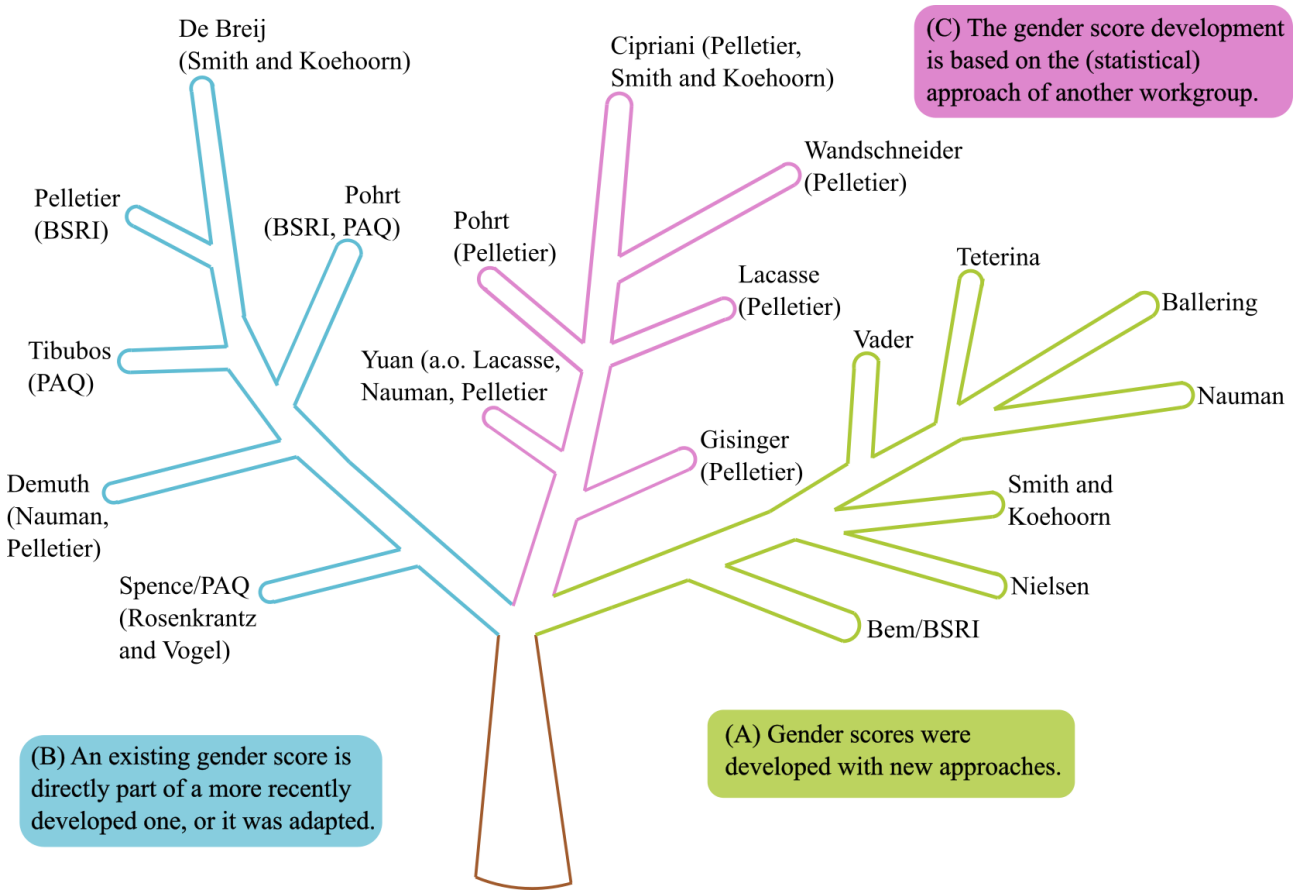
Interrelations

Based on the extracted data, we performed further analysis to better understand the currently available gender scores. We noticed early on that direct dependencies exist between gender scores. Therefore, we extracted and modeled different types of dependencies between gender scores (see Figure 2). This

XSL•FO
RenderX

analysis allowed us to investigate how scores were developed and whether improvements were made to existing gender scores.

Figure 2. Interrelations between gender scores, where the first 1 or 2 author names identify the work and the developed gender score (see Tables 2 and 3) [9-11,17,19,20,29-41]: (A) gender scores developed with completely new approaches and methods are shown in green; (B) the included or adapted gender scores are in parentheses in the blue branch; (C) gender scores developed based on previous methods and statistical approaches, which are shown in parentheses, are shown in pink. BSRI: Bem Sex Role Inventory; PAQ: Personal Attributes Questionnaire.



One relation between gender scores captures cases in which an existing gender score was used directly or adapted in another score (see Figure 2B). For instance, the gender score from Pelletier et al [20] includes the BSRI [19]. de Breij et al [33] adapted the score from Smith and Koehoorn [34] but changed "caring for children" to "informal caregiving" as the data were available in their dataset (see Figure 2B).

Moreover, development relationships (see Figure 2C) reflect when one gender score was developed based on the approach, for instance the statistical model, of another research group. For example, Gisinger et al [35], with their statistical model of principal component analysis and logistic regression analysis, were inspired by the approach of Pelletier et al [20].

Applicability on Clinical Routine Data

To assess whether gender scores are applicable on existing data and whether they can be integrated in clinical routine data, we used the technical specification of the MII Core Data Set - Module "Person" as a representative sample against which to evaluate the gender scores (see Figure 1). The MII CDS [14] (see the Background section) is being collected across all 38 university hospitals in Germany. It contains interoperable, consolidated data items with a high degree of standardization in line with international standards to make collaborative data analysis practicable and efficient.

The CDS [42] consists of basic and extension modules that are subject to iterative refinement. The basic modules are generic, and the extension modules are relevant for specific medical fields and disciplines such as oncology or intensive care. We used the *Person Module*, one of the early basic modules, which contains demographic information such as name, date of birth, address, and administrative gender as stated in official documents. The CDS has a comprehensive description of how the data are structured and how the data can be accessed, which allows us to directly compare it against gender scores. The information is available and can be compared based on a sample dataset [43]. Overall, this makes the CDS a suitable choice for studying how gender scores can be applied in realistic clinical data collection.

Retrospective scores required that we compared the variables contained in the gender scores against the variables available in routine clinical data. To perform this comparison, we systematically analyzed all variables contained in the sample dataset of the MII CDS against the variables included in the statistical models of the gender scores.

Ethical Considerations

We did not collect any personal data, so ethical approval was not needed.

XSL•FO
RenderX

Results

The full results of the review are presented in Multimedia Appendix 1 to maintain brevity in the main text.

Scoping Review

Our review results are categorized into prospective development (Tables 1–2) and retrospective development (Table 3) and between gender scores (Tables 2–3) and gender identity measures (Table 1). Articles that overlapped with the already published review by Ballering et al [26] are marked in the tables.

Assessment of Applicability and Practicability

Social Gender, Gender Identity, and Prospective and Retrospective Scores

We distinguished between articles targeting gender identity and articles targeting social gender. We further classified them into prospective and retrospective development and application.

The 4 identified gender identity measures (see Table 1) are prospectively applicable and could be included in future studies for further data collection regarding gender identity. Participants need to be asked explicitly which gender they belong to; this information cannot be analyzed from existing data.

Moreover, we found 18 gender scores to measure social gender in different ways. Of these, 7 are prospective gender scores (see Table 2) that could provide guidance on how to apply (social) gender to a study from the outset. Of the 18 scores, 11 are retrospective (see Table 3) and could be applied on already existing data.

Cohort Type and Size

The background of the cohorts is of importance for the practicability. A score contains different variables, depending on whether it was developed on data from population studies or on patient data. The majority of datasets are based on population studies and thus on specific information that is typically collected in these studies. For a complete comparison of cohorts, see our results in Multimedia Appendix 1.

Only 2 gender scores were developed exclusively based on patient data. Teterina et al [37] focused on patients who entered the emergency department with traumatic brain injuries and published an innovative approach to distinguishing gender by diagnostic codes. These diagnostic codes also describe the background of the accident or incident, enabling conclusions to be drawn about gender. Cipriani et al [36] focused on the effects of gender on emergency patients with psychological crisis. Therefore, they used variables assessing childhood trauma, education, sleep quality, and living situation.

Moreover, the article by McGuire et al [27] assessed gender identity during transition through the collection of patient and population data.

However, all of these studies were restricted by the specific patient cohort group and focused on specific conditions. This hampers their applicability on the general public and their usefulness as broadly applied gender scores for routine clinical data.

In addition, the used population cohorts were neither comprehensive nor inclusive. Many studies used a cohort with a specific age group [11,20,29,31,38,39] or with a specific background like being retired [29] or being a worker [10,33]. This is a major weakness for generalizing results, as a gender score based on a limited cohort is not representative of society as a whole and therefore lacks comprehensiveness.

Furthermore, the majority of cohorts used for developing gender scores are based in the same geographic region, primarily Europe and North America. Only one gender identity measure was used on a cohort in Oceania [12]. We could not find a single gender score publication with a cohort from Asia, South America, or Africa. As a result, the developed scores are not representative for every cohort group and are not universally applicable.

With respect to the size of the cohorts, which forms the basis for the statistical development of retrospective gender scores and influences the reliability of statistical models, we noted that some works used huge cohorts or datasets with n>100,000 [35,37,40] and the outstanding cohort size of approximately 700,000 individuals in the study by Smith and Koehoorn [34]. Other gender scores worked with smaller cohort sizes of n<1000 [19,32,33].

Usability

We examined the validity of scores, meaning how representative the work is regarding gender. Gender can be different in each society, so it is a major task to implement a gold standard. Some research groups used sex as the dependent variable and, therefore, gender as a predictor for sex (eg, [9-11,20]). This approach can be problematic since gender should be considered beyond biological sex.

Statistical regression approaches can be systematically validated by using a control set to test the generalization of the regression model (eg, by systematic evaluation on a dedicated validation set [20,36,37,40]). Researchers who developed a number of scores using the approach by Pelletier et al [20] did not specify whether they performed a systematic validation of their developed scores [9,29,35]. Nauman et al [11] did not use a validation, and Yuan et al [39] reported a performance value without describing the validation setting they used.

3120,2922

Exclusively theory-driven gender scores, where variable scaling is determined by human experts, were not evaluated since there is no gold standard for gender [33,34,38]. Looking at usability, referring to how well established a score already is, some authors did not mention any practical application of their scores [31], whereas other scores were often cited, adapted, and reapplied [19,20,32,34]. Although highly cited scores seem to be widely accepted and used, they might still not be well suited to represent gender. The most cited score was the BSRI, which is also part of other gender scores (eg, [20,29]; see Figure 2), and was often criticized and considered outdated [22].

Regarding practicability, we considered the scope and complexity of questionnaires and how universally applicable variables are on existing data. Some variables were very specific

XSL•FO
RenderX

for the used cohort, whereas others frequently appeared in similar studies. Most gender identity measures are well-suited for clinical data collection, being concise and easy to understand, except for the score from McGuire et al [27], which involves 23 questions and focuses on gender transition, making it less suitable. BSRI [19] and PAQ [32] are lengthy prospective gender scores, same as other gender scores containing them. Tibubos et al [30] developed a shorter form of the PAQ with only 8 variables. We also listed different forms of data collection, from self-reporting questionnaires (eg, [20,34]) to interviews (eg, [13,33,42]). The length of forms could be especially relevant if participants complete a self-reporting questionnaire. Interviews tend to be more time-consuming for the staff conducting them.

Regarding retrospective scores, the length does not have such a big impact, but some scores are much shorter than the majority [34,39]. A minimalistic score can be a trade-off between prediction quality and the number of necessary variables. This was, for instance, investigated by Ballering et al [40], who compared the performance between gender scores including 9 and 85 predictive variables.

Implementation

Only 3 research groups made their analysis code publicly available (*code published*) and provided reproducible gender scores [9,17,40].

Two studies published their data [17] or used publicly available data [39] (*data published*), which are necessary to evaluate their conclusion. All other works did not publish data, which might be justified by data protection regulations. However, anonymized data could still be helpful for recreating a study.

Statistical parameters were published in most retrieved papers (*statistical parameters published*), some more complete than others. However, in the majority of cases, our assessment showed that reconstructing a gender score from the provided parameters would be challenging.

We identified one paper that did not publish the gender score (*not published*) [31]. They adapted the scores from Nauman et al [11] and Pelletier et al [20] to form a new gender score, but they did not provide any specific information regarding their methodology.

Overall, repeatability was made difficult by many of the research groups, and the use of existing scores was made complicated. A systematic overview of the reproducibility of each study is given by our results in Multimedia Appendix 1.

Interrelations

We registered different dependencies and relationships between gender scores and visualized them (see Figure 2). Of the retrieved articles, 7 were developed with totally new approaches or methods (see Figure 2A).

Other scores adapted or included content-related parts of already existing scores (see Figure 2B). For example, Pelletier et al [20] included the BSRI masculinity and femininity scores [19]. Spence et al [32] were inspired by Rosenkrantz et al [41] for the development of the PAQ in 1975, and thereupon, Tibubos et al [30] developed a short form of the PAQ in 2022. It is notable that the BSRI [19] and PAQ [32] each form the basis of 2 other scores.

Sometimes research groups implemented the statistical or methodological approach of another team (see Figure 2C). For instance, the statistical approach of Pelletier et al [20] was frequently used as the basis for current gender scores, for example, by Wandschneider et al [9], Lacasse et al [10], Gisinger et al [35], Pohrt et al [29], Cipriani et al [36], and Yuan et al [39].

Applicability on Clinical Routine Data

Evaluation Overview

We tested all retrieved scores and their published variables for whether they are available in the MII CDS. Initially, we planned to only test retrospective scores developed on clinical data. However, we found that this would limit the number of possible scores too much.

The evaluation of the applicability on clinical routine data required a careful and considerate approach, as variable names are often ambiguous and it is usually not possible to match them directly. This can, for instance, be seen in variables related to care work, which include “responsibility for caring for other people in the household” [29], “dinner is always prepared by someone else” [40], “caregiver strain” [17], “taking care of sick people” [38], “informal caregiving” [33], and “responsibility for caring for children” [34]. Even though some gender scores might be well described and could be used for other datasets, transferring the variables appropriately to a different dataset proved to be complex.

Applicability of Gender Scores on the MII CDS

In the comparison between the variables contained in gender scores and within the MII CDS, we used a two-step approach: We first tried to match the exact gender score variables against the MII CDS, and in case this remained unsuccessful, we extended the search to similar or generic terms.

We found that no variables of any gender score are covered in the MII CDS. This finding is surprising since a high number of variables covered in gender scores, such as occupation, level of education, civil status, children, household composition and size, and social support, are part of social history, which is routine clinical information.

Our principle finding regarding this question was that no gender score can be applied upon the MII CDS. One exception is the score from Teterina et al [37], which includes diagnostic codes and disease-specific data, which could be matched against the MII CDS. This score, however, is highly specific and not usable as a generalized approach. Currently, the MII CDS only encodes the administrative gender as male, female, undifferentiated, or diverse [43]. Administrative gender refers to the gender that is recorded or recognized in official documents (eg, the ID card). It may differ from the birth-assigned sex.

Overall, most scores were applied on population studies or included in studies from the outset. These gender scores frequently include personality traits and psychological

XSL•FO
RenderX

assessments (eg, stress, loneliness scales), which are not part of routine clinical practice and, therefore, also not part of the MII CDS.

To summarize, there are no fitting variables for the MII CDS, and it is impossible to apply an existing gender score to the dataset in order to include gender-specific analysis on clinical routine data.

Discussion

Principal Findings

This work identified and reviewed state-of-the-art gender scores and gender identity measures. These were systematically assessed regarding their applicability and practicability in health research. Last, we tested their applicability on routine clinical data using the MII CDS as a reference dataset.

We found that no gender score is applicable on the MII CDS, because the variables required for the gender scores are not part of the CDS. However, many variables that are commonly included in gender scores are also assessed during a patient's clinical stay. For instance, variables regarding personal background (eg, living with your partner or children, professional field, or self-employment) are often asked during a clinic stay to verify if specific therapies or treatments are feasible. Physicians need to know if patients have responsibilities at home, if they have support, and if they have a stable social environment. A person's background is gender-dependent and influences not only clinical routine but also research data and therapy guidelines (eg, risk factors dependant on lifestyle). This is underlined by the information in Multimedia Appendix 2, which presents examples of current social history documentation practices in a German hospital compared against typical variables used in gender scores. Several variables related to gender scores can be directly inferred from these documentations; however, they are currently not included in the MII CDS. We therefore argue that the current lack of applicability of gender scores on clinical routine data results from missing structured data collection in or extraction methods from clinical information systems. Since the MII CDS is an ongoing effort to standardize research data from routine clinical practice, we recommend that the Person Module of the MII CDS should be updated and expanded to create a sufficient knowledge base for gender health research. Specifically, we formulated 4 action steps that will enable gender-specific analyses (see Figure 1): analyze social history data, collect social history data in a machine-readable form, integrate a 2-step gender identity measure during patient admission, and develop custom gender score approaches for clinical data.

Recommended Action Steps

Analyze Social History Data

Social history is part of routine clinical data collected for many patients. However, the captured information is oftentimes not structured. It can be found in the doctor's notes or additional study data if patients are included in clinical trials. As a first step toward evaluating what information is currently available, we suggest analyzing the full-text information in the patient records using natural language processing and large language models for data extraction. It will be interesting to learn what conclusions and predictions one could draw from the available information and how this information corresponds to the described gender scores. Typical social anamnesis texts are given in Multimedia Appendix 2, showcasing the low information content. A structured analysis of social history data is still ongoing, but the preliminary check of the information content already indicates that the practice of documentation must be significantly improved.

Collect Social History Data in a Machine-Readable Form

We recommend that key components of patients' social histories — such as occupation, education, civil status, number of children, household composition and size, social support — be collected in a structured, machine-readable format. Instead of recording social history solely as narrative text in physician reports, clinical information systems should provide individual fields for each variable, allowing clinicians to select or enter information directly. This approach would reduce the effort required for data processing and aggregation, enable semantic annotation, and facilitate automated export and comparison across datasets.

Several national-level initiatives already exemplify such practices. Scandinavian countries, for instance, collect structured social and demographic data that can be integrated with health care data [44,45]. Similarly, the Canadian Institute for Health Information promotes the standardized collection of social and behavioral data across health care settings, including variables such as family composition, living arrangements, social support, and socioeconomic status. Existing research has also demonstrated how such information can be effectively captured using discrete, machine-readable fields rather than free-text formats [46].

A widely recognized standard that could be further adapted for this purpose is HL7 FHIR [47], which provides a structured framework for encoding social history elements in clinical IT systems. Future work should build on these existing resources — as well as our overview for gender scores — to develop a system that supports the structured capture of variables necessary for calculating gender scores. Although it is important to build future work upon existing standards, ideally, the analysis of clinical data items determining gender should be conducted before designing structured data capture forms for patient social history.

Integrate a 2-Step Gender Identity Measure During Patient Admission

Moreover, to include as many gender dimensions as possible, it would be beneficial to collect more gender-specific data. A promising approach would be to implement the 2-step gender identity measure [13] with open-ended answers during patient admission. Implementing a prospective approach, with minimal effort, would provide valuable insights into gender identity in everyday clinical practice. Although the DIVERGesTOOL [48] was originally developed as an extension of the 2-step approach for the German research context, it may also be well suited for use in patient admission settings. It adds a third question

XSL•FO
RenderX

regarding whether differences in sex development have ever been medically diagnosed.

However, collecting such data must be approached with care, because asking gender-related questions at patient admission can involve sensitive issues and carries potential risks. For transgender, nonbinary, or gender-nonconforming individuals, disclosing gender identity may raise concerns about discrimination or biased treatment. Patients frequently withhold this information due to previous negative experiences and a lack of trust in health care providers [49,50]. Inadequate staff training and poorly designed forms can lead to misgendering or exclusion. To ensure respectful and safe care, it is essential to provide staff with proper training, use inclusive language, and implement supportive documentation systems.

Develop Custom Gender Score Approaches for Clinical Data

Researchers should think beyond the common gender score approaches and develop a gender score that is suitable to clinical data and not only population studies.

An innovative approach was carried out by Teterina et al [37], who used diagnostic codes of patients with traumatic brain injuries in Canadian emergency departments. Researchers should try similar approaches on general — nondisease-specific — clinical data with diagnostic or medication variables. Furthermore, developing a minimalistic retrospective score could prove advantageous in balancing prediction quality with the number of required variables, making the scores applicable even when limited information is available (eg, [40]).

Even though gender is gaining importance in health research, it is still far from being implemented. Our investigation shows that a gold standard is missing that accounts for variations across social groups (eg, including geographical regions, age, time). Due to the lack of a gold standard, some research groups use gender as a predictor for sex when developing gender scores, which contradicts the purpose of introducing gender in research — to move beyond biological sex and gather more detailed information. Solely theory-driven approaches, where experts determine the used variables and their scaling, cannot be validated.

Last, repeatability is a major challenge when aiming at implementing existing gender scores. Due to inconsistent research group practices, existing gender scores are often described only theoretically without providing detailed variables, questions, or complete information on variable weighting. We therefore urge authors to provide complete documentation and executable models when developing gender scores.

To achieve the necessary progress, every inclusion of gender into health research is beneficial, even if it might not be comprehensive or sufficient to cover all aspects.

Limitations

Our collection of gender scores is based on a scoping review conducted to provide a broad overview of the topic. One limitation of this approach is that only one reviewer conducted the full-text screening and the categorization of the retrieved articles. However, during data extraction, more than 30% (7/22, 32%) of the publications were initially double-checked by a second domain expert [23]. The reviewers were in full agreement on all reviewed articles; thus, a single reviewer continued the data extraction. Moreover, our review does not claim to provide a complete list of all existing gender scores but does provide general statements on the current state of gender scores and their applicability in health research.

Moreover, the MII CDS might not be fully representative to test against population-based gender scores. Most gender scores were developed on population or specific patient data, while the MII CDS contains routine clinical data. However, it is important to include gender not only in population research but also in clinical research. This is why we selected the MII CDS as a target dataset for routine clinical data to test the practicability in this setting. As a result, we were able to show that gender scores are not yet a feasible instrument for routine clinical data and are able to make recommendations on how to update and extend the MII CDS in the future.

Comparison With Prior Work

Ballering et al [26] highlighted that gender scores should be the minimum effort in epidemiological studies, but the community should go further than that because the retrospective application of a gender score implies a lack of gender consideration from the outset of the study (eg, the study design). Our findings found several shortcomings with currently available retrospective gender scores. Therefore, our results suggest that self-reported gender assessments are more precise and better suited to assess gender, confirming the theory of Ballering et al [26].

Moreover, we can confirm the findings of Ballering et al [26], specifically that the most common variables in gender scores are related to occupation, income, education, civil status, caregiving and household responsibilities, and ways of spending (leisure) time. Furthermore, our results showed that, even though several of these variables are collected in everyday clinic work, they are not processed and stored in a machine-readable way.

Conclusions

Considering sex and gender enhances equity and research quality. Despite the high importance of the topic, it remains challenging to include gender into health research for several reasons. Ethical implications such as properly defining the concept of “gender” as opposed to “sex,” identifying appropriate ways of asking about a patient’s gender, and raising awareness of the importance with nurses and doctors are nontechnical reasons why the health gender gap remains. Technical implementation is hampered by the absence of a generally applicable clinical gender score. Structured, computer-processable data and metadata about gender-specific aspects in a patient’s social history is also lacking.

In this work, we identified, categorized, and systematically assessed state-of-the-art gender scores from epidemiology and clinical studies. We evaluated their applicability and practicability on a German national research dataset for routine clinical practice, the MII CDS. We found that gender cannot be predicted based on the MII CDS, even though several of the frequently used variables are part of routine clinical practice in Germany.

XSL•FO
RenderX

However, we see an urgent need to include gender-relevant information to the MII CDS in order to narrow the gap between routinely collected clinical data and available research data (see Multimedia Appendix 2). Therefore, further work is necessary to enable gender-specific analysis and to routinely collect more gender-specific data in clinics, for example, during patient admission.

Despite different approaches for assessing gender, no standardized and validated gender score that could be used retrospectively in clinical research exists.

Our study is limited to the German clinical landscape, and our evaluation of possible scores is based on a scoping review. Future investigations into the literature and common practices at clinics inside and outside Germany might give further insights and add to the action items suggested in this work.

Funding

This work was funded by the German Federal Ministry of Education andResearch as part of the project “Inclusive Excellence in Medicine” (01FP23G10A,01FP23G10B) and the “Network University Medicine” (01KXX2121).

Data Availability

Further data underlying our publication are available in Multimedia Appendix 1.

Authors' Contributions

Conceptualization: LS, SS, EK, DW
Data curation: LS
Formal analysis: LS
Funding acquisition: SS, EK, DW
Investigation: LS
Methodology: LS, SS, EK, DW
Supervision: SS, EK, DW
Validation: HB, SS, EK, DW
Visualization: LS
Writing – original draft: LS, DW
Writing – review & editing: HB, DL, SS, EK, DW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed information from the assessment of the retrieved gender scores.
[XLSX File, 18 KB - medinform_v14i1e74162_app1.xlsx]

Multimedia Appendix 2
Three patients showcasing the current social history documentation in clinical practice in a German hospital. Footnotes and color coding indicate the comparison with typical variables used in gender scores. The text examples have been constructed for illustration.
[PNG File, 303 KB - medinform_v14i1e74162_app2.png]

References

1. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. Nature New Biol 2019 Nov;575(7781):137-146. [doi: 10.1038/s41586-019-1657-6] [Medline: 31695204]
2. Arcand M, Bilodeau-Houle A, Juster RP, Marin MF. Sex and gender role differences on stress, depression, and anxiety symptoms in response to the COVID-19 pandemic over time. Front Psychol 2023;14:1166154. [doi: 10.3389/fpsyg.2023.1166154] [Medline: 37207028]
3. Opinion: intersexuality. German Ethics Council. 2012. URL: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/englisch/opinion-intersexuality.pdf [accessed 2025-11-22]
4. Gender and health. WHO. 2024. URL: https://www.who.int/health-topics/gender [accessed 2024-09-13]
5. Oertelt-Prigione S. Chapter 32 - the operationalization of gender in medicine. In: Legato MJ, editor. Principles of Gender-Specific Medicine 2023:503-512. [doi: 10.1016/B978-0-323-88534-8.00017-1]
6. Geller SE, Koch AR, Roesch P, Filut A, Hallgren E, Carnes M. The more things change, the more they stay the same: a study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. Acad Med 2018 Apr;93(4):630-635. [doi: 10.1097/ACM.0000000000002027] [Medline: 29053489]

XSL•FO
RenderX

7. Potluri T, Engle K, Fink AL, Vom Steeg LG, Klein SL. Sex reporting in preclinical microbiological and immunological research. MBio 2017 Nov 14;8(6):e01868-17. [doi: 10.1128/mBio.01868-17] [Medline: 29138308]
8. Shah K, McCormack CE, Bradbury NA. Do you know the sex of your cells? Am J Physiol Cell Physiol 2014 Jan 1;306(1):C3-18. [doi: 10.1152/ajpcell.00281.2013] [Medline: 24196532]
9. Wandschneider L, Sauzet O, Razum O, Miani C. Development of a gender score in a representative German population sample and its association with diverse social positions. Front Epidemiol 2022;2:914819. [doi: 10.3389/fepid.2022.914819] [Medline: 38455329]
10. Lacasse A, Pagé MG, Choinière M, et al. Conducting gender-based analysis of existing databases when self-reported gender data are unavailable: the GENDER Index in a working population. Can J Public Health 2020 Apr;111(2):155-168. [doi: 10.17269/s41997-019-00277-2] [Medline: 31933236]
11. Nauman AT, Behlouli H, Alexander N, et al. Gender score development in the Berlin Aging Study II: a retrospective approach. Biol Sex Differ 2021 Jan 18;12(1):15. [doi: 10.1186/s13293-020-00351-2] [Medline: 33461607]
12. Fraser G, Bulbulia J, Greaves LM, Wilson MS, Sibley CG. Coding responses to an open-ended gender measure in a New Zealand national sample. J Sex Res 2020 Oct;57(8):979-986. [doi: 10.1080/00224499.2019.1687640] [Medline: 31729890]
13. Lagos D, Compton D. Evaluating the use of a two-step gender identity measure in the 2018 General Social Survey. Demography 2021 Apr 1;58(2):763-772. [doi: 10.1215/00703370-8976151] [Medline: 33834217]
14. Ammon D, Kurscheidt M, Buckow K, et al. Arbeitsgruppe Interoperabilität: Kerndatensatz und Informationssysteme für Integration und Austausch von Daten in der Medizininformatik-Initiative. Bundesgesundheitsbl 2024 Jun;67(6):656-667. [doi: 10.1007/s00103-024-03888-4]
15. Clayton JA, Tannenbaum C. Reporting sex, gender, or both in clinical research? JAMA 2016 Nov 8;316(18):1863-1864. [doi: 10.1001/jama.2016.16405] [Medline: 27802482]
16. Thibaut F. Gender does matter in clinical research. Eur Arch Psychiatry Clin Neurosci 2017 Jun;267(4):283-284. [doi: 10.1007/s00406-017-0797-7] [Medline: 28389888]
17. Nielsen MW, Stefanick ML, Peragine D, et al. Gender-related variables for health research. Biol Sex Differ 2021 Feb 22;12(1):23. [doi: 10.1186/s13293-021-00366-3] [Medline: 33618769]
18. Morgan R, Yin A, Kalbarczyk A, et al. Reconsidering tools for measuring gender dimensions in biomedical research. Biol Sex Differ 2024 Nov 25;15(1):96. [doi: 10.1186/s13293-024-00663-7] [Medline: 39587652]
19. Bem SL. The measurement of psychological androgyny. J Consult Clin Psychol 1974 Apr;42(2):155-162. [doi: 10.1037/h0036215] [Medline: 4823550]
20. Pelletier R, Ditto B, Pilote L. A composite measure of gender and its association with risk factors in patients with premature acute coronary syndrome. Psychosom Med 2015 Jun;77(5):517-526. [doi: 10.1097/PSY.0000000000000186] [Medline: 25984818]
21. Neufang S, Li F, Akhrif A, Beyan OD. Toward a fair, gender-debiased classifier for the diagnosis of attention deficit/hyperactivity disorder- a machine-learning based classification study. BMC Med Inform Decis Mak 2025 Aug 5;25(1):290. [doi: 10.1186/s12911-025-03126-0] [Medline: 40764552]
22. Donnelly K, Twenge JM. Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: a cross-temporal meta-analysis. Sex Roles 2017 May;76(9-10):556-565. [doi: 10.1007/s11199-016-0625-y]
23. Schindler L, Beelich H, Röll S, Katsari E, Stracke S, Waltemath D. Applicability of retrospective and prospective gender scores for clinical and health data: protocol for a scoping review. JMIR Res Protoc 2025 Jan 20;14(1):e57669. [doi: 10.2196/57669] [Medline: 39832171]
24. Horstmann S, Schmechel C, Palm K, Oertelt-Prigione S, Bolte G. The operationalisation of sex and gender in quantitative health-related research: a scoping review. Int J Environ Res Public Health 2022 Jun 18;19(12):7493. [doi: 10.3390/ijerph19127493] [Medline: 35742742]
25. Miani C, Wandschneider L, Niemann J, Batram-Zantvoort S, Razum O. Measurement of gender as a social determinant of health in epidemiology-a scoping review. PLoS ONE 2021;16(11):e0259223. [doi: 10.1371/journal.pone.0259223] [Medline: 34731177]
26. Ballering AV, Olde Hartman TC, Rosmalen JGM. Gender scores in epidemiological research: methods, advantages and implications. Lancet Reg Health Eur 2024 Aug;43:100962. [doi: 10.1016/j.lanepe.2024.100962] [Medline: 38989448]
27. McGuire JK, Beek TF, Catalpa JM, Steensma TD. The Genderqueer Identity (GQI) scale: measurement and validation of four distinct subscales with trans and LGBQ clinical and community samples in two countries. Int J Transgend 2019;20(2-3):289-304. [doi: 10.1080/15532739.2018.1460735] [Medline: 32999614]
28. Lombardi E, Banik S. The utility of the two-step gender measure within trans and cis populations. Sex Res Social Policy 2016 Sep;13(3):288-296. [doi: 10.1007/s13178-016-0220-6] [Medline: 34956416]
29. Pohrt A, Kendel F, Demuth I, et al. Differentiating sex and gender among older men and women. Psychosom Med 2022 Apr 1;84(3):339-346. [doi: 10.1097/PSY.0000000000001056] [Medline: 35149636]
30. Tibubos AN, Otten D, Beutel ME, Brähler E. Validation of the Personal Attributes Questionnaire-8: gender expression and mental distress in the German population in 2006 and 2018. Int J Public Health 2022;67:1604510. [doi: 10.3389/ijph.2022.1604510] [Medline: 35370535]

XSL•FO
RenderX

31. Demuth I, Banszerus V, Drewelies J, et al. Cohort profile: follow-up of a Berlin Aging Study II (BASE-II) subsample as part of the GendAge study. BMJ Open 2021 Jun 23;11(6):e045576. [doi: 10.1136/bmjopen-2020-045576] [Medline: 34162642]
32. Spence JT, Helmreich R, Stapp J. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. J Pers Soc Psychol 1975 Jul;32(1):29-39. [doi: 10.1037/h0076857] [Medline: 1206468]
33. de Breij S, Huisman M, Boot CRL, Deeg DJH. Sex and gender differences in depressive symptoms in older workers: the role of working conditions. BMC Public Health 2022 May 21;22(1):1023. [doi: 10.1186/s12889-022-13416-1] [Medline: 35597949]
34. Smith PM, Koehoorn M. Measuring gender when you don't have a gender measure: constructing a gender index using survey data. Int J Equity Health 2016 May 28;15(1):82. [doi: 10.1186/s12939-016-0370-4] [Medline: 27233478]
35. Gisinger T, Azizi Z, Alipour P, et al. Sex and gender aspects in diabetes mellitus: focus on access to health care and cardiovascular outcomes. Front Public Health 2023;11:1090541. [doi: 10.3389/fpubh.2023.1090541] [Medline: 36817907]
36. Cipriani E, Samson-Daoust E, Giguère CE, et al. A step-by-step and data-driven guide to index gender in psychiatry. PLoS ONE 2024;19(1):e0296880. [doi: 10.1371/journal.pone.0296880] [Medline: 38271402]
37. Teterina A, Zulbayar S, Mollayeva T, Chan V, Colantonio A, Escobar M. Gender versus sex in predicting outcomes of traumatic brain injury: a cohort study utilizing large administrative databases. Sci Rep 2023 Oct 27;13(1):18453. [doi: 10.1038/s41598-023-45683-2] [Medline: 37891419]
38. Vader SS, Lewis SM, Verdonk P, Verschuren WMM, Picavet HSJ. Masculine gender affects sex differences in the prevalence of chronic health problems - The Doetinchem Cohort Study. Prev Med Rep 2023 Jun;33:102202. [doi: 10.1016/j.pmedr.2023.102202] [Medline: 37223572]
39. Yuan J, Sang S, Pham J, Kong WJ. Gender modifies the association of cognition with age-related hearing impairment in the Health and Retirement Study. Front Public Health 2021;9:751828. [doi: 10.3389/fpubh.2021.751828] [Medline: 34976918]
40. Ballering AV, Bonvanie IJ, Olde Hartman TC, Monden R, Rosmalen JGM. Gender and sex independently associate with common somatic symptoms and lifetime prevalence of chronic disease. Soc Sci Med 2020 May;253:112968. [doi: 10.1016/j.socscimed.2020.112968]
41. Rosenkrantz P, Bee H, Vogel S, Broverman I. Sex-role stereotypes and self-concepts in college students. J Consult Clin Psychol 1968 Jun;32(3):287-295. [doi: 10.1037/h0025909] [Medline: 5666128]
42. The Medical Informatics Initiative's core data set. Medical Informatics Initiative Germany. URL: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [accessed 2025-03-03]
43. MII core data set - datasets. Art Decor. URL: https://art-decor.org/art-decor/decor-datasets--mide [accessed 2025-03-03]
44. Ludvigsson JF, Almqvist C, Bonamy AKE, et al. Registers of the Swedish total population and their use in medical research. Eur J Epidemiol 2016 Feb;31(2):125-136. [doi: 10.1007/s10654-016-0117-y] [Medline: 26769609]
45. Pedersen CB. The Danish civil registration system. Scand J Public Health 2011 Jul;39(7 Suppl):22-25. [doi: 10.1177/1403494810387965] [Medline: 21775345]
46. Health inequalities data tool. Government of Canada. URL: https://health-infobase.canada.ca/health-inequalities/Indicat [accessed 2025-08-08]
47. HL7/fhir-sdoh-clinicalcare. GitHub. URL: https://github.com/HL7/fhir-sdoh-clinicalcare [accessed 2025-08-08]
48. Horstmann S, Schmechel C, Becher E, Oertelt-Prigione S, Palm K, Bolte G. DIVERGesTOOL – Entwicklung einer Toolbox zur Erfassung geschlechtlicher Vielfalt in der quantitativen Gesundheitsforschung. Bundesgesundheitsbl 2024 Sep;67(9):1054-1061. [doi: 10.1007/s00103-024-03915-4]
49. Banerjee SC, Staley JM, Alexander K, Walters CB, Parker PA. Encouraging patients to disclose their lesbian, gay, bisexual, or transgender (LGBT) status: oncology health care providers' perspectives. Transl Behav Med 2020 Oct 8;10(4):918-927. [doi: 10.1093/tbm/iby105] [Medline: 30476333]
50. Koch A, Rabins M, Messina J, Brennan-Cook J. Exploring the challenges of sexual orientation disclosure among lesbian, gay, bisexual, transgender, queer individuals. J Nurse Pract 2023 Nov;19(10):104765. [doi: 10.1016/j.nurpra.2023.104765]

Abbreviations

BSRI: Bem Sex Role Inventory
CDS: core dataset
MII: Medical Informatics Initiative
PAQ: Personal Attributes Questionnaire

XSL•FO
RenderX

Edited by A Castonguay; submitted 19.Mar.2025; peer-reviewed by B Hoyt, H Anandan; revised version received 24.Aug.2025; accepted 02.Sep.2025; published 08.Jan.2026.

Please cite as:
Schindler L, Beelich H, Katsari E, Liprandi D, Stracke S, Waltemath D
Applicability of Existing Gender Scores for German Clinical Research Data: Scoping Review and Data Mapping
JMIR Med Inform 2026;14:e74162
URL: https://medinform.jmir.org/2026/1/e74162
doi:10.2196/74162

© Lea Schindler, Hilke Beelich, Elpiniki Katsari, Daniele Liprandi, Sylvia Stracke, Dagmar Waltemath. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

AI-Enabled Customer Relationship Management Platforms for Patient Services in Health Care, Early Lessons From Governance, and Program-Level Outcomes

Anup Kant Gupta[1], MBA

Strategic Management & Leadership, Kelley School of Business, Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Anup Kant Gupta, MBA
Strategic Management & Leadership
Kelley School of Business
Indiana University Bloomington
107 S Indiana Ave
Bloomington, IN, 47405
United States
Phone: 1 (812) 855 4848
Email: anupgupta8@gmail.com

Abstract

This research letter summarizes early lessons from 4 enterprise implementations of artificial intelligence–enabled customer relationship management platforms in health care and describes governance practices associated with improvements in affordability, adherence, and access at program level.

(JMIR Med Inform 2026;14:e83564) doi:10.2196/83564

KEYWORDS

artificial intelligence; health care customer relationship management; salesforce; patient affordability; patient adherence; patient access; digital transformation; governance

Introduction

Health systems continue to face pressure to improve patient outcomes while safeguarding affordability and equitable access [1-4]. Electronic health records support clinical workflows but are not designed for proactive patient outreach or service coordination [5]. Artificial intelligence (AI)-enabled customer relationship management (CRM) platforms support automated eligibility checks, outreach, and risk alerts [6-8]. This research letter summarizes early descriptive insights from enterprise implementations of AI-enabled CRM platforms.

Methods

Program Selection and Sampling Frame

A case-informed thematic analysis was conducted across 4 large health care programs. These were the largest CRM-based patient service deployments accessible to the author during this period. Programs were included if they had live AI-enabled CRM workflows for affordability, adherence, or access for at least 12 months. Each program operated independently with different teams, geographies, and product lines. Data reflect operations recorded between January 2019 and March 2024 across the 4 programs. All outcomes reflect a descriptive analysis of the operational dashboards and do not infer causal relationships.

Data Sources and Definitions

The following three data sources were used, each aggregated at program level.

1. Governance documents summarizing decision logs, compliance checkpoints, and escalation patterns.
2. Stakeholder feedback was gathered through routine program review meetings and documented in standard internal templates.
3. Internal dashboards tracking key operational indicators. These dashboards were deidentified, contained no patient level data, and were part of routine program monitoring.

Program level outcome metrics were defined as follows.

1. Adoption: proportion of active CRM users among provisioned patient service staff within a measurement window.
2. Affordability efficiency: average time for co-pay or financial assistance verification.

XSL•FO
RenderX

3. Therapy initiation time: average days between benefit verification and therapy start.
4. Discontinuation rate: proportion of enrolled patients who stopped therapy during a defined period.

Baseline values came from preimplementation operations, and follow-up reflected the first stable post launch period. Findings are observational.

Ethical Considerations

This work used only aggregated, deidentified operational dashboards and program documents. No patient-level identifiable data or human subject interaction occurred. Institutional review board approval was not required.

Results

Theme 1: Patient Outcome Alignment and Sponsorship

Programs that centered goals on patient experience and outcomes such as therapy initiation time or affordability enrollment sustained stronger executive sponsorship. Programs framed as IT upgrades struggled to maintain alignment.

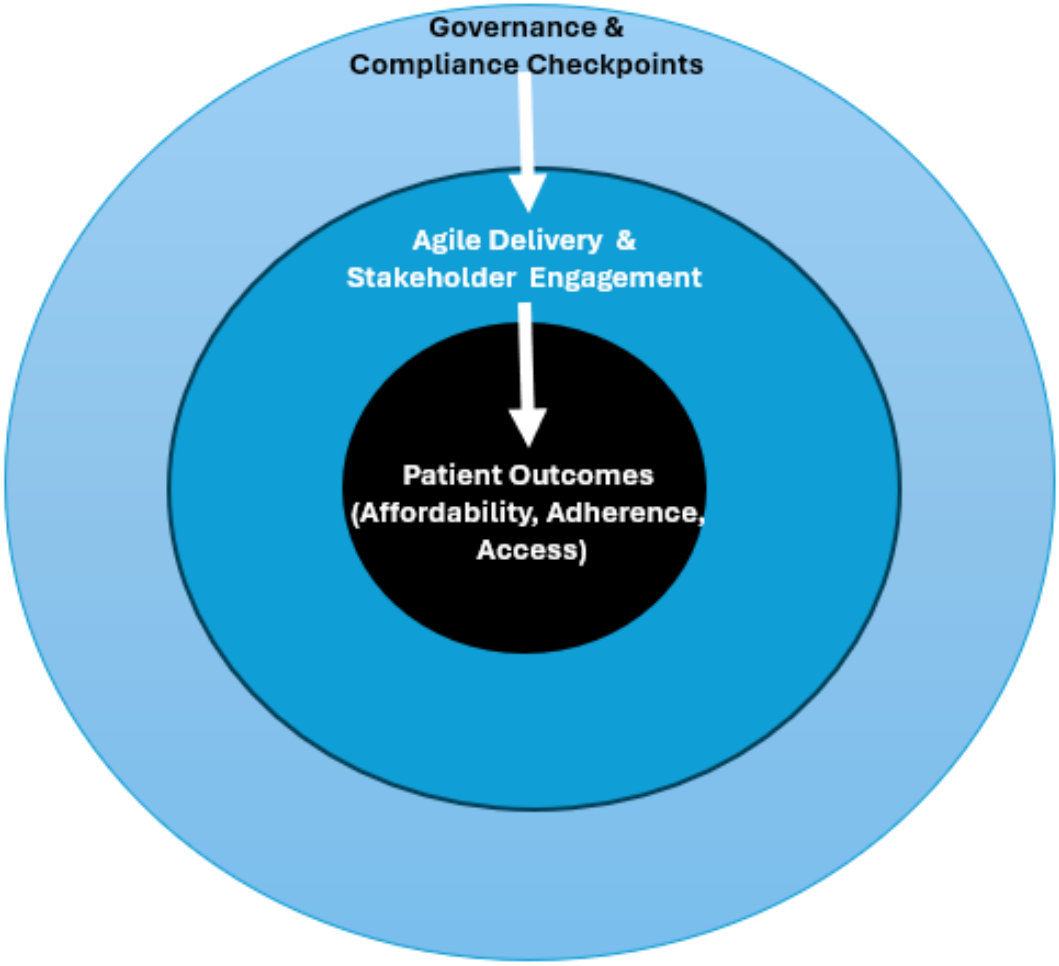
Theme 2: Continuous Engagement and Adoption

High adoption correlated with structured engagement of patient service teams and compliance officers. Regular feedback cycles supported adoption above 85% vs below 60% in minimally engaged programs [2].

Theme 3: Hybrid Governance Improved Delivery Efficiency

All 4 programs used a hybrid model that combined Agile sprints with scheduled compliance reviews. This approach reduced backlog resolution time by about 30%. Figure 1 summarizes the governance structure.

Figure 1. Hybrid governance framework for artificial intelligence–enabled customer relationship management. The framework illustrates three concentric layers: (1) patient outcomes (affordability, adherence, access) at the core, (2) agile delivery and stakeholder engagement in the middle layer, and (3) governance and compliance checkpoints in the outer layer. Arrows depict how governance structures enable engagement, which in turn drives improved patient outcomes.



XSL•FO
RenderX

Theme 4: Program-Level Improvements in Affordability, Adherence, and Access

Changes reflect aggregated dashboards and are not causal. Examples include a reduction in co-pay verification time from 4 days to 3 days (25% faster), a 12% reduction in therapy discontinuation rates, and a shortening of therapy initiation time from 20 days to 17 days (15% faster) (Table 1).

Table 1. Program level customer relationship management–enabled improvements.

Patient service	Artificial intelligence functionality	Outcome	Example (anonymized)
Affordability	Automated co-pay verification	Reduced processing time (4 days to 3 days, a 25% improvement)	Fortune 10 rollout (USA)
Adherence	Predictive risk alerts	12% reduction in therapy discontinuation	Consulting-led program
Access	Automated prior authorization	Therapy initiation shortened (20 days to 17 days, a 15% improvement)	Enterprise-wide program

Discussion

Summary of Findings

Across 4 large-scale implementations, AI-enabled CRM platforms supported improvements in affordability, adherence, and access through workflow automation, risk identification, and coordinated service tasks. The most consistent predictors of success were early alignment with patient-centered outcomes, continuous stakeholder engagement, and hybrid governance structures.

Interpretation and Limitations

Results reflect aggregated, unaudited operational dashboards and cannot establish causality [9,10]. Findings describe program-level improvements observed post implementation and may not generalize to organizations with different scales, regulatory environments, or infrastructure maturity.

Future Directions

Future studies should incorporate multi-site comparative designs, validated outcomes, and integration with digital therapeutics and device streams to strengthen evidence for CRM-enabled patient service transformation.

Conflicts of Interest

None declared.

References

1. Digital Health and Innovation. Global Strategy on Digital Health 2020–2025. Geneva, Switzerland: World Health Organization; 2025.
2. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption, a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
3. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019 Mar 20;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
5. Adler-Milstein J, Holmgren A, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
6. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
7. Obermeyer Z, Emanuel EJ. Predicting the future, big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219. [doi: 10.1056/nejmp1606181]
8. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
9. Wang Y, Kung L, Wang W, Cegielski C. An integrated big data analytics-enabled transformation model: Application to health care. Information & Management 2018 Jan;55(1):64-79. [doi: 10.1016/j.im.2017.04.001]
10. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol 2016 Jun;13(6):350-359. [doi: 10.1038/nrcardio.2016.42] [Medline: 27009423]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

CRM: customer relationship management

Edited by A Benis; submitted 05.Sep.2025; peer-reviewed by E Oluwagbade, A Okoye; comments to author 22.Oct.2025; revised version received 18.Nov.2025; accepted 01.Jan.2026; published 02.Feb.2026.

Please cite as:
Gupta AK
AI-Enabled Customer Relationship Management Platforms for Patient Services in Health Care, Early Lessons From Governance, and Program-Level Outcomes
JMIR Med Inform 2026;14:e83564
URL: https://medinform.jmir.org/2026/1/e83564
doi:10.2196/83564
PMID:41481793

©Anup Kant Gupta. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of AI Chatbot Responses to a Standardized Patient Query on Myelin Oligodendrocyte Glycoprotein Antibody–Associated Disease: Cross-Sectional Content Analysis

Meryem Tuba Sönmez[1], MD; Mehmet Fatih Yetkin[2], MD, Prof Dr; Duygu Arslan Mehdiyev[1], MD; Nazlı Durmaz Çelik[3], MD, Prof Dr; Merve Bahar Ercan[4], MD; Pınar Öztürk[5], MD; Yeşim Eylev Akboğa[6], MD; Emine Rabia Koç[7], MD, Prof Dr; Semra Mungan[8], MD, Prof Dr

[1]Neurology Clinic, Ankara Etlik City Hospital, Ankara, Türkiye

[2]Department of Neurology, Faculty of Medicine, Erciyes University, Gevher Nesibe Hospital, Prof. Dr. Turhan Feyzioğlu Street No. 42, Kayseri, Melikgazi, Türkiye

[3]Department of Neurology, Eskişehir Osmangazi University, Eskişehir, Türkiye

[4]Department of Neurology, Gazi University, Ankara, Türkiye

[5]Neurology Clinic, Ankara Yıldırım Beyazıt University Yenimahalle Training and Research Hospital, Ankara, Türkiye

[6]Neurology Clinic, Sincan Research and Training Hospital, Ankara, Türkiye

[7]Department of Neurology, Bursa Uludag University, Bursa, Türkiye

[8]Department of Neurology, University of Health Sciences Ankara City Hospital, Ankara, Türkiye

Corresponding Author:
Mehmet Fatih Yetkin, MD, Prof Dr
Department of Neurology, Faculty of Medicine, Erciyes University, Gevher Nesibe Hospital, Prof. Dr. Turhan Feyzioğlu Street No. 42, Kayseri, Melikgazi, Türkiye

Abstract

Background: Large language model–based chatbots are increasingly used by the public to access medical information. Although these tools can improve access and convenience, their quality, clarity, and transparency remain uncertain for rare and diagnostically complex neurological conditions, such as myelin oligodendrocyte glycoprotein antibody–associated disease (MOGAD).

Objective: This study aimed to evaluate the scientific quality, understandability, citation transparency, and readability of responses generated by widely used artificial intelligence chatbot platforms to a standardized, patient-centered query on MOGAD.

Methods: We conducted a cross-sectional content analysis using the query, "What is MOGAD, and how is MOGAD treated?" Ten widely accessible chatbot platforms were queried once on the same day in new sessions. Responses were anonymized and independently evaluated by 7 blinded neurologists using DISCERN (treatment-related scientific quality), Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), and the Web Resource Rating (WRR; citation transparency). Readability was assessed using the Flesch-Kincaid Grade Level (FKGL) and Coleman-Liau Index, and word count was recorded. Platforms were compared by functional orientation and the access model. Mann-Whitney *U* and Kruskal-Wallis tests with Dunn post hoc tests were used. Interrater reliability was assessed using intraclass correlation coefficients.

Results: Significant differences were observed across platforms for DISCERN, PEMAT-P, and WRR scores (all $P<.001$). Search-focused platforms achieved higher understandability than conversation-focused platforms (median PEMAT-P 52.6, IQR 47.4-54 vs 46.7, IQR 42-47.3; $P=.04$), whereas conversation-focused platforms had higher WRR scores (median 26.8, IQR 19.6-26.8 vs 19.6, IQR 19.6-25.9; $P=.001$). DISCERN scores did not differ significantly by functional orientation ($P=.11$). Paid-access platforms outperformed free-access platforms in DISCERN (median 42, IQR 36-45 vs 33, IQR 23.8-41.3; $P<.001$), PEMAT-P (median 52.6, IQR 46-54 vs 46, IQR 26.3-47.4; $P=.002$), and WRR (median 26.8, IQR 23.2-26.8 vs 10.7, IQR 3.57-19.6; $P<.001$). However, no statistically significant differences were observed between paid and free platforms in response length (median word count 336, IQR 271-369 vs 206, IQR 116-294; $P=.11$) or readability metrics. FKGL scores were comparable between paid and free outputs (median 17.54, IQR 16.6-18.4 vs 17.56, IQR 16.5-17.6; $P=.61$), and Coleman-Liau Index values similarly showed no significant difference by access model (median 21.30, IQR 20.6-22.3 vs 21.71, IQR 20.9-22.1; $P=.91$). Readability remained limited: all outputs exceeded recommended public health readability thresholds (FKGL≥8). High interrater agreement was observed (intraclass correlation coefficient=0.902 for DISCERN, 0.887 for WRR, and 0.838 for PEMAT-P).

Conclusions: Artificial intelligence chatbot responses to a patient-centered MOGAD query varied substantially in scientific quality, understandability, transparency, and readability. Search-focused systems were more understandable, whereas conversation-focused systems showed greater citation transparency. Paid-access platforms achieved higher quality and transparency

XSL•FO
RenderX

scores, without differences in readability or response length. All outputs exceeded recommended public health readability thresholds. These findings highlight the need for context-sensitive evaluation of chatbot outputs in rare and clinically complex conditions such as MOGAD.

(JMIR Med Inform 2026;14:e81720) doi:10.2196/81720

KEYWORDS

artificial intelligence; large language models; chatbots; patient education; myelin oligodendrocyte glycoprotein antibody–associated disease; readability; health information quality; citation transparency

Introduction

Artificial intelligence (AI) is rapidly reshaping health care, influencing how patients access, interpret, and interact with medical information. One of the most visible applications of this technology is the rise of large language models (LLMs), now widely available through chatbot interfaces. These systems can generate fluent, natural-sounding responses to health-related questions, offering the promise of more accessible and personalized communication. However, the consistency, accuracy, and clarity of their responses remain uncertain, especially when the topic involves rare or diagnostically challenging conditions [1-4].

Myelin oligodendrocyte glycoprotein antibody–associated disease (MOGAD) is one such condition. It is a rare, immune-mediated disorder of the central nervous system that can present in a variety of ways, including optic neuritis, transverse myelitis, and acute disseminated encephalomyelitis [5]. The publication of international diagnostic criteria in 2023 has improved the ability to distinguish MOGAD from related disorders such as multiple sclerosis and aquaporin-4 antibody-positive neuromyelitis optica spectrum disorder [6]. Despite this progress, clinical awareness of MOGAD remains limited, and many patients and caregivers turn to online resources in search of reliable information [7,8].

In this context, AI-powered chatbots have emerged as promising tools for bridging knowledge gaps, especially for people affected by rare diseases. These platforms can offer quick, conversational responses to health-related questions, simulating interaction with a knowledgeable guide. Yet most prior studies evaluating chatbot performance have focused on common medical conditions [9,10]. Very few have investigated how these tools respond to queries about complex and less-recognized diseases like MOGAD, and even fewer have used multidimensional, validated instruments to assess the quality, clarity, and transparency of the information provided [11,12]. Furthermore, studies involving blinded expert evaluations and structured comparisons across different chatbot platforms remain limited [13,14].

To address these gaps, this study systematically evaluated the responses of ten widely used AI chatbot platforms to a standardized, patient-centered query regarding MOGAD. Using established instruments, we assessed the quality, comprehensibility, citation transparency, and readability of each response. Chatbot performance was also compared based on access model (free or paid) and functional design (chat-based or search-based). All responses were independently evaluated by experienced neurologists who were blinded to the source of each answer. This study aims to provide insight into the reliability and educational value of chatbot-generated medical content in the context of rare diseases and to inform best practices for the safe and effective use of AI in digital health communication.

Methods

Study Design, Query Selection, and Chatbot Sampling

This cross-sectional content analysis was designed to evaluate the quality, readability, comprehensibility, and source transparency of medical information generated by AI chatbot platforms in response to a standardized, patient-centered query related to MOGAD. Only publicly accessible chatbot outputs were used, and no human participants, personal health information, or clinical interactions were involved.

To identify the most relevant public-facing search terms related to MOGAD, a 5-year global search trend analysis (from May 20, 2020, to May 20, 2025) was conducted using Google Trends. The term “MOGAD” exhibited a marked increase in popularity, reflecting growing public awareness. Among the top ten co-occurring search terms, those most directly related to the disease included: “mogad disease,” “mog,” “what is mogad,” “mogad treatment,” “mogad symptoms,” and “mogad radiology.” Other common terms such as “nmosd,” “ms,” “nmo,” and “adem” were excluded because they were not specific to MOGAD.

Based on these findings, a standardized, patient-oriented question was formulated to reflect a realistic information-seeking behavior of individuals affected by MOGAD. The final query selected for evaluation was: “What is MOGAD, and how is MOGAD treated?”

This question was submitted on May 24, 2025, to ten widely accessible AI chatbot platforms. These platforms were purposefully chosen to represent diversity in model architecture, developer affiliation, and access modality (ie, free vs paid). The chatbot platforms evaluated included Copilot (Microsoft), Claude 3.5 and Claude 3.7 (Anthropic), Perplexity Pro (Perplexity AI), Grok 3 (xAI), Gemini Advanced and Gemini 2.5 Pro (Google), ChatGPT-4 and ChatGPT-4.5 (OpenAI), and DeepSeek (DeepSeek AI).

For consistency and ecological validity, each chatbot was queried only once in a newly initiated session. No follow-up prompts, rephrased questions, or clarifications were used. During data collection, all internet-browsing and retrieval features embedded within chatbot interfaces (eg, “Browse with Bing”

XSL•FO
RenderX

or comparable web access modes) were disabled. Responses were generated in newly initiated sessions using a single standardized prompt, without follow-up queries, external link clicks, or activation of plugins or tools, and the first unedited output was captured verbatim and included in the evaluation dataset.

Data Collection and Blinding Procedures

The study team consisted of 9 individuals. Two experienced neurologists (MTS and MFY) were responsible for the study design. MTS conducted a standardized training session outlining the study rationale and the evaluation instruments, while MFY remained blinded throughout the assessment phase and served as one of the independent raters.

NDÇ, who was not involved in the study design or evaluation, submitted all chatbot queries and anonymized the resulting outputs by assigning random numeric codes (1-10). These coded responses were then distributed to the evaluators without revealing the originating platforms.

Seven neurologists (MTS, MFY, DAM, MBE, PÖ, YEA, and SM), all blinded to chatbot identities, independently evaluated the anonymized responses. This group included MFY and 6 additional experts, each with at least 5 years of experience in diagnosing and managing demyelinating disorders. MTS and NDÇ were not blinded because they were directly responsible for preparing, coding, and distributing the response dataset. For transparency and reproducibility, chatbot platform names are shown in Multimedia Appendix 1; however, these identifiers were removed from the materials provided to blinded expert raters during the scoring process.

Evaluation Instruments

Each chatbot response was evaluated in its original, fully formatted, and verbatim form as generated at the time of data capture. DISCERN, Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), and WRR scoring were conducted exclusively on these unedited outputs. Validated instruments and standardized measures were applied across four domains:

1. Scientific quality: DISCERN Tool, a 16-item instrument with a 5-point Likert scale (range: 16 - 80), used to evaluate clarity, risk-benefit balance, citation of evidence, and overall reliability of treatment-related information. Higher scores indicate higher quality [15].
2. Comprehensibility: PEMAT-P (understandability domain) developed by the U.S. Agency for Healthcare Research and Quality (AHRQ), this tool measures the clarity of language, sentence structure, layout, and logical flow of information. Scores are expressed as percentages. Although PEMAT-P items are scored binarily, the overall percentage score derived from the 7 expert ratings was used as a continuous variable for the calculation of interrater reliability (intraclass correlation coefficient [ICC]) [16].
3. Source transparency: Web Resource Rating (WRR) assesses citation practices, author disclosure, source traceability, and update recency. Final scores are expressed as percentages [17]. In this study, WRR scoring was applied to the response content as presented to the user at the time of data capture. References were evaluated based on their visibility to the reader, accessibility, and explicit association with specific statements in the response text. Both hyperlink-based references and text-based citations were eligible for scoring. Hyperlinks were recorded as accessible references, whereas text-based citations were evaluated according to whether the source could be clearly identified and traced by the reader.
4. Readability: Coleman-Liau Index (CLI) and Flesch-Kincaid Grade Level (FKGL). CLI is based on character and sentence length [18]; FKGL considers word syllables and sentence structure [19]. Both metrics estimate the US grade level required to understand the text. Scores <8 indicate content appropriate for general public health communication [12]. Word counts were also recorded. For readability analysis only, URL strings, embedded hyperlinks, and reference lists were computationally removed prior to calculating word count, CLI, and FKGL in order to prevent artificial inflation of character and word counts. This preprocessing step was applied solely for quantitative readability assessment and did not affect expert scoring or the evaluation dataset.

Instrument Licensing and Accessibility

All instruments used, DISCERN, PEMAT-P, CLI, FKGL, and WRR, are freely accessible, open-source tools. DISCERN and PEMAT-P are provided by Oxford University and AHRQ, respectively. CLI and FKGL use publicly available formulas, and WRR is distributed under a Creative Commons license. This ensures transparency, reproducibility, and ethical adherence in the methodology.

Chatbot Categorization by Function and Access

For subgroup-level comparisons, chatbot platforms were classified according to 2 independent parameters: functional orientation and access model.

Functional orientation was defined based on the primary design characteristics of the chatbots. Platforms emphasizing natural dialogue, contextual flow, and conversational interaction were categorized as conversation-focused and included ChatGPT-4, ChatGPT-4.5, Claude 3.5, Claude 3.7, and Grok 3. In contrast, platforms optimized for information retrieval, structured responses, and search-enhanced outputs were designated as search-focused, comprising Copilot, DeepSeek, Gemini Advanced, Gemini 2.5 Pro, and Perplexity Pro.

The access model was defined based on the user entry cost at the time of data collection. Chatbots available without subscription or payment were categorized as free-access platforms: ChatGPT-4, Claude 3.5, Copilot, and DeepSeek. Platforms requiring a subscription or premium tier to access advanced versions were classified as paid-access, including ChatGPT-4.5, Claude 3.7, Gemini Advanced, Gemini 2.5 Pro, Perplexity Pro, and Grok 3.

Qualitative Assessment of Content Depth

In addition to quantitative scoring, the 7 expert raters provided structured qualitative notes regarding specific content gaps (eg, absence of specific drug classes and lack of clinical nuance) for

XSL•FO
RenderX

each response. These observations were aggregated to identify recurring thematic deficiencies across platforms.

Statistical Analysis

Descriptive statistics were used to summarize chatbot performance across all predefined evaluation domains. For each AI platform, median (IQR) values and full ranges were calculated for 3 primary quality metrics: DISCERN (evidence-based treatment information), PEMAT-P (understandability), and WRR (reference transparency). Additionally, readability and response length were assessed using CLI, FKGL, and total word count.

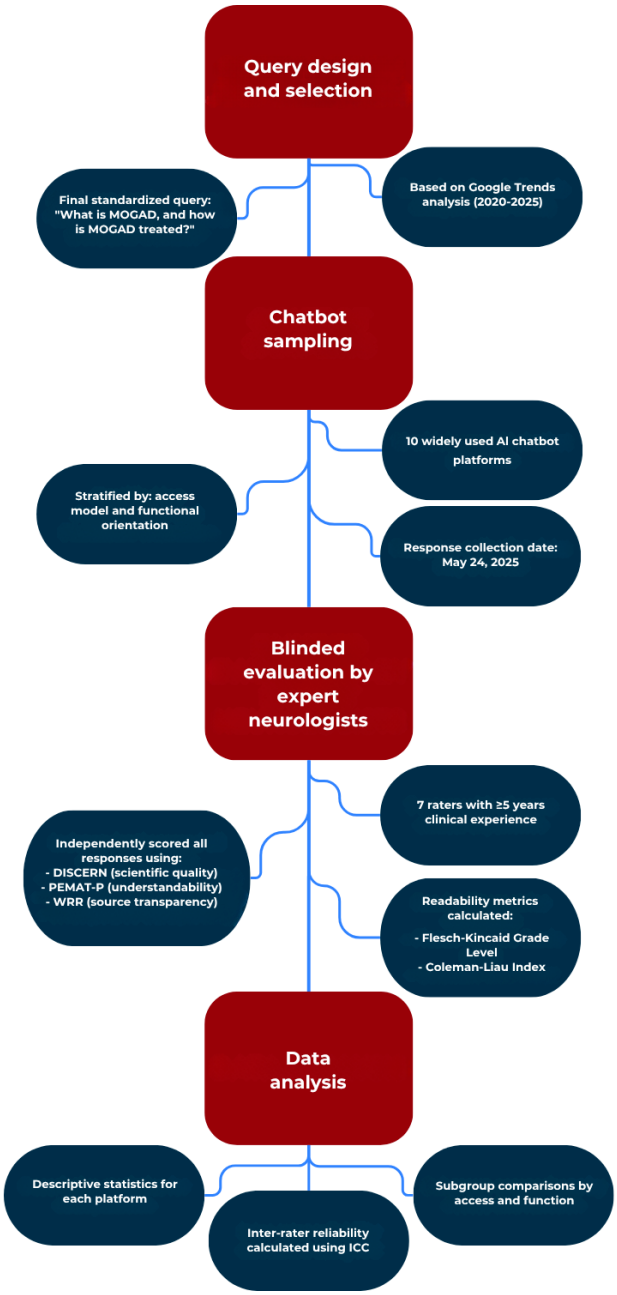
Chatbots were further stratified by functional orientation (conversation-focused vs search-focused) and access model (free vs paid). To compare performance across these subgroups, nonparametric inferential statistics were applied. Specifically, the Mann-Whitney *U* test was used to compare median scores for DISCERN, PEMAT-P, and WRR between the 2 subgroups in each classification. Given the small subgroup sample sizes (n=5 per group), exact 2-sided tests were performed to ensure accurate *P* value estimation. Differences among multiple chatbot platforms were assessed using the Kruskal-Wallis test. For statistically significant Kruskal-Wallis results, pairwise comparisons were performed using the Dunn post hoc test with Bonferroni correction to control for the family-wise error rate. A *P* value less than .05 was considered indicative of statistical significance for all comparisons.

To assess interrater reliability, ICCs were computed for the 3 primary metrics. DISCERN and WRR scores were treated as continuous variables. Although PEMAT-P items are scored binarily, the overall percentage score derived from the 7 expert ratings was used as a continuous variable for the calculation of ICC, using a 2-way random-effects model with absolute agreement (ICC [1,2]). ICC values were interpreted according to conventional thresholds: values less than 0.50 indicate poor reliability, 0.50 - 0.75 indicate moderate reliability, 0.75 - 0.90 indicate good reliability, and values equal to or greater than 0.90 indicate excellent reliability [20].

All statistical analyses were conducted using R software (version 4.4.3; R Core Team, R Foundation for Statistical Computing) and IBM SPSS Statistics (version 25.0). The study workflow is illustrated in Figure 1, which outlines the sequential methodology involving standardized query formulation, chatbot selection, anonymized response collection, blinded expert evaluation, and statistical analysis.

XSL•FO
RenderX

Figure 1. Stepwise workflow of the study evaluating artificial intelligence chatbot responses to a standardized myelin oligodendrocyte glycoprotein antibody–associated disease query. The process includes query design, chatbot selection, anonymized response collection, expert evaluation using validated tools, readability assessment, and statistical analysis. AI: artificial intelligence; ICC: intraclass correlation coefficient; MOGAD: myelin oligodendrocyte glycoprotein antibody–associated disease; PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; WRR: Web Resource Rating.



Ethical Considerations

This study did not involve human participants, patient data, biological specimens, or any clinical intervention. All analyzed material consisted exclusively of publicly available, nonidentifiable text generated by AI chatbot platforms in response to a standardized informational query. The neurologists who served as blinded expert raters were not research participants, and no personal or identifiable information about them was collected, stored, or analyzed. Therefore, institutional ethics committee review and informed consent were not required. This determination is consistent with internationally recognized principles for research ethics, including the World Medical Association's Declaration of Helsinki. In addition, under national research ethics regulations in Türkiye, ethical review is required primarily for studies involving identifiable human participants, human data, or clinical interventions; therefore, this study does not fall within the scope requiring institutional ethics approval [21].

Results

Overview of Chatbot Performance

Descriptive statistics for chatbot performance across key quality, readability, and transparency indicators are summarized in Table 1. Statistically significant differences were observed among platforms for DISCERN, PEMAT-P, and WRR scores (all $P<.001$, Kruskal-Wallis test).

XSL•FO
RenderX

Table . Descriptive statistics of chatbot responses regarding myelin oligodendrocyte glycoprotein antibody–associated disease information quality, readability, and transparency[a].

Metric	ChatGPT-4	ChatGPT-4.5	Claude 3.7	Claude 3.5	Copilot	DeepSeek	Gemini 2.5 pro	Gemini advanced	Grok 3	Perplexity pro	*P* value
DISCERN[b], median (IQR)	38 (38-41)	36 (33-38)	36 (34-36)	24 (23.5-29)	23 (20-24.5)	48 (43.5-52)	45 (43-45.5)	44 (43-44.5)	50 (48-52)	40 (39-44)	<.001
PEMAT-P[c], median (IQR)	47.3 (46.7-47.3)	46 (46-47.3)	46 (42-50.7)	33 (29.7-33)	17 (17-21.7)	52.6 (50-54)	54 (53.3-58.5)	63 (57.8-63)	46 (46-50)	52.6 (52.6-53.3)	<.001
WRR[d], median (IQR)	23.2 (19.6-26.8)	23.2 (21.5-26.8)	26.8 (26.8-26.8)	10.7 (3.6-10.7)	2.6 (3.6-3.6)	19.6 (19.6-19.6)	23.2 (19.6-25)	23.2 (21.4-23.2)	32.1 (26.8-32.1)	26.8 (26.8-28.5)	<.001
Word count	293	260	210	110	118	295	369	419	369	303	—[e]
CLI[f]	21.42	21.10	23.16	22.41	19.22	22.00	22.53	21.49	20.23	20.48	—
FKGL[g]	17.62	16.90	18.59	17.50	13.47	17.63	18.17	18.54	16.54	16.55	—

[a]Values are presented as median (IQR) for DISCERN, Patient Education Materials Assessment Tool for Printable Materials, and Web Resource Rating, and as single values for readability indices and word count. Chatbot evaluation comparisons for applicable variables across independent groups were performed using the Kruskal-Wallis test.

[b]DISCERN indicates the quality of treatment-related information.

[c]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[d]WRR: Web Resource Rating.

[e]Not applicable.

[f]CLI: Coleman-Liau Index.

[g]FKGL: Flesch-Kincaid Grade Level.

DISCERN scores, reflecting the quality of treatment-related information, demonstrated notable variability across chatbots. Grok 3 exhibited the highest median DISCERN score (median 50, IQR 48-52), followed by DeepSeek (median 48, IQR 43.5-52) and Gemini 2.5 Pro (median 45, IQR 43-45.5). Intermediate median scores were observed for ChatGPT-4 (median 38, IQR 38.0-41), ChatGPT-4.5 (median 36, IQR 33-38), and Claude 3.7 (median 36, IQR 34-36). In contrast, lower median DISCERN scores were recorded for Claude 3.5 (median 24, IQR 23.5-29) and Copilot (median 23, IQR 20-24.5).

Understandability, as assessed using the PEMAT-P, varied across chatbots, with median scores ranging from 17 (IQR 17-21.7) for Copilot to 63 (IQR 57.8-63) for Gemini Advanced. Among free-access platforms, Copilot exhibited the lowest median PEMAT-P score (median 17, IQR 17-21.7), followed by Claude 3.5 (median 33, IQR 29.7-33).

WRR scores, reflecting citation transparency, varied across chatbots. The highest median WRR values were observed for Grok 3 (median 32.1, IQR 26.8-32.1), Claude 3.7 (median 26.8, IQR 26.8-26.8), and Perplexity Pro (median 26.8, IQR 26.8-28.5). In contrast, lower median WRR scores were recorded for Copilot (median 3.6, IQR 3.6-3.6) and Claude 3.5 (median 10.7, IQR 3.6-10.7).

In terms of readability, CLI values ranged from 19.22 (Copilot) to 23.16 (Claude 3.7), while FKGL scores varied from 13.47 (Copilot) to 18.59 (Claude 3.7). All FKGL values exceeded the recommended Grade 8 threshold for public health communication. Word count ranged from 110 words (Claude 3.5) to 419 words (Gemini Advanced).

A visual summary of each chatbot's multidimensional performance across 6 metrics is provided in Figure 2, illustrating differences in content quality, transparency, readability, and response length.

Figure 3 displays pairwise comparisons using the Dunn post hoc test following Kruskal-Wallis analysis, further emphasizing the performance variability among platforms.

The comparative readability assessment of responses generated across the 10 evaluated AI chatbot platforms is graphically presented in Figure 4.

XSL•FO
RenderX

Figure 2. Multidimensional performance profiles of ten artificial intelligence chatbot platforms in response to a standardized myelin oligodendrocyte glycoprotein antibody–associated disease query. Radar plots illustrate chatbot performance across six key metrics: DISCERN (scientific quality), Patient Education Materials Assessment Tool for Printable Materials (understandability), Web Resource Rating (source transparency), word count (response length), Flesch-Kincaid Grade Level, and Coleman-Liau Index (readability). The top row shows interrater variability for expert-rated metrics (DISCERN, Patient Education Materials Assessment Tool for Printable Materials, and Web Resource Rating), while the bottom row presents readability and length scores derived from single-response outputs. PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; WRR: Web Resource Rating.

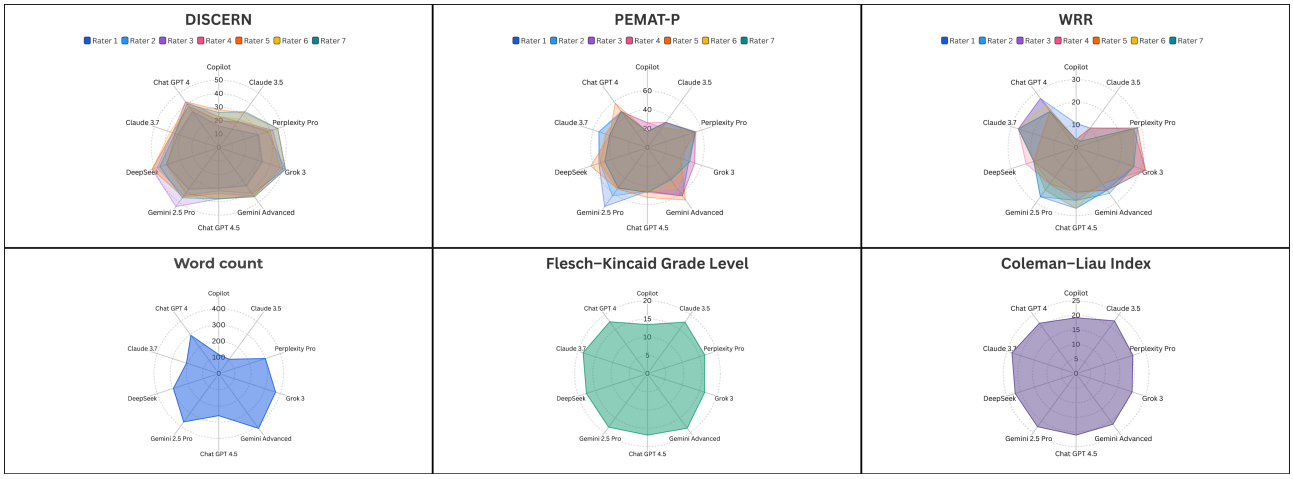
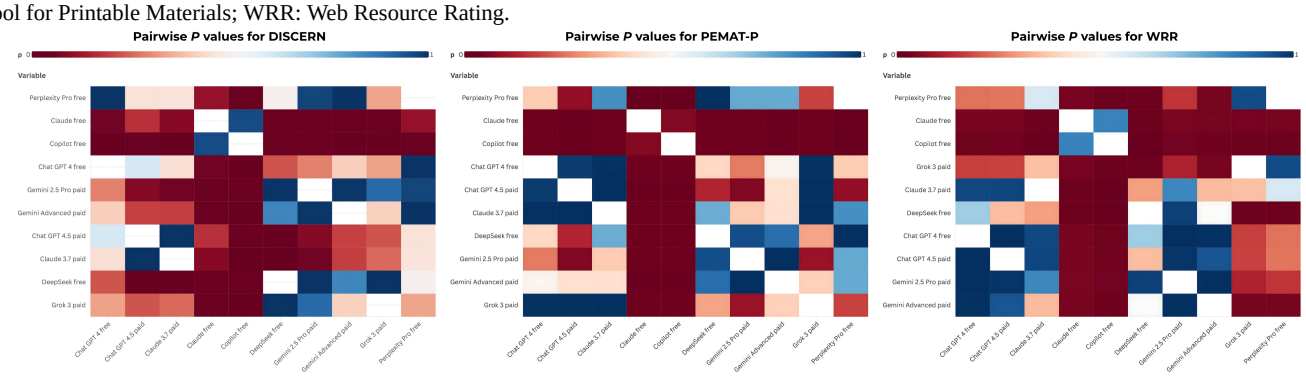
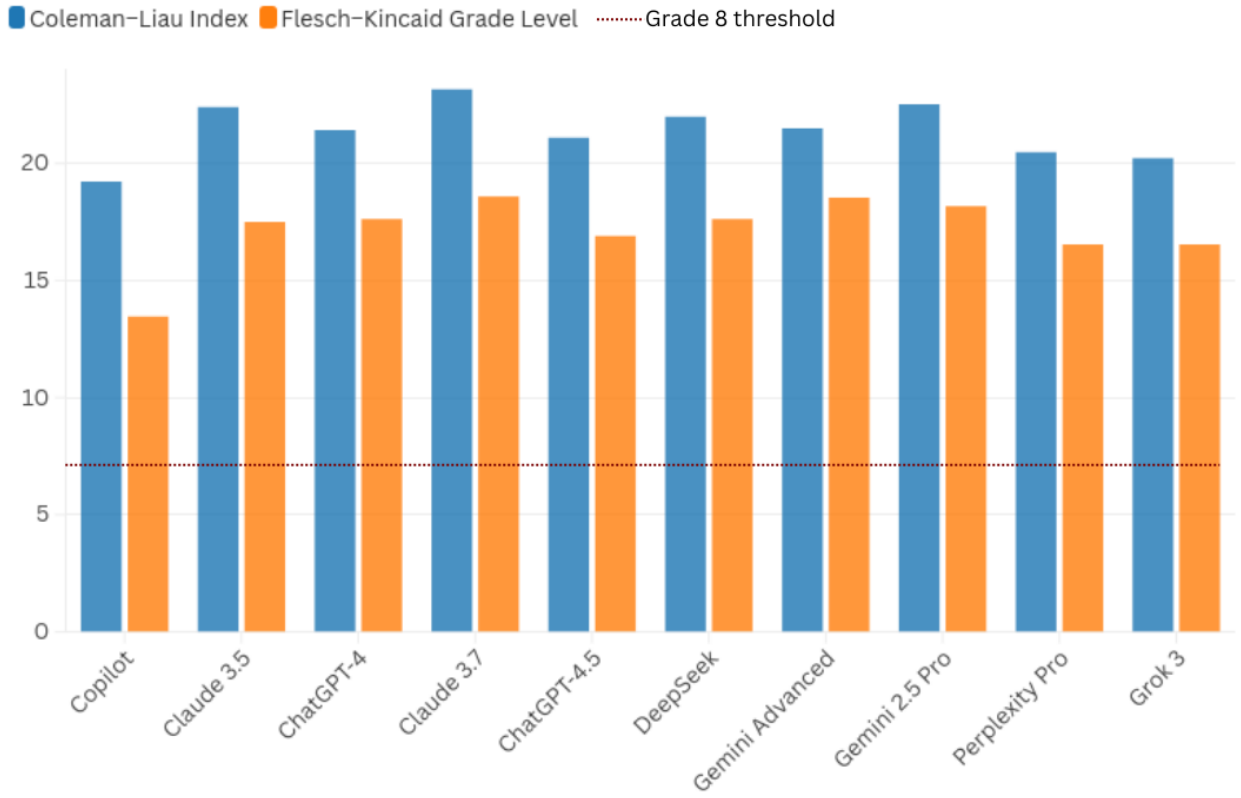


Figure 3. Pairwise statistical comparisons of ten artificial intelligence chatbot platforms across three expert-rated performance metrics (DISCERN, Patient Education Materials Assessment Tool for Printable Materials, and Web Resource Rating). This figure presents three separate heatmaps illustrating the results of the Dunn post hoc test with Bonferroni correction, which was applied following the overall significant Kruskal-Wallis test results (P<.001) across the ten evaluated chatbot platforms (n=10 platforms, assessed by seven independent expert neurologists). The heatmaps compare all possible pairs of the ten platforms based on their median scores for DISCERN (scientific quality of treatment information), Patient Education Materials Assessment Tool for Printable Materials (understandability), and Web Resource Rating (source transparency). The X-axis and Y-axis represent the ten unique artificial intelligence chatbot platforms compared in this study. The color intensity in each cell represents the adjusted P value of the pairwise comparison. Deep red tones (higher P values) indicate no statistically significant difference between the two compared platforms (P>.05), while deep blue tones (lower P values) denote a statistically significant difference in performance (P<.05). These heatmaps are essential for identifying the specific platform pairings responsible for the high overall performance variability observed in the expert-rated metrics. PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; WRR: Web Resource Rating.



XSL•FO
RenderX

Figure 4. Combined readability scores of chatbot responses based on Coleman-Liau Index and Flesch-Kincaid Grade Level. Comparative readability scores for the ten artificial intelligence platforms using Coleman-Liau Index (blue bars) and Flesch-Kincaid Grade Level (orange bars). Scores are plotted against the recommended maximum Grade 8 threshold (red dashed line) for public health communication. CLI: Coleman-Liau Index; FKGL: Flesch-Kincaid Grade Level.



Comparison Between Conversation-Focused and Search-Focused Chatbots

Performance metrics stratified by chatbot functionality are summarized in Table 2. Although search-focused models showed higher median DISCERN scores (median 44, IQR 36.5-45.5) than conversation-focused models (median 38, IQR 32.5-39), this difference did not reach statistical significance (*P*=.11).

Table . Comparison of quality, readability, and transparency metrics by functionality[a].

Metric	Conversation-focused, median (IQR)	Search-focused, median (IQR)	*P* value
DISCERN	38 (32.5-39)	44 (36.5-45.5)	.11
PEMAT-P[b]	46.7 (42-47.3)	52.6 (47.4-54)	.04
WRR[c]	26.8 (19.6-26.8)	19.6 (19.6-25.9)	.001
Word count	260 (210-293)	303 (295-369)	.21
CLI[d]	21.42 (21.10-22.41)	21.49 (20.48-22)	.75
FKGL[e]	17.50 (16.90-17.62)	17.63 (16.55-18.17)	.92

[a]Values are presented as median (IQR). DISCERN refers to the quality of treatment-related information. Patient Education Materials Assessment Tool for Printable Materials indicates understandability. Web Resource Rating reflects transparency and citation of external sources. The Coleman-Liau Index measures textual complexity; higher scores indicate more complex language. Flesch-Kincaid Grade Level represents the U.S. school grade level required to comprehend the text; higher scores denote reduced accessibility for the general public. Conversation-focused chatbots: ChatGPT-4, ChatGPT-4.5, Claude 3.7, Claude 3.5, Grok 3. Search-focused chatbots: Gemini Advanced, Gemini 2.5 Pro, Perplexity Pro, Copilot, DeepSeek. Group comparisons were performed using the Mann-Whitney *U* test.

[b]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[c]WRR: Web Resource Rating.

[d]CLI: Coleman-Liau Index.

[e]FKGL: Flesch-Kincaid Grade Level.

XSL•FO
RenderX

Search-focused platforms also showed higher median PEMAT-P scores compared with conversation-focused platforms (median 52.6, IQR 47.4-54 vs 46.7, IQR 42.0-47.3). Although this difference reached statistical significance (*P*=.04), the small subgroup sample size (n=5 per group) warrants cautious interpretation. In contrast, WRR scores were higher among conversation-focused platforms than search-focused platforms (median 26.8, IQR 19.6-26.8 vs 19.6, IQR 19.6-25.9; *P*=.001).

Median word count was greater among search-focused platforms (median 303, IQR 295-369) than conversation-focused platforms (median 260, IQR 210-293), but this difference was not statistically significant (*P*=.21).

No statistically significant differences were observed between functional orientations for readability metrics, including CLI (median 21.49, IQR 20.48-22 vs 21.42, IQR 21.10-22.41; *P*=.75) and Flesch-Kincaid Grade Level (median 17.63, IQR 16.55-18.17 vs 17.50, IQR 16.90-17.62; *P*=.92).

Comparison Between Paid and Free Chatbots

Chatbots were also compared based on their access model, distinguishing between paid-access platforms and free-access platforms. Comparative performance metrics are summarized in Table 3.

Table . Comparison of quality, readability, and transparency metrics by access type[a].

Metric	Paid, median (IQR)	Free, median (IQR)	*P* value
DISCERN	42 (36-45)	33 (23.8-41.3)	<.001
PEMAT-P[b]	52.6 (46-54)	46 (26.3-47.4)	.002
WRR[c]	26.8 (23.2-26.8)	10.7 (3.57-19.6)	<.001
Word count	336 (271-369)	206 (116-294)	.11
CLI[d]	21.30 (20.6-22.3)	21.71 (20.9-22.1)	.91
FKGL[e]	17.54 (16.6-18.4)	17.56 (16.5-17.6)	.61

[a]Values are presented as median (IQR). DISCERN reflects the quality of evidence-based treatment information. Patient Education Materials Assessment Tool for Printable Materials indicates understandability based on the Patient Education Materials Assessment Tool. Web Resource Rating assesses transparency and use of external references. The Coleman-Liau Index evaluates textual complexity; higher values indicate more difficult language. Flesch-Kincaid Grade Level denotes the U.S. school grade level required for comprehension; higher scores correspond to decreased readability. Paid chatbots: Perplexity Pro, ChatGPT-4.5, Claude 3.7, Gemini Advanced, Gemini 2.5 Pro, Grok 3. Free chatbots: ChatGPT-4, Claude 3.5, Copilot, DeepSeek. Group comparisons were performed using the Mann-Whitney *U* test.

[b]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[c]WRR: Web Resource Rating.

[d]CLI: Coleman-Liau Index.

[e]FKGL: Flesch-Kincaid Grade Level.

Overall content quality, as assessed by the DISCERN total score was significantly higher among paid chatbots than free chatbots (median 42, IQR 36-45 vs 33, IQR 23.8-41.3, respectively; *P*<.001). Similarly, understandability scores measured using the PEMAT-P tool were greater in the paid group (median 52.6, IQR 46-54) compared with the free group (median 46, IQR 26.3-47.4; *P*=.002).

Transparency, evaluated using the WRR, differed markedly between access models. Paid chatbots achieved substantially higher WRR scores than free chatbots (median 26.8, IQR 23.2-26.8 vs 10.7, IQR 3.57-19.6; *P*<.001), reflecting more frequent inclusion of external references and source information.

Median word count did not differ significantly between access models, although paid chatbots generated numerically longer responses than free chatbots (median 336, IQR 271-369 vs 206, IQR 116-294; *P*=.11).

For readability metrics, no statistically significant differences were observed between paid and free platforms. FKGL scores were comparable between paid and free platforms (median 17.54, IQR 16.6-18.4 vs 17.56, IQR 16.5-17.6; *P*=.61). Similarly, CLI values did not differ significantly between groups (median 21.30, IQR 20.6-22.3 vs 21.71, IQR 20.9-22.1; *P*=.91).

The comparative strengths and limitations of each AI chatbot platform evaluated in this study are summarized in Table 4, highlighting the variability in treatment quality, citation transparency, comprehensibility, and readability across different models.

XSL•FO
RenderX

Table . Summary of strengths and limitations of artificial intelligence chatbot platforms in response to a standardized myelin oligodendrocyte glycoprotein antibody–associated disease-related patient query based on DISCERN, Patient Education Materials Assessment Tool for Printable Materials, Web Resource Rating, Flesch-Kincaid Grade Level, and Coleman-Liau Index assessments.

AI[a] chatbots	Strengths	Limitations
ChatGPT-4	Freely accessible; offers moderate citation visibility (WRR[b]); designed for conversational interaction	Demonstrates only moderate performance in treatment quality (DISCERN); suboptimal understandability; exceeds recommended readability thresholds (FKGL[c] >8)
ChatGPT-4.5	Strong citation practices; consistently high-quality content regarding treatment information (DISCERN)	Textual density compromises accessibility; readability metrics exceed public health communication standards
Claude 3.5	Open-access platform; conversationally fluent interface	Among the lowest in treatment quality (DISCERN) and source transparency (WRR); limited medical detail and citation support
Claude 3.7	Integrates citations effectively; maintains coherent conversational flow	Despite improved content structure, textual complexity remains high; readability concerns persist
Grok 3	Highest overall performance across DISCERN and WRR; provides comprehensive and well-referenced responses	Subscription required; lengthy and lexically dense outputs may reduce usability for non-expert audiences
Perplexity Pro	Search-augmented design; uses structured formatting (eg, bullet points); relatively strong performance in understandability	Exhibits high linguistic complexity (FKGL and CLI)[d]; demonstrates RAG[e] errors, including acronym confusion, resulting in the inclusion of clinically irrelevant citations
Gemini Advanced	Achieved the highest score in understandability (PEMAT-P)[f]; benefits from search-enhanced architecture	Moderate citation transparency; content complexity remains above recommended readability thresholds
Gemini 2.5 Pro	Search-optimized structure; favorable DISCERN performance; clearly organized responses	Dense language and elevated FKGL scores limit accessibility; moderate citation performance
Copilot	Freely accessible; concise responses	Performed lowest in both understandability (PEMAT-P) and transparency (WRR) in this study; superficial content
DeepSeek	Freely accessible; balanced performance profile; favorable DISCERN scores; search-focused output generation	Moderate understandability and citation transparency; readability metrics exceed recommended limits for patient education

[a]AI: artificial intelligence.

[b]WRR: Web Resource Rating.

[c]FKGL: Flesch-Kincaid Grade Level.

[d]CLI: Coleman-Liau Index.

[e]RAG: retrieval-augmented generation.

[f]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

Interrater Reliability Analysis

To assess the consistency of expert evaluations across chatbot-generated outputs, interrater reliability was evaluated using ICC. A 2-way random-effects model with absolute agreement (ICC [1,2]) was applied for continuous scores. The ICC for DISCERN total scores was 0.902, indicating excellent agreement among evaluators. Similarly, WRR scores demonstrated strong reliability with an ICC of 0.887. The PEMAT-P understandability scores also showed a high level of consistency, yielding an ICC of 0.838. These results suggest robust interrater agreement across all evaluated dimensions.

These findings provide a detailed snapshot of current chatbot performance across multiple dimensions and serve as a foundation for interpreting their implications in the broader context of digital health communication.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to systematically evaluate chatbot-generated responses on MOGAD using a validated, multidimensional assessment framework. Across ten widely used platforms, we found statistically significant differences in scientific quality, understandability, and source transparency as measured by DISCERN, PEMAT-P, and WRR (P<.001 for all instruments). These differences were evident both between paid and free models and among chatbots with similar functional orientations. Paid-access platforms

XSL•FO
RenderX

generally demonstrated higher scientific quality and citation transparency. Grok 3, a subscription-based conversational system, consistently produced the highest scores, whereas free-access platforms such as Copilot and Claude 3.5 showed lower performance. However, performance was not determined solely by access model, and marked individual variability across platforms was evident. Importantly, all chatbots generated content that exceeded recommended public health readability thresholds according to FKGL and CLI metrics, indicating a uniformly high level of linguistic complexity across platforms. No statistically significant differences were observed between paid and free platforms in readability (FKGL and CLI) or response length. Interrater reliability among the 7 blinded neurologists was high (ICC=0.902 for DISCERN, 0.887 for WRR, and 0.838 for PEMAT-P), supporting the robustness of our evaluation protocol.

Taken together, these results demonstrate marked variability in the clarity, scientific reliability, and source transparency of AI-generated medical information on MOGAD. Notably, improvements in scientific quality and transparency did not consistently translate into greater readability or easier comprehension for general audiences. Given that patients and caregivers may rely on such tools when navigating a rare and clinically complex neuroimmunological disorder, the observed differences across platforms underscore the importance of systematically evaluating chatbot outputs before integrating them into patient-facing contexts.

Comparison With Prior Work

The performance differences we observed between paid and free access models are largely consistent with prior research. In our study, subscription-based platforms produced higher scores in scientific quality (DISCERN, $P<.001$), understandability (PEMAT-P, $P=.002$), and source transparency (WRR, $P<.001$). These findings align with studies reporting that premium systems tend to deliver more comprehensive and clinically aligned health information [12,22,23]. Such advantages may reflect differences in model architecture, update frequency, and overall technological capacity in favor of paid platforms [9].

However, higher cost does not necessarily translate into superior clarity or usability. Previous research has shown that some free access models can perform strongly, particularly in common and well-determined clinical topics. A striking example is found in the work of Şahin et al [13], who demonstrated that Copilot, the lowest scoring model across all metrics in our study, achieved the highest DISCERN score in an independent evaluation of erectile dysfunction content. This contrast illustrates how strongly chatbot performance can depend on clinical context and the formulation of the question.

Topic-dependent variability is a recurring theme in broader literature. Carlson et al [10] reported a wide range of performance across chatbots responding to vasectomy-related queries, with no platform demonstrating consistent superiority. Similar heterogeneity in accuracy, clarity, and transparency has been documented across studies involving sarcoidosis, chronic rhinosinusitis, appendicitis, benign anorectal disorders, and dental implantology [11,23-26]. In line with these observations, our analysis shows that variability becomes even more pronounced in MOGAD, a rare and phenotypically heterogeneous neuroimmunological disorder. The limited availability of high-quality open-access resources on MOGAD may further contribute to narrower or less comprehensive responses generated by certain models.

Across published evaluations of AI chatbots, a model's functional orientation plays a central role in shaping the structure, clarity, and reliability of the medical information it produces. Studies conducted in areas such as sarcoidosis, chronic rhinosinusitis, appendicitis, and dental implantology consistently show that search-focused systems tend to generate more structured and easily navigable content that incorporates bullet points, subheadings, and hierarchical formatting [11,23-25]. In contrast, conversation-focused models typically provide a smoother narrative flow and stronger contextual coherence, allowing users to follow complex medical explanations more intuitively [13,24,27]. Our findings reflect this broader pattern: search-oriented platforms achieved higher PEMAT-P scores, while conversational systems demonstrated stronger WRR performance by integrating citations in a manner that was more consistent with the surrounding text.

Previous research also highlights a recurring vulnerability among search-focused platforms that rely on retrieval-augmented generation. Although these models often supply a larger number of citations, their references may not consistently align with the clinical context, and some may be only partially related to the queried topic [28]. In our study, this limitation was most clearly illustrated by Perplexity Pro, which achieved high WRR scores yet produced retrieval-related errors, including acronym confusion between "MOGAD" and unrelated terms such as "Monad." This resulted in the inclusion of clinically irrelevant sources, including pharmaceutical products and conditions unrelated to neuroimmunology.

This finding underscores an important methodological limitation of the WRR instrument when applied to AI-generated content. WRR captures the visibility and quantity of cited web resources, but it does not assess the clinical appropriateness or accuracy of those citations. As a result, high transparency scores may coexist with retrieval errors that carry potential patient safety implications, particularly in search-based chatbots.

Conversely, while conversation-focused systems tend to produce more coherent narratives, they may be more prone to hallucinating mechanistic or treatment-related details when faced with topics that are uncertain or poorly represented in the literature [29]. Collectively, these findings indicate that each functional orientation carries distinct strengths and limitations. For rare and clinically heterogeneous conditions such as MOGAD, effective patient communication may require a balanced approach that draws on the structural advantages of search-focused models and the contextual strengths of conversational systems.

In our study, readability emerged as a consistent limitation across all platforms. As illustrated in Figure 4, every chatbot generated FKGL and CLI scores exceeding the recommended Grade 8 threshold for public health communication, indicating that the linguistic complexity of these outputs may pose

XSL•FO
RenderX

challenges for users with limited health literacy. Similar findings have been reported in previous evaluations, where chatbot-generated medical information frequently exceeded recommended reading levels [11,12,23,24,27]. Taken together, the evidence suggests that elevated readability demands constitute a structural characteristic of current language models rather than a condition-specific exception.

For MOGAD, the implications of this burden are likely amplified. As a rare and clinically heterogeneous neuroimmunological disorder, MOGAD requires nuanced explanations of diagnostic criteria, relapse patterns, and treatment approaches, which may already be difficult even for medically literate users. When this level of clinical complexity is expressed through dense vocabulary and abstract phrasing, the cognitive load on patients and caregivers increases substantially. This interaction between disease complexity and high readability demands may heighten the risk of misunderstanding, false reassurance, or overly simplified self-directed interpretations. Ensuring that chatbot-generated information is linguistically accessible, therefore, remains essential for improving the safety and usability of AI-based medical communication in neuroimmunology.

Limitations

This study has several limitations. First, the analysis was based on a single standardized patient query, which strengthened internal validity but limited generalizability to the more varied and interactive ways patients seek information in real-world settings. Second, only English-language outputs were evaluated, so performance in other languages or multilingual models remains uncertain. Third, chatbot responses were collected on a single day to minimize temporal drift, yet LLMs evolve rapidly, and their performance may change with future updates. Fourth, although validated instruments such as DISCERN, PEMAT-P, and WRR provided a structured basis for comparison, these tools were developed for traditional static educational materials and may not capture the dynamic and adaptive features of AI-generated content. Finally, we did not perform a formal qualitative or thematic analysis, meaning that nuanced impressions about clinical depth or emphasis were not systematically assessed.

Future Directions

Future research should adopt evaluation designs that better reflect how individuals seek health information in real-world settings. Because this study used a single standardized question, future work would benefit from multiscenario approaches that incorporate sequential prompts, follow-up queries, and patient-specific contexts. Such designs may offer a more realistic understanding of how chatbots manage evolving information needs and uncertainty, particularly in complex neuroimmunological disorders such as MOGAD.

There is also a need for tools specifically developed to assess AI-generated medical content. While DISCERN, PEMAT-P, and WRR remain valuable, they were designed for static patient education materials and do not fully capture the adaptive and conversational nature of LLMs. Future metrics should evaluate hallucination risk, citation quality, internal consistency, and clinical safety. They should also integrate usability testing with patients and caregivers to clarify how tone, framing, and readability influence understanding.

Multilingual and culturally sensitive assessments represent another important direction. As chatbots expand globally, their performance across different languages, health systems, and literacy levels should be examined to identify potential disparities in access to reliable online health information. Longitudinal evaluations will also be important, as rapid updates to model architecture may lead to meaningful changes in accuracy and transparency over time.

Finally, incorporating qualitative methods such as thematic or discourse-level analysis may provide deeper insight into how chatbot responses structure complex concepts, communicate uncertainty, and present risk–benefit information. Blending qualitative insights with quantitative scoring may support the development of safer and more clinically meaningful AI tools for individuals seeking information about MOGAD and other rare neurological disorders.

Conclusions

This study provides a multidimensional evaluation of AI chatbot responses to a patient-centered MOGAD query and demonstrates marked variability across platforms in scientific quality, understandability, transparency, and readability. Paid-access models achieved significantly higher scores in scientific quality, understandability, and citation transparency, whereas search-focused systems showed modest but statistically significant advantages in understandability, highlighting a trade-off rather than uniform superiority across model types. Importantly, all evaluated chatbots produced content exceeding recommended readability thresholds, indicating persistent barriers to linguistic accessibility, particularly for patients and caregivers navigating a rare and clinically complex disorder.

These findings underscore that strong citation practices or higher transparency scores do not necessarily guarantee clinical relevance or usability, and that current evaluation instruments may incompletely capture risks specific to AI-generated medical content. For rare diseases such as MOGAD, ensuring safe and patient-centered use of chatbots will require assessment frameworks that better account for readability, contextual accuracy, and potential retrieval-related errors, alongside traditional measures of quality and transparency.

Funding

The authors declare that no funds, grants, or other support were received during the preparation or conduct of this study.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. The source material consisted of chatbot responses obtained from publicly accessible platforms and anonymized prior to expert evaluation. No human participant data were collected or analyzed in this study.

Authors' Contributions

MTS and MFY conceived the study and planned its overall design. DAM contributed to the development of the study methodology. NDC submitted all chatbot queries, collected the raw outputs, and anonymized the responses. MFY, DAM, MBE, PÖ, YEA, ERK, and SM participated in the evaluation process as blinded expert raters. Data analysis and interpretation were performed by MTS and MFY. The manuscript was drafted by MTS and MFY and critically revised by all authors for important intellectual content. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Chatbot responses.
[PDF File, 1435 KB - medinform_v14i1e81720_app1.pdf]

References

1. Giansanti D. The chatbots are invading us: a map point on the evolution, applications, opportunities, and emerging problems in the health domain. Life (Basel) 2023 May 5;13(5):1130. [doi: 10.3390/life13051130] [Medline: 37240775]
2. Kaur P, Kaur R, Kaur G, Mittal K. Artificial intelligence in medical information retrieval: a word of caution. Transfus Clin Biol 2025 May;32(2):253-254. [doi: 10.1016/j.tracli.2024.11.003] [Medline: 39547543]
3. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or pandora's box? JAMA Intern Med 2023 Jun 1;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
4. Erkan A, Koc A, Barali D, et al. Can patients with urogenital cancer rely on artificial intelligence chatbots for treatment decisions? Clin Genitourin Cancer 2024 Dec;22(6):102206. [doi: 10.1016/j.clgc.2024.102206] [Medline: 39236508]
5. Moseley CE, Virupakshaiah A, Forsthuber TG, Steinman L, Waubant E, Zamvil SS. MOG CNS autoimmunity and MOGAD. Neurol Neuroimmunol Neuroinflamm 2024 Sep;11(5):e200275. [doi: 10.1212/NXI.0000000000200275] [Medline: 38996203]
6. Banwell B, Bennett JL, Marignier R, et al. Diagnosis of myelin oligodendrocyte glycoprotein antibody-associated disease: international MOGAD panel proposed criteria. Lancet Neurol 2023 Mar;22(3):268-282. [doi: 10.1016/S1474-4422(22)00431-8] [Medline: 36706773]
7. Sechi E, Cacciaguerra L, Chen JJ, et al. Myelin Oligodendrocyte glycoprotein antibody-associated disease (MOGAD): a review of clinical and MRI features, diagnosis, and management. Front Neurol 2022;13:885218. [doi: 10.3389/fneur.2022.885218] [Medline: 35785363]
8. Stefan KA, Ciotti JR. MOG antibody disease: nuances in presentation, diagnosis, and management. Curr Neurol Neurosci Rep 2024 Aug;24(8):219-232. [doi: 10.1007/s11910-024-01344-z] [Medline: 38805147]
9. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
10. Carlson JA, Cheng RZ, Lange A, et al. Accuracy and readability of artificial intelligence chatbot responses to vasectomy-related questions: public beware. Cureus 2024 Aug;16(8):e67996. [doi: 10.7759/cureus.67996] [Medline: 39347335]
11. Yassa A, Ayad O, Cohen DA, et al. Search for medical information for chronic rhinosinusitis through an artificial intelligence chatbot. Laryngoscope Investig Otolaryngol 2024 Oct;9(5):e70009. [doi: 10.1002/lio2.70009] [Medline: 39257728]
12. Durmaz Engin C, Karatas E, Ozturk T. Exploring the Role of ChatGPT-4, BingAI, and Gemini as virtual consultants to educate families about retinopathy of prematurity. Children (Basel) 2024 Jun 20;11(6):750. [doi: 10.3390/children11060750] [Medline: 38929329]
13. Şahin MF, Ateş H, Keleş A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis. J Med Syst 2024 Apr 3;48(1):38. [doi: 10.1007/s10916-024-02056-0] [Medline: 38568432]
14. Collins CE, Giammanco PA, Guirgus M, et al. Evaluating the quality and readability of generative artificial intelligence (AI) chatbot responses in the management of achilles tendon rupture. Cureus 2025 Jan;17(1):e78313. [doi: 10.7759/cureus.78313] [Medline: 40034889]

XSL•FO
RenderX

15. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
16. The patient education materials assessment tool (PEMAT) and user's guide. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html [accessed 2026-03-19]
17. Dobbins M, Watson S, Read K, Graham K, Yousefi Nooraie R, Levinson AJ. A tool that assesses the evidence, transparency, and usability of online health information: development and reliability assessment. JMIR Aging 2018 May 7;1(1):e3. [doi: 10.2196/aging.9216] [Medline: 31518240]
18. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol 1975;60(2):283-284. [doi: 10.1037/h0076540]
19. Kincaid JP, Fishburne, Jr. RP, Rogers RL, et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. : Naval Technical Training Command, Millington; 1975 URL: https://apps.dtic.mil/sti/citations/ADA006655 [accessed 2026-03-19]
20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
21. Yükseköğretim kurumları bilimsel araştırma ve yayın etiği yönergesi [Web page in Turkish]. Council of Higher Education (YÖK). URL: https://www.yok.gov.tr/documents/documents/68fb63b8b2511.pdf [accessed 2026-03-19]
22. Patel VR, Tsai AK, Blaes A, Parikh RB, Adamson AS, Gupta A. Association between price and appropriateness of cancer-related medical information generated by chat-based artificial intelligence models. JCO Clin Cancer Inform 2023 Jun;7:e2300041. [doi: 10.1200/CCI.23.00041] [Medline: 37327463]
23. Wu Y, Zhang Y, Xu M, Jinzhi C, Xue Y, Zheng Y. Effectiveness of various general large language models in clinical consensus and case analysis in dental implantology: a comparative study. BMC Med Inform Decis Mak 2025 Mar 26;25(1):147. [doi: 10.1186/s12911-025-02972-2] [Medline: 40140812]
24. Yetkin NA, Baran B, Rabahoğlu B, Tutar N, Gülmez İ. Evaluating the reliability and quality of sarcoidosis-related information provided by AI chatbots. Healthcare (Basel) 2025 Jun 5;13(11):11. [doi: 10.3390/healthcare13111344] [Medline: 40508957]
25. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc 2024 May;38(5):2887-2893. [doi: 10.1007/s00464-024-10739-5] [Medline: 38443499]
26. Maron CM, Emile SH, Horesh N, Freund MR, Pellino G, Wexner SD. Comparing answers of ChatGPT and Google Gemini to common questions on benign anal conditions. Tech Coloproctol 2025 Jan 26;29(1):57. [doi: 10.1007/s10151-024-03096-x] [Medline: 39864043]
27. Esmailpour H, Rasaie V, Babaee Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. BMC Oral Health 2025 Apr 15;25(1):574. [doi: 10.1186/s12903-025-05965-9] [Medline: 40234820]
28. Sohn J, Park Y, Yoon C, et al. Rationale-guided retrieval augmented generation for medical question answering. arXiv. Preprint posted online on Nov 1, 2024. [doi: 10.48550/arXiv.2411.00300]
29. Salles A, Farisco M. Neuroethics and AI ethics: a proposal for collaboration. BMC Neurosci 2024 Aug 29;25(1):41. [doi: 10.1186/s12868-024-00888-7] [Medline: 39210267]

Abbreviations

AHRQ: U.S. Agency for Healthcare Research and Quality
AI: artificial intelligence
CLI: Coleman-Liau Index
FKGL: Flesch-Kincaid Grade Level
ICC: intraclass correlation coefficient
LLM: large language model
MOGAD: myelin oligodendrocyte glycoprotein antibody–associated disease
PEMAT-P: Patient Education Materials Assessment Tool for Printable Material
WRR: Web Resource Rating

XSL•FO
RenderX

Edited by A Coristine; submitted 02.Aug.2025; peer-reviewed by A Erkan, MV Kayra, P Bhatnagar; revised version received 24.Feb.2026; accepted 24.Feb.2026; published 29.Apr.2026.

Please cite as:
Sönmez MT, Yetkin MF, Mehdiyev DA, Çelik ND, Ercan MB, Öztürk P, Akboğa YE, Koç ER, Mungan S
Evaluation of AI Chatbot Responses to a Standardized Patient Query on Myelin Oligodendrocyte Glycoprotein Antibody–Associated Disease: Cross-Sectional Content Analysis
JMIR Med Inform 2026;14:e81720
URL: https://medinform.jmir.org/2026/1/e81720
doi:10.2196/81720

© Meryem Tuba Sönmez, Mehmet Fatih Yetkin, Duygu Arslan Mehdiyev, Nazlı Durmaz Çelik, Merve Bahar Ercan, Pınar Öztürk, Yeşim Eylev Akboğa, Emine Rabia Koç, Semra Mungan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Model Automated Extraction of Clinical Signs and Symptoms From Emergency Department Reports for Machine Learning Prediction Models: Development and Validation Study

Anoeska Schipper[1,2], MSc; Peter Belgers[3], MSc; Rory David O'Connor[4], MSc, MD; Lieke van de Wouw[4], MSc, MD; Luc Builtjes[1], MSc; Joeran S Bosma[1], MSc; Ron Kusters[2,5], PhD; Steef Kurstjens[2,6], PhD; Matthieu Rutten[1,7], MD, PhD; Bram van Ginneken[1], PhD

[1]Diagnostic Image Analysis Group, Medical Imaging Department, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

[2]Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands

[3]Psychiatry Department, Radboud University Medical Center, Nijmegen, The Netherlands

[4]Emergency Department, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands

[5]Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, The Netherlands

[6]Laboratory of Clinical Chemistry and Laboratory Medicine, Dicoon BV, Canisius-Wilhelmina Ziekenhuis, Nijmegen, The Netherlands

[7]Medical Imaging Department, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands

Corresponding Author:
Bram van Ginneken, PhD
Diagnostic Image Analysis Group, Medical Imaging Department, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

Abstract

Background: Most clinically relevant information in emergency department (ED) visits is documented in free text, limiting reuse for research and clinical decision support. Despite growing interest in large language model (LLM)–based feature extraction, very few studies have examined it directly on ED reports. Existing work has mainly addressed binary tasks and rarely evaluated their impact on downstream prediction models. Furthermore, evidence for small multilingual LLMs remains limited, especially for underrepresented languages such as Dutch. Locally deployable LLMs could enable automated feature extraction for decision support systems without increasing physician workload.

Objective: We aim to evaluate whether a small open-source LLM (Qwen 2.5:14B) can automatically extract 16 clinical signs and symptoms from ED reports and use these as input for an appendicitis prediction model. LLM performance under minimal and optimized 0-shot prompts was assessed against researcher annotations (reference standard) and physician annotations.

Methods: This retrospective study used 336 ED reports from patients presenting with acute abdominal pain to a Dutch teaching hospital (2016-2023). One hundred reports were randomly selected to develop a minimal and an optimized 0-shot prompt strategy. The remaining 236 reports, reserved for evaluation, were annotated by 2 ED physicians and processed by the LLM to extract 16 signs and symptoms, covering binary, multiclass, and multilabel classification tasks. These features were used as input to the HIVE (History, Intake, Vitals, Examination) appendicitis prediction model. LLM extraction accuracy, sensitivity, and specificity were measured against the researcher's (reference standard) and physician annotations. The HIVE model's area under the receiver operating characteristic curve was evaluated using LLM-extracted vs physician-annotated features.

Results: Among 336 ED reports from patients with acute abdominal pain (median age 41, IQR 22 - 62 years, 205/336, 61% female), 50% (167/336) had appendicitis. The LLM achieved weighted average accuracies of 0.910 (95% CI (0.018) with minimal prompts and 0.929 (95% CI ±0.016) with optimized prompts, vs 0.961 (95% CI ±0.012) and 0.951 (95% CI ±0.015) for physicians. Corresponding HIVE model area under the receiver operating characteristic curves were 0.871 (95% CI ±0.019) and 0.911 (95% CI ±0.014) with LLM inputs under the minimal and optimized prompts, compared to 0.917 (95% CI ±0.015) and 0.924 (95% CI ±0.018) for physician inputs.

Conclusions: A small locally deployable multilingual LLM can approach physician-level accuracy in extracting structured binary, multiclass, and multilabel clinical data from free-text Dutch ED reports, while preserving patient privacy, interpretability, and statistical transparency for downstream diagnostic modeling.

(JMIR Med Inform 2026;14:e81500) doi:10.2196/81500

XSL•FO
RenderX

KEYWORDS

large language models; natural language processing; electronic health records; machine learning; predictive modeling; health informatics; emergency medicine

Introduction

In the emergency department (ED), clinical information is primarily documented as free-text. While this documentation style aligns with routine clinical practice, it presents challenges for reusing data in research and decision support applications [1-5]. Structured templates in the electronic health record (EHR) system offer a common workaround, but impose extra workload on already time-pressed ED physicians and are rarely adopted in practice [6,7]. Automated feature extraction using natural language processing (NLP) offers a more sustainable alternative. To date, extracting clinical data from ED reports has primarily relied on long short-term memory and BERT (Bidirectional Encoder Representations From Transformers) models, which perform reasonably well but require extensive preprocessing, large manually annotated corpora, and model retraining [8-10].

Generative pretrained large language models (LLMs) present an alternative, as they can be optimized using only prompting, making them more accessible for downstream tasks. While these LLMs have proved effective in extracting diverse features from radiology and pathology reports [11-13], their use in automated feature extraction from ED reports is limited. Unlike radiology or pathology reports, ED reports contain extensive abbreviations, succinct phrasing, and domain-specific terminology, presenting unique challenges to NLP. Recent work has begun to explore LLM-based information extraction in emergency medicine. McMurry et al [14] demonstrated LLM-based identification of binary respiratory symptoms using 0-shot prompting and multicenter validation, while Bejan et al [15] phenotyped symptomatic kidney-stone presentations using a range of prompting and fine-tuning strategies. Gao et al [16] used LLM-generated severity scores, supported by in-context learning and retrieval-augmented generation, to improve early triage predictions. Collectively, recent studies demonstrate the potential of LLM-based feature extraction from ED reports. However, these efforts primarily addressed binary outcomes or limited symptom sets, or did not investigate the downstream impact on diagnostic model performance. Prior work also relied mainly on large LLMs or Chinese bilingual LLMs, leaving open whether smaller multilingual LLMs can perform reliably in underrepresented languages [17]. This is particularly relevant for the Dutch, where the inherent complexity of ED reports is compounded by limited representation in LLM training corpora.

This study addresses these gaps by evaluating whether a small multilingual LLM (Qwen 2.5:14B) can automatically extract binary, multiclass, and multilabel clinical features from Dutch ED reports and provide reliable inputs for a downstream prediction model. Our focus is a clinically relevant use case: acute abdominal pain (AAP). Extracted features are used as input for the HIVE (History, Intake, Vitals, Examination) model, a previously established appendicitis prediction model based on 16 contributing clinical signs and symptoms [18]. These features were originally annotated manually, a process that is labor-intensive and difficult to scale. This study evaluates whether an LLM can automate this extraction and produce comparable inputs by developing and validating two 0-shot prompting strategies. Secondary outcomes included comparison to physician annotations and assessing the appendicitis prediction model using LLM-extracted vs physician-annotated features. We hypothesize that, with appropriate prompting, a small locally run multilingual LLM could achieve near-expert accuracy in feature extraction and maintain comparable predictive models' performance, supporting the feasibility of a scalable, privacy-preserving workflow for decision support systems at the ED.

Methods

Data Collection

Data from 350 patients with AAP were retrospectively collected at Jeroen Bosch Hospital, a Dutch teaching hospital, between July 2016 and January 2023. Patient inclusion criteria and cohort construction are detailed in our previous study [18]. This dataset included 167 appendicitis cases with 169 other AAP presentations. Among other AAP presentations, those suspected of appendicitis were balanced with those having nonspecific or other AAP causes based on initial ED assessments by triage nurses or referring physicians. Cases lacking sufficient medical history or physical examination data or missing more than 70% (n=14) of vital signs were excluded, resulting in 336 cases. Each case contained triage and intake data, vital signs, and the medical history and physical examination sections. Further cohort details are in Supplemental Tables 1A-E of our previous publication [18]. No exclusions were made based on age, comorbidities, medication use, or symptom presentation. Data were extracted and pseudonymized using CTcue (IQVIA Nederland B.V.).

Study Design

All 336 ED reports were independently annotated by 2 lead researchers to establish the reference standard for 16 clinical signs and symptoms, as previously described [18] (Multimedia Appendix 1). Interrater agreement, assessed on a random sample of 80 reports, demonstrated high reliability, with average Krippendorff α values of 0.93 for binary features and 0.95 for the multiclass feature, and an average Jaccard similarity of 0.76 for multilabel features. The 16 features were derived from our previous HIVE model, in which XGBoost (Extreme Gradient Boosting) and SHAP (Shapley Additive Explanations) analyses were used to identify features contributing more than 95% of the model's predictive output (Multimedia Appendix 2). Their discriminative performance was previously assessed using the area under the receiver operating characteristic curve (AUROC) to determine the optimal feature set. The final selection comprised 11 medical history and 5 physical examination features, covering 8 binary, 1 multiclass, and 7 multilabel outputs.

LLM-based feature extraction using a 0-shot prompting strategy was chosen over conventional tokenization-based text

XSL•FO
RenderX

classification because the goal was to extract clinically meaningful variables and integrate them with structured vital signs and intake information directly available from the EHR. Traditional token-based NLP methods focus on identifying or classifying text spans and cannot reliably capture negation, multilabel outputs, or nuanced clinical context, and typically provide token-level outputs that obscure the contribution of specific clinical findings. This reduces interpretability and alters feature-importance patterns (eg, via SHAP), limiting transparency and alignment with the structured design of the HIVE model.

For the present analysis, 100 reports were used to optimize two 0-shot prompting strategies for extracting these features using an LLM. Zero-shot prompting directly instructs the LLM without examples [19]. The remaining 236 reports were reserved to evaluate manual labeling by 2 ED physicians and automated extraction by the LLM (Figure 1A).

These reports were also used for the HIVE appendicitis prediction model. Of the 336 reports, 268 had previously been labeled by the lead researchers and were used for model training and tuning. The remaining 68 reports, drawn from the 236 physician-labeled and LLM-processed reports for evaluation, served as 4 identical validation sets (Figure 1B). This validation set was selected before model development, contained no cases used in LLM prompt construction or hyperparameter tuning, and matched the validation size used in our previous study [18]. This ensured a fully independent and consistent benchmark for assessing the effect of LLM-based feature extraction on downstream model performance. Given the modest dataset size (n=336) and single-center design, this study aimed to evaluate feasibility rather than generalizable performance. The results should therefore be interpreted as proof-of-concept.

XSL•FO
RenderX

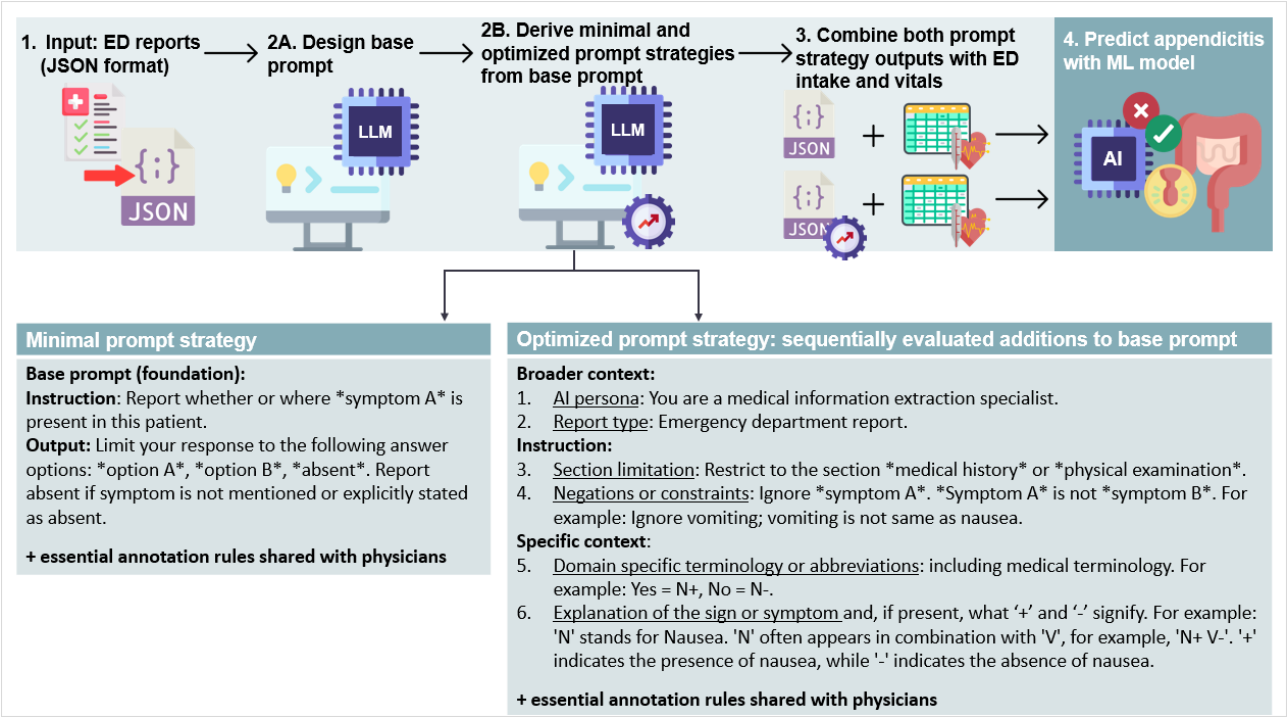
Figure 1. Study design overview comparing feature extraction strategies for the HIVE appendicitis prediction model. (A) Flowchart illustrating the experimental setup for comparing automated (LLM-based) and manual (physician-labeled) feature extraction strategies. All four experiments were conducted on the same set of 236 evaluation cases. (B) Flowchart outlining the HIVE model development and validation process. The model was trained and tuned on a separate, manually labeled development set (n=268), and validated on 68 cases drawn from the same pool of 236 cases used in panel (A). The same trained model was used across all 4 validation experiments; only the feature extraction method varied for the identical set of 68 validation cases. ED: emergency department; HIVE: History, Intake, Vitals, Examination; LLM: large language model; ML: machine learning.

XSL•FO
RenderX

Workflow and Prompting Strategies

The primary objective of this study was to evaluate the LLM's automated extraction of 16 clinical signs and symptoms from ED reports, all of which contributed to our ML (HIVE) model (Multimedia Appendix 2). To that end, a 4-step workflow was implemented (Figure 2). First, ED reports were compiled into a JSON file. Next, a base prompt was designed, consisting of an instruction specifying which symptom or sign to extract and an output with a predefined set of answer options. From this base prompt, two 0-shot prompting strategies were derived. The minimal prompt strategy extended the base prompt with a limited set of essential clarifications, identical to the annotation rules provided to ED physicians to ensure equal comparison. The optimized prompt strategy further integrated additional elements aimed at steering performance, including broader context (artificial intelligence [AI] persona and report type), additional instructions (section limitations, and negations or constraints of symptoms), and specific context (domain-specific terminology or abbreviations, and symptom explanations). To quantify the contribution of individual elements, an ablation study was conducted at the start of the study on a development set of 100 ED reports. Each element was sequentially added to the base prompt, and performance was compared against the researcher's annotations. Only elements that improved performance were included in the optimized prompt. Both prompts used the same annotation rules; any differences between them are due to the prompt design, not to different instructions. All prompts and reports were provided to the LLM in Dutch. The LLM processed each report from the development (n=100) and evaluation sets (n=236), producing structured output in JSON format. The JSON format consisted of a pseudonymized patient identifier, the name of the target symptom, and the predefined answer options of the target symptom (eg, {"uid": "82547948A6A5E91B93B2BAACAE3F943508FD7EFA," "abdominal pain location": ["right lower quadrant," "diffuse"]}). Third, features extracted from 68 of 236 evaluation reports were combined with structured intake data and vital signs. Finally, these datasets were used for ML model validation. The remaining 268 cases, annotated by the research team, were used for developing the ML model.

Figure 2. Modular workflow for extracting clinical signs and symptoms using a local LLM, followed by appendicitis prediction. ED reports in JSON format (1) were processed using a base prompt (0-shot) (2A) containing an instruction and predefined output format. From this base, two prompting strategies were derived (2B): a minimal prompt, which included essential clarifications also provided to ED physicians for equal comparison, and an optimized prompt, integrating additional elements that were sequentially evaluated through an ablation study. Feature outputs from both strategies were combined with ED intake and vital-sign data (3) and entered into the HIVE ML model to predict appendicitis (4). AI: artificial intelligence; ED: emergency department; HIVE: History, Intake, Vitals, Examination; LLM: large language model; ML: machine learning.



LLM Selection

In December 2024 and January 2025, four popular LLMs available on the Ollama platform were evaluated using 4 representative features varying in complexity and output type. Selected models, Mistral-Nemo 12B Instruct [20], Qwen2.5:14B [21], DeepSeek R1:14B distilled from Qwen2.5 [22], and Gemma2:9B [23], were chosen for their multilingual capabilities, input length, and compact size (<10 GB), allowing deployment on a 12 GB GPU laptop without requiring connection to a computational cluster, which may be impractical in ED settings.

After evaluation, Alibaba's Qwen2.5:14B Instruct was selected based on performance on the development set (n=100). All LLMs were assessed on the same data, prompts, and settings (eg, temperature=0.1), supporting deterministic output [24]. Ollama was chosen for its capability to facilitate local LLM deployment. The underlying LLM extraction tool for this study is available in a repository (10.6084/m9.figshare.28931030)

XSL•FO
RenderX

and on GitHub [25] with a complete walkthrough to adopt the framework and tooling, and incorporates components from a previously published data extraction repository [26].

Postprocessing of LLM Outputs

LLM outputs were checked against the predefined answer options. To mimic daily practice, where manual postprocessing is infeasible, responses outside these options were removed. Occasionally, the LLM returned descriptive terms accurately copied from the ED report but not listed among the predefined answers. These nonconforming terms were removed, and the feature was labeled as "not reported." All performance metrics presented reflect this postprocessing step.

Reader Study–Manual Labeling

A reader study was conducted to compare LLM outputs, under the minimal and optimized prompt strategies, with manual labeling of 2 ED physicians using the same evaluation set (n=236). Each ED report was presented in its original format and independently reviewed by 2 ED physicians, one with 2 years of residency experience and one with 5 years of postqualification experience. Physicians labeled 16 clinical signs and symptoms using an Excel (Microsoft Corp) spreadsheet. Predefined answer options matched those given to the LLM and were provided as drop-down menus: single-choice for binary and multiclass features, and multiple-choice for multilabel features. If a sign or symptom was absent or not reported, the field was left blank.

Statistical Analysis

LLM- and physician-labeled features were evaluated using specificity, sensitivity, and accuracy. The HIVE model's performance was assessed via AUROCs. For multiclass and multilabel features, LLM performance metrics were calculated for each positive class and averaged over all positive classes. 95% CIs were estimated through bootstrapping with replacement.

ML Model Development and Validation

An XGBoost (version 2.1.1) algorithm was selected to estimate the probability of appendicitis, following the approach described in our previous study [18,27]. Referred to as the HIVE model, it incorporates ED intake information, vital signs, medical history, and physical examination features. As explained in this study's design, only routinely measured vital signs and intake information, combined with medical history and physical examination features that contributed most to our previous HIVE model, were included in this model. Following the same set-up, 336 cases were divided into 268 cases for training or tuning (researcher-labeled), and 68 validation cases relabeled by the LLM, under 2 prompting strategies, and by 2 ED physicians. This allowed comparison against the original reference annotations.

For model tuning, repeated stratified 10-fold cross-validation was used to preserve class distributions. Categorical parameters were encoded using CatBoost (version 2.6.3) into numerical representations derived from training data [28,29]. Hyperparameters were optimized to maximize AUROC using Bayesian optimization with Optuna (version 3.6.1) over 100 trials [30]. The optimized hyperparameters were a learning rate (η) of 0.012, a maximum tree depth of 3, a minimum sum of weights in a child of 4, a subsample ratio of features for each tree of 0.64, and 156 boosting rounds. Class weights were not adjusted, given that the dataset was approximately balanced (167 appendicitis vs 169 other AAP causes). Feature contributions to model predictions were quantified using TreeSHAP (version 0.46.0). To examine whether downstream performance depended on model choice, we trained a random forest on the same 236 researcher-labeled cases and evaluated it on the same 68-case validation set using the LLM-extracted features. Preprocessing matched the HIVE pipeline, with numeric missing values imputed using an Iterative Imputer (max_iter=10) for compatibility with the random forest model. A standard configuration was used (n_estimators=500, max_depth=None, bootstrap=True). The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist was followed to ensure methodological transparency (Checklist 1), and all code including the LLM extraction tool are available via figshare (10.6084/m9.figshare.28931030).

Ethical Considerations

This study was conducted according to the Declaration of Helsinki and Guidelines for Good Clinical Practice. The execution of this retrospective observational study of patient records was approved by the local review board of the Jeroen Bosch Hospital (number 2023.11.22.02), and was judged by the Medical Research Ethics Committee Brabant, which waived this study to be subject to the regulations of the Dutch Medical Research Involving Human Subjects Act, including a waiver of informed written consent and a consent for publication (Medical Research Ethics Committee number NW2024-05) [31].

Data were extracted from the EHR using CTcue (IQVIA) and handled in a pseudonymized manner. No directly identifiable patient information was accessed by the researchers. Data extraction and analysis were performed in a secure environment in accordance with institutional and European and national data protection regulations. Participants received no compensation.

Results

Overview

Among 336 patients with AAP (median age 41, IQR 22 - 62 years, 205/336, 61% were female), 50% (167/336) had appendicitis (Table 1, Table 2). Reports were authored by various physicians over 7 years. Feature prevalence ranged from 6% (n=20) for pollakiuria to 91% (n=303) for pain location (Table 3).

XSL•FO
RenderX

Table . Categorical patient characteristics (n=336). Structured EHR[a] data (intake and vital signs).

	Feature	Section	Class	Values, n (%)
1	Sex	Intake	Female	205 (61.0)
2	Transport	Intake		
			Ambulance	32 (9.5)
			Own transport	303 (90.2)
			Other	1 (0.0)
3	Referrer	Intake		
			Primary care physician	219 (65.2)
			Self-referral	98 (29.2)
			Hospital	9 (2.7)
			Ambulance	5 (1.5)
			Other facility	1 (0.0)
			Not reported	4 (1.2)
4	EMV[b]	Vital signs		201 (59.8)
5	Q-SOFA[c]	Vital signs		103 (30.7)

[a]EHR: electronic health record.

[b]EMV: Eye Opening, Best Motor Response, Best Verbal Response.

[c]Q-SOFA: quick sepsis-related organ failure assessment

Table . Numeric patient characteristics (n=336). Structured EHR[a] data (intake and vital signs).

	Feature	Section	Median (IQR)	Values, n (%)
1	Age	Intake	41 (22 - 62)	336 (100.0)
2	Prior ED[b] visits with AAP[c]	Intake	1 (1-1)	336 (100.0)
3	Pain rating	Intake	6 (4-8)	291 (86.6)
4	Diastolic arterial pressure (mm Hg)	Vital signs	79 (71 - 88)	334 (99.4)
5	Systolic arterial pressure (mm Hg)	Vital signs	130 (118 - 144)	326 (97.0)
6	Mean arterial pressure (mm Hg)	Vital signs	96 (88 - 105)	316 (94.0)
7	Respiratory rate (rpm)	Vital signs	16 (15 - 19)	132 (39.3)
8	Heart rate (bpm)	Vital signs	85 (74 - 98)	316 (94.0)
9	SIRS[d]	Vital signs	0 (0 - 1)	189 (56.3)
10	Oxygen saturation	Vital signs	98 (97 - 100)	334 (99.4)
11	Temperature (°C)	Vital signs	37.1 (36.6 - 37.5)	333 (99.4)

[a]EHR: electronic health record.

[b]ED: emergency department.

[c]AAP: acute abdominal pain.

[d]SIRS: systematic inflammatory response syndrome.

XSL•FO
RenderX

Table . Prevalence of 16 clinical features in ED[a] reports (medical history and physical examination; n=336).

	Feature	Section	Classification	Class	Class prevalence, n (%)	Feature prevalence, n (%)
1	Abdominal pain location	Physical examination	Multilabel	Right lower quadrant	222 (67.0)	301 (89.6)
				Left lower quadrant	32 (9.5)	
				Right upper quadrant	28 (8.3)	
				Left upper quadrant	9 (2.7)	
				Epigastric region	18 (5.4)	
				Hypogastric region	11 (3.3)	
				Periumbilical	27 (8.0)	
				Diffuse	21 (6.3)	
2	Nausea	Medical history	Binary		177 (52.7)	177 (52.7)
3	Development of complaints	Medical history	Multilabel	Increase	132 (39.2)	150 (44.6)
				Decrease	40 (11.9)	
4	Onset of pain	Medical history	Multiclass	Acute	56 (16.7)	63 (18.8)
				Gradual	7 (2.1)	
5	McBurney's sign	Physical examination	Binary		118 (35.1)	118 (35.1)
6	Rebound tenderness	Physical examination	Binary		100 (29.7)	100 (29.7)
7	Anorexia	Medical history	Binary		122 (36.3)	122 (36.3)
8	Pain migration to the right lower quadrant	Medical history	Binary		77 (22.9)	77 (22.9)
9	Fever	Medical history	Binary		70 (20.8)	70 (20.8)
10	Abdominal inspection	Physical examination	Multilabel	Adipose	44 (13.1)	131 (39.0)
				Scars	22 (6.5)	
				Distended abdomen	25 (7.4)	
				No abnormalities	49 (14.6)	
11	Pollakiuria	Medical history	Binary		20 (6.0)	20 (6.0)
12	Stool consistency	Medical history	Multilabel	Diarrhea	39 (11.6)	257 (76.5)
				Loose	62 (18.5)	
				Normal	154 (45.8)	
				Stiff or obstipation	31 (9.2)	
13	Pain location	Medical history	Multilabel	Right lower quadrant	166 (49.4)	303 (91.2)
				Left lower quadrant	64 (19.0)	
				Right upper quadrant	29 (8.6)	
				Left upper quadrant	18 (5.4)	
				Epigastric region	34 (10.1)	
				Hypogastric region	9 (2.7)	
				Periumbilical	71 (21.1)	
				Diffuse	25 (7.4)	

XSL•FO
RenderX

	Feature	Section	Classification	Class	Class prevalence, n (%)	Feature prevalence, n (%)
				Flank right	8 (2.4)	
				Flank left	4 (1.2)	
				Flanks both sides	3 (1.0)	
14	Pain manifestation	Medical history	Multilabel	Attack-wise	71 (21.1)	180 (53.6)
				Continuous	138 (41.1)	
15	Nature of pain	Medical history	Multilabel	Aching	46 (13.7)	122 (36.3)
				Cramping	34 (10.1)	
				Stabbing	66 (19.6)	
				Burning	2 (0.5)	
16	Palpation tenderness (supple)	Physical examination	Binary		236 (70.2)	236 (70.2)

[a]ED: emergency department.

Ablation Study

An ablation study on the development set (n=100) quantified the contribution of individual prompt elements (Tables 4-6). Sequentially adding each element to the base prompt showed the largest performance gains for negation or constraint, domain-specific context, and explanation, while section limitation, AI persona, and report type produced smaller and less consistent improvements across features (Tables 5 and 6). These findings are incorporated into the composition of the optimized prompts (Table 4). Some clarifying elements, particularly negation or constraint elements, were also integrated into the minimal prompt to ensure an equal comparison with manual annotations by ED physicians. Certain elements were not applicable to specific features (eg, when no negation or explanation was possible) and are therefore left blank in Tables 5 and 6. Table 7 lists the minimal prompts used to extract each of the 16 features; these incorporate the same annotation rules that were provided to ED physicians. Table 8 presents the corresponding optimized prompts, which integrate the prompt elements identified in the ablation study as improving extraction performance. Multimedia Appendix 3 presents all prompt elements evaluated for each feature in the ablation study.

XSL•FO
RenderX

Table . Ablation study (development set, n=100). Performance of base, minimal, and optimized prompting strategies.

	Prompting strategies			Base (instruction + output)			Minimal			Optimized		
	Feature	Section	Classes	Specificity ±CI	Sensitivity ±CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI
1	Abdominal pain location	Physical examination	8	0.93 ± 0.02	0.78 ± 0.07	0.91 ± 0.02	0.95 ± 0.02	0.78 ± 0.07	0.93 ± 0.02	0.96 ± 0.01	0.89 ± 0.06	0.95 ± 0.01
2	Nausea	Medical history	1	0.85 ± 0.11	0.67 ± 0.12	0.76 ± 0.08	0.92 ± 0.07	0.67 ± 0.12	0.80 ± 0.07	0.96 ± 0.05	0.94 ± 0.07	0.95 ± 0.04
3	Development of complaints	Medical history	2	0.89 ± 0.05	0.86 ± 0.09	0.88 ± 0.05	0.89 ± 0.05	0.86 ± 0.09	0.88 ± 0.05	0.91 ± 0.04	0.86 ± 0.09	0.90 ± 0.04
4	Onset of pain	Medical history	2	0.64 ± 0.07	0.94 ± 0.08	0.67 ± 0.07	0.64 ± 0.07	0.94 ± 0.08	0.67 ± 0.07	0.67 ± 0.07	0.89 ± 0.14	0.69 ± 0.06
5	McBurney's sign	Physical examination	1	0.80 ± 0.10	0.97 ± 0.04	0.86 ± 0.07	0.91 ± 0.07	0.94 ± 0.07	0.92 ± 0.05	0.94 ± 0.05	0.97 ± 0.04	0.95 ± 0.04
6	Rebound tenderness	Physical examination	1	0.93 ± 0.06	1.00 ± 0.00	0.95 ± 0.04	0.96 ± 0.05	1.00 ± 0.00	0.97 ± 0.03	__[a]	__[a]	__[a]
7	Anorexia	Medical history	1	0.98 ± 0.02	0.95 ± 0.06	0.97 ± 0.03	0.98 ± 0.02	0.95 ± 0.06	0.97 ± 0.03	__[a]	__[a]	__[a]
8	Pain migration to right lower quadrant	Medical history	1	0.90 ± 0.07	0.95 ± 0.07	0.91 ± 0.05	0.90 ± 0.07	0.95 ± 0.07	0.91 ± 0.05	__[a]	__[a]	__[a]
9	Fever	Medical history	1	0.82 ± 0.08	0.95 ± 0.07	0.85 ± 0.07	0.83 ± 0.08	0.95 ± 0.07	0.86 ± 0.07	__[a]	__[a]	__[a]
10	Abdominal inspection	Physical examination	4	1.00 ± 0.00	0.88 ± 0.09	0.98 ± 0.01	1.00 ± 0.00	0.88 ± 0.09	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.03	0.99 ± 0.01
11	Pollakiuria	Medical history	1	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03	0.98 ± 0.03	1.00 ± 0.00	0.98 ± 0.03
12	Stool consistency	Medical history	4	0.97 ± 0.02	0.67 ± 0.10	0.91 ± 0.03	0.97 ± 0.02	0.67 ± 0.10	0.91 ± 0.03	0.96 ± 0.02	0.87 ± 0.07	0.94 ± 0.02
13	Pain location	Medical history	11	0.94 ± 0.01	0.70 ± 0.08	0.91 ± 0.02	0.95 ± 0.01	0.79 ± 0.07	0.93 ± 0.01	0.96 ± 0.01	0.83 ± 0.07	0.94 ± 0.01
14	Pain manifestation	Medical history	2	0.77 ± 0.07	0.78 ± 0.10	0.77 ± 0.06	0.77 ± 0.07	0.78 ± 0.10	0.77 ± 0.06	0.94 ± 0.04	0.78 ± 0.10	0.90 ± 0.04
15	Nature of pain	Medical history	4	0.96 ± 0.02	0.95 ± 0.06	0.96 ± 0.02	0.96 ± 0.02	0.95 ± 0.06	0.96 ± 0.02	0.98 ± 0.02	0.95 ± 0.06	0.97 ± 0.02
16	Palpation tenderness (supple)	Physical examination	1	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	__[a]	__[a]	__[a]

[a]No prompt element improved performance compared with the minimal prompt; the minimal prompt was retained.

XSL•FO
RenderX

Table . Ablation study (development set, n=100). Incremental effect of individual prompt elements on extraction performance relative to the base prompt.

	Prompt element[a]	AI[b] persona			Report type			Section limitation		
	Feature	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI
1	Abdominal pain location	0.93 ± 0.02[c]	0.79 ± 0.07[c]	0.91 ± 0.02[c]	0.94 ± 0.02	0.75 ± 0.08	0.91 ± 0.02	0.95 ± 0.02	0.75 ± 0.08	0.92 ± 0.02
2	Nausea	0.88 ± 0.09[c]	0.65 ± 0.12[c]	0.77 ± 0.08[c]	0.90 ± 0.09[c]	0.65 ± 0.12[c]	0.78 ± 0.08[c]	0.83 ± 0.11	0.69 ± 0.12	0.76 ± 0.08
3	Development of complaints	0.84 ± 0.06	0.86 ± 0.09	0.84 ± 0.05	0.83 ± 0.06	0.86 ± 0.09	0.83 ± 0.05	0.71 ± 0.07	0.82 ± 0.11	0.74 ± 0.06
4	Onset of pain	0.62 ± 0.07	0.94 ± 0.08	0.65 ± 0.07	0.62 ± 0.07	0.94 ± 0.08	0.65 ± 0.07	0.57 ± 0.07	1.00 ± 0.00	0.60 ± 0.07
5	McBurney's sign	0.80 ± 0.10	0.97 ± 0.04	0.86 ± 0.07	0.80 ± 0.10	0.97 ± 0.04	0.86 ± 0.07	0.75 ± 0.11	0.97 ± 0.04	0.83 ± 0.07
6	Rebound tenderness	0.93 ± 0.06	1.00 ± 0.00	0.95 ± 0.04	0.91 ± 0.06	1.00 ± 0.00	0.94 ± 0.04	0.91 ± 0.06	1.00 ± 0.00	0.94 ± 0.04
7	Anorexia	0.98 ± 0.02	0.95 ± 0.06	0.97 ± 0.03	0.98 ± 0.02	0.92 ± 0.08	0.96 ± 0.03	0.98 ± 0.02	0.95 ± 0.06	0.97 ± 0.03
8	Pain migration to the right lower quadrant	0.68 ± 0.10	0.95 ± 0.07	0.74 ± 0.08	0.73 ± 0.10	0.95 ± 0.07	0.78 ± 0.08	0.87 ± 0.07	1.00 ± 0.00	0.90 ± 0.05
9	Fever	0.82 ± 0.08	0.95 ± 0.07	0.85 ± 0.07	0.82 ± 0.08	0.95 ± 0.07	0.85 ± 0.07	0.82 ± 0.08	0.95 ± 0.07	0.85 ± 0.07
10	Abdominal inspection	0.99 ± 0.01	0.91 ± 0.08	0.98 ± 0.01	0.99 ± 0.01	0.91 ± 0.08	0.98 ± 0.01	0.99 ± 0.01	0.88 ± 0.09	0.98 ± 0.01
11	Pollakiuria	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03
12	Stool consistency	0.97 ± 0.02	0.66 ± 0.10	0.90 ± 0.03	0.98 ± 0.01	0.67 ± 0.10	0.91 ± 0.03	0.96 ± 0.02	0.74 ± 0.09	0.91 ± 0.03
13	Pain location	0.93 ± 0.02	0.69 ± 0.08	0.91 ± 0.02	0.94 ± 0.02	0.73 ± 0.08	0.91 ± 0.02	0.94 ± 0.01[c]	0.76 ± 0.07[c]	0.92 ± 0.02[c]
14	Pain manifestation	0.74 ± 0.07	0.81 ± 0.10	0.77 ± 0.06	0.77 ± 0.07	0.78 ± 0.10	0.77 ± 0.06	0.63 ± 0.08	0.76 ± 0.11	0.67 ± 0.06
15	Nature of pain	0.98 ± 0.02[c]	0.95 ± 0.06[c]	0.97 ± 0.02[c]	0.91 ± 0.03	0.95 ± 0.06	0.92 ± 0.03	0.88 ± 0.03	0.98 ± 0.03	0.89 ± 0.03
16	Palpation tenderness (supple)	__[d]	__[d]	__[d]	__[d]	__[d]	__[d]	__[d]	__[d]	__[d]

[a]Each prompt element shows the isolated incremental effect compared to the base prompt.

[b]AI: artificial intelligence.

[c]Element retained in the optimized prompt configuration.

[d]Prompt element was not applicable and evaluated for that specific variable.

XSL•FO
RenderX

Table . Ablation study (development set, n=100). Incremental effect of individual prompt elements on extraction performance relative to the base prompt.

	Prompt element[a]	Negation or constraint			Domain-specific context			Explanation		
	Feature	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI	Specificity ± CI	Sensitivity ± CI	Accuracy ± CI
1	Abdominal pain location	0.95 ± 0.02[bc]	0.78 ± 0.07[bc]	0.93 ± 0.02[bc]	0.94 ± 0.02[c]	0.87 ± 0.06[c]	0.93 ± 0.02[c]	0.94 ± 0.02[c]	0.78 ± 0.07[c]	0.92 ± 0.02[c]
2	Nausea	0.92 ± 0.07[bc]	0.67 ± 0.12[bc]	0.80 ± 0.07[bc]	0.87 ± 0.09[c]	0.81 ± 0.10[c]	0.84 ± 0.07[c]	0.90 ± 0.09[c]	0.85 ± 0.09[c]	0.88 ± 0.06[c]
3	Development of complaints	0.91 ± 0.04[bc]	0.86 ± 0.09[bc]	0.90 ± 0.04[bc]	0.78 ± 0.07	0.90 ± 0.09	0.81 ± 0.05	—[d]	—[d]	—[d]
4	Onset of pain	0.67 ± 0.07[c]	0.89 ± 0.14[c]	0.69 ± 0.06[c]	0.62 ± 0.07	1.00 ± 0.00	0.65 ± 0.07	0.59 ± 0.07	0.94 ± 0.08	0.62 ± 0.06
5	McBurney's sign	0.91 ± 0.07[bc]	0.94 ± 0.07[bc]	0.92 ± 0.05[bc]	0.91 ± 0.07[c]	0.94 ± 0.07[c]	0.92 ± 0.05[c]	0.73 ± 0.11	0.94 ± 0.07	0.81 ± 0.08
6	Rebound tenderness	0.96 ± 0.05[bc]	1.00 ± 0.00[bc]	0.97 ± 0.03[bc]	—[d]	—[d]	—[d]	0.93 ± 0.06	1.00 ± 0.00	0.95 ± 0.04
7	Anorexia	—[d]	—[d]	—[d]	0.95 ± 0.06	0.97 ± 0.04	0.96 ± 0.03	0.95 ± 0.06	0.97 ± 0.04	0.96 ± 0.03
8	Pain migration to the right lower quadrant	0.72 ± 0.10	0.95 ± 0.07	0.77 ± 0.08	—[d]	—[d]	—[d]	0.87 ± 0.07	1.00 ± 0.00	0.90 ± 0.05
9	Fever	0.78 ± 0.10	0.91 ± 0.11	0.81 ± 0.08	0.83 ± 0.08[bc]	0.95 ± 0.07[bc]	0.86 ± 0.07[bc]	0.74 ± 0.10	0.95 ± 0.07	0.79 ± 0.08
10	Abdominal inspection	1.00 ± 0.00	0.88 ± 0.09	0.98 ± 0.01	0.99 ± 0.01[c]	0.98 ± 0.03[c]	0.99 ± 0.01[c]	0.99 ± 0.01	0.88 ± 0.09	0.98 ± 0.01
11	Pollakiuria	—[d]	—[d]	—[d]	0.98 ± 0.03[c]	1.00 ± 0.00[c]	0.98 ± 0.03[cd]	0.98 ± 0.03	0.50 ± 0.33	0.95 ± 0.03
12	Stool consistency	—[d]	—[d]	—[d]	0.97 ± 0.02[c]	0.81 ± 0.08[c]	0.94 ± 0.02[c]	0.97 ± 0.02[c]	0.73 ± 0.09[c]	0.92 ± 0.03[c]
13	Pain location	0.95 ± 0.01[bc]	0.79 ± 0.07[bc]	0.93 ± 0.01[bc]	0.94 ± 0.01[c]	0.78 ± 0.07[c]	0.93 ± 0.02[c]	—[d]	—[d]	—[d]
14	Pain manifestation	0.92 ± 0.05[c]	0.75 ± 0.11[c]	0.87 ± 0.05[c]	0.72 ± 0.07	0.83 ± 0.09	0.75 ± 0.06	0.79 ± 0.07[c]	0.80 ± 0.10[c]	0.80 ± 0.05[c]
15	Nature of pain	0.96 ± 0.02	0.95 ± 0.06	0.96 ± 0.02	—[d]	—[d]	—[d]	—[d]	—[d]	—[d]
16	Palpation tenderness (supple)	—[d]	—[d]	—[d]	—[d]	—[d]	—[d]	—[d]	—[d]	—[d]

[a]Each prompt element shows the isolated incremental effect compared to the base prompt.

[b]Element retained in the minimal prompt configuration. (Additional instruction shared with physician)

[c]Element retained in the optimized prompt configuration.

[d]Prompt element was not applicable and evaluated for that specific variable.

XSL•FO
RenderX

Table . Minimal 0-shot prompts as presented to the LLM.[a]

	Feature	Prompt[b]
1	Abdominal pain location	Report where abdominal pain is present in this patient. If multiple locations of abdominal pain are mentioned, report the most painful location (PM[c]).[d] Limit your answer to the following options: “right lower quadrant,” “right upper quadrant,” “left lower quadrant,” “left upper quadrant,” “entire upper abdomen,” “entire lower abdomen,” “entire left abdomen,” “entire right abdomen,” “periumbilical,” “epigastric,” “hypogastric,” “diffuse,” or “absent.” Report “absent” if abdominal pain is not mentioned or explicitly absent.
2	Nausea	Report whether this patient has nausea based on the last status in the report. Ignore vomiting. Vomiting is not the same as nausea.[d] Limit your answer to the following options: “yes” or “no.” Report “no” if nausea is not mentioned or explicitly absent.
3	Development of complaints	Report whether there is an increase or decrease in the development of complaints after the onset in this patient. A decrease in the development of complaints occurring after the administration of pain medication must be ignored and should not be reported as a decrease.[d] Limit your answer to the following options: “increase,” “decrease,” or “not reported.” If none of the mentioned options are explicitly stated in the text, the answer should be “not reported.”
4	Onset of pain	Report whether there is a description of how the onset of pain started in this patient. Limit your answer to the following options: “acute,” “gradual,” or “not reported.” If none of the mentioned options are explicitly stated in the text, the answer should be “not reported.”
5	McBurney’s sign	Report whether this patient has abdominal pain at McBurney’s point. If McBurney’s sign is dubious, report “no.”[d] Limit your answer to the options: “yes” or “no.” Report “no” if abdominal pain at McBurney's point is not mentioned or explicitly absent.
6	Rebound tenderness	Report whether this patient has rebound tenderness. Ignore contralateral rebound tenderness.[d] Limit your answer to the options: “yes” or “no.” Report “no” if rebound tenderness is not mentioned or explicitly absent.
7	Anorexia	Report whether this patient has anorexia. Limit your answer to the options: “yes” or “no.” Report "no” if anorexia is not mentioned or explicitly absent.
8	Pain migration to the right lower quadrant	Report whether the abdominal pain has migrated to the right lower quadrant in the patient. Limit your answer to the options: “yes” or “no.” Report “no” if migration to the right lower quadrant is not mentioned or explicitly absent

XSL•FO
RenderX

	Feature	Prompt[b]
9	Fever	Report whether this patient has a fever. Limit your answer to the options: “yes” or “no.” Note that the term may not be mentioned exactly as written. Yes=recorded elevated temperature, feeling of fever, measured fever, feverish, temperature or “T” of 38°C or higher. No=temperature or “T” of 37.9°C or lower, no fever, not feverish, no elevated temperature.[d] Report “no” if fever is not mentioned or explicitly absent.
10	Abdominal inspection	Report whether any abnormalities are noted in the inspection of the abdomen for this patient. Limit your answer to the following options: “adipose,” “distended abdomen,” “scars,” “no abnormalities,” or “not reported.” If none of the mentioned options are explicitly stated in the text, answer “not reported.”
11	Pollakiuria	Report whether this patient has pollakiuria. Limit your answer to the options: “yes” or “no.” Report “no” if pollakiuria is not mentioned or explicitly absent.
12	Stool consistency	Report whether stool consistency is described for this patient. Limit your answer to the following options: “diarrhea,” “loose,” “normal,” “stiff/obstipation,” or “not reported.” If none of the mentioned options are explicitly stated in the text, answer “not reported.”
13	Pain location	Report where the pain initially started in this patient. Ignore pain resulting from displacement, migration, or radiation. Ignore abdominal tenderness. Ignore findings from the physical examination.[d] Limit your answer to the following options: “right lower quadrant,” “right upper quadrant,” “left lower quadrant,” “left upper quadrant,” “entire upper abdomen,” “entire lower abdomen,” “entire left abdomen,” “entire right abdomen,” “around the navel,” “epigastric,” “hypogastric,” “diffuse,” “bilateral flanks,” “right flank” “left flank,” and “absent.” Report “absent” if the location where the pain initially started is not mentioned.
14	Pain manifestation	Report whether this patient has a specific abdominal pain pattern. Limit your answer to the following options: “attack wise,” “continuous,” or “not reported.” If none of the mentioned options are explicitly stated in the text, answer “not reported.”
15	Nature of pain	Report whether the nature of pain is described for this patient. Limit your answer to the following options: “stabbing,” “aching,” “cramping,” “burning,” or “not reported.” If none of the mentioned options are explicitly stated in the text, answer “not reported.”
16	Palpation tenderness	Report whether this patient has a supple abdomen upon palpation. Limit your answer to the options: “yes” or “no.” Report “no” if a supple abdomen upon palpation is not mentioned or explicitly absent.

[a]LLM: large language model.

[b]All prompts and emergency department reports were provided to the large language model in Dutch and have been translated into English for readability.

[c]PM: punctum maximum.

[d]Annotation rules shared with or instructed to emergency department physicians for manual labeling and included in the minimal prompt.

XSL•FO
RenderX

Table . Optimized 0-shot prompts as presented to the LLM.[a]

	Feature	Prompt[b]
1	Abdominal pain location	You are a medical information extraction specialist.[c] Report where abdominal pain is present in this patient. If multiple locations of abdominal pain are mentioned, report the most painful location (punctum maximum/PM).[d] Limit your answer to the following options: "right lower quadrant," "right upper quadrant," "left lower quadrant," "left upper quadrant," "entire upper abdomen," "entire lower abdomen," "entire left abdomen," "entire right abdomen," "periumbilical," "epigastric," "hypogastric," "diffuse," or "absent." Note that these terms may not be mentioned exactly as listed. Right lower quadrant=RLQ, RLA, McBurney, right iliac fossa. Right upper quadrant=RUQ, Murphy. Left lower quadrant=LLQ, LLA, left iliac fossa. Left upper quadrant=LUQ. Entire right abdomen=right hemiabdomen. Entire left abdomen=left hemiabdomen. Periumbilical=around the navel, paraumbilical, supraumbilical. Epigastric=stomach region. Hypogastric=uterus, suprapubic, above pubis. Diffuse=entire abdomen. Absent=no abdominal pain. During the physical examination, the physician presses on all these locations and describes where pain is felt.[e] Report "absent" if pain is not mentioned or explicitly absent.
2	Nausea	You are a medical information extraction specialist.[c] Your task is to report from the emergency department report whether[c] this patient has nausea based on the last status in the report. Ignore vomiting. Vomiting is not the same as nausea.[d] Limit your answer to the following options: "yes" or "no." Note that these terms may not be mentioned exactly as listed. Yes=N+, N+. No=N−, N−. "N" stands for nausea. "N" often appears in combination with "V," for example, "N+ V−." "+" means nausea is present. "−" means nausea is absent.[e] Report "no" if nausea is not mentioned or explicitly absent.
3	Development of complaints	Report whether there is an increase or decrease in the development of complaints after the onset in this patient. A decrease in the development of complaints occurring after the administration of pain medication must be ignored and should not be reported as a decrease.[d] Ignore migration and radiation of abdominal pain. These do not count as an increase in the progression of abdominal pain complaints.[f] Limit your answer to the following options: "increase," "decrease," or "not reported." If none of the mentioned options are explicitly stated in the text, the answer should be "not reported."
4	Onset of pain	Report whether there is a description of how the onset of pain started in this patient. Ignore the timing of onset and focus only on how the complaints began.[f] Limit your answer to the following options: "acute," "gradual," or "not reported." If none of the mentioned options are explicitly stated in the text, the answer should be "not reported."

XSL•FO
RenderX

	Feature	Prompt[b]
5	McBurney’s sign	Report whether this patient has abdominal pain at McBurney's point. If McBurney's sign is dubious, report “no.”[d] Limit your answer to the options: “yes” or “no.” Note that the term may not be mentioned exactly as listed. Yes=McBurney +, PM McBurney, punctum maximum McBurney. No=McBurney −, no pain at McBurney's point, not at McBurney.[e] Report “no” if abdominal pain at McBurney's point is not mentioned or explicitly absent.
6	Rebound tenderness	__[g]
7	Anorexia	__[g]
8	Pain migration to right lower quadrant	__[g]
9	Fever	__[g]
10	Abdominal inspection	Report whether any abnormalities are noted in the inspection of the abdomen for this patient. Limit your answer to the following options: “adipose,” “distended abdomen,” “scars,” “no abnormalities,” or “not reported.” Note that the terms may not be mentioned exactly as listed: adipose=obese, adiposity. Distended abdomen=bloated, distended, tense. Scar=Pfannenstiel.[e] If none of the mentioned options are explicitly stated in the text, answer “not reported.”
11	Pollakiuria	Report whether this patient has pollakiuria. Limit your answer to the options: “yes” or “no.” Note that the term may not be mentioned exactly as listed: Yes=urinating in small amounts, frequent urination, urinating more often than normal, pollakiuria +. No=pollakiuria −.[e] Report “no” if pollakiuria is not mentioned or explicitly absent.
12	Stool consistency	Report whether the stool consistency is described for this patient. Limit your answer to the following options: “diarrhea,” “loose,” “normal,” “stiff/obstipation,” or “not reported.” Note that the terms may not be mentioned exactly as listed: diarrhea=watery, liquid stool, extremely loose, explosive. Normal=na (no abnormalities), no notable findings, no deviations, no diarrhea or obstipation. Constipation=difficult passage. Stool may also be referred to as defecation or def.[e] If none of the mentioned options are explicitly stated in the text, answer “not reported.”

XSL•FO
RenderX

	Feature	Prompt[b]
13	Pain location	Report where the pain initially started in this patient. Limit yourself to the medical history section.[c] Ignore pain resulting from displacement, migration, or radiation. Ignore abdominal tenderness. Ignore findings from the physical examination.[d] Limit your answer to the following options: "right lower quadrant," "right upper quadrant," "left lower quadrant," "left upper quadrant," "entire upper abdomen," "entire lower abdomen," "entire left abdomen," "entire right abdomen," "around the navel," "epigastric," "hypogastric," "diffuse," "bilateral flanks," "right flank," "left flank," "absent." Note that the terms may not be mentioned exactly as listed: right lower quadrant=RLQ, RLA. Right upper quadrant=RUQ. Left lower quadrant=LLQ, LLA. Left upper quadrant=LUQ. Around the navel=periumbilical, para-umbilical. Entire right abdomen=right hemiabdomen. Entire left abdomen=left hemiabdomen. Entire lower abdomen=lower part of the abdomen. Hypogastric=uterus, suprapubic, above pubis. Epigastric=stomach region. Diffuse=entire abdomen.[e] Report "absent" if the location where the pain initially started is not mentioned.
14	Pain manifestation	Report whether this patient has a specific abdominal pain pattern. Ignore descriptions of acute, sudden, or stabbing abdominal pain, as these are not specific pain patterns.[f] Limit your answer to the following options: "attack wise," "continuous," or "not reported." Note that the terms may not be mentioned exactly as written: attack wise=worsening in waves, peak periods, fluctuating, intermittent, variable in intensity, flare-ups, colicky pain, on and off.[e] If none of the mentioned options are explicitly stated in the text, answer "not reported."
15	Nature of pain	You are a medical information extraction specialist.[c] Report whether the nature of pain is described for this patient. Limit your answer to the following options: "stabbing," "aching," "cramping," "burning," or "not reported." If none of the mentioned options are explicitly stated in the text, answer "not reported."
16	Palpation Tenderness	—[g]

[a]LLM: large language model.

[b]All prompts and emergency department reports were provided to the large language model in Dutch and have been translated to English for readability.

[c]Broader context: (1) artificial intelligence persona or (2) report type. These are prompt elements included in the optimized prompt.

[d]Annotation rules shared with or instructed to emergency department physicians for manual labeling and included in the minimal prompt.

[e]Specific context: (5) domain-appropriate terminology or abbreviations, or (6) explanation of the symptom.

[f]Instruction: (3) section limitation or (4) negations or constraints.

[g]No prompt element improved performance compared with the minimal prompt; the minimal prompt was retained.

Minimal vs Optimized Prompt Strategies

Table 9 presents the feature-level performance of the LLM under both prompt strategies, evaluated against a reference standard based on lead researcher annotations, shown alongside manual labeling by ED physicians on the evaluation set (n=236). Using the minimal prompt strategy, the LLM achieved an average specificity of 0.929 (95% CI ±0.020), sensitivity of 0.810 (95% CI ±0.058), and accuracy of 0.910 (95% CI ±0.018). Lower feature prevalence in the ED reports did not consistently result in lower LLM accuracy (Table 3). When applying the optimized prompt strategy, performance improved across all metrics; specificity increased to 0.943 (95% CI ±0.017), sensitivity to 0.859 (95% CI ±0.055), and accuracy to 0.929 (95% CI ±0.016). The 5 largest gains in accuracy were observed for nausea, pain manifestation, onset of pain, pain location, and

XSL•FO
RenderX

abdominal pain location. No gains appeared for rebound tenderness, anorexia, pain migration to the right lower quadrant, fever, and palpation tenderness. Optimized prompts were particularly beneficial for features involving domain-specific terminology, abbreviations, or overlapping definitions, such as nausea and vomiting.

XSL•FO
RenderX

Table . Extraction performance of the LLM[a], under both prompting strategies, vs ED[b] physicians on the evaluation set (n=236). Each row represents a clinical feature with the number of positive answer options (classes).

		Minimal 0-shot			Optimized 0-shot			ED physician 1			ED physician 2			Comparisons[c]		
Feature	Classes[d]	Spec.[e] ± CI	Sens.[f] ± CI	Acc.[g] ± CI	Spec. ± CI	Sens. ± CI	Acc. ± CI	Spec. ± CI	Sens. ± CI	Acc. ± CI	Spec. ± CI	Sens. ± CI	Acc. ± CI	Δ Acc.[h]	Δ Acc.[i]	Δ Acc.[j]
Abdominal pain location	8	0.96 ± 0.01	0.76 ± 0.05	0.93 ± 0.01	0.96 ± 0.01	0.87 ± 0.04	0.95 ± 0.01	0.99 ± 0.01	0.94 ± 0.03	0.98 ± 0.01	0.96 ± 0.01	0.78 ± 0.05	0.93 ± 0.01	0.02	−0.04	0.01
Nausea	1	0.93 ± 0.05	0.74 ± 0.08	0.82 ± 0.05	0.97 ± 0.03	0.91 ± 0.05	0.94 ± 0.03	0.82 ± 0.07	0.95 ± 0.04	0.89 ± 0.04	1.00 ± 0.00	0.95 ± 0.04	0.98 ± 0.02	0.11	0.05	−0.04
Development of complaints	2	0.88 ± 0.03	0.84 ± 0.06	0.87 ± 0.03	0.91 ± 0.03	0.81 ± 0.07	0.88 ± 0.03	0.99 ± 0.01	0.76 ± 0.07	0.93 ± 0.02	0.98 ± 0.02	0.74 ± 0.08	0.92 ± 0.02	0.01	−0.05	−0.03
Onset of pain	2	0.62 ± 0.08	0.91 ± 0.04	0.65 ± 0.04	0.65 ± 0.04	0.85 ± 0.10	0.67 ± 0.04	0.99 ± 0.01	0.69 ± 0.13	0.96 ± 0.02	0.99 ± 0.01	0.84 ± 0.10	0.98 ± 0.01	0.03	−0.29	−0.31
McBurney's sign	1	0.97 ± 0.03	1.00 ± 0.00	0.98 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.98 ± 0.03	0.98 ± 0.02	0.99 ± 0.01	0.98 ± 0.03	0.99 ± 0.02	0.01	0.00	0.00
Rebound tenderness	1	0.99 ± 0.01	0.97 ± 0.03	0.97 ± 0.02	0.99 ± 0.01[k]	0.97 ± 0.03[k]	0.97 ± 0.02[k]	1.00 ± 0.00	0.91 ± 0.06	0.98 ± 0.02	0.98 ± 0.02	0.84 ± 0.09	0.94 ± 0.03	0.00	0.00	0.03
Anorexia	1	0.97 ± 0.03	0.83 ± 0.08	0.92 ± 0.04	0.97 ± 0.0[k]	0.83 ± 0.0[k]	0.92 ± 0.04[k]	0.98 ± 0.02	0.81 ± 0.08	0.92 ± 0.03	0.99 ± 0.01	0.59 ± 0.01	0.85 ± 0.04	0.00	0.00	0.07
Pain migration to right lower quadrant	1	0.88 ± 0.05	0.89 ± 0.08	0.88 ± 0.04	0.88 ± 0.05[k]	0.89 ± 0.08[k]	0.88 ± 0.04[k]	0.97 ± 0.03	0.86 ± 0.09	0.95 ± 0.03	0.97 ± 0.03	0.89 ± 0.08	0.95 ± 0.03	0.00	−0.06	−0.07
Fever	1	0.92 ± 0.04	0.77 ± 0.11	0.89 ± 0.04	0.92 ± 0.04[k]	0.77 ± 0.11[k]	0.89 ± 0.04[k]	0.94 ± 0.03	0.92 ± 0.07	0.93 ± 0.03	0.92 ± 0.04	0.98 ± 0.03	0.93 ± 0.03	0.00	−0.05	−0.05
Abdominal inspection	4	0.98 ± 0.01	0.93 ± 0.05	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.05	0.98 ± 0.01	0.99 ± 0.01	0.91 ± 0.06	0.98 ± 0.01	1.00 ± 0.00	0.89 ± 0.06	0.99 ± 0.01	0.00	−0.00	−0.01
Pollakiuria	1	1.00 ± 0.00	0.57 ± 0.25	0.98 ± 0.02	1.00 ± 0.00	0.86 ± 0.18	0.99 ± 0.01	1.00 ± 0.00	0.46 ± 0.23	0.97 ± 0.01	1.00 ± 0.01	0.69 ± 0.23	0.98 ± 0.02	0.02	0.02	0.01
Stool consistency	4	0.98 ± 0.01	0.70 ± 0.06	0.92 ± 0.02	0.96 ± 0.02	0.83 ± 0.05	0.93 ± 0.02	0.96 ± 0.02	0.75 ± 0.06	0.91 ± 0.02	0.99 ± 0.01	0.73 ± 0.02	0.94 ± 0.01	0.01	0.02	−0.01
Pain location	11	0.95 ± 0.01	0.75 ± 0.05	0.93 ± 0.01	0.97 ± 0.01	0.80 ± 0.05	0.95 ± 0.01	0.98 ± 0.01	0.85 ± 0.04	0.96 ± 0.01	0.99 ± 0.01	0.71 ± 0.05	0.95 ± 0.01	0.02	−0.02	−0.01

XSL•FO
RenderX

		Minimal 0-shot			Optimized 0-shot			ED physician 1			ED physician 2			Comparisons[c]		
Pain manifestation	2	0.75 ± 0.05	0.85 ± 0.06	0.78 ± 0.04	0.93 ± 0.03	0.79 ± 0.06	0.88 ± 0.03	0.98 ± 0.01	0.87 ± 0.05	0.95 ± 0.02	0.94 ± 0.02	0.91 ± 0.04	0.93 ± 0.02	0.11	−0.06	−0.05
Nature of pain	4	0.95 ± 0.02	0.92 ± 0.05	0.95 ± 0.01	0.95 ± 0.02	0.92 ± 0.05	0.94 ± 0.02	1.00 ± 0.00	0.91 ± 0.05	0.99 ± 0.01	1.00 ± 0.00	0.82 ± 0.05	0.98 ± 0.05	0.00	−0.05	−0.04
Palpation tenderness supple	1	0.96 ± 0.06	0.96 ± 0.03	0.96 ± 0.02	0.96 ± 0.06[k]	0.96 ± 0.03[k]	0.96 ± 0.02[k]	0.93 ± 0.07	0.98 ± 0.02	0.97 ± 0.02	0.94 ± 0.00	0.97 ± 0.07	0.96 ± 0.01	0.00	0.00	0.00
Weighted average[l]	45	0.93 ± 0.02	0.81 ± 0.06	0.91 ± 0.02	0.94 ± 0.02	0.86 ± 0.05	0.93 ± 0.02	0.98 ± 0.01	0.86 ± 0.05	0.96 ± 0.01	0.98 ± 0.01	0.79 ± 0.06	0.95 ± 0.02			

[a]LLM: large language model.

[b]ED: emergency department.

[c]Δ Accuracy columns show pairwise accuracy differences. Δ values indicate improved accuracy of optimized prompts compared with minimal prompts or those of physicians.

[d]Each class is evaluated as a separate binary classification (present vs absent), meaning the number of binary decisions equals twice the number of classes.

[e]Spec.: specificity.

[f]Sens.: sensitivity.

[g]Acc.: accuracy.

[h]Accuracy difference (Δ) between minimal vs optimized prompts.

[i]Accuracy difference (Δ) between emergency department physician 1 vs optimized prompts.

[j]Accuracy difference (Δ) between emergency department physician 2 vs optimized prompts.

[k]Optimized prompts were retained only when they improved performance on the development set; otherwise, the minimal prompt was used).

[l]Weighted averages account for the number of classes per feature. CIs (95% CI) were estimated using bootstrapping.

LLM vs Physician Labeling

Both physicians achieved higher weighted average specificity and accuracy on the evaluation set than the LLM under either prompting strategy. Physician 1 reached a specificity of 0.978 (95% CI ±0.011), sensitivity of 0.859 (95% CI ±0.015), and accuracy of 0.961 (95% CI ±0.012), while physician 2 achieved a specificity of 0.980 (95% CI ±0.008), sensitivity of 0.793 (95% CI ±0.056), and accuracy of 0.951 (95% CI ±0.015; Table 2).

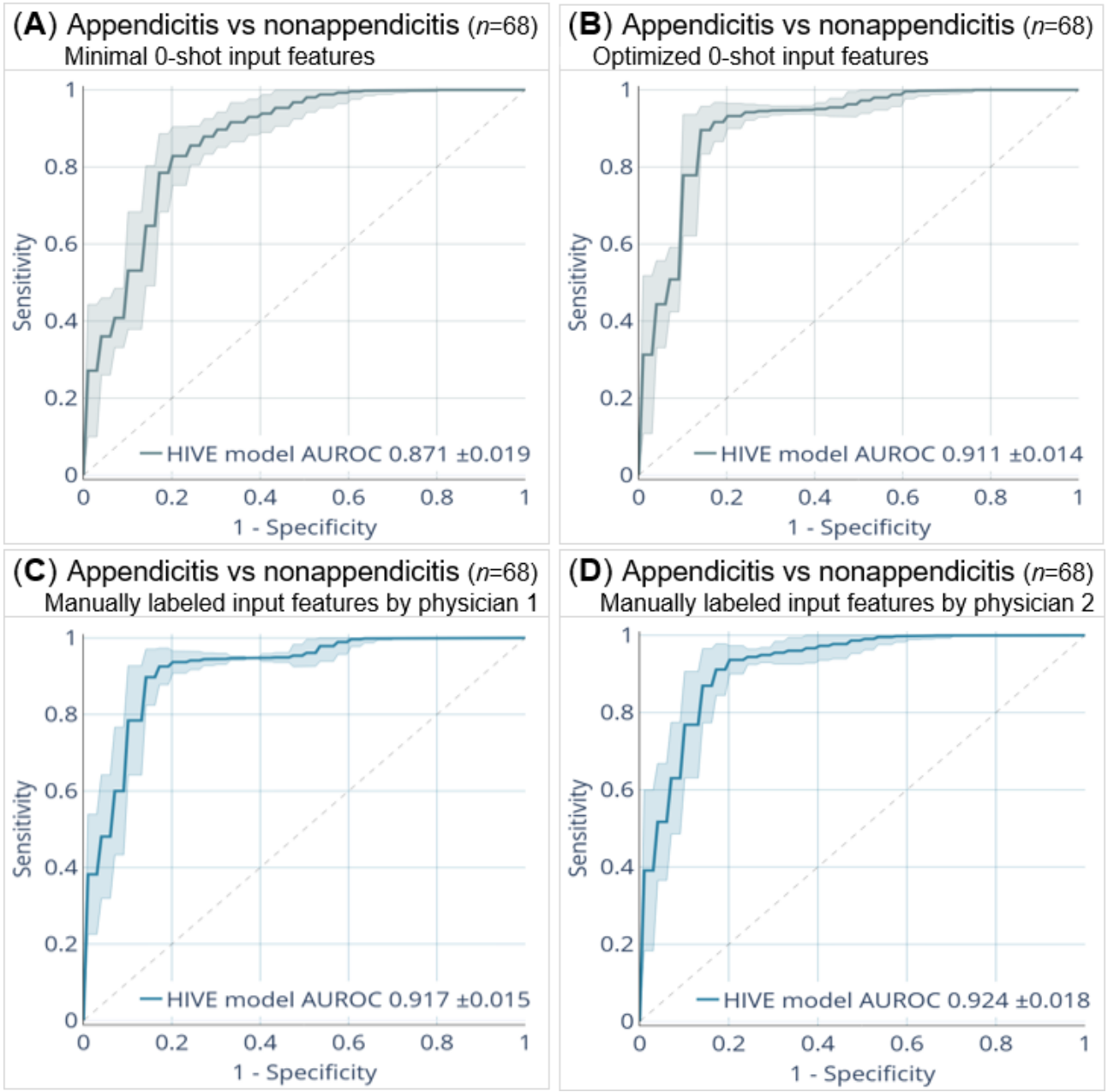
At the feature level, the LLM using the optimized prompt matched or exceeded the accuracy of physician 1 for 8 of 16 features, and for physician 2 for 9 of 16. Notably, both physicians outperformed the LLM on the onset of pain, migration to the right lower quadrant, and pain manifestation. For the remaining features, physician performance was slightly higher, with marginal accuracy differences. All feature-level outputs for this comparison are available in Multimedia Appendix 4.

HIVE Model Performance Using LLM vs Physicians' Inputs

To compare predictive performance using LLM- vs physician-labeled input features, the HIVE model was evaluated for its ability to distinguish appendicitis from other AAP causes on the validation set (n=68; Figure 3). Using LLM-extracted features, the HIVE model achieved an AUROC of 0.871 (95% CI ±0.019) with the minimal prompt, and 0.911 (95% CI ±0.014) with the optimized prompt. In comparison, AUROCs were 0.917 (95% CI ±0.015) using features labeled by physician 1, and 0.924 (95% CI ±0.018) using those from physician 2.

XSL•FO
RenderX

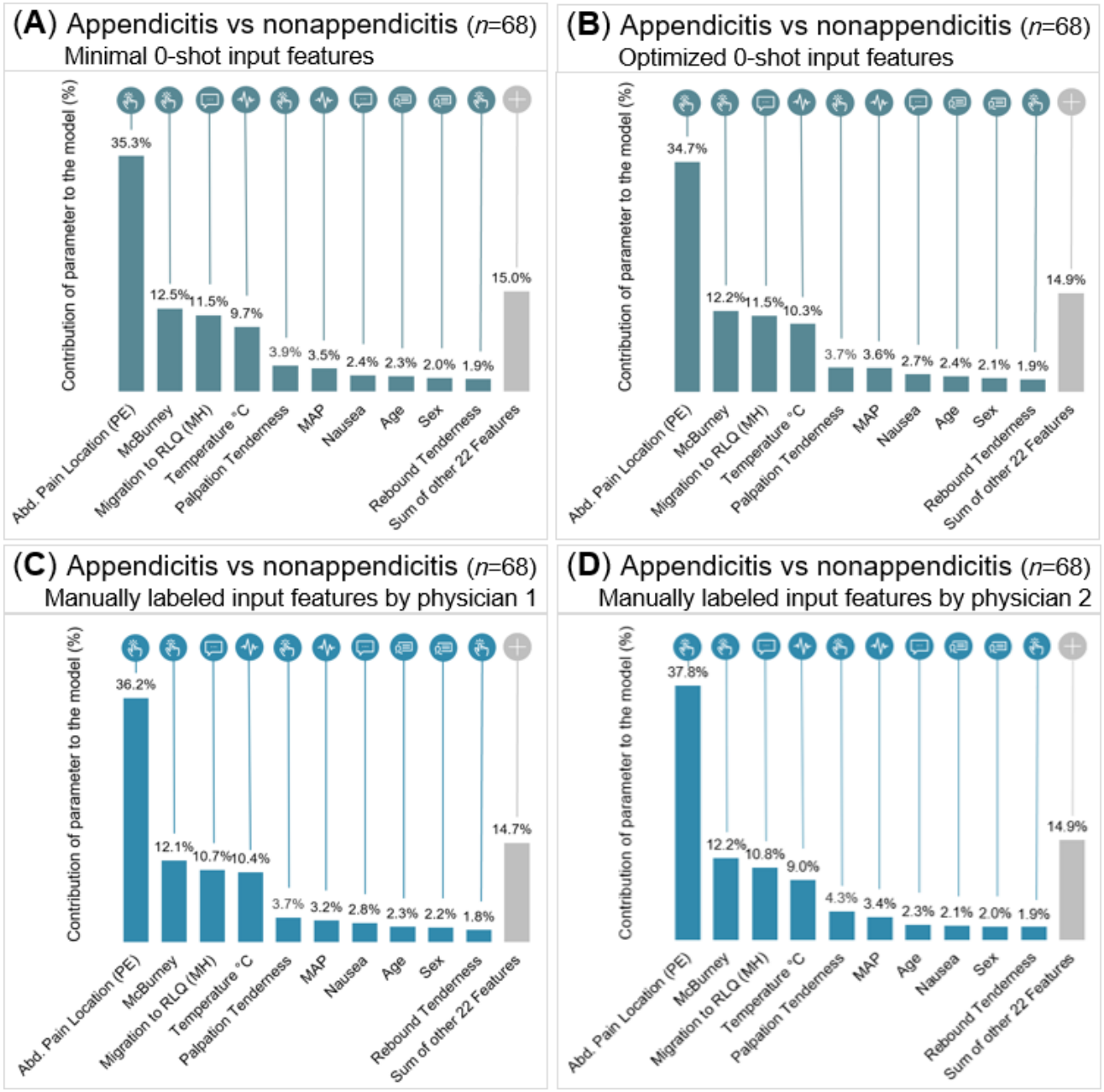
Figure 3. AUROC of the HIVE model for predicting appendicitis vs other AAP causes (ie, no appendicitis) in the validation set (n=68). (A,B) HIVE model performance using LLM-extracted features obtained through the minimal and optimized 0-shot prompting, respectively. (C,D) HIVE model performance using manually labeled features by ED physician 1 and physician 2, respectively. 95% CIs. AAP: acute abdominal pain; AUROC: area under the receiver operating characteristic curves; HIVE: History, Intake, Vitals, Examination; LLM: large language model.



SHAP analysis demonstrated consistent feature importance across all input types (Figure 4). Features are displayed in descending order of their mean absolute SHAP contributions. The overall SHAP distributions were nearly identical, with the top 10 features cumulatively contributing 85% to 85.9% of the model output (full list in Multimedia Appendix 5). Although the minimal 0-shot prompt achieved a lower AUROC of 0.871 (95% CI ±0.019), its feature importance pattern closely resembled those of the optimized and physician-labeled inputs.

XSL•FO
RenderX

Figure 4. Feature contributions to the HIVE model across 4 different input types on the validation set (n=68). Contributions of each feature to the model are SHAP (Shapley Additive Explanations) values scaled and plotted as percentage contributions to the prediction. (A,B) Feature contributions to the HIVE model performance using LLM-extracted features obtained through the minimal and optimized 0-shot prompting, respectively. (C,D) Feature contributions to the HIVE model performance using manually labeled features by ED physician 1 and physician 2, respectively. Abd.: abdominal; ED: emergency department; HIVE: History, Intake, Vitals, Examination; LLM: large language model; MAP: mean arterial pressure; MH: medical history; PE: physical examination; RLQ: right lower quadrant.



The random forest performed similarly to the XGBoost-based HIVE model, with AUROCs of 0.909 (95% CI ±0.13) using the optimized prompt strategy and 0.883 (95% CI ±0.13) using the minimal prompt strategy (Multimedia Appendix 6).

Error Analysis LLM

To examine misclassifications, 8 evaluation cases from the 4 lowest-performing features, pain manifestation, complaint development, pain onset, and abdominal-pain location, were reviewed to illustrate diverse error types.

The first 4 cases in Figure 5 highlight challenges with multilabel outputs, where the LLM was expected to report a combination of applicable labels but instead provided only a single answer. Specifically, in cases 1 and 2, pain manifestation evolved from continuous to attack-wise, yet the LLM recognized only 1 label. Similarly, in cases 3 and 4, the LLM missed fluctuating symptom development over time. In Figure 6, cases 5 and 6 illustrate 2 reports lacking explicit mention of the onset of pain; the LLM inferred labels such as “acute” or “gradual,” whereas reference was “not reported.” In case 7, the ED report described a punctum maximum located 4 cm to the right of the umbilicus. While this was labeled as “periumbilical” in the reference standard, the LLM returned “right lower quadrant.” In case 8, despite instructions to report only the primary pain location, the LLM provided both “entire right abdomen” and “diffuse” for a

XSL•FO
RenderX

case describing diffuse abdominal pain with pronounced tenderness in the right hemiabdomen.

Each panel presents a snippet from an ED report alongside the corresponding LLM output. Words highlighted in magenta within the ED report indicate incorrect (false negative or false positive) or missed interpretations by the LLM, while words in green represent correct (true positive) interpretations. On the right-hand side, the LLM's JSON output displays the unique patient identifier, the target clinical feature (magenta), and the extracted output (green). All prompts and ED reports were provided to the LLM in Dutch and have been translated into English for readability.

Figure 5. Case analysis of 8 ED reports across 4 features from the evaluation set. ED: emergency department; LLM: large language model.

XSL•FO
RenderX

Figure 6. Case analysis of 8 ED reports across 4 features from the evaluation set. ED: emergency department; LLM: large language model.

Discussion

Conventional ML prediction models typically require structured input, posing a significant barrier to clinical integration, especially in EDs where physicians predominantly use free-text reports. This study demonstrates that a locally deployable LLM can automatically extract clinically relevant signs and symptoms from unstructured text, including binary, multiclass, and multilabel features. Using an optimized prompt strategy, the LLM reliably extracted features from free-text ED reports, enabling the appendicitis prediction model to achieve diagnostic performance comparable to that based on manually labeled features by ED physicians. This performance was consistent with that achieved in our previous study [18], underscoring the reliability of the LLM in this context.

Numerous ML prediction models have relied on clinical signs and symptoms to support early disease detection and guide ED workflows. While NLP techniques are commonly used to structure this data, each comes with limitations. Conventional techniques, such as bag-of-words, term frequency-inverse document frequency, and phrase skip-gram, rely on word frequency or co-occurrence, with limited ability to capture contextual meaning [32-36]. As a result, they often miss subtle language cues, such as symptom affirmation, negation, or uncertainty, which are essential in clinical interpretation. Topic modeling can uncover broad textual themes but may not align well with specific clinical concepts [33]. Moreover, these techniques are unsupervised, meaning they do not rely on labeled data, which makes their output more difficult to interpret and control. Deep learning models used in this context (eg, long short-term memory or BERT) understand the meaning of words in a broader context [37-40] but require fine-tuning to handle domain-specific language, large annotated datasets, and are challenged by imbalanced data distributions, limiting their performance and generalizability in real-world settings [41,42].

XSL•FO
RenderX

A small, locally hosted LLM for feature extraction can be deployed with modest requirements, offering advantages in data security and independence from external infrastructure. In comparison with conventional NLP or deep learning models, the LLM used in this study required neither extensive preprocessing nor fine-tuning nor large annotated datasets. Instead, it successfully extracted relevant features directly from raw text using relatively simple 0-shot prompting strategies. Importantly, the LLM was used solely for feature extraction; diagnostic predictions were generated by a separate ML prediction model. This modular approach preserves interpretability and statistical transparency, including calibrated probabilities and CIs. SHAP analysis confirmed that the model's internal reasoning remained stable across input types: AUROC differences primarily reflected feature extraction accuracy rather than alterations in model reasoning. A random forest benchmark produced similar AUROCs to the XGBoost algorithm, indicating that the predictive signal in the LLM-extracted features is consistent across ML algorithms and not dependent on a specific model choice. Although the approach is lightweight and easy to deploy, real-world use still requires seamless EHR integration and fast inference. Modern EHR systems increasingly provide vendor APIs and Python-based toolkits for AI integration, but embedding such workflows into routine EHR practice remains a significant challenge [43]. Given the diagnostic complexity of appendicitis [44-46], this workflow could function as a safety net by flagging high-risk presentations and supporting diagnostic decisions. Larger multicenter evaluations, including physician performance with and without AI assistance, will be needed to determine real-world value.

The ablation study showed that adding specific, symptom-related context, including explanatory cues, domain-specific terminology, and negation or constraint handling, most consistently improved extraction performance. These elements enhanced the LLM's ability to interpret synonyms and abbreviations commonly used in Dutch ED reports (eg, "misselijkheid," "nausea," or "N+" for nausea present) and to distinguish related but distinct concepts (eg, ignore vomiting when assessing nausea). In contrast, broader context elements such as report type, AI persona, or section limitation contributed little to performance. The LLM generalized well across features, classes, and prevalence levels, without requiring balancing strategies, and the close agreement with physician annotations suggests that more complex prompting strategies (eg, few-shot or chain-of-thought) add limited value for this task.

Despite its strengths, error analysis revealed that the LLM struggled with multilabel features, particularly when symptoms evolve over time. Clearly and consistently described features such as nausea or stool consistency were extracted more reliably than the onset of pain, which the LLM may have inferred from the surrounding context. In some cases, the LLM failed to prioritize the most prominent symptom, possibly due to reliance on textual similarity over clinical meaning, as observed in determining abdominal pain location. The LLM occasionally returned descriptive terms that matched the ED report but not the prescribed answer options. Some discrepancies also reflected inherent ambiguities between LLM output and manual annotations. Importantly, the LLM did not hallucinate nonexistent features outside the instructed features and answer options. These findings highlight the potential of LLM-based feature extraction, but also have limitations in handling time discrepancies or ambiguity, which are common challenges in clinical narratives.

This study has several limitations. Only 1 LLM (Qwen 2.5:14B) was evaluated, and performance may differ across models with different architectures or training corpora. As LLMs evolve, future LLMs with improved language understanding or instruction-tuned capabilities may outperform current results and reduce the need for prompt optimization. Nonetheless, given the complexity of ED reports, including abbreviations, medical jargon, and ambiguous phrasing, a structured prompting framework, as presented in this study, will likely remain necessary to ensure consistency and clinically accurate outputs. This study focused on Dutch-language ED reports, reflecting the clinical and linguistic context of our setting. While the prompting framework may be transferable, prompt content must be tailored to language-specific clinical conventions [47]. Translating Dutch reports would not serve as a valid proxy for cross-lingual evaluation because it measures translation quality rather than the LLM's ability to process native clinical text. Although most open-source multilingual LLMs perform best in English, Chinese, and Spanish [17,48,49], the LLM in this study achieved strong results in Dutch, suggesting that carefully designed prompts can support effective use in underrepresented languages. Despite the single-center scope, the dataset spans 7 years and includes reports from many ED physicians with varying documentation styles, supporting robustness, but external validation across linguistic and clinical contexts is needed [50]. Finally, this study focused on extracting signs and symptoms, not numeric values as laboratory results or vital signs, which are typically stored in structured EHR fields.

This study demonstrates that a locally deployable LLM can automatically extract clinical signs and symptoms from free-text Dutch ED reports with performance comparable to manual labeling by two ED physicians using a structured prompt framework, even in a less widely represented language. This approach has the potential to support scalable implementation of our ML prediction models across clinical contexts, without extensive human annotation or computational resources. Ultimately, it may inform future development of real-time integration of prediction models into clinical workflows, facilitating more scalable, transparent, and privacy-preserving decision-support systems.

XSL•FO
RenderX

Acknowledgments

ChatGPT 4o (OpenAI) was used to improve grammar and spelling. We would like to express our gratitude to Harm Geraerdts, MSc, for his contribution to the publication of this paper. We thank Kim Jie, PhD, MD, for contributions to the conceptual framework.

Funding

The authors received financial support from Radboudumc, Jeroen Bosch Hospital and Health~Holland (LSHM20103) for this research, authorship, and publication of this paper. The funders had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

Information and the raw data are available from the corresponding author upon reasonable request. The underlying code for the large language model data extraction tool, the HIVE (History, Intake, Vitals, Examination) model, as well as the saved HIVE model, is available in a figshare repository (10.6084/m9.figshare.28931030).

Authors' Contributions

All authors have accepted responsibility for the entire content of this paper and approved its submission. AS, PB, MR, and BvG initiated and designed this study. AS collected the patient data. AS and PB created the reference standard and designed the features; KJ (Kim Jie, MD, PhD) and RDoC reviewed these features for annotation in a previous study. LvdW and RDoC manually annotated the emergency department reports in the reader study. JSB contributed to the analysis of large language model–based feature extraction. AS and PB developed and finetuned the prompts. AS analyzed the data and developed the History, Intake, Vitals, Examination prediction models for appendicitis prediction. All authors provided input and improved this paper. SK, RK, MR, and BvG supervised this project. AS wrote the paper with input from all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Development of structured features from free-text ED (emergency department) reports using researcher annotations.
[DOCX File, 34 KB - medinform_v14i1e81500_app1.docx]

Multimedia Appendix 2
Feature contributions to the HIVE (History, Intake, Vitals, Examination) model from the previous study's validation set (n=68).
[DOCX File, 44 KB - medinform_v14i1e81500_app2.docx]

Multimedia Appendix 3
Elements evaluated in the prompt ablation study for each feature.
[XLSX File, 20 KB - medinform_v14i1e81500_app3.xlsx]

Multimedia Appendix 4
All outputs from the LLM (large language model) and ED (emergency department) physicians' annotations (n=236).
[XLSX File, 109 KB - medinform_v14i1e81500_app4.xlsx]

Multimedia Appendix 5
Full list of feature contributions using SHAP (Shapley Additive Explanations) values scaled as percentage contributions to the HIVE (History, Intake, Vitals, Examination) model on the validation set (n=68).
[DOCX File, 39 KB - medinform_v14i1e81500_app5.docx]

Multimedia Appendix 6
AUROC (area under the receiver operating characteristic curves) of the HIVE (History, Intake, Vitals, Examination) random forest model for predicting appendicitis vs other AAP (acute abdominal pain) causes in the validation set (n=68).
[DOCX File, 120 KB - medinform_v14i1e81500_app6.docx]

Checklist 1
TRIPOD Checklist.

XSL•FO
RenderX

[DOCX File, 67 KB - medinform_v14i1e81500_app7.docx]

References

1. Kachman MM, Brennan I, Oskvarek JJ, Waseem T, Pines JM. How artificial intelligence could transform emergency care. Am J Emerg Med 2024 Jul;81:40-46. [doi: 10.1016/j.ajem.2024.04.024] [Medline: 38663302]
2. Wu H, Toti G, Morley KI, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 1;25(5):530-537. [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
3. Ben-Haim G, Yosef M, Rowand E, et al. Combination of machine learning algorithms with natural language processing may increase the probability of bacteremia detection in the emergency department: a retrospective, big-data analysis of 94,482 patients. Digit Health 2024;10:20552076241277673. [doi: 10.1177/20552076241277673] [Medline: 39291149]
4. Mohanty SD, Lekan D, McCoy TP, Jenkins M, Manda P. Machine learning for predicting readmission risk among the frail: explainable AI for healthcare. Patterns (N Y) 2022 Jan 14;3(1):100395. [doi: 10.1016/j.patter.2021.100395] [Medline: 35079714]
5. Lee SH. Natural language generation for electronic health records. NPJ Digit Med 2018;1:63. [doi: 10.1038/s41746-018-0070-0] [Medline: 30687797]
6. Sonoo T, Iwai S, Inokuchi R, Gunshin M, Kitsuta Y, Nakajima S. Embedded-structure template for electronic records affects patient note quality and management for emergency head injury patients: an observational pre and post comparison quality improvement study. Medicine (Baltimore) 2016 Oct;95(40):e5105. [doi: 10.1097/MD.0000000000005105] [Medline: 27749590]
7. Moy AJ, Hobensack M, Marshall K, et al. Understanding the perceived role of electronic health records and workflow fragmentation on clinician documentation burden in emergency departments. J Am Med Inform Assoc 2023 Apr 19;30(5):797-808. [doi: 10.1093/jamia/ocad038] [Medline: 36905604]
8. Lee S, Lee J, Park J, et al. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. Am J Emerg Med 2024 Mar;77:29-38. [doi: 10.1016/j.ajem.2023.11.063] [Medline: 38096637]
9. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 1;26(4):364-379. [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
11. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med 2024 May 1;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
12. Bigolin Lanfredi R, Mukherjee P, Summers RM. Enhancing chest X-ray datasets with privacy-preserving large language models and multi-type annotations: a data-driven approach for improved classification. Med Image Anal 2025 Jan;99:103383. [doi: 10.1016/j.media.2024.103383] [Medline: 39546982]
13. Dorfner FJ, Jürgensen L, Donle L, et al. Comparing commercial and open-source large language models for labeling chest radiograph reports. Radiology 2024 Oct;313(1):e241139. [doi: 10.1148/radiol.241139] [Medline: 39470431]
14. McMurry AJ, Phelan D, Dixon BE, et al. Large language model symptom identification from clinical text: multicenter study. J Med Internet Res 2025 Jul 31;27:e72984. [doi: 10.2196/72984] [Medline: 40743494]
15. Bejan CA, Reed AM, Mikula M, et al. Large language models improve the identification of emergency department visits for symptomatic kidney stones. Sci Rep 2025 Jan 28;15(1):3503. [doi: 10.1038/s41598-025-86632-5] [Medline: 39875475]
16. Gao H, Wang K, Yuan Y, et al. A large language model based pipeline for extracting information from patient complaint and anamnesis in clinical notes for severity assessment. Sci Rep 2025;15(1):25345. [doi: 10.1038/s41598-025-07649-4]
17. Hasan MA, Tarannum P, Dey K, et al. Do large language models speak all languages equally? a comparative study in low-resource settings. arXiv. Preprint posted online on Aug 5, 2024. [doi: 10.48550/arXiv.2408.02237]
18. Schipper A, Belgers P, O'Connor R, et al. Machine-learning based prediction of appendicitis for patients presenting with acute abdominal pain at the emergency department. World J Emerg Surg 2024 Dec 23;19(1):40. [doi: 10.1186/s13017-024-00570-7] [Medline: 39716296]
19. Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. arXiv. Preprint posted online on Feb 8, 2022. [doi: 10.48550/arXiv.2109.01652]
20. Mistral NeMo. Mistral AI. 2024 Jul 18. URL: https://mistral.ai/news/mistral-nemo [accessed 2026-04-07]
21. Qwen. Qwen2.5 technical report. arXiv. Preprint posted online on Jan 3, 2025. [doi: 10.48550/arXiv.2412.15115]
22. DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMS via reinforcement learning. arXiv. Preprint posted online on Jan 4, 2026. [doi: 10.48550/arXiv.2501.12948]
23. Gemma Team. Gemma 2: improving open language models at a practical size. arXiv. Preprint posted online on Oct 2, 2024. [doi: 10.48550/arXiv.2408.00118]
24. Blackwell RE, Barry J, Cohn AG. Towards reproducible LLM evaluation: quantifying uncertainty in LLM benchmark scores. arXiv. Preprint posted online on Jun 27, 2025. [doi: 10.48550/arXiv.2410.03492]

XSL•FO
RenderX

25. LLM-data-extractor-from-reports-ollama-. GitHub. URL: https://github.com/aschipper/LLM-data-extractor-from-reports-ollama- [accessed 2026-04-07]
26. Llm_extractinator-v0.4.2. Zenodo. URL: https://zenodo.org/records/15089764 [accessed 2026-04-07]
27. XGBoost documentation. DMLC XGBoost. URL: https://xgboost.readthedocs.io [accessed 2026-04-07]
28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Proc 32nd Conf Neural Inf Process Syst NeurIPS 2018 2018 [FREE Full text]
29. CatBoost Encoder. URL: https://contrib.scikit-learn.org/category_encoders/catboost.html [accessed 2026-04-07]
30. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework (version 361). arXiv. Preprint posted online on Jul 25, 2019. [doi: 10.1145/3292500.3330701]
31. IQVIA netherlands. URL: https://www.iqvia.com/nl-nl/locations/netherlands
32. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. J Am Med Inform Assoc 2022 Jan 29;29(3):559-575. [doi: 10.1093/jamia/ocab236] [Medline: 34897469]
33. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711. [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
34. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016 May;69:53-61. [doi: 10.1016/j.artmed.2016.03.003] [Medline: 27085847]
35. Choi A, Kim C, Ryoo J, et al. A pediatric emergency prediction model using natural language process in the pediatric emergency department. Sci Rep 2025 Jan 28;15(1):3574. [doi: 10.1038/s41598-025-87161-x] [Medline: 39875462]
36. Nunez JJ, Leung B, Ho C, Bates AT, Ng RT. Predicting the survival of patients with cancer from their initial oncology consultation document using natural language processing. JAMA Netw Open 2023 Feb 1;6(2):e230813. [doi: 10.1001/jamanetworkopen.2023.0813] [Medline: 36848085]
37. Shankar R, Bundele A, Mukhopadhyay A. Natural language processing of electronic health records for early detection of cognitive decline: a systematic review. npj Digit Med 2025 Mar 1;8(1):133. [doi: 10.1038/s41746-025-01527-z] [Medline: 40025194]
38. Zhang X, Wang Y, Jiang Y, Pacella CB, Zhang W. Integrating structured and unstructured data for predicting emergency severity: an association and predictive study using transformer-based natural language processing models. BMC Med Inform Decis Mak 2024 Dec 5;24(1):372. [doi: 10.1186/s12911-024-02793-9] [Medline: 39633370]
39. Chen MC, Huang TY, Chen TY, Boonyarat P, Chang YC. Clinical narrative-aware deep neural network for emergency department critical outcome prediction. J Biomed Inform 2023 Feb;138:104284. [doi: 10.1016/j.jbi.2023.104284] [Medline: 36632861]
40. Iscoe M, Socrates V, Gilson A, et al. Identifying signs and symptoms of urinary tract infection from emergency department clinical notes using large language models. Acad Emerg Med 2024 Jun;31(6):599-610. [doi: 10.1111/acem.14883] [Medline: 38567658]
41. Wu H, Wang M, Wu J, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. npj Digit Med 2022 Dec 21;5(1):186. [doi: 10.1038/s41746-022-00730-6] [Medline: 36544046]
42. Japkowicz N, Stephen S. The class imbalance problem: a systematic study1. IDA 2002;6(5):429-449. [doi: 10.3233/IDA-2002-6504]
43. Masayoshi K, Hashimoto M, Yokoyama R, Toda N, Uwamino Y, Fukuda S, et al. EHR-MCP: real-world evaluation of clinical information retrieval by large language models via model context protocol. arXiv. Preprint posted online on Sep 19, 2025. [doi: 10.48550/arXiv.2509.15957]
44. Mahajan P, Basu T, Pai CW, et al. Factors associated with potentially missed diagnosis of appendicitis in the emergency department. JAMA Netw Open 2020 Mar 2;3(3):e200612. [doi: 10.1001/jamanetworkopen.2020.0612] [Medline: 32150270]
45. Drake FT, Flum DR. Improvement in the diagnosis of appendicitis. Adv Surg 2013;47:299-328. [doi: 10.1016/j.yasu.2013.03.003] [Medline: 24298858]
46. Gelpke K, Hamminga JTH, van Bastelaar JJ, et al. Reducing the negative appendectomy rate with the laparoscopic appendicitis score; a multicenter prospective cohort and validation study. Int J Surg 2020 Jul;79:257-264. [doi: 10.1016/j.ijsu.2020.04.041] [Medline: 32387211]
47. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inf Assoc 2011;18(2):181-186. [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
48. Nemkova P, Adhikari A, Pearson M, Sadu VK, Albert MV. Cross-lingual stability and bias in instruction-tuned language models for humanitarian NLP. arXiv. Preprint posted online on Oct 26, 2025. [doi: 10.48550/arXiv.2510.22823]
49. Vanroy B. Language resources for Dutch large language modelling. arXiv. Preprint posted online on Dec 20, 2023. [doi: 10.48550/arXiv.2312.12852]
50. Intrator Y, Halfon M, Goldenberg R, et al. Breaking the language barrier: can direct inference outperform pre-translation in multilingual LLM applications? Proc 2024 Conf North Am Chapter Assoc Comput Linguist 2024:829-844. [doi: 10.18653/v1/2024.naacl-short.75]

XSL•FO
RenderX

Abbreviations

AAP: acute abdominal pain
AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations From Transformers
ED: emergency department
EHR: electronic health record
HIVE: History, Intake, Vitals, Examination
LLM: large language model
NLP: natural language processing
SHAP: Shapley Additive Explanations
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: Extreme Gradient Boosting

Edited by A Benis; submitted 30.Jul.2025; peer-reviewed by H Xu, H Li, R Luo; revised version received 06.Mar.2026; accepted 06.Mar.2026; published 30.Apr.2026.

Please cite as:
Schipper A, Belgers P, O'Connor RD, van de Wouw L, Builtjes L, Bosma JS, Kusters R, Kurstjens S, Rutten M, van Ginneken B
Large Language Model Automated Extraction of Clinical Signs and Symptoms From Emergency Department Reports for Machine Learning Prediction Models: Development and Validation Study
JMIR Med Inform 2026;14:e81500
URL: https://medinform.jmir.org/2026/1/e81500
doi:10.2196/81500

© Anoeska Schipper, Peter Belgers, Rory David O'Connor, Lieke van de Wouw, Luc Builtjes, Joeran S Bosma, Ron Kusters, Steef Kurstjens, Matthieu Rutten, Bram van Ginneken. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Approval of AI-Based Medical Devices in China From 2020 to 2025: Retrospective Analysis

Lingli Zhang[1], PhD; Jianzhou Yan[1,2], PhD

[1]School of International Pharmaceutical Business, China Pharmaceutical University, 639 Longmian Avenue, Jiangning District, Nanjing, China
[2]Research Center of National Drug Policy and Ecosystem, China Pharmaceutical University, Nanjing, China

Corresponding Author:
Jianzhou Yan, PhD
School of International Pharmaceutical Business, China Pharmaceutical University, 639 Longmian Avenue, Jiangning District, Nanjing, China

Abstract

Background: Artificial intelligence–based medical devices (AIMDs) have emerged as transformative technologies in modern health care. However, comprehensive analysis of recent approval trends and characteristics of AIMDs in China remains limited.

Objective: This study aimed to provide an up-to-date overview of AIMDs approved in China up to June 2025.

Methods: We conducted a search of the Drugdataexpy database to identify AIMDs approved up to June 30, 2025, using artificial intelligence–related keywords in the "structural composition" and "intended use" fields. After manual verification and exclusion of non-AIMDs, we collected key characteristics, including name, manufacturer, approval date, risk class, clinical evaluation pathway, medical specialty, data source, review pathway, and algorithm type. Statistical analysis encompassed descriptive statistics and trend analysis. We used the Fisher exact test and Pearson chi-square test to assess the associations between risk class and categorical variables.

Results: A total of 154 AIMDs were identified since the first approval in 2020, with annual approvals increasing from 9 in 2020 to 45 in 2024 (a 49.53% compound annual growth rate), although the 20 approvals in the first half of 2025 suggest a potential moderation in pace. Most AIMDs (123/154, 79.9%) were categorized as class III, and the risk class was significantly associated with approval year (P=.03), manufacturer location (P=.03), and medical specialty (P=.004). Of the 123 class III devices, 19 (15.4%) were approved through innovation review, and 2 (1.6%) each were approved through priority and emergency approval. Deep learning was the dominant algorithm (143/154, 92.9%). Radiology dominated the field (106/154, 68.8%), with computed tomography serving as the primary data source (96/154, 62.3%), particularly for applications in pulmonary nodule detection and cardiovascular assessment. Clinical trials were the primary evaluation pathway for 76.6% (118/154) of all AIMDs. This approach was predominant for class III devices (116/123, 94.3%), whereas most class II devices (21/31, 67.7%) used a clinical exemption pathway. Market concentration was evident, with the top 4 manufacturers accounting for 38.3% (59/154) of all approvals and geographically clustered in major innovation hubs such as Beijing, Shanghai, Shenzhen, and Hangzhou.

Conclusions: China's AIMD ecosystem is experiencing growth, heavily focused on radiology and computed tomography–based solutions within a risk-proportionate regulatory framework. The market is characterized by significant manufacturer and geographic concentration.

(JMIR Med Inform 2026;14:e85538) doi:10.2196/85538

KEYWORDS

medical device; artificial intelligence; AI; artificial intelligence–based medical device; AI-based medical device; medical device approval; regulatory science; China; deep learning; radiology

Introduction

Artificial intelligence (AI) has emerged as a transformative force in modern health care, offering unprecedented capabilities in data analysis, pattern recognition, and clinical decision support [1,2]. Among its many applications, AI-based medical devices (AIMDs) have gained substantial traction, particularly in medical imaging, diagnostics, and predictive analytics [3]. These technologies promise to enhance diagnostic accuracy; reduce physician workload; and, ultimately, improve patient outcomes [4].

Worldwide, regulatory authorities have responded to the rapid development of AIMDs by establishing frameworks to ensure their safety, efficacy, and clinical utility [5-8]. In the United States, the Food and Drug Administration approved 691 AI- and machine learning (ML)–based medical devices by October 2023, with most cleared through the 510(k) pathway and concentrated in radiology and cardiology [9,10]. Similarly,

XSL•FO
RenderX

Europe has witnessed a substantial increase in CE-marked AIMDs, with approval numbers growing significantly since 2015 [11].

China, as a major player in both AI innovation and health care digitalization, has also made significant strides in this domain [12,13]. The National Medical Products Administration (NMPA) approved its first AIMD in 2020. Since then, the number of approvals has grown steadily, reaching 59 by mid-2023, all categorized as class III [14]. In China, medical devices are categorized into 3 classes based on increasing risk: class I (low risk), class II (medium risk), and class III (high risk). This classification dictates the regulatory pathway for market approval. Class III devices, deemed to pose the highest risk, require centralized registration approval from the NMPA itself and are subject to the most stringent review of clinical evidence. Class II devices may receive approval from provincial-level medical product administrations. AIMDs in China are typically categorized as class II or class III. The classification is determined by the algorithm's maturity and its role in clinical decision-making. Specifically, AIMDs with lower algorithm maturity are managed as class III devices if they are intended for auxiliary decision-making. Conversely, if they are intended for non–decision support tasks, they are managed as class II devices. In China, the clinical evaluation pathway for medical devices consists of 3 types: clinical trials, clinical comparison with similar products, and exemption from clinical evaluation. The clinical exemption pathway relies on the *Catalogue of Medical Devices Exempt From Clinical Evaluation* issued by the NMPA [15].

In recent years, the NMPA has actively refined and optimized its review processes for AIMDs. Key efforts include issuing evaluation guidelines, clarifying regulatory classifications, developing clinical evaluation guidance for specific product types, and introducing comprehensive support measures spanning the entire product life cycle [16-19]. These advancements have spurred a new wave of submissions and approvals, fostering a more structured and efficient ecosystem for AIMD market entry. Despite this progress, the literature to date primarily documents AIMD approvals in China up to 2023. There remains a lack of comprehensive analyses covering the most recent developments, particularly those in 2024 and 2025—a period likely marked by regulatory maturation and increased approvals. This study adopted innovation diffusion theory [20] as its primary analytical lens. This theory posits that the adoption rate and pattern of an innovation are influenced by its perceived attributes, including relative advantage, compatibility, complexity, trialability, and observability, within a specific social system.

Therefore, this study aimed to provide an up-to-date analysis of AIMDs approved in China as of June 2025. By examining trends in approval numbers, device characteristics, and regulatory pathways, it may offer insights into the evolving AIMD landscape in China and its implications for future health care innovation.

Methods

Data Source and Search Strategy

To identify AIMDs approved up to June 30, 2025, we searched the Drugdataexpy database [21]. As this is a Chinese-language database, the keyword search was performed using the Chinese equivalents of the following English-language terms in the database fields for "structural composition" and "indications" or "intended use": "deep learning," "machine learning," "neural network," "artificial intelligence," or "algorithm." The NMPA's official website, while providing free access to regulatory announcements, lacks keyword-based filtering capabilities in its search function, thereby limiting efficient retrieval of AIMDs.

This study included AIMDs regulated under the Guidelines for Registration Review of Artificial Intelligence Medical Devices [16], encompassing stand-alone software as a medical device and software in a medical device (SiMD) where AI is a core component. After manual review of full-text descriptions, we excluded records that did not meet the AIMD definition. We also removed entries corresponding solely to routine administrative renewals without changes to the device's core AI software or intended use. Substantive approvals for major changes or modifications to existing devices (eg, the addition of new AI functionality) were retained. Details of the search strategy and process are provided in Multimedia Appendix 1. To minimize omissions, we reran the search for each retained product name to capture devices manufactured by other companies. Our results were cross-validated against the AIMDs approved before July 2023 as listed in the study by Liu et al [14]. Ambiguous entries were checked against the corresponding NMPA evaluation reports or the manufacturers' websites.

Data Extraction

For each eligible AIMD, we extracted the approval number, device name, manufacturer, approval date, structural composition, intended use, risk class, clinical evaluation pathway (clinical trials, clinical comparison with similar products, or exemption from clinical evaluation), and review pathway (innovative, priority, or emergency special review vs standard). From these fields, we derived the city of the manufacturer, medical specialty, body area, and data source of the AIMD. The medical specialty for each AIMD was assigned based on the device's registered intended use and its core function. For devices with applications relevant to multiple specialties, the specialty was determined by the primary clinical context and decision support role stated in the indications for use. The body area for each device was functionally classified into nonoverlapping categories based on the primary anatomical structure and type of clinical analysis specified in the device's intended use: "heart" referred to devices targeting cardiac muscle, chambers, or electrophysiology (eg, electrocardiogram analysis); "cardiovascular" referred to devices focused on the vascular system (eg, computed tomography [CT] or magnetic resonance angiography); "lung" referred to devices analyzing lung parenchyma (eg, nodule detection); and "thorax" referred to devices assessing nonparenchymal structures within the thoracic cavity (eg, rib fracture detection). Algorithm type was determined through a tiered review: it was first identified from

XSL•FO
RenderX

the "structural composition" and "intended use" fields. If not specified there, the full registration dossier was consulted for explicit technical descriptions.

To facilitate downstream data processing, minor lexical variants in device names referring to the same product—such as "medical image" vs "imaging"—were standardized to a single, consistent terminology. For the analysis of market concentration, "manufacturer" referred to the ultimate parent corporate entity. AIMDs approved under the legal names of subsidiaries belonging to the same corporate group were aggregated under the parent company's name. A customized extraction form was piloted on 10 AIMDs and then applied independently by 2 reviewers (LZ and JY). Any disagreements were resolved through discussion until consensus was reached.

Statistical Analysis

Descriptive statistics were used to summarize the extracted variables and assess temporal trends. We calculated the compound annual growth rate for annual approvals from 2020 to 2024 and applied the Mann-Kendall trend test to assess the significance of the growth trend. For AIMDs that used multiple data sources, each modality was counted separately in the subsequent analysis of data source distribution. Associations between categorical variables (approval year, medical specialty, and manufacturer's city) and AIMD risk class were evaluated. The Fisher exact test was used for analyses involving variables with low expected cell counts (approval year and medical specialty), whereas the Pearson chi-square test was used for other comparisons. All tests were 2 sided, with $P<.05$ indicating statistical significance. Analyses were conducted using SPSS Statistics for Windows (version 27.0; IBM Corp).

Ethical Considerations

This study was a secondary analysis of publicly available regulatory records retrieved from a database. No data from human participants or identifiable personal information were used, and no primary data collection was conducted. The data are anonymized and publicly accessible. In accordance with the institutional policy of China Pharmaceutical University, ethics board approval was not required for this study. This study complies with China's Data Security Law [22] and Personal Information Protection Law [23].

Results

Overview of AIMD Approvals in China

As of June 30, 2025, a total of 154 AIMDs had been approved in China since the initial authorization in 2020. Figure 1 shows the inclusion flowchart for AIMDs approved in China. All devices were intended for use by health care professionals, with none designed for direct patient use. The annual approval numbers exhibited a clear upward trend from 2020 to 2024: 9 in 2020, a total of 16 in 2021, a total of 26 in 2022, a total of 38 in 2023, and 45 in 2024. From 2020 to 2024, the compound annual growth rate was 49.53%, and the trend test confirmed a statistically significant increasing trend in annual approvals ($P<.001$). Data for the first half of 2025 show 20 approvals, indicating that the trend may be entering a phase of moderation or plateau (Figure 2). Detailed information on these AIMDs sorted by approval date is provided in Multimedia Appendix 2.

Among the 154 AIMD approvals, 15 device types had at least 3 approvals (Table 1). Pulmonary nodule CT image auxiliary detection software ranked first with 14 (9.1%) approvals. It was followed by coronary artery CT angiography image vessel stenosis auxiliary assessment software with 8 (5.2%) approvals. Coronary artery CT fractional flow reserve calculation software and intracranial hemorrhage CT image auxiliary triage software tied for third, each with 7 (4.5%) approvals. Diabetic retinopathy fundus image auxiliary diagnosis software and pneumonia CT image auxiliary triage and assessment software each had 6 (3.9%) approvals. Pediatric hand x-ray image auxiliary bone age assessment software had 4 (2.6%) approvals. A total of 8 additional products each obtained 3 (1.9%) approvals, spanning various applications from CT image processing and electrocardiogram analysis to specialized endoscopic, angiographic, and x-ray–based diagnostic tools.

XSL•FO
RenderX

Figure 1. Inclusion flowchart for approved artificial intelligence (AI)–based medical devices (AIMDs) in China (2020 - 2025).

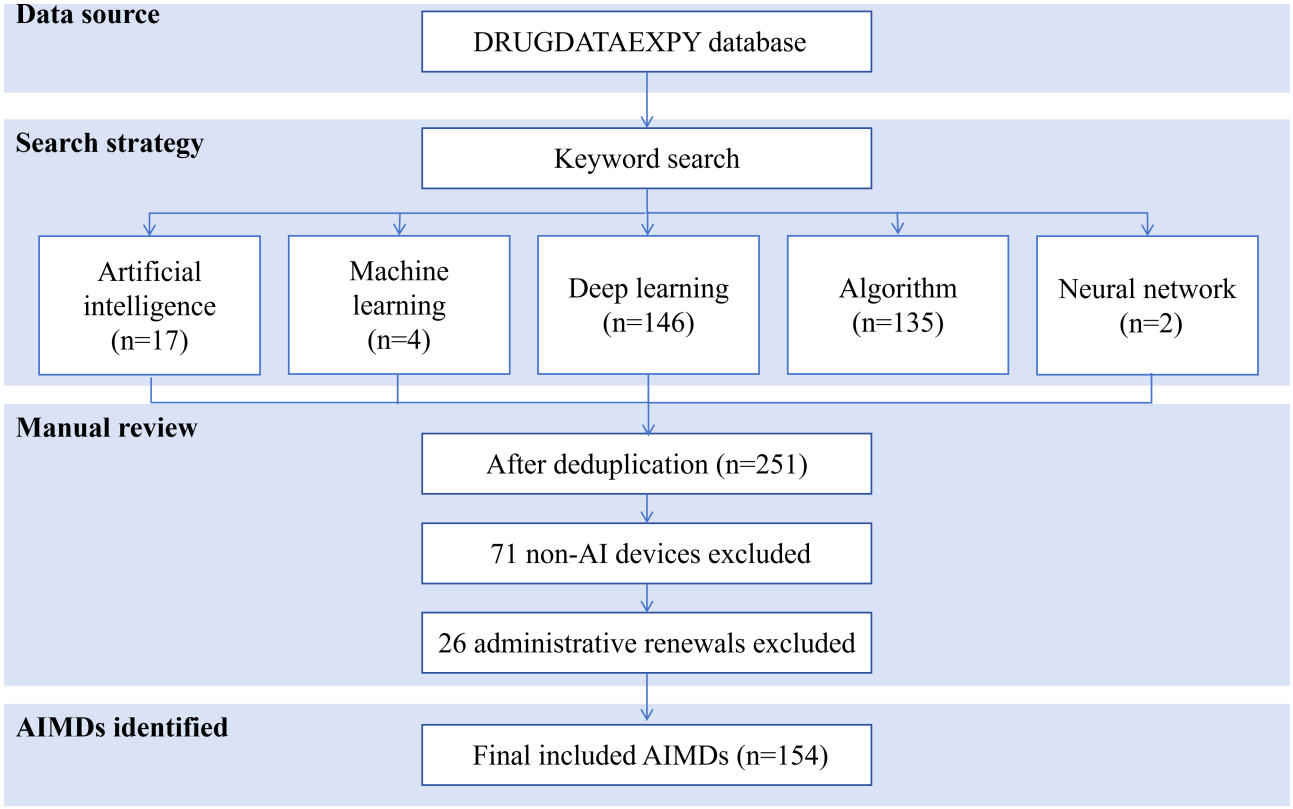
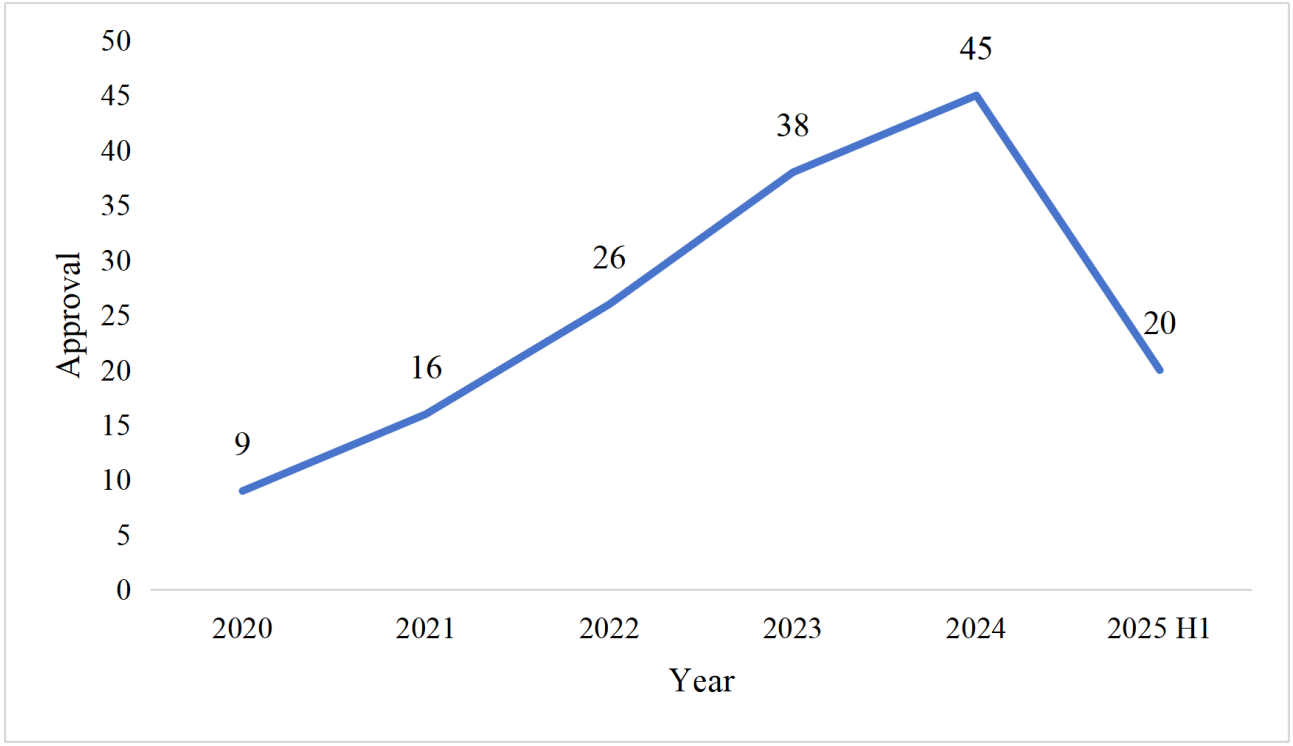


Figure 2. Trends in annual approvals of artificial intelligence–based medical devices in China (2020-2025; n=154). H1: first half of the year.



XSL•FO
RenderX

Table . Artificial intelligence–based medical devices with ≥3 approvals in China (2020 - 2025; n=154).

Device description	Approvals, n (%)
Pulmonary nodule CT[a] image auxiliary detection software	14 (9.1)
Coronary artery CT angiography image vessel stenosis auxiliary assessment software	8 (5.2)
Coronary artery CT fractional flow reserve calculation software	7 (4.5)
Intracranial hemorrhage CT image auxiliary triage software	7 (4.5)
Diabetic retinopathy fundus image auxiliary diagnosis software	6 (3.9)
Pneumonia CT image auxiliary triage and assessment software	6 (3.9)
Pediatric hand x-ray image auxiliary bone age assessment software	4 (2.6)
CT image processing software	3 (1.9)
Dynamic electrocardiogram analysis software	3 (1.9)
Electrocardiogram analysis software	3 (1.9)
Electronic lower gastrointestinal endoscopy image auxiliary detection software for intestinal polyps	3 (1.9)
Head and neck CT angiography image auxiliary assessment software	3 (1.9)
Breast x-ray image auxiliary detection software	3 (1.9)
Medical image processing software	3 (1.9)
Rib fracture CT image auxiliary detection software	3 (1.9)

[a]CT: computed tomography.

Data Source of AIMDs

As shown in Table 2, CT was the predominant data source for these AIMDs, identified in 62.3% (96/154) of approvals. This markedly exceeded other data sources, including x-ray (18/154, 11.7%), magnetic resonance imaging (12/154, 7.8%), endoscopy (9/154, 5.8%), electrocardiography (9/154, 5.8%), and fundus photography (9/154, 5.8%).

Table . Data sources used in approved artificial intelligence–based medical devices in China (2020 - 2025). Devices using multiple data sources were counted under each relevant modality. Therefore, the sum of counts exceeds the total number of devices (n=154).

Data source	Devices, n (%)
CT[a]	96 (62.3)
X-ray	18 (11.7)
MRI[b]	12 (7.8)
Electrocardiogram	9 (5.8)
Endoscopy	9 (5.8)
Fundus photography	9 (5.8)
Microscopic images	6 (3.9)
Ultrasound	3 (1.9)
Polysomnography	1 (0.6)

[a]CT: computed tomography.

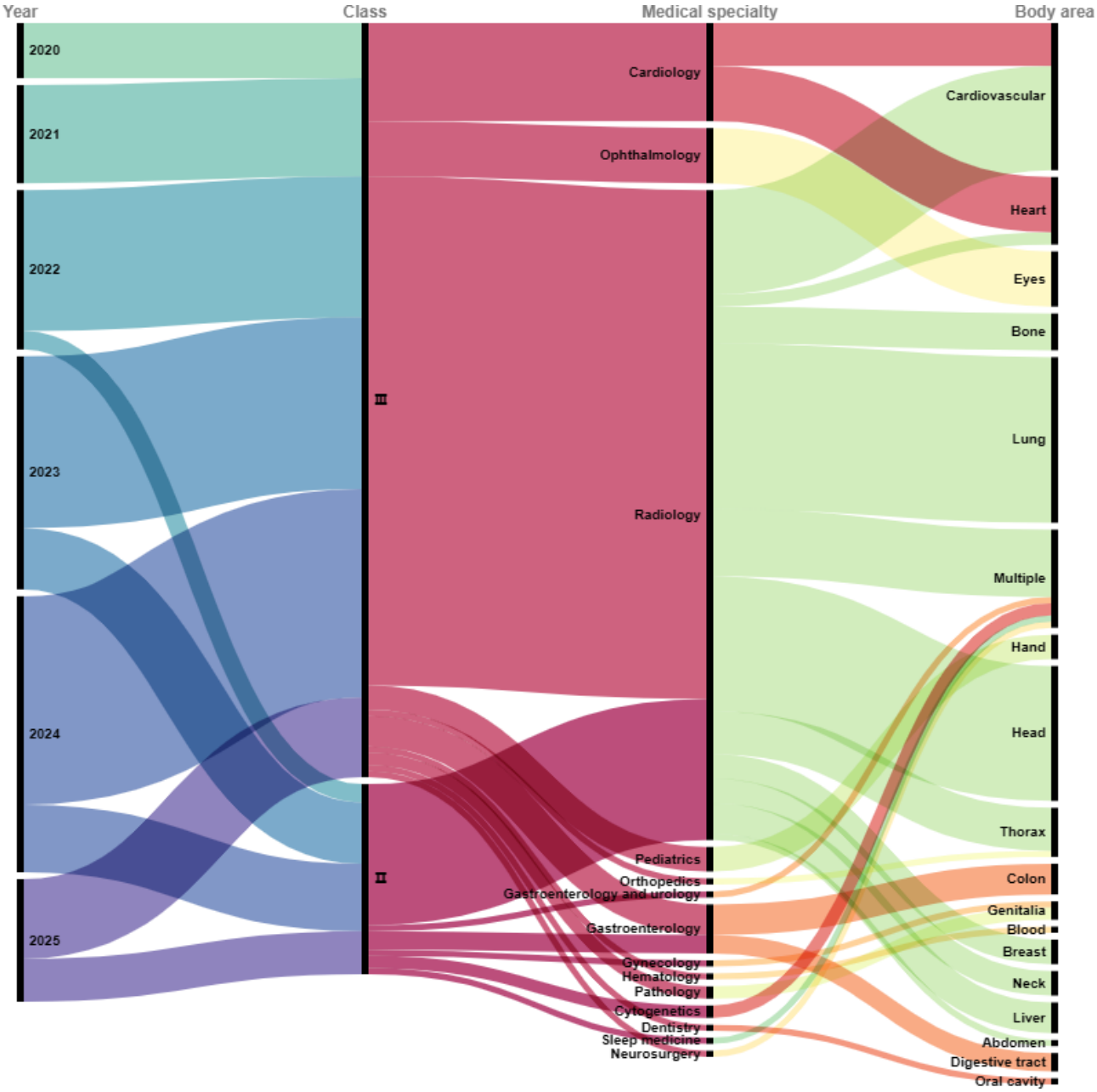
[b]MRI: magnetic resonance imaging.

Risk Class and Correlation Analysis

The vast majority of AIMDs (123/154, 79.9%) were categorized as class III, with 20.1% (31/154) categorized as class II (Figure 3). Significant associations were found between risk class and 3 key variables: the year of approval (P=.03), the manufacturer's city (P=.03), and the medical specialty (P=.004). In 2020 and 2021, only class III AIMDs were approved. After 2022, the share of class II AIMDs rose each year, from 11.5% (3/26) in 2022 to 35% (7/20) in 2025. The proportion of class III AIMDs was markedly higher in Hangzhou (16/16, 100%) and Beijing (47/54, 87%) than in other cities. The distribution of risk classes varied substantially across medical specialties. Ophthalmology and cardiology exhibited a 100% (16/16) class III rate, whereas 21.7% (23/106) of AIMDs in radiology were class II.

XSL•FO
RenderX

Figure 3. Sankey diagram of approved artificial intelligence–based medical devices in China (2020-2025).



Review Pathway and Algorithm

Of the 154 AIMDs, 23 (14.9%) gained market access through national-level special review pathways, which were exclusively for class III devices. Specifically, of the 123 class III devices, 19 (15.4%) were approved via innovation review, 2 (1.6%) were approved via priority review, and 2 (1.6%) were approved via emergency approval, constituting a combined 18.7% (23/123) of all class III approvals. The remaining 85.1% (131/154) of the devices followed the standard review process. Of the 154 AIMDs, 143 (92.9%) used deep learning architectures; 3 (1.9%) used classic ML methods; and for the remaining 8 (5.2%) devices, the algorithmic approach was not disclosed in the available registration materials.

Clinical Evaluation Pathways

Analysis of the clinical evaluation pathways for the 154 AIMDs (Table 3) showed that most (n=118, 76.6%) submitted clinical trial reports as their primary evidence. The remaining devices were approved either through an exemption from clinical evaluation (n=23, 14.9%) or via a clinical comparison with similar products already available on the market (n=13, 8.4%).

Of the 123 class III devices, 116 (94.3%) used clinical trials, whereas clinical comparison with similar products and clinical exemption were less common (n=5, 4.1% and n=2, 1.6% of devices, respectively). In contrast, among the 31 class II devices, clinical exemption was the most frequent pathway (n=21, 67.7%), followed by clinical comparison with similar products (n=8, 25.8%) and clinical trials (n=2, 6.5%).

XSL•FO
RenderX

Table . Clinical evaluation pathways for approved artificial intelligence–based medical devices in China (2020 - 2025; n=154).

	Total, n (%)	Class III devices (n=123), n (%)	Class II devices (n=31), n (%)
Clinical trial	118 (76.6)	116 (94.3)	2 (6.5)
Exemption from clinical evaluation	23 (14.9)	2 (1.6)	21 (67.7)
Clinical comparison with similar products	13 (8.4)	5 (4.1)	8 (25.8)

Medical Specialty and Body Area

The 154 AIMDs spanned 14 medical specialties. Radiology dominated with 68.8% (n=106) of approvals, followed by cardiology (n=16, 10.4%), ophthalmology (n=9, 5.8%), and gastroenterology (n=9, 5.8%). The remaining AIMDs covered pediatrics, pathology, cytogenetics, sleep medicine, orthopedics, neurosurgery, hematology, gynecology, urology, and dentistry (Table 4).

The most common body areas for these AIMDs were the lung (27/154, 17.5%), the cardiovascular system (24/154, 15.6%), and the head (22/154, 14.3%), with the heart accounting for 7.1% (11/154). Additionally, 10.4% (16/154) of the AIMDs were used across multiple body areas (Table 5).

Table . Distribution of artificial intelligence–based medical devices by medical specialty in China (2020 - 2025; n=154).

Medical specialty	Devices, n (%)
Radiology	106 (68.8)
Cardiology	16 (10.4)
Ophthalmology	9 (5.8)
Gastroenterology	9 (5.8)
Pediatrics	4 (2.6)
Others	10 (6.5)

Table . Distribution of artificial intelligence–based medical devices by target body area in China (2020 - 2025; n=154).

Body area	Devices, n (%)
Lung	27 (17.5)
Cardiovascular system	24 (15.6)
Head	22 (14.3)
Multiple	16 (10.4)
Heart	11 (7.1)
Eyes	9 (5.8)
Thorax	8 (5.2)
Bone	6 (3.9)
Liver	5 (3.2)
Colon	5 (3.2)
Neck	4 (2.6)
Hand	4 (2.6)
Breast	4 (2.6)
Others	9 (5.8)

Manufacturer and Geographic Distribution

We identified 67 manufacturers in total, with a high degree of market concentration. The top 4 companies alone produced 38.3% (59/154) of the AIMDs: Yukun and United Imaging led with 11% (17/154) of the devices each, followed by Infervision with 9.1% (14/154) and Deepwise with 7.1% (11/154). In contrast, 48 manufacturers had only 1 or 2 products each. Manufacturers with more than 2 approved AIMDs are shown in Table 6.

Geographically, of the 154 approved AIMDs, 54 (35.1%) were from manufacturers headquartered in Beijing, 32 (20.8%) were from Shanghai, 18 (11.7%) were from Shenzhen, and 16 (10.4%) were from Hangzhou (Table 7). Thus, these 4 cities accounted for 77.9% (n=120) of all the approved AIMDs. Beijing was home to Yukun and Infervision, Shanghai was home to United

XSL•FO
RenderX

Imaging, and Hangzhou was home to Deepwise, which were precisely the top 4 companies listed above. Shenzhen hosted a broader ecosystem, including Carewell, Tencent Smart Healthcare, RuiXin, and Edan Instruments, but none brought more than 10 AIMDs to market.

Table . Manufacturers with ≥2 approved artificial intelligence–based medical devices in China (2020‐2025; n=154).

Manufacturer	Devices, n (%)
Yukun	17 (11.0)
United Imaging	17 (11.0)
Infervision	14 (9.1)
Deepwise	11 (7.1)
Yizhun	5 (3.2)
SenseTime	4 (2.6)
Carewell	4 (2.6)
Tongxin	3 (1.9)
Tencent Smart Healthcare	3 (1.9)
Siemens	3 (1.9)
RuiXin	3 (1.9)
Neusoft	3 (1.9)
Longwood Valley Medical	3 (1.9)
InnoVision	3 (1.9)
Huiyi Huiying	3 (1.9)
Edan Instruments	3 (1.9)
YiTu	2 (1.3)
Visionary Intelligence	2 (1.3)
Suzhou Volexcloud	2 (1.3)
Shukun	2 (1.3)
Keya Medical	2 (1.3)
Shenzhen Dr. Brain Technologies	2 (1.3)
BioMind	2 (1.3)

XSL•FO
RenderX

Table . Cities of manufacturers of approved artificial intelligence–based medical devices in China (2020 - 2025; n=154).

Cities	Devices, n (%)
Beijing	54 (35.1)
Shanghai	32 (20.8)
Shenzhen	18 (11.7)
Hangzhou	16 (10.4)
Suzhou	4 (2.6)
Xiamen	4 (2.6)
Lishui	4 (2.6)
Shenyang	3 (1.9)
Kunming	3 (1.9)
Changsha	2 (1.3)
Wuhan	2 (1.3)
Guangzhou	2 (1.3)
Chengdu	2 (1.3)
Zhengzhou	1 (0.6)
Wuxi	1 (0.6)
Nanning	1 (0.6)
Nanjing	1 (0.6)
Nanchang	1 (0.6)
Longyan	1 (0.6)
Hefei	1 (0.6)
Guiyang	1 (0.6)

Discussion

Principal Findings

This study provided an overview of AIMDs in China up to June 2025. A total of 154 AIMDs were approved, and annual approvals showed an upward trend. Most of these AIMDs (123/154, 79.9%) were class III and relied on deep learning. They were concentrated in radiology, particularly CT-based; produced by a small number of manufacturers; and located in a few cities.

China did not approve its first AIMD until 2020, later than both the United States and Europe, where approvals for AIMDs occurred as early as 2015 [10] despite variations in the definitions of AIMDs. The number of AIMD approvals in China remains lower than that in the United States and Europe. However, we have observed a rapid increase in AIMD approvals in China, reflecting technological advancements and a supportive regulatory environment.

Our analysis, extending the timeline to mid-2025, built on and updated the foundational work of Liu et al [14]. While confirming the continued dominance of radiology and class III devices previously reported, our data revealed new dynamics. Specifically, we observed a marked increase in the absolute number of approvals and a significant shift with the emergence of class II AIMDs after 2022—a trend that was not fully evident in the earlier data. Furthermore, the sustained market concentration among top manufacturers and in key geographic hubs suggests that these early structural characteristics consolidated rather than dispersed as the market grew. These comparisons highlight the rapid evolution of this field and underscore the value of longitudinal tracking to inform regulatory and market strategies. It should be noted that our counts for 2021 to 2022 were slightly higher than those reported by Liu et al [14]. We identified additional AIMDs, including image processing software and CT fractional flow reserve calculation software. A primary reason for this discrepancy is a fundamental difference in data scope. Our dataset encompassed both centrally approved class III devices and provincially approved class II devices, whereas Liu et al [14] primarily focused on class III approvals granted by the NMPA. Additionally, differences can be attributed to distinct search strategies: our approach incorporated AI-related keywords—such as “deep learning” and “algorithm”—within the “structural composition” and “intended use” fields, whereas their search relied on device-naming conventions, using terms such as “computer-assisted detection” and “computer-assisted triage.”

A prime example of an AIMD category with significant clinical penetration in China is pulmonary nodule CT image auxiliary detection software. This category had 14 approved variants from multiple manufacturers, making it the most widespread AIMD type. According to manufacturer data, United Imaging has deployed its solution in over 2000 hospitals nationwide. Their

XSL•FO
RenderX

clinical evaluations indicate that the software can increase the detection rate of pulmonary nodules by approximately 32% and improve radiologists' reading efficiency by approximately 26% compared with conventional methods [24].

Regarding risk classification, most Chinese AIMDs are designated as class III, whereas all US AIMDs are cleared as class II [25] and Europe has only approximately 1% of AIMDs in class III [11]. The large share of class III in China suggests that many AIMDs were treated as high risk and required the strictest review. From 2020 to 2021, only class III AIMDs were observed; from 2022 onward, this pattern shifted, with a steady influx of class II products. The observed change can be attributed to the *Principles for the Classification Defining of AI-Based Medical Software Products* issued by the NMPA in 2021 [26], which applies to both software as a medical device and SiMD. For SiMD, the assigned risk class applies to the product as a whole, reflecting a holistic regulatory assessment of the AI functionality's clinical role within the device. This guideline allowed AIMDs with mature algorithms that provided only non–real-time assistance—such as measurement, marking, or workflow optimization—to be placed in class II. Previously, any software bearing AI technology had been categorized as class III. The significant association between risk classification and the manufacturer's city likely reflects regional policy support and the clustering of expertise in cities such as Beijing and Hangzhou, where high-risk AIMD development appears concentrated.

Our analysis further revealed that the vast majority of approved AIMDs were supported by premarket clinical evidence through clinical trials or clinical comparison with similar products. Exemption from clinical evaluation was granted for only 23 AIMDs, 21 (91.3%) of which were class II and predominantly consisted of image processing software. This pattern reflects the practical implementation of a risk-based regulatory framework: higher-risk AIMDs generally necessitate more robust premarket clinical validation, while pathways such as exemption or comparison are more frequently applied to lower-risk AIMDs to facilitate efficient market access. Nevertheless, reliance on exemptions and comparative assessments for lower-risk approvals raises concerns about the sufficiency and transparency of premarket evidence. To strengthen the regulatory framework, we recommend clarifying exemption eligibility criteria and tightening the evidentiary standards for demonstrating equivalence.

We did not identify any safety-related recalls among the 154 AIMDs approved as of June 2025. A recent study of Food and Drug Administration–authorized AI and ML devices found that 5.2% were linked to adverse events and 5.8% were recalled, primarily due to software issues [27]. The lack of similar public reports in China may stem from its later market start and differences in data transparency. Accordingly, establishing a robust postmarket surveillance system is essential to detect risks arising from algorithm or workflow changes, and it requires commitments from regulators and manufacturers to ensure the ongoing safety of AIMDs.

Radiology dominated China's AIMD landscape, with an emphasis on CT-based applications, reflecting alignment with the nation's health care needs and its existing technological infrastructure. This focus is evident in the concentration of devices targeting pulmonary nodules and cardiovascular conditions, both representing substantial disease burden. Pulmonary nodules are clinically significant as early indicators of lung cancer, which remains the most common cancer type and the leading cause of cancer death in China [28]. Similarly, cardiovascular diseases rank as the leading cause of overall mortality, accounting for approximately 2 out of every 5 deaths [29].

The high proportion of radiology-related AIMDs approved in China aligns with the trend observed in the United States [25]. Medical imaging modalities such as CT, magnetic resonance imaging, and x-ray generate structured, digital data well suited for AI applications. Among these modalities, CT images have formed the basis for most AIMDs developed in China. The widespread availability of CT scanners in Chinese health care facilities, together with relatively uniform imaging protocols, has accelerated the growth of AI diagnostic software based on CT.

Market concentration emerged as another notable finding, with a limited number of manufacturers accounting for a large share of approved AIMDs. Notably, these dominant manufacturers were primarily headquartered in major cities such as Beijing, Shanghai, and Hangzhou. This market dominance appears to stem from the convergence of several reinforcing factors. These include the concentration of technical and regulatory expertise, preferential access to high-quality medical data through partnerships with leading hospitals, and the availability of substantial venture capital in these regions [30,31].

The observed characteristics of China's AIMD market, specifically its pronounced concentration in clinical specialties and geographic regions, were interpreted through the lens of innovation diffusion theory. First, radiology dominated the field, especially CT-based applications, illustrating innovations with high compatibility through seamless integration into digital imaging workflows and a clear relative advantage in improving efficiency during high-volume screenings, which promotes rapid early adoption. Second, the geographic concentration of approvals in Beijing, Shanghai, Shenzhen, and Hangzhou reflects the social system effect. These megacities concentrate critical resources for diffusion, such as policy incentives, venture capital, and technical talent, forming powerful hubs that accelerate local development and initial market penetration of AIMDs.

The observed approval trends could not be fully understood without considering China's evolving regulatory framework. Our data showed sustained growth consistent with the aims of recent NMPA reforms. Notably, the 2025 measures to optimize full life cycle supervision and support the innovative development of advanced medical devices explicitly aimed to accelerate reviews of advanced devices, optimize premarket processes, and strengthen postmarket surveillance. This policy shift provided essential context for interpreting market dynamics and indicated a regulatory environment oriented toward facilitating efficient innovation while ensuring safety throughout a product's life cycle. Furthermore, the emphasis on fostering

XSL•FO
RenderX

high-level industrial clusters offered a lens to interpret the pronounced geographic concentration of approvals observed in this study. The synergy between concentrated innovation resources in hubs such as Beijing and Shanghai and proactive regulatory support appears to be a defining feature of China's AIMD ecosystem.

Limitations

This study has several limitations. First, the identification of an AIMD depended on the explicit mention of AI techniques in its functional components or intended use. Devices that did not explicitly state the use of AI might have been missed despite cross-checking ambiguous cases with review reports and websites. Second, classifying medical specialties was challenging because many devices spanned multiple clinical domains. We assigned a single primary specialty, which may underrepresent secondary applications. Third, although we standardized device names by merging minor variations that referred to the same product, some inconsistencies may persist, underscoring the need for unified naming conventions.

Conclusions

This analysis indicated that AIMDs in China are experiencing rapid growth. The regulatory landscape is dominated by radiology applications, with CT-based solutions for pulmonary nodules and cardiovascular conditions especially prominent. There is also a high degree of market concentration among leading manufacturers.

Acknowledgments

The authors thank the Drugdataexpy database for providing complimentary access to data on medical devices approved in China.

Funding

This work was funded by the Fundamental Research Funds for the Central Universities (grant 2632025RWPY04) and the National Natural Science Foundation of China (grant 72304150).

Data Availability

The original data supporting this study are publicly available from the Drugdataexpy database. The processed dataset generated and analyzed during this study is provided in Multimedia Appendix 2.

Authors' Contributions

Conceptualization: LZ, JY
Data curation: LZ
Formal analysis: LZ
Funding acquisition: LZ
Methodology: LZ, JY
Project administration: JY
Resources: JY
Software: JY
Supervision: JY
Validation: JY
Visualization: LZ
Writing—original draft: LZ
Writing—review and editing: JY
Both authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy and process for identifying artificial intelligence–based medical devices.
[DOCX File, 13 KB - medinform_v14i1e85538_app1.docx]

Multimedia Appendix 2
Detailed information of approved artificial intelligence–based medical devices in China.
[DOCX File, 67 KB - medinform_v14i1e85538_app2.docx]

References

1. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]

XSL•FO
RenderX

2. Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. Int J Med Inform 2024 Aug;188:105474. [doi: 10.1016/j.ijmedinf.2024.105474] [Medline: 38733640]
3. Meder B, Asselbergs FW, Ashley E. Artificial intelligence to improve cardiovascular population health. Eur Heart J 2025 May 21;46(20):1907-1916. [doi: 10.1093/eurheartj/ehaf125] [Medline: 40106837]
4. Faiyazuddin M, Rahman SJQ, Anand G, et al. The impact of artificial intelligence on healthcare: a comprehensive review of advancements in diagnostics, treatment, and operational efficiency. Health Sci Rep 2025 Jan;8(1):e70312. [doi: 10.1002/hsr2.70312] [Medline: 39763580]
5. Niemiec E. Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? Digit Health 2022 Mar 30;8:20552076221089079. [doi: 10.1177/20552076221089079] [Medline: 35386955]
6. Tang D, Xi X, Li Y, Hu M. Regulatory approaches towards AI medical devices: a comparative study of the United States, the European Union and China. Health Policy 2025 Mar;153:105260. [doi: 10.1016/j.healthpol.2025.105260] [Medline: 39951854]
7. Chen WP, Teng WG, Kuo CB, et al. Regulatory insights from 27 years of artificial intelligence/machine learning-enabled medical device recalls in the United States: implications for future governance. JMIR Med Inform 2025 Jul 11;13:e67552. [doi: 10.2196/67552] [Medline: 40644609]
8. Reddy S. Global harmonization of artificial intelligence-enabled software as a medical device regulation: addressing challenges and unifying. Mayo Clin Proc Digit Health 2024 Dec 24;3(1):100191. [doi: 10.1016/j.mcpdig.2024.100191] [Medline: 40207007]
9. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020 Sep 11;3:118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
10. Joshi G, Jain A, Araveeti SR, Adhikari S, Garg H, Bhandari M. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. Electronics (Basel) 2024 Jan 24;13(3):498. [doi: 10.3390/electronics13030498]
11. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203. [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
12. Han Y, Ceross A, Bergmann J. Regulatory frameworks for AI-enabled medical device software in China: comparative analysis and review of implications for global manufacturer. JMIR AI 2024 Jul 29;3:e46871. [doi: 10.2196/46871] [Medline: 39073860]
13. Zhang S, Li Y, Liu W, et al. A decade of review in global regulation and research of artificial intelligence medical devices (2015–2025). Front Med (Lausanne) 2015 Jul 17;12:1630408. [doi: 10.3389/fmed.2025.1630408] [Medline: 40747091]
14. Liu Y, Yu W, Dillon T. Regulatory responses and approval status of artificial intelligence medical devices with a focus on China. NPJ Digit Med 2024 Sep 18;7(1):255. [doi: 10.1038/s41746-024-01254-x] [Medline: 39294318]
15. Announcement no. 19 of 2025 from the National Medical Products Administration on issuing the list of medical devices exempt from clinical evaluation. National Medical Products Administration. 2025. URL: https://www.nmpa.gov.cn/xxgk/ggtg/ylqxggtg/ylqxqtggtg/20250513095643145.html [accessed 2026-01-21]
16. Guidelines for registration review of artificial intelligence medical devices (no. 8, 2022). Center for Medical Device Evaluation. 2022. URL: https://www.cmde.org.cn//flfg/zdyz/zdyzwbk/20220309091014461.html [accessed 2025-09-11]
17. Guiding principles for clinical evaluation and registration review of artificial intelligence-assisted detection medical devices (software) (no. 38, 2023). Center for Medical Device Evaluation. 2023. URL: https://www.cmde.org.cn//flfg/zdyz/zdyzwbk/20231116154535183.html [accessed 2025-09-11]
18. Announcement no. 101 of 2023 by the National Medical Products Administration on adjusting certain contents of the “medical device classification catalog”. National Medical Products Administration. 2023. URL: https://www.nmpa.gov.cn/directory/web/nmpa/ylqx/ylqxggtg/20230817153633135.html [accessed 2025-09-17]
19. Announcement no. 63 of 2025 issued by the National Medical Products Administration on measures to optimize the whole life cycle supervision and support the innovative development of high-end medical devices. National Medical Products Administration. 2025. URL: https://www.nmpa.gov.cn/xxgk/ggtg/ylqxggtg/ylqxqtggtg/20250703163951182.html [accessed 2025-09-11]
20. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003.
21. DRUGDATAEXPY database. URL: https://db.yaozh.com/ [accessed 2026-02-20]
22. Data security law of the People’s Republic of China. The National People’s Congress of the People’s Republic of China. 2021. URL: http://www.npc.gov.cn/englishnpc/c2759/c23934/202112/t20211209_385109.html [accessed 2026-02-18]
23. Personal information protection law of the People’s Republic of China. The National People’s Congress of the People’s Republic of China. 2021. URL: http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm [accessed 2026-02-18]
24. CT lung nodule intelligent screening system: assisting in lung cancer screening. United Imaging Intelligence. URL: https://www.uii-ai.com/product/1.html [accessed 2026-02-17]

XSL•FO
RenderX

25. Zhu S, Gilbert M, Chetty I, Siddiqui F. The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: an analysis of the characteristics and intended use. Int J Med Inform 2022 Sep;165:104828. [doi: 10.1016/j.ijmedinf.2022.104828] [Medline: 35780651]
26. Announcement no. 47 of 2021 issued by the National Medical Products Administration on the guiding principles for the classification and definition of artificial intelligence medical software products. National Medical Products Administration. 2021. URL: https://www.nmpa.gov.cn/ylqx/ylqxggtg/20210708111147171.html [accessed 2025-09-29]
27. Lin JC, Jain B, Iyer JM, et al. Benefit-risk reporting for FDA-cleared artificial intelligence-enabled medical devices. JAMA Health Forum 2025 Sep 5;6(9):e253351. [doi: 10.1001/jamahealthforum.2025.3351] [Medline: 41004181]
28. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl) 2021 Mar 17;134(7):783-791. [doi: 10.1097/CM9.0000000000001474] [Medline: 33734139]
29. Writing Committee of the Report on Cardiovascular Health and Diseases in China. Report on cardiovascular health and diseases in China 2022: an updated summary. Biomed Environ Sci 2023 Aug 20;36(8):669-701. [doi: 10.3967/bes2023.106] [Medline: 37711081]
30. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. Phys Med 2021 Mar;83:1-8. [doi: 10.1016/j.ejmp.2021.02.011] [Medline: 33657513]
31. Pinsky MR, Bedoya A, Bihorac A, et al. Use of artificial intelligence in critical care: opportunities and obstacles. Crit Care 2024 Apr 8;28(1):113. [doi: 10.1186/s13054-024-04860-z] [Medline: 38589940]

Abbreviations

AI: artificial intelligence
AIMD: artificial intelligence–based medical device
CT: computed tomography
ML: machine learning
NMPA: National Medical Products Administration
SiMD: software in a medical device

Edited by A Coristine; submitted 09.Oct.2025; peer-reviewed by PT Chen, VK Rathi, WG Teng; revised version received 29.Jan.2026; accepted 30.Jan.2026; published 18.Mar.2026.

Please cite as:
Zhang L, Yan J
Approval of AI-Based Medical Devices in China From 2020 to 2025: Retrospective Analysis
JMIR Med Inform 2026;14:e85538
URL: https://medinform.jmir.org/2026/1/e85538
doi:10.2196/85538

© Lingli Zhang, Jianzhou Yan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tailoring Discharge Summaries to Health Care Providers' Needs (Part 1 of the Framework and Implementation of AI Tools Project): User-Centered Design Approach

Mieke Deschepper[1], MSci, PhD; Helga Rogge[1], MD; Mathias Syx[1], MSCE; Kirsten Colpaert[1,2,3], MD, PhD

[1]Data Science Institute, Ghent University Hospital, Corneel Heymanslaan 10, Ghent, Belgium

[2]Department of Intensive Care, Data Science Institute, Ghent University Hospital, Ghent, Belgium

[3]Department of Internal Medicine and Pediatrics, Ghent University, Ghent, Belgium

Corresponding Author:
Mieke Deschepper, MSci, PhD
Data Science Institute, Ghent University Hospital, Corneel Heymanslaan 10, Ghent, Belgium

Abstract

Background: Medical discharge letters are critical for continuity of care but often lack clarity and personalization, making it difficult for health care providers to retrieve essential information. While large language models (LLMs) offer potential for automating summary generation, their effectiveness depends heavily on the quality and contextual relevance of the prompts used.

Objective: The objective of this study was to develop and describe a human-centered, replicable framework for creating individualized prompts that guide LLMs in generating summaries tailored to the specific needs of health care providers.

Methods: A multidisciplinary workshop was conducted at Ghent University Hospital with 26 health care providers from 5 institutions, including hospitals and general practitioner networks. Participants brainstormed ideal summary formats, generating 170 ideas categorized into themes such as "structure," "medical history," "medication," and "follow-up." These insights informed the development of a 110-item structured questionnaire distributed to 33 participants. Responses were used to generate personalized and generic prompts, refined using the context, objective, style, tone, audience, and response (CO-STAR) framework.

Results: "Structure and layout" (40/170, 23.5%) and "follow-up" (27/170, 15.9%) were the most emphasized categories in the workshop. The questionnaire confirmed the importance of the "follow-up" and "medical history" sections. Prompts were generated per participant and by health care type, incorporating frequently selected responses. The CO-STAR framework was applied to improve prompt clarity and alignment with clinical expectations. Communication emerged as a new category during the workshop and was universally valued in the questionnaire.

Conclusions: This study presents a novel, systematic approach to prompt engineering in clinical artificial intelligence applications. By translating qualitative input into structured, individualized prompts, the framework is designed to improve the usability and relevance of artificial intelligence–generated summaries. It proposes a scalable approach for integrating human-centered design into LLM deployment in health care, with the aim of supporting more accurate, context-aware clinical documentation. However, these outcomes remain to be empirically validated; this study is limited to the design and implementation of a human-centered prompt construction pipeline in a specific multicenter setting.

(JMIR Med Inform 2026;14:e80613) doi:10.2196/80613

KEYWORDS

large language models; clinical artificial intelligence integration; clinical AI integration; prompt engineering; discharge summaries; human-centered design; artificial intelligence; AI

Introduction

Background

Medical discharge letters are essential communication tools, yet their length, complexity, and variability often hinder efficient information retrieval by health care professionals, who frequently encounter difficulties in identifying key elements such as medication changes, follow-up recommendations, and significant diagnostic findings. Moreover, the informational needs of a general practitioner, a surgeon, an intensivist, and other health care providers could differ substantially.

This complexity not only hinders continuity of care but also introduces risk. In recent years, large language models (LLMs) have shown promise in helping extract, summarize, or restructure clinical documents such as medical discharge letters [1]. However, evaluating the quality and safety of such applications remains a human-centered task requiring careful planning and expert validation.

XSL•FO
RenderX

Numerous studies have evaluated the use of LLMs for generating medical discharge letters [2-4]. Despite the growing sophistication of these models, human evaluation remains indispensable for assessing the clinical accuracy, completeness, and safety of the generated outputs. Physicians, in particular, play a critical role in validating whether artificial intelligence (AI)–generated summaries address the nuanced informational needs of health care providers [2,3]. As Tam et al [5] emphasize, a structured planning phase is essential before initiating human evaluation. This phase begins with clearly defining the evaluation objectives and the specific tasks to be assessed. However, while the final evaluation is typically conducted by clinicians, the development of the prompts that guide LLMs is often carried out by engineers or a single physician. This disconnect may result in prompts that do not fully reflect the diverse and context-specific requirements of end users [2-4,6,7].

Effective medical prompt engineering is emerging as a critical skill for health care professionals navigating the integration of AI into clinical practice. As Meskó [8] emphasizes, the rise of LLMs such as ChatGPT represents a paradigm shift in health care, offering unprecedented opportunities for enhancing clinical workflows, patient communication, and decision-making. However, realizing these benefits depends heavily on the clinician's ability to craft precise, context-aware prompts that guide AI systems toward generating accurate and relevant outputs. Patil et al [9] further highlight that well-designed prompts—grounded in medical knowledge and aligned with evidence-based guidelines—are essential for eliciting clinically meaningful responses from AI tools. They argue that prompt engineering should be recognized as a core competency in medical education, enabling physicians to harness AI effectively while maintaining ethical standards and minimizing bias. These perspectives underscore that prompt engineering is not solely a technical task. However, at present, it is neither a skill that all health care professionals possess nor one that they have time to develop in depth. Moreover, as generative AI continues to evolve, its evaluation should be approached as an ongoing process, recognizing that performance, reliability, and clinical relevance may shift over time and require continuous monitoring and adaptation. Furthermore, due to the variability in ChatGPT's responses, evaluating its performance in health care requires repeated testing to ensure reliability. Without incorporating repetition, research findings may be biased and lack credibility [10].

In our project, Framework & Implementation of AI Tools (FRAIT), we aim to develop AI-driven solutions that are tailored to the specific needs of health care providers. The first phase of the project adopts a design thinking approach to design a prompt that accurately reflects the unique requirements and expectations of individual health care professionals. This involves close collaboration to understand their workflows and information needs and the specific clinical contexts in which they operate. Once this tailored prompt is established, the second phase will involve evaluating its effectiveness by applying it to Dutch medical discharge letters, assessing how well it extracts or summarizes relevant information. In this paper, we concentrate on the initial phase: constructing a prompt that is both meaningful and practical for the health care provider, ensuring that it aligns with their clinical and administrative objectives.

Objectives

The aim of this study was to develop a comprehensive process that starts by identifying the needs of health care providers regarding the ideal format for medical discharge letter summaries and results in creating an individualized prompt to generate such summaries effectively.

Methods

Part 1: Workshop

Overview

Health care providers, primarily physicians, were recruited from 5 different institutions: 3 hospitals (Ghent University Hospital, AZ Oudenaarde, and AZ Sint-Lucas Ghent) and 2 general practitioner circles (Huisartsenvereniging Ghent and Huisartsenkring Schelde Leie). The target recruitment range was 20 to 30 health care providers.

The workshop took place on Wednesday, October 16, 2024, at Ghent University Hospital. The workshop followed a 3-part structure consistent with design thinking principles: an energizer to foster engagement, a brainstorming session to elicit ideas, and a feedback round to collaboratively refine and prioritize outputs.

A set of 7 questions, each with 3 possible responses (blue, yellow, and green; Table S1 in Multimedia Appendix 1) was provided to assess the level of LLM use. Results were presented as numbers and percentages. Data were collected through 2 different persons manually counting the groups and a validation based on pictures (Figure S1 in Multimedia Appendix 1).

Following an introduction to the project, the participants were organized into 6 groups, each encompassing a diverse array of institutions and specializations. We used the digital platform Padlet [11] to conduct this brainstorming session, which allowed us to document all steps in the process. The participants had the option to enter their ideas on an available laptop or through their smartphones, which did not require any log-in credentials or other personal information.

The main question for the brainstorming session was the following:

> *What does my ideal generated summary from a medical discharge letter look like? Eg: use of bullet points, Eg: include history. Take us through your thinking: what should be included? All ideas are welcome – personal preference counts. No consensus required, the groups are for cross-pollination / inspiration.*

The emphasis was placed on both layout and content (eg, "I would like to see the accurate medical history, and I would like it to be highlighted").

All ideas were free text, although they required a label or category. The available labels at the start of the brainstorming session were "history," "current problems," "medication,"

XSL•FO
RenderX

"decision," "follow-up," "structure and layout," and "other" or "I don't know." New categories could be added during the brainstorming session.

A preliminary feedback session was held after 15 minutes to discuss the label "other" in more detail for refinement. Subsequently, an additional 15-minute brainstorming session was scheduled.

Statistics

A word cloud was used to visualize and cluster tokens and words based on their frequency within various ideas. The number of ideas by group and category was reported as frequencies and percentages, means and SDs, and medians and IQRs.

During the workshop, results were presented by category as well as the most common words within each category. The full text of the ideas for these specific words was provided to facilitate further discussion within the group if any clarification was required.

Part 2: Questionnaire

The ideas generated from the brainstorming session were systematically categorized. Two independent researchers (MD and HR) conducted data cleaning to eliminate duplicates and segment the ideas into smaller, manageable components. A structured table was developed (Table S2 in Multimedia Appendix 1) where each row represented a primary idea and subsequent columns contained subquestions. Each idea was assigned to a predefined category. For multiple-choice questions, the possible answers were maintained in a separate table (Table S3 in Multimedia Appendix 1). Using this structured table, a complete list of questions was generated. Additionally, we developed a prompt for each potential response in this questionnaire in Dutch, as well as an automated English translation. The questionnaire was designed using Microsoft Forms, with the embedded Microsoft Copilot used to streamline the automation process.

Initially, some general questions were added regarding the type of health care provider and the number of years in practice. An open-ended question for suggestions was included at the end. This questionnaire was distributed to all participants of the workshop and an additional group of general practitioners who will be involved in the ongoing process of the FRAIT project.

The results of the questionnaire were combined with the predetermined prompt for each answer, resulting in a unique prompt for each participant. Additionally, a generic prompt was created overall and by type of health care provider. This generic prompt selected the most frequently chosen answer for each question. For multiple-choice questions, an option must be present in at least one-third of the answers.

To transform the workshop ideas into a structured questionnaire and prompt templates, we applied a systematic multistep process. Each free-text idea was carefully reviewed and split into the smallest meaningful units, with each segment representing a single, actionable suggestion. For example, a statement such as "I want the summary to highlight medication changes and provide a clear follow-up plan" was divided into 2 distinct ideas: highlighting medication changes and providing a clear follow-up plan. Segmentation was guided by the principle that each unit should be assignable to a single category and actionable in prompt design.

Segments were then assigned to predefined categories, such as "medical history," "medication," "follow-up," and "structure and layout." If a segment fit multiple categories, it was assigned to all relevant categories (multilabel coding). New categories were created when multiple participants raised themes not covered by existing categories.

The coding process was iterative and collaborative. Two researchers independently segmented and categorized all ideas. Rather than aiming for consensus or exclusion of divergent ideas, we used a structured framework (Tables S2 and S3 and Figure S2 in Multimedia Appendix 1) to ensure that all contributions from the workshop were systematically included. MD constructed the initial framework by segmenting and categorizing all ideas, whereas HR independently reviewed the results for completeness and accuracy. Any ambiguities or uncertainties were discussed, but the primary objective was the comprehensive inclusion of all perspectives, not convergence on a single interpretation. This approach maximizes methodological transparency and reproducibility, as every idea generated during the workshop is traceable through the coding and questionnaire development pipeline.

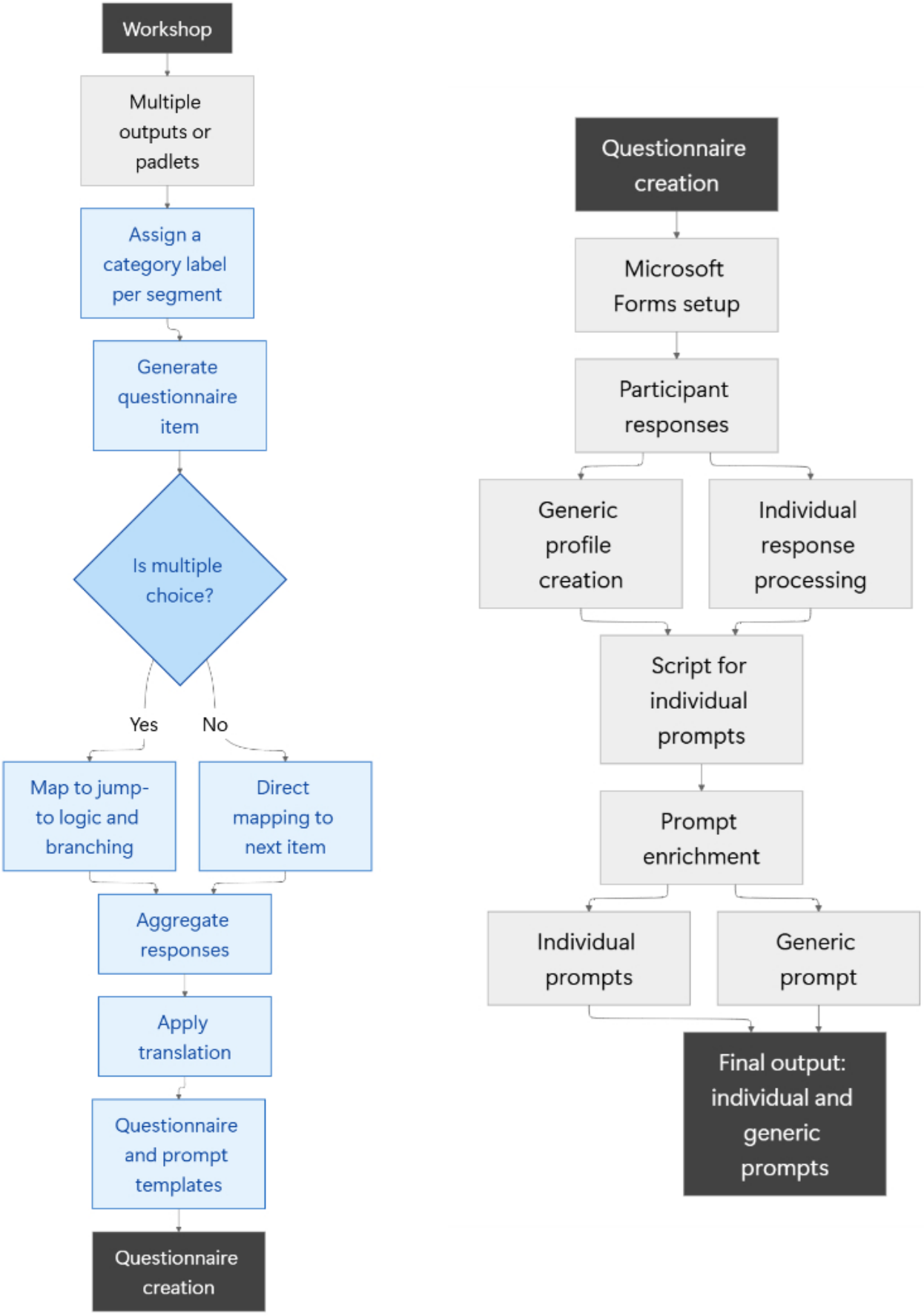
Two Python scripts (Python Software Foundation) were created to automate data processing and prompt generation. The first script constructs a dynamic questionnaire by normalizing text, analyzing response frequencies, and reshaping data for optimal mapping between questions and answers. The second script translates completed responses into individualized and generic AI prompts using index-based mapping and neural machine translation for Dutch-to-English conversion. Both scripts aggregate and profile responses to ensure that generated prompts reflect the most common preferences among health care providers, supporting transparent and reproducible prompt engineering.

Finally, we enhanced all prompts by incorporating segments before and after the original prompts, adhering to the context, objective, style, tone, audience, and response (CO-STAR) framework [12-14]. To operationalize the CO-STAR framework, we incorporated both participant input and established clinical communication standards during the prompt enrichment phase. Workshop and questionnaire participants were asked to articulate their preferences regarding the presentation of discharge summaries, including aspects such as clarity, formality, and professional perspective. These responses informed the initial drafting of prompts. Figure 1 illustrates the complete process.

We compared responses in the questionnaire for each category of health care provider, ensuring that there were at least 5 responses per type. Our primary emphasis was on content questions, not layout questions. Because our sample size was limited, we restricted our analysis to descriptive statistics only.

XSL•FO
RenderX

Figure 1. Overview of the full process flow: from multidisciplinary idea generation and categorization, through questionnaire development and automation, to the final context, objective, style, tone, audience, and response framework–optimized prompt for clinical discharge summaries.



XSL•FO
RenderX

Ethical Considerations

All approvals were granted by the ethical committee of Ghent University Hospital (ONZ-2024 - 0304). The ethical committees of AZ Sint-Lucas Ghent and AZ Oudenaarde were informed of the study. Each participating organization, both hospitals and general practitioner circles, handled recruitment themselves, using their own internal communication channels and procedures to invite participants. The recruitment letter was included with our submission to the ethical committee. Informed consent was obtained from all participants before the start of the workshop (Figure S3 in Multimedia Appendix 1).

To ensure participant anonymity and data privacy, all responses collected during the workshop and questionnaire phases were anonymized before analysis. No personally identifiable information was linked to the responses, and participants were given the option to contribute via digital platforms without requiring log-in or personal credentials. Data were stored securely and accessed only by authorized research team members. These measures were implemented in accordance with institutional guidelines and approved by the relevant ethical committees.

Participants received compensation for the time they spent contributing to the project. This compensation acknowledged the substantial effort required to evaluate the summaries and attendance to the workshop, to ensure that clinicians were remunerated for their professional expertise and time commitment. No additional financial or material incentives were provided beyond this time-based compensation.

Results

Part 1: Workshop

A total of 26 health care providers were recruited for this workshop (Tables S4 and S5 in Multimedia Appendix 1). However, 7.7% (2/26) of the individuals were not present during the first half hour and did not respond to the initial questions. Overall, 5 groups were formed: 4 groups consisting of 5 participants each and 1 group consisting of 6 participants.

Table S1 in Multimedia Appendix 1 lists the questions at the start of the workshop and outlines the profiles of the participants. Figure S4 in Multimedia Appendix 1 presents the outcomes of tokenizing all ideas, which are visualized using a word cloud. Figure S5 in Multimedia Appendix 1 also contains an example of the initial concepts as presented to the participants from one of the groups.

Following a 15-minute feedback session and a plenary discussion on topics categorized as “other,” a new “communication” category was established based on requests from 2 separate groups, demonstrating responsiveness to participant input. Through an iterative process of coding and review, all ideas, including those applicable to multiple categories, were systematically preserved.

At the conclusion of the workshop, a comprehensive compilation of 170 ideas generated by all groups was assembled. The complete list, along with corresponding group names, can be found in Table S6 in Multimedia Appendix 1. The “structure and layout” category contained the greatest number of ideas (40/170, 23.5%), whereas only 6.5% (11/170) of the ideas were left unlabeled by participants. “Communication,” introduced during the second round, constituted the smallest category (Table 1).

The finalized coded table served as the foundation for developing a 110-item questionnaire using branching logic for multiple-choice questions. This systematic approach facilitated the creation of both individualized and generic prompt templates.

Table . Distribution of workshop-generated ideas by category, illustrating the thematic focus and relative emphasis placed by health care providers on different aspects of discharge summary design.

Category	Participant group	Ideas (N=170), n (%)	Ideas (n), mean (SD)	Ideas (n), median (IQR)
Communication	4	8 (4.7)	2 (1.2)	2 (1-3)
Current problems	5	20 (11.8)	4 (2)	5 (3-5)
Decision	5	19 (11.1)	3.8 (1.9)	3 (3-4)
Follow-up	5	27 (15.9)	5.4 (1.1)	5 (5-6)
Medical history	5	24 (14.1)	4.8 (3.8)	4 (3-5)
Medication	5	21 (12.4)	4.2 (3.3)	3 (2-6)
Other or “I don’t know”	4	11 (6.5)	2.8 (1.5)	2 (2-3)
Structure and layout	5	40 (23.5)	8 (2.5)	8 (7-8)

Part 2: Questionnaire

After refining the ideas, a structured table was created as the foundation for the questionnaire. These 2 files, along with their interpretations (available only in Dutch), are included in Tables S2 and S3 and Figure S2 in Multimedia Appendix 1. We provide a Python script that can be used to generate a list of questions and the appropriate “jump-to” points where necessary starting from these files. The resulting 110 questions were uploaded to Microsoft Forms, ready to be answered by the participants. The complete list of questions is available in Table S7 in Multimedia Appendix 1, and a (translated) snippet of these questions can be found in Figure 2. In the same script, when creating questions with multiple possible answers, the corresponding prompt for each answer was also added.

XSL•FO
RenderX

Figure 2. Snippet from the resulting questionnaire list translated into English demonstrating the mapping of provider needs to specific questions and corresponding artificial intelligence prompt template.

Number	Label	Section	Answer	Jump to question	Prompt	Question type
27	Do you want to see Coagulation risk in the History section?	Medical History	['Yes', 'No']	[28, 29]	['Show the coagulation risk in section History Section', '']	item
28	What formatting should be used for this?	Medical History	['None', 'Underlined', 'Italic', 'Bold', 'Strikethrough', 'Fluorescent marking']	[29, 29, 29, 29, 29, 29]	['Coagulation risk: No specific formatting', 'Coagulation risk: Underline the text', 'Coagulation risk: Place the text in italic', 'Coagulation risk: Place the text in bold', 'Coagulation risk: Strike out the text', 'Coagulation risk: Use fluorescent marking']	layout
29	Do you want to see Deep Vein Thrombosis in the History section?	Medical History	['Yes', 'No']	[30, 31]	['Show in section History De Diep Venous Thrombosis', '']	item
30	What formatting should be used for this?	Medical History	['None', 'Underlined', 'Italic', 'Bold', 'Strikethrough', 'Fluorescent marking']	[31, 31, 31, 31, 31, 31]	['Deep Vein Trombosis: No specific formatting', 'Deep Vein Trombosis: Underline the text', 'Deep Vein Trombosis: Place the text in italic', 'Deep Vein Trombosis: Place the text in bold', 'Deep Vein Trombosis: Strike out the text', 'Deep Vein Trombosis: Use fluorescent marking']	layout
31	Do you want to see New diagnoses in the History section?	Medical History	['Yes', 'No']	[32, 33]	['Show the new diagnoses in section History', '']	item
32	What formatting should be used for this?	Medical History	['None', 'Underlined', 'Italic', 'Bold', 'Strikethrough', 'Fluorescent marking']	[33, 33, 33, 33, 33, 33]	['New diagnosis: No specific formatting', 'Nieuwe diagnosen: Underline the text', 'New diagnosis: Place the text in italic', 'New diagnosis: Place the text in bold', 'New diagnosis: Strike out the text', 'New diagnosis: Use fluorescent marking']	layout
33	Do you want to see Allergies in the History section?	Medical History	['Yes', 'No']	[34, 38]	['Show the allergies in section History', '']	item
34	Want to see the type of reaction?	Medical History	['Yes', 'No']	[35, 35]	['Show the reaction on the allergy.', 'Do not show the reaction on the allergy.']	item
35	What allergies do you want to see?	Medical History	['All Allergies', 'Latex', 'Penicillin', 'New', 'Life-Threatening']	[36, 36, 36, 36, 36]	['Show all allergies', 'Show latex allergy', 'Show penicillin allergy', 'Show new allergies', 'Show life-threatening allergies']	item
36	Which list do you want to see?	Medical History	['Table', 'Bullet points - all', 'Bullet points - < 10', 'Prose - short keywords', 'Prose - Full sentences']	[37, 37, 37, 37, 37]	['Allergies: Place this in table form', 'Allergies: Use bullet points', 'Allergies: Use a maximum of 10 bullet points', 'Allergies: Use short keywords', 'Allergies: Use full sentences']	layout
37	What formatting should be used for this?	Medical History	['None', 'Underlined', 'Italic', 'Bold', 'Strikethrough', 'Fluorescent marking']	[38, 38, 38, 38, 38, 38]	['Allergieën: No specific formatting', 'Allergieën: Underline the text', 'Allergieën: Place the text in italic', 'Allergieën: Place the text in bold', 'Allergieën: Strike out the text', 'Allergieën: Use fluorescent marking']	layout
38	Do you want to see section Medication in your summary?	Medication	['Yes', 'No']	[39, 53]	['Show the Medication section in the summary', 'Do not show the medication section in the summary']	section
39	What medication do you want to see?	Medication	['None', 'All Medication', 'Medication on Admission', 'Medication on	[40, 40, 40, 40, 40, 40, 40]	['Show the none', 'Show the all medication', 'Show the medication upon admission', 'Show the medication upon dismissal', 'Show the changed	item

Table S8 in Multimedia Appendix 1 shows the number of questions for each section. “Medical history” and “admission history” constituted the largest parts of the questionnaire.

Most health care providers (30/33, 90.9%) were physicians, with over 81.8% (27/33) having more than 5 years of experience (Figure S6 in Multimedia Appendix 1). The “follow-up” section was highly valued by all participants, whereas the “examination” section received the least interest (Figure S7 in Multimedia Appendix 1).

An example of the resulting individualized and final generic prompt can be found in Figure S8 in Multimedia Appendix 1.

The implementation of the CO-STAR framework produced prompts that effectively reflected both health care provider preferences and established professional standards. The “style” component was defined as “written from the perspective of a medical professional,” and the “tone” element specified the use of formal language. These parameters were selected based on aggregated participant feedback, which underscored the importance of professionalism and clarity in clinical communication. Consequently, the final prompts incorporated user-driven insights alongside best practices, ensuring that AI-generated summaries were both contextually relevant and professionally suitable. A comprehensive overview of the adjustments made to the CO-STAR framework is provided in Figure 3. The prompt structure was enhanced with an introductory segment and a concluding section, whereas the central portion, labeled “<Individual preference>,” reflects findings from the questionnaire.

Table 2 shows the differences in the 10 most common preferences regarding discharge summaries between general practitioners and specialists. General practitioners tend to seek additional details regarding medication and follow-up, whereas specialists more frequently request information pertaining to admission history and examinations.

XSL•FO
RenderX

Figure 3. Illustration of the context, objective, style, tone, audience, and response framework application in prompt engineering: integrating health care provider preferences and clinical standards to optimize artificial intelligence–generated discharge summary prompts.

Color code: C=context - O=objective - S=style - T=tone - A=audience - R=response

"This text is a medical discharge letter (the report of a hospital admission). I would like a summary in Dutch. The intention is that I, as a healthcare provider, can use this letter to properly follow up the patient. The summary must be a maximum of one page long and clearly structured. Use formal language based on the expertise of a physician. Start with the patient's name, date of birth and the name of the hospital. All dates must be in the format day/month/year. If a requested part is not in the letter, bundle the missing parts at the bottom per section in one sentence and use 'Not mentioned' as an indication. “

<Individual preferences>

“We would like to provide some additional explanation: The name of the sections should be translated like: General = Algemeen, Medical History = Voorgeschiedenis, Medication = Medicatie, Examinations = Onderzoeken, Admission history = Opnameverloop, Follow-up = Follow-up, Decision = Besluit. If new diagnoses are requested in the history, this means that you add the new diagnoses that were made during admission to the history. By interventions we mean operations. The admission process describes the path that the patient takes during the admission, the evolution of the condition and the steps that are taken during the admission in chronological order. Here we also expect to report the administration of blood transfusions, radiotherapy or chemotherapy, the performance of diagnostic examinations such as biopsies, endoscopies, the placement of feeding tubes, etc. In Medical imaging / Technical examinations we mean by the protocol all actions that are performed during this examination and all findings that are made in this. If changes in medication are requested, we mean changes in dose, frequency and form and whether medication has been stopped completely or new medication has been started. To do this, make a comparison between the medication on admission on the one hand and the medication on discharge and the medication at the therapy proposal on the other. Medication that has remained the same is not mentioned in the case of changed medication. If the question is about the risk of clotting, mention blood clots/thrombosis and embolisms and the use of blood-thinning medication. Also mention here if you find anything about cerebral infarction/CVA or heart attack/AMIAMI. Clotting disorders should also be mentioned here. Communication is about discussing the condition and the further policy with the patient and his family. If a section should not be mentioned, leave it out of the summary completely. The information in the summary should come from the original letter. An exception to this is if ICD codes or ICPC codes are mentioned. When asking for ICD codes, the ICD-10 code should be mentioned next to the diagnosis, as can be found at https://icd10be.health.belgium.be/. When asking for ICPC codes, you can find them at https://browser.icpc-3.info/. If a patient has come from another hospital, mention this under general.”

XSL•FO
RenderX

Table . Comparison of discharge summary preferences between general practitioners (GPs) and specialists, highlighting the top 20 content areas with the greatest differences in selection rates. Options with negative percentages were favored by specialists, whereas those with positive percentages were preferred by GPs.

Section	Preference	Difference between GPs and specialists (%)
Admission history	Showing acquisition of the interventions performed	−25.3
Admission history	Showing consultations	−14.0
Examinations	Showing the protocol	−13.3
Medical history	Showing life-threatening allergies	−13.3
Examinations	Showing laboratory test results	−12.6
Admission history	Showing main diagnosis	−11.6
Medical history	Showing reaction to allergy	−11.6
Admission history	Showing new diagnoses	−9.8
Examinations	Showing clinical investigation	−8.6
Examinations	Showing medical imaging or technical investigations	−8.6
Follow-up	Showing inclusion in specific care processes	26.8
Follow-up	Showing message for GP	18.7
General	Showing whether the patient is palliative	16.6
Follow-up	Showing the institutions involved	16.6
Medication	Showing changed medication (dose, frequency, and shape)	15.1
Medication	Showing new medication	14.4
Follow-up	Showing in section “follow-up” the communication	13.2
Follow-up	Showing multidisciplinary follow-up (social and home help)	13.2
Medication	Showing stopped medication	12.5
Medication	Showing duration of medication use	12.3

Discussion

Principal Findings

From a single workshop with motivated health care providers, it was possible to generate an automated, individualized prompt tailored to their respective needs, enabling the creation of an ideal medical discharge summary. The process began with the identification of key areas that required attention, such as medical history, medication details, and follow-up procedures. Participants provided input on what specific information should be included in each section, ensuring that the final generated summary would be comprehensive and tailored to the needs of health care professionals. To the best of our knowledge, this is the first study to begin with a multidisciplinary workshop involving physicians from diverse specialties and translate their input into tailored AI prompts.

The collaboration between specialists and general practitioners resulted in an exchange of ideas, leading to a broader perspective from both groups. This synergy allowed for a more holistic approach to the creation of the summary, taking into account both the specialized and general aspects of patient information. At the same time, it became clear that reflecting on one’s ideal summary is not always easy—despite it being a personal matter. Sparring with others did not necessarily lead to adopting their viewpoints but proved valuable in helping participants critically reassess and refine their own preferences.

Dividing the scope into layout and content made the resulting ideas more concrete. By separating these 2 elements, participants were able to focus more effectively on each aspect, leading to clearer and more actionable suggestions. This also leads to a clearer prompt for further evaluation possibilities (in future work), providing a solid foundation for future improvements and refinements based on feedback from health care professionals.

Within the workshop, numerous ideas were generated. The open question at the end—“Do you have any additional comments or suggestions regarding this questionnaire? Are there aspects that you think are missing or not sufficiently addressed? Please note them below.”—received responses primarily on practical issues from only a few participants.

During the workshop, it was observed that all groups contributed an equal number of ideas. Most ideas were related to the

XSL•FO
RenderX

structure of the generated summary, primarily due to the existence of a single category for the broader theme. The "follow-up" category encompassed the highest number of ideas, although there was significant overlap with other categories. The "communication" category, introduced during the workshop, predominantly contained ideas that pertained to the "follow-up" category. This trend was also evident in the questionnaire, where the "communication" section was the only one indicated as desirable by all participants.

The "medical history" and "admission history" sections often included diverse content, resulting in numerous questions from participants. The "medication" and "follow-up" sections also encompassed a wide range of topics, which necessitated their inclusion as multiple-choice questions. This strategy allowed for a targeted approach, meaning that only 1 detailed question on content was asked for each selected topic (eg, "What information do you want to see for the medication? 'Start date,' 'Duration,' 'Dose,' 'Time of intake,' 'Frequency,' 'Form e.g. oral,' 'Show drug substance name,' 'Clinical effect,' 'Criteria for stopping/changing specific medication,' 'Reason for starting, stopping, changing'").

Such a detailed and structured approach ensures that all relevant aspects are thoroughly explored and addressed, providing a comprehensive framework for summarizing medical discharge letters efficiently and effectively.

Comparison With Prior Work

The most significant refinement aligned with the CO-STAR prompt engineering framework was introduced in the "response" section, with a particular emphasis on clarifying the definitions of key terminology embedded within the prompt. This adjustment is crucial for enhancing the model's interpretability and ensuring consistent semantic alignment between the prompt and the generated summary. As demonstrated in the work on SARA [13], a generative AI for legal summarization, precise control over terminology and density within prompts directly influences the coherence and legal validity of the summaries produced. Similarly, the Institute of Electrical and Electronics Engineers Computer Society's International Conference on Computers, Software, and Applications 2024 study [12] on automatic smart contract generation underscores the importance of terminological precision in the response phase to mitigate ambiguity and improve automation reliability. Furthermore, the insights from Teo [14] from the GPT-4 prompt engineering competition highlight how nuanced refinements in prompt structure—especially in the response section—can significantly elevate performance in competitive and real-world applications.

Our findings align with broader trends in health informatics that emphasize the centrality of user needs and adaptive information delivery. For instance, psychoeducational mobile health interventions for parents of small-for-gestational-age fetuses have shown how adaptive, user-centered systems can enhance engagement and outcomes [15]. The ADEPT framework further illustrates how optimizing data flow in dynamic, real-world environments supports effective, personalized care [16]. By translating qualitative input from health care providers into structured, individualized prompts, our framework extends these principles to the domain of clinical AI, ensuring that LLM-generated discharge summaries are not only technically robust but also contextually relevant and responsive to end user needs.

Limitations

An important aspect of this framework's broader impact is its adaptability beyond its original Dutch-speaking setting. Future efforts should focus on translation and cultural adjustment of all materials, questionnaires, and prompts used in workshops. This will require close collaboration with language specialists and domain specialists to ensure accurate terminology, workflows, and compliance. Owing to its modular design, the framework allows for easy alignment of categories and prompts with various health care systems' specific summarization needs. This methodology is applicable across medical specialties by engaging domain experts during assessment and prompt development, ensuring that specialty-specific information needs are addressed. By taking these steps, the framework will become more valuable and help develop AI tools that are effective, equitable, and well-suited to diverse clinical contexts.

Despite aiming for comprehensiveness, the balance between breadth and depth sometimes required trade-offs, prioritizing clarity and efficiency over detailed customization. For instance, preferences were gauged using a multiple-choice question such as "What do you want to see..." followed by only a single question about layout preferences for each item.

This initial prompt is a first step in the improvement in the quality and relevance of AI-generated summaries. It is evident that further refinement is necessary. This ongoing fine-tuning will involve a meticulous evaluation of the generated summaries, taking into account detailed feedback from health care professionals. Such iterative adjustments aim to enhance the precision and contextual accuracy of the summaries, ensuring that they meet the nuanced needs of diverse medical specializations.

Future Steps

The subsequent phase will involve evaluating the generic and individual prompts, with the assessment being developed collaboratively with the health care providers. The prompt will be broken down into smaller components, generating binary questions regarding layout and more specific inquiries related to content accuracy (eg, if an outcome is incorrect, does it constitute a hallucination?). A collection of representative medical discharge letters will be made available for this purpose. Upon completion of the evaluation, it will likely be necessary to review the prompt to identify opportunities for further optimization and integrate additional content informed by the evaluation findings.

Conclusions

This study presents a novel, systematic framework for developing individualized prompts to summarize medical discharge letters grounded in the real-world needs of health care providers. A key strength of this approach lies in the deliberate separation between layout and content preferences, which proved highly effective in capturing both structural and informational expectations.

XSL•FO
RenderX

By translating qualitative input into structured, personalized prompts, this framework aims to improve the relevance and usability of AI-generated summaries and serve as a replicable model for diverse clinical and research settings, although empirical validation remains necessary. The framework supports applications in data collection, data analysis, and personalized feedback generation. This study is limited to the design and implementation of a human-centered prompt construction pipeline in a specific multicenter setting; direct evaluation of AI-generated outputs and broader scalability will be addressed in future work.

Acknowledgments

The authors would like to express their sincere gratitude to the entire team at the Data Science Institute of Ghent University Hospital, as well as to the management, program participants, and coordinators representing the participating organizations. Artificial intelligence was used within the project to enhance efficiency: Figure 1 was created with assistance from Claude (Anthropic). The questionnaire was created in Microsoft Forms using Microsoft Copilot under the Microsoft Office license of Ghent University Hospital.

Funding

The Framework & Implementation of AI Tools project was funded by the Belgian government through the Data Innovation project.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files (Multimedia Appendix 1). The supporting Python scripts are available in a public GitHub repository [17]. The script frait_part1_questions.py generates a dynamic questionnaire for the Framework & Implementation of AI Tools system. Its primary output is an organized set of questions designed to assist health care providers in selecting relevant sections and details for inclusion in a medical summary. The script frait_part1_prompt.py translates completed questionnaire responses into artificial intelligence prompts, which are used for medical document summarization.

Authors' Contributions

Conceptualization: MD, MS, KC
Data curation: MD
Formal analysis: MD
Funding acquisition: MD, MS, KC
Investigation: MD, HR
Methodology: MD, MS
Project administration: MD, MS, HR
Resources: MD, MS, KC
Software: MD
Supervision: MS, KC
Validation: KC
Visualization: MD, HR
Writing—original draft: MD
Writing—review and editing: MD, HR, MS, KC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of all supplementary materials, including workshop energizer results, input questionnaire structure, informed consent template, participant characteristics, word-cloud analysis, padlet examples, workshop output (in Dutch), full questionnaire, and questionnaire results.
[PDF File, 1061 KB - medinform_v14i1e80613_app1.pdf]

Checklist 1
STROBE checklist.
[PDF File, 137 KB - medinform_v14i1e80613_app2.pdf]

References

XSL•FO
RenderX

1. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
2. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025 Jul 1;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
3. Koh MCY, Ngiam JN, Oon JEL, Lum LHW, Smitasin N, Archuleta S. Using ChatGPT for writing hospital inpatient discharge summaries - perspectives from an inpatient infectious diseases service. BMC Health Serv Res 2025 Feb 10;25(1):221. [doi: 10.1186/s12913-025-12373-w] [Medline: 39924512]
4. Wals Zurita AJ, Miras Del Rio H, Ugarte Ruiz de Aguirre N, et al. The transformative potential of large language models in mining electronic health records data: content analysis. JMIR Med Inform 2025 Jan 2;13:e58457. [doi: 10.2196/58457] [Medline: 39746191]
5. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024 Sep 28;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
6. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform 2025 Jan 19;32(1):e101139. [doi: 10.1136/bmjhci-2024-101139] [Medline: 39832824]
7. Croxford E, Gao Y, Pellegrino N, et al. Development and validation of the provider documentation summarization quality instrument for large language models. J Am Med Inform Assoc 2025 Jun 1;32(6):1050-1060. [doi: 10.1093/jamia/ocaf068] [Medline: 40323321]
8. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
9. Patil R, Heston TF, Bhuse V. Prompt engineering in healthcare. Electronics (Basel) 2024;13(15):2961. [doi: 10.3390/electronics13152961]
10. Zhu L, Mou W, Hong C, et al. The evaluation of generative AI should include repetition to assess stability. JMIR Mhealth Uhealth 2024 May 6;12:e57978. [doi: 10.2196/57978] [Medline: 38688841]
11. Visual collaborationfor creative workand education. URL: https://padlet.com/ [accessed 2026-02-18]
12. Napoli EA, Barbàra F, Gatteschi V, Schifanella C. Leveraging large language models for automatic smart contract generation. In: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC): IEEE; 2024. [doi: 10.1109/COMPSAC61105.2024.00100]
13. Bomfim FD, Neto JA, Bezerra Filho G, et al, editors. SARA - a generative AI for legal process summarization based on chain of density prompt engineering. In: Intelligent Systems: Springer; 2025. [doi: 10.1007/978-3-031-79032-4_26]
14. Teo S. How I won Singapore's GPT-4 prompt engineering competition. Medium. 2023 Dec 29. URL: https://medium.com/data-science/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41 [accessed 2026-02-14]
15. Balderas-Díaz S, Rodríguez-Fórtiz MJ, Garrido JL, Bellido-González M, Guerrero-Contreras G. A psycho-educational intervention programme for parents with SGA foetuses supported by an adaptive mHealth system: design, proof of concept and usability assessment. BMC Med Inform Decis Mak 2022 Nov 11;22(Suppl 4):291. [doi: 10.1186/s12911-022-02036-9] [Medline: 36357878]
16. Balderas-Díaz S, Guerrero-Contreras G, editors. ADEPT framework: optimizing data flow in dynamic environments for ambient assisted living. In: Ambient Intelligence: Springer; 2025. [doi: 10.1007/978-3-031-83117-1_12]
17. FRAIT (framework & implementation of AI tools). GitHub. URL: https://github.com/miekedeschepperUZG/FRAIT [accessed 2026-02-18]

Abbreviations

AI: artificial intelligence
CO-STAR: context, objective, style, tone, audience, and response
FRAIT: Framework & Implementation of AI Tools
LLM: large language model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Benis; submitted 14.Jul.2025; peer-reviewed by G Guerrero-Contreras, J Declerck, S Patel; revised version received 12.Dec.2025; accepted 01.Jan.2026; published 04.Mar.2026.

Please cite as:
Deschepper M, Rogge H, Syx M, Colpaert K
Tailoring Discharge Summaries to Health Care Providers' Needs (Part 1 of the Framework and Implementation of AI Tools Project): User-Centered Design Approach
JMIR Med Inform 2026;14:e80613
URL: https://medinform.jmir.org/2026/1/e80613
doi:10.2196/80613

© Mieke Deschepper, Helga Rogge, Mathias Syx, Kirsten Colpaert. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 4.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating GPT-4 Responses on Scars or Keloids for Patient Education: Large Language Model Evaluation Study

Mingjun Rao, MD; Tang Xiujun, MSc; Wang Haoyu, BA

Department of Plastic Surgery, Guizhou Provincial People's Hospital, 83 Zhongshan East Road, Nanming District, Guiyang, China

Corresponding Author:
Tang Xiujun, MSc
Department of Plastic Surgery, Guizhou Provincial People's Hospital, 83 Zhongshan East Road, Nanming District, Guiyang, China

Abstract

Background: Scars and keloids impose significant physical and psychological burdens on patients, often leading to functional limitations, cosmetic concerns, and mental health issues such as anxiety or depression. Patients increasingly turn to online platforms for information; however, existing web-based resources on scars and keloids are frequently unreliable, fragmented, or difficult to understand. Large language models such as GPT-4 show promise for delivering medical information, but their accuracy, readability, and potential to generate hallucinated content require validation for patient education applications.

Objective: This study aimed to systematically evaluate GPT-4's performance in providing patient education on scars and keloids, focusing on its accuracy, reliability, readability, and reference quality.

Methods: This study involved collecting 354 questions from Reddit communities (r/Keloids, r/SCAR, and r/PlasticSurgery), covering topics including treatment options, pre- and postoperative care, and psychological impacts. Each question was input into GPT-4 in independent sessions to mimic real-world patient interactions. Responses were evaluated using multiple tools: the Patient Education Materials Assessment Tool-Artificial Intelligence for understandability and actionability, DISCERN-AI for treatment information quality, the Global Quality Scale for overall information quality, and standard readability metrics (Flesch Reading Ease score, and Gunning Fog Index). Three plastic surgeons used the Natural Language Assessment Tool for Artificial Intelligence to rate the accuracy, safety, and clinical appropriateness, while the Reference Evaluation for Artificial Intelligence tool validated references for reference hallucination, relevance, and source quality. We conducted the same analysis to assess the quality of GPT-4–generated content in response to questions from 3 medical websites.

Results: GPT-4 demonstrated high accuracy and reliability. The Patient Education Materials Assessment Tool-Artificial Intelligence showed 75.5% understandability, DISCERN-AI rated responses as "good" (26.3/35), and the Global Quality Scale score was 4.28 of 5. Surgeons' evaluations averaged 3.94 to 4.43 out of 5 across dimensions (accuracy 3.9, SD 0.7; safety 4.3, SD 0.8; clinical appropriateness 4.4, SD 0.5; actionability 4.1, SD 0.8; and effectiveness 4.1, SD 0.8). Readability analyses indicated moderate complexity (Flesch Reading Ease Score: 50.13; Gunning Fog Index: 12.68), corresponding to a 12th-grade reading level. Reference Evaluation for Artificial Intelligence identified 11.8% (383/3250) hallucinated references, while 88.2% (2867/3250) of references were real, with 95.1% (2724/2867) from authoritative sources (eg, government guidelines and the literature). The overall results about questions from medical websites were consistent with the answers to Reddit questions.

Conclusions: GPT-4 has serious potential as a patient education tool for scars and keloids, offering reliable and accurate information. However, improvements in readability (to align with sixth to eighth grade standards) and reduction of reference hallucinations are essential to enhance accessibility and trustworthiness. Future large language model optimizations should prioritize simplifying medical language and strengthening reference validation mechanisms to maximize clinical utility.

(JMIR Med Inform 2026;14:e78838) doi:10.2196/78838

KEYWORDS

scar; keloid; GPT-4; patient education; generative AI; generative artificial intelligence; readability; large language model

Introduction

Scars and keloids are common skin healing outcomes [1], often causing discomfort during the proliferative phase [2]. When located on visible areas such as the face, they can severely impact a patient's appearance, leading to psychological distress such as low self-esteem, anxiety, and depression, which may further hinder social interactions and career development [3]. Scars near joints may cause functional limitations, while perineal scars can result in long-term complications such as dyspareunia and dysmenorrhea [4,5]. Many patients urgently seek to improve both the aesthetic and functional aspects of scars or keloids. However, treatments often require multimodal approaches over weeks to months, making high patient adherence crucial.

XSL•FO
RenderX

Consequently, access to accurate, timely, and comprehensive management information is critical for patients to better understand treatment expectations, options, prognosis, and potential complications [6-9].

Currently, patients increasingly rely on internet-based health information [10]. This trend faces multifaceted challenges, including information overload, variability in source credibility and content accuracy, and the health literacy required to understand the contents [11,12]. Notably, many online resources on scars or keloids are often unreliable, fragmented, or difficult to understand, failing to tackle the fundamental needs of patients with scars or keloids [13].

In recent years, artificial intelligence (AI) tools powered by large language models (LLMs), such as GPT-4 (used by ChatGPT), have demonstrated significant potential in delivering medical information [14]. GPT-4's capacity to generate natural language responses through interactive conversions could aid users in understanding intricate medical concepts, treatment, and management strategies, positioning it as a potentially valuable alternative to traditional search engines for accessing knowledge associated with scars or keloids [15-17].

It is reported that 52% of US adults have used LLMs, and GPT-4, as a leading LLM, receives over 5 billion monthly visits. In total, 39% of LLM users have used LLMs for health care queries [18]. Despite the increasing use of LLMs in health care [19-21], there remains a research gap, and it is currently unclear whether GPT-4 can generate high-quality patient education content related to scars and keloids. Thus, we conducted this study to comprehensively assess the use of GPT-4 in keloid and scar patient education by performing a multidimensional evaluation (encompassing accuracy, reliability, readability, and reference quality) of GPT-4 responses.

Methods

Study Objective

This study aimed to investigate the potential of GPT-4 to provide reliable, accurate, readable, and actual medical information for patients with scars or keloids. To achieve this, we used GPT-4 (OpenAI) to evaluate its accuracy, reliability, readability, and hallucinations in answering questions related to treatments of scars or keloids.

Question Collection

The research questions were manually collected by the authors from Reddit. First, the authors analyzed all posts on the "Hot" page (the most popular and recently active posts) of the r/Keloids subreddit, all posts on the r/SCAR subreddit, and all posts on the r/PlasticSurgery subreddit as of April 6, 2025. We extracted the main text of each post that included the keywords "Scar" or "Keloid" and organized the data using Microsoft Excel. Second, one author (MR) excluded the posts that contained no questions, were duplicates, or had repeated questions. In addition, the same author (MR) performed the initial classification of the questions. To ensure reliability and minimize bias, another author (WH) independently reviewed the process. Consistency between the two authors' classifications was confirmed through discussion. This data collection approach has been adopted in previous Reddit-based research [22]. Furthermore, we adopted 49 questions about keloids or scars from 3 medical websites.

Ethical Considerations

The institutional review board of the People's Hospital of Guizhou Province, affiliated with Guizhou University, deemed this study exempt from ethics approval.

Quality Assessment

Overview

Each question was input individually into GPT-4. Following previous research protocols, a new chat session was initiated for each question to avoid context contamination and to simulate real-world patient interactions [23]. Consistent with real-world activities using GPT-4, no specialized prompt instructions were appended to the question inputs. The contents generated by ChatGPT-4 were evaluated using a modified version of existing health information quality assessment tools.

Patient Education Materials Assessment Tool-AI Tool

The Patient Education Materials Assessment Tool (PEMAT) was used to assess the understandability and actionability of ChatGPT-generated content. The original PEMAT includes 17 items for understandability and 7 for actionability. Since all contents generated by ChatGPT are pure text, the PEMAT was simplified to 8 items for understandability and 3 items for actionability (Multimedia Appendix 1). Each item received 1 point if it met the standard, and scores were reported as percentages. A score of 70% or greater was recorded as a "pass" per PEMAT guidelines [24].

DISCERN-AI Tool

The DISCERN standard, a previously validated tool to help health care consumers and professionals evaluate the quality of treatment information, was adapted for ChatGPT-generated content. Since all contents generated by ChatGPT are pure text, 7 items (questions 3 - 9 from the 15-item DISCERN tool) were selected and scored on a 1 to 5 scale (Multimedia Appendix 2). Each output was rated as follows: very poor (7 - 12 points), poor (13 - 17 points), fair (18 - 23 points), good (24 - 28 points), and excellent (29 - 35 points) [24,25].

Global Quality Scale

The Global Quality Scale (GQS) is a 5-point Likert scale used to evaluate information quality and the flow and ease of use of information. The scores range from 1 (low quality) to 5 (high quality), while scores of 4 or 5 indicated high-quality outputs, a score of 3 was considered moderate quality and scores of 1 or 2 were categorized as low quality.

Readability Assessment

The readability of the ChatGPT-generated content was evaluated using several established readability formulas, including Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, Coleman-Liau Index, and Simple Measure of Gobbledygook (SMOG). Each output was copied into Microsoft Word and analyzed via the Readable website [26]. The Flesch Reading Ease score ranges from 0 to 100, and higher scores

XSL•FO
RenderX

indicate greater readability. A score between 60 and 70 corresponds to reading levels of grades 8 and 9 and is generally understandable by the average adult. The Gunning Fog Index and Flesch-Kincaid Grade Level are used to estimate sentence complexity; the scores represent the years of formal education required to understand the contents. For example, a score of 12 implies the output is suitable for readers at the 12th-grade level. The Coleman-Liau Index is similar to the Gunning Fog Index and Flesch-Kincaid Grade Level but uses character counts instead of syllables, making it more suitable for languages where syllable counts may not accurately reflect complexity. The SMOG Index measures syllable density, often used to assess health information materials. A score of 12 indicates that the material is suitable for readers at the 12th-grade level or higher.

Natural Language Assessment Tool for Artificial Intelligence

Three experienced plastic surgeons independently reviewed each GPT-4–generated content using a specially developed Natural Language Assessment Tool for Artificial Intelligence (NLAT-AI) [24]. Using this tool, we assessed accuracy, safety, appropriateness, actionability, and effectiveness. Each output was rated using a 5-point Likert scale (1=strongly disagree, 5=strongly agree; Multimedia Appendix 3). All results were summarized using descriptive statistics.

Reference Evaluation for AI

Given known issues of LLM hallucination (ie, generating plausible but nonexistent references), a brief evaluation tool, Reference Evaluation for AI, was developed to analyze references provided in ChatGPT-generated content [27]. Each reference was verified through direct links or a Google search. The tool assessed (1) reference hallucination (whether references were real or fabricated), (2) relevance and consistency between references and AI output, and (3) source quality (based on the authority of the issuing institution or organization, such as government guidelines, health care organizations, or scientific research; Multimedia Appendix 4).

Results

Question Collection and Classification

A total of 507 posts were identified and analyzed (posts from the r/Keloids subreddit: n=193, 38.1%; posts from the r/Keloids subreddit: n=211, 41.6%; and posts from the r/Scar subreddit: n=103, 20.3%). After removing posts that merely shared information or were duplicates, 354 unique questions were obtained. The questions were categorized into 16 groups based on their contents (Table 1). Furthermore, we obtained 49 questions from 3 medical websites that included 38 unique questions (Table S1 in Multimedia Appendix 5).

Table . Questions on scars or keloids from Reddit (N=354).

Question group	Questions, n (%)
Questions on other respects	28 (7.9)
Questions on other treatments for scars or keloids	4 (1.1)
Questions on common treatments for scars or keloids	46 (13)
Questions on trauma-related scars or keloids	16 (4.5)
Questions on psychological issues caused by scars or keloids	9 (2.5)
Questions on at-home scar or keloid care	3 (0.8)
Questions on preoperative scar or keloid consultation	37 (10.5)
Questions on postoperative scar or keloid consultation	55 (15.5)
Questions on selection of treatments for scars or keloids	80 (22.6)
Questions on impact of scars or keloids on daily life	2 (0.6)
Questions on scar or keloid symptoms	7 (2)
Questions on scar camouflage	6 (1.7)
Questions on the impact of nutrition on scars or keloids	3 (0.8)
Questions on choosing physicians for scar or keloid treatment or related costs	32 (9)
Questions on old scars	14 (4)
Questions on scar or keloid prevention	12 (3.4)

Evaluation of GPT-4–Generated Content

GPT-4 generated content that provided a wide range of medically accurate information. Using the PEMAT-AI, DISCERN-AI, and GQS patient education material evaluation tools, the output of GPT-4 was assessed, with all tools indicating high scores. The overall understandability score using PEMAT-AI easily surpassed the 70% threshold for acceptability (mean 75.5%, SD 12.2%). The DISCERN-AI tool resulted in an overall rating of “good” quality (mean 26.3, SD 3.4), with all 16 groups of questions rated as “good.” The GQS score averaged 4.3 out of 5 (SD 0.8), categorizing the outputs as high quality. More details are shown in Table S1 in Multimedia Appendix 6. Intraclass correlation coefficient (ICC) for

XSL•FO
RenderX

PEMAT-AI, DISCERN-AI, and GQS were 0.73, 0.69, and 0.78, respectively (Table S2 in Multimedia Appendix 5). The results of the ICC demonstrated high reliability of the evaluation tools.

Plastic Surgeons' Evaluation via the NLAT-AI Tool

Using the NLAT-AI tool, 3 independent plastic surgeons evaluated the GPT-4–generated content. All dimensions of the contents received scores above the neutral midpoint of 3 on a 5-point Likert scale. The overall average scores for each dimension were as follows: accuracy 3.9 (SD 0.7), safety 4.3 (SD 0.8), appropriateness 4.4 (SD 0.5), actionability 4.1 (SD 0.8), and effectiveness 4.1 (SD 0.8). More detailed descriptive statistics for each question are presented in Table S2 in Multimedia Appendix 6. Internal validity tests showed an ICC of 0.76 (Table S2 in Multimedia Appendix 5), indicating high reliability.

Readability Assessment

The results of the readability assessments indicated that the GPT-4–generated content was "difficult to read." The average Flesch Reading Ease score was 50.1 (SD 8.1), which is considered moderately difficult. The Gunning Fog Index averaged 12.7 (SD 3.3), and the Flesch-Kincaid Grade Level was 12.4 (SD 2.5), indicating that the text was at a high school level (approximately suitable for individuals aged 16 - 17 years). The Coleman-Liau Index averaged 12.8 (SD 2.6), and the SMOG Index averaged 11.3 (SD 3.16). More detailed evaluation results are shown in Table S3 in Multimedia Appendix 6.

Reference Evaluation for AI Assessment

Most of the references provided in GPT-4's output effectively supported the content. A total of 88.2% (2867/3250) of the references were from actual sources (actual websites or academic papers), while 383 hallucinated references were identified. Among these 2867 real references, 2746 (95.8%) references effectively supported the content. In addition, a total of 95.1% (2724/2867) of the real references were from authoritative sources (government guidelines, health care organizations, or scientific research). More detailed evaluation results are shown in Table S4 in Multimedia Appendix 6.

The Assessment of Questions From Websites

The evaluation results of GPT-4 responses to website-sourced questions were broadly consistent with those from Reddit-derived questions across all assessments (Tables S3-S6 in Multimedia Appendix 1).

Discussion

Principal Findings

This is the first study to assess the overall quality of ChatGPT responses to real-world questions from Reddit about keloids or scars. The results revealed that the content generated by GPT-4 was generally comprehensive and aligned with current medical guidelines and the literature. Using several assessment tools, as well as plastic surgeons' evaluations, the scores were robust, and the plastic surgeons' evaluations were largely positive. The overall results indicate that GPT-4–generated content is reliable, accurate, safe, and actionable, despite there being room for improvement in terms of readability and hallucination.

Over 80% of dermatology outpatients obtain medical information through social media or the internet, with 47% considering it an important source of information [28]. Although patients have access to a wealth of information, studies evaluating the quality of online health information have identified significant deficiencies [29]. As for scars and keloids, the information available to patients contains a lot of low-quality content. A previous study assessing 88 websites related to "burn scars" showed that most of the commercial websites provided information of moderate to poor quality [13]. In contrast, LLMs provide a broad range of fundamentally accurate information and real-time dynamic interactions compared to traditional webpages [30,31]. As a leading LLM, GPT-4 exhibits certain advantages over other LLMs and has demonstrated top-tier performance across diverse evaluations in health care. In answering questions from the American Board of Surgery In-Training Examination, GPT-4 achieved an accuracy rate comparable to that of Copilot, while significantly outperforming Gemini [32]. In other fields of clinical medicine, GPT-4 also attained superior performance relative to other LLMs [33,34]. However, in a substantial number of evaluative scenarios, the performance of GPT-4 did not yield statistically significant differences when compared with Copilot or Gemini. Collectively, the performance of GPT-4 currently represents the best capability of LLMs.

In our study, experienced plastic surgeons evaluated the outputs of GPT-4, confirming that the contents were reliable and accurate. The accuracy of GPT-4 in patient education has also been studied in other clinical contexts (eg, rhinoplasty, sleep apnea, and prostate cancer) where it demonstrated high accuracy and strong reliability [24,35,36]. Such high accuracy and reliability suggest that LLMs such as GPT-4 can effectively address clinical questions from patients with scars or keloids, serving as a valuable auxiliary tool in clinical medicine.

Despite GPT-4's significant potential in responding to keloid or scar patient queries, its outputs commonly had high reading difficulty. Our study revealed that the average reading level of GPT-4–generated content was at a high school level. The results suggest that ChatGPT does not always meet the comprehension needs of all patients. The relatively low readability of GPT-4 can hinder accessibility for certain socioeconomic populations with limited health literacy [37]. Among the latest generation of young adults in the United States, up to 13% have not graduated from high school. This rate reaches 20% among people of color (including African Americans and Native Americans) [38], who are also identified as high-risk groups for developing malignant scars [39]. Due to poor readability, GPT-4 has apparent barriers in its application among these populations [40,41]. To enhance the utility of LLMs for populations with lower educational attainment, it is recommended that developers consider training specialized LLMs based on datasets with good readability [42]. Biomedical text can be simplified through hyperparameter substitution techniques, improving patient understanding [43]. In addition, structured prompting can also contribute to enhancing readability [44].

XSL•FO
RenderX

Moreover, our study also revealed the existence of hallucination, where GPT-4 cited nonexistent references or websites. Fabricated references not only mislead readers and distort their understanding of keloid or scar but also—given the presence of numerous seemingly authoritative yet false information sources—may lead patients to overtrust the content generated by GPT-4 [45,46]. Given the presence of hallucinations, specific clinical diagnosis and treatment must rely on clinicians; LLMs can only serve as auxiliary tools. To address the hallucination issue in LLMs, it is recommended that developers effectively apply retrieval-augmented generation to retrieve documents from an external corpus (such as academic library systems), as this can significantly reduce the hallucinations [47-49]. Integrating external, structured knowledge sources (such as knowledge graphs, databases, or other domain-specific resources) into LLMs can also help ensure that LLMs produce responses with fewer hallucinations [50]. Furthermore, prompt engineering can mitigate hallucination by improving the reasoning capabilities [51].

GPT-4 can provide comprehensive and generally accurate information, which can further assist patients with keloids or scars in accessing timely and precise information. However, current LLMs exhibit limitations, such as hallucinations and relatively low readability; therefore, they are not recommended as the sole source of information for patients. Limited by the lack of clinical background in current LLMs; the insufficient ability to process audio, image, and video information; limited ability to access academic libraries; and the noninterpretability of black box algorithms, current LLMs still require further development to be adapted for applications in health care [52]. The AI agent, as a promising approach, can extend the capabilities of LLMs by enabling them to use external tools, plan and execute multistep tasks, as well as interact dynamically [53]. Multimodal LLM is promising to process text (eg, clinical notes and user-input questions), medical images (eg, photos and computed tomography scans), and videos (eg, treatment procedures) provided by patients, which will more effectively assist patients and health care providers in clinical practice about keloid and scar management [54].

Limitations

Most of the questions collected from Reddit were posts from patients who had not yet sought medical care. Consequently, the questions posed may be biased toward pretreatment information needs, as fewer questions were reported during the treatment phase. This may compromise the generalizability of GPT-4's evaluation across different patient care stages. In addition, Reddit users are concentrated in the age group of 18 to 49 years, with an average age of 23 years, and the majority are aged under 30 years. Thus, the data collected from Reddit clearly fails to represent the middle-aged and older population [55]. Relying solely on Reddit posts for data collection introduces demographic selection bias.

In terms of assessment tools, the qualitative assessment conducted by experienced plastic surgeons was inherently at risk of bias, given the surgeons' attitudes toward the use of GPT-4. Nevertheless, they provided valuable insights owing to their in-depth understanding of scar and keloid education materials. Furthermore, exploratory assessment tools (DISCERN-AI, PEMAT-AI, and NLAT-AI) were used in this study, while their validity requires further testing. LLMs differ from traditional printed educational materials in that their responses to repeated queries of the same question are generated instantaneously and may vary. Currently, existing assessment tools lack the ability to detect such variability in LLM outputs when the same question is posed multiple times [56]. Furthermore, content generated by LLMs is often conveyed with excessive certainty, as these models lack the ability to accurately express information involving uncertainties. Providing definitive answers to such uncertain content may mislead patients, yet current assessment scales fail to evaluate this critical limitation [57,58]. Further research is needed to develop specific tools to enable more robust evaluation of LLM output quality.

Conclusions

Our analysis found that GPT-4 provided high-quality responses to real-world questions related to scars and keloids, suggesting its potential as a useful patient education tool in scar and keloid treatment. The GPT-4 outputs were generally reliable and accurate but need improvement, primarily in readability and hallucinations.

Funding

This study is funded by the Talent Fund of Guizhou Provincial People's Hospital (awarded to MR; [2023]-30).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient Education Materials Assessment Tool for Artificial Intelligence for evaluating the understandability (8 items) and actionability (3 items) of artificial intelligence–generated patient education text.
[DOCX File, 18 KB - medinform_v14i1e78838_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
DISCERN-AI tool (7 core items) for assessing artificial intelligence–generated treatment information quality, with 1-3 scoring for each item (relevance, source clarity, date transparency, balance, additional support, uncertainty acknowledgment, and overall quality).
[DOCX File, 21 KB - medinform_v14i1e78838_app2.docx]

Multimedia Appendix 3
Natural Language Assessment Tool for Artificial Intelligence assessment framework: 5 domains (accuracy, safety, appropriateness, actionability, and effectiveness).
[DOCX File, 24 KB - medinform_v14i1e78838_app3.docx]

Multimedia Appendix 4
Reference Evaluation for AI assessment criteria: 3-item tool for validating artificial intelligence–generated references (real or fabricated, content support, and authoritative source status) on large language model reference hallucinations.
[DOCX File, 20 KB - medinform_v14i1e78838_app4.docx]

Multimedia Appendix 5
Supplementary tables for health care website–derived scar or keloid questions: includes 38 unique questions from 3 medical websites (Table S1), intraclass correlation coefficient values for assessment tools (Table S2), and artificial intelligence output evaluation scores (Patient Education Materials Assessment Tool for Artificial Intelligence, DISCERN-AI, Global Quality Scale, Natural Language Assessment Tool for Artificial Intelligence, readability, and reference quality) for website questions (Tables S3-S6).
[DOCX File, 23 KB - medinform_v14i1e78838_app5.docx]

Multimedia Appendix 6
Supplementary tables for Reddit-derived scar or keloid questions: includes subcategory-specific artificial intelligence output evaluation scores for all 16 question groups (Patient Education Materials Assessment Tool for Artificial Intelligence, DISCERN-AI, or Global Quality Scale: Table S1; Natural Language Assessment Tool for Artificial Intelligence: Table S2; and readability metrics: Table S3) and subcategory-specific reference evaluation results (Reference Evaluation for AI) for 3250 total cited references (Table S4), plus overall aggregate scores for all assessments.
[DOCX File, 25 KB - medinform_v14i1e78838_app6.docx]

References

1. Stoica AE, Grumezescu AM, Hermenean AO, Andronescu E, Vasile BS. Scar-free healing: current concepts and future perspectives. Nanomaterials (Basel) 2020 Oct 31;10(11):2179. [doi: 10.3390/nano10112179] [Medline: 33142891]
2. Xiao Y, Sun Y, Zhu B, et al. Risk factors for hypertrophic burn scar pain, pruritus, and paresthesia development. Wound Repair Regen 2018 Mar;26(2):172-181. [doi: 10.1111/wrr.12637] [Medline: 29719102]
3. Hsieh JC, Maisel-Campbell AL, Joshi CJ, Zielinski E, Galiano RD. Daily quality-of-life impact of scars: an interview-based foundational study of patient-reported themes. Plast Reconstr Surg Glob Open 2021 Apr;9(4):e3522. [doi: 10.1097/GOX.0000000000003522] [Medline: 33868874]
4. Schouten H, Nieuwenhuis M, van der Schans C, Niemeijer A, van Zuijlen P. Considerations in determining the severity of burn scar contractures with focus on the knee joint. J Burn Care Res 2023 Jul 5;44(4):810-816. [doi: 10.1093/jbcr/irad016] [Medline: 36752774]
5. Woodward AP, Matthews CA. Outcomes of revision perineoplasty for persistent postpartum dyspareunia. Female Pelvic Med Reconstr Surg 2010 Mar;16(2):135-139. [doi: 10.1097/SPV.0b013e3181cc8702] [Medline: 22453161]
6. Waibel JS, Waibel H, Sedaghat E. Scar therapy of skin. Facial Plast Surg Clin North Am 2023 Nov;31(4):453-462. [doi: 10.1016/j.fsc.2023.06.005] [Medline: 37806679]
7. Gomolin T, Cline A, Ginsberg D, Safai B. Scar tissue I wish you saw: patient expectations regarding scar treatment. J Cosmet Dermatol 2021 Sep;20(9):2739-2742. [doi: 10.1111/jocd.13945] [Medline: 33434326]
8. Cho SB, Ryu DJ, Lee SJ, et al. Scar characteristics and treatment expectations: a survey of 589 patients. J Cosmet Laser Ther 2009 Dec;11(4):224-228. [doi: 10.3109/14764170903341723] [Medline: 19951193]
9. Andrews N, Jones LL, Moiemen N, et al. Below the surface: parents' views on the factors that influence treatment adherence in paediatric burn scar management - a qualitative study. Burns 2018 May;44(3):626-635. [doi: 10.1016/j.burns.2017.09.003] [Medline: 29031888]
10. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav 2016 May;58:240-248. [doi: 10.1016/j.chb.2016.01.007]
11. Battineni G, Baldoni S, Chintalapudi N, et al. Factors affecting the quality and reliability of online health information. Digit Health 2020;6:2055207620948996. [doi: 10.1177/2055207620948996] [Medline: 32944269]

XSL•FO
RenderX

12. Khaleel I, Wimmer BC, Peterson GM, et al. Health information overload among health consumers: a scoping review. Patient Educ Couns 2020 Jan;103(1):15-32. [doi: 10.1016/j.pec.2019.08.008] [Medline: 31451363]
13. Bohacek L, Gomez M, Fish JS. An evaluation of internet sites for burn scar management. J Burn Care Rehabil 2003;24(4):246-251. [doi: 10.1097/01.BCR.0000075844.04297.D9] [Medline: 14501424]
14. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 2024 Mar;245:108013. [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
15. Neha F, Bhati D, Shukla DK, Amiruzzaman M. ChatGPT: transforming healthcare with AI. AI 2024;5(4):2618-2650. [doi: 10.3390/ai5040126]
16. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023 Dec;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
17. Iqbal U, Lee LTJ, Rahmanti AR, Celi LA, Li YCJ. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? J Am Med Inform Assoc 2024 May 20;31(6):1341-1347. [doi: 10.1093/jamia/ocae067] [Medline: 38578616]
18. Lee R. Close encounters of the AI kind: the increasingly human-like way people are engaging with language models.: Elon University URL: https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report [accessed 2025-11-26]
19. Gencer G, Gencer K. Large language models in healthcare: a bibliometric analysis and examination of research trends. J Multidiscip Healthc 2025;18:223-238. [doi: 10.2147/JMDH.S502351] [Medline: 39844924]
20. Kumar D, Sood SK, Rawat KS. Empowering elderly care with intelligent IoT-driven smart toilets for home-based infectious health monitoring. Artif Intell Med 2023 Oct;144:102666. [doi: 10.1016/j.artmed.2023.102666] [Medline: 37783534]
21. Kumar D, Rawat KS, Sood SK. Revolution of artificial intelligence and IOT in healthcare: a keyword co-occurrence network analysis using CiteSpace. In: Shukla AK, Thakur DG, Arabkoohsar A, editors. Recent Advances in Mechanical Engineering 2023:231-237. [doi: 10.1007/978-981-99-2349-6_20]
22. Wang J, Patel P, Jagdeo J. An analysis of keloid patient questions on Reddit. Wound Repair Regen 2024;32(2):164-170. [doi: 10.1111/wrr.13160] [Medline: 38372454]
23. Campbell DJ, Estephan LE, Sina EM, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid 2024 Mar;34(3):371-377. [doi: 10.1089/thy.2023.0491] [Medline: 38010917]
24. Gibson D, Jackson S, Shanmugasundaram R, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res 2024 Aug 14;26:e55939. [doi: 10.2196/55939] [Medline: 39141904]
25. Cassidy JT, Baker JF. Orthopaedic patient information on the world wide web: an essential review. J Bone Joint Surg Am 2016 Feb 17;98(4):325-338. [doi: 10.2106/JBJS.N.01189] [Medline: 26888683]
26. Readable. Our Readability Checker helps you to communicate clearly. URL: https://readable.com [accessed 2026-02-24]
27. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
28. AlGhamdi KM, Almohideb MA. Internet use by dermatology outpatients to search for health information. Int J Dermatol 2011 Mar;50(3):292-299. [doi: 10.1111/j.1365-4632.2010.04705.x] [Medline: 21342162]
29. Zhang Y, Kim Y. Consumers' evaluation of web-based health information quality: meta-analysis. J Med Internet Res 2022 Apr 28;24(4):e36463. [doi: 10.2196/36463] [Medline: 35482390]
30. Monib WK, Qazi A, Mahmud MM. Exploring learners' experiences and perceptions of ChatGPT as a learning tool in higher education. Educ Inf Technol 2025 Jan;30(1):917-939. [doi: 10.1007/s10639-024-13065-4]
31. Jin Y, Liu H, Zhao B, Pan W. ChatGPT and mycosis- a new weapon in the knowledge battlefield. BMC Infect Dis 2023 Oct 27;23(1):731. [doi: 10.1186/s12879-023-08724-9] [Medline: 37891532]
32. Sanli AN, Tekcan Sanli DE, Karabulut A. Can American Board of Surgery in Training Examinations be passed by large language models? Comparative assessment of Gemini, Copilot, and ChatGPT. Am Surg 2025 Nov;91(11):1923-1929. [doi: 10.1177/00031348251341956] [Medline: 40353502]
33. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. Comparison of Gemini Advanced and ChatGPT 4.0's performances on the Ophthalmology Resident Ophthalmic Knowledge Assessment Program (OKAP) examination review question banks. Cureus 2024 Sep;16(9):e69612. [doi: 10.7759/cureus.69612] [Medline: 39421095]
34. Tan YZ, Nah SA, Saw SN, Rajandram R, Ong TA. Evaluating the performance of artificial intelligence chatbots in answering urology questions derived from guidelines or board examinations: a systematic review. Urol Sci 2025:10. [doi: 10.1097/us9.0000000000000089]
35. Lim B, Seth I, Kah S, et al. Using generative artificial intelligence tools in cosmetic surgery: a study on rhinoplasty, facelifts, and blepharoplasty procedures. J Clin Med 2023 Oct 14;12(20):6524. [doi: 10.3390/jcm12206524] [Medline: 37892665]
36. Incerti Parenti S, Bartolucci ML, Biondi E, et al. Online patient education in obstructive sleep apnea: ChatGPT versus Google search. Healthcare (Basel) 2024 Sep 5;12(17):1781. [doi: 10.3390/healthcare12171781] [Medline: 39273804]
37. Toiv A, Saleh Z, Ishak A, et al. Digesting digital health: a study of appropriateness and readability of ChatGPT-generated gastroenterological information. Clin Transl Gastroenterol 2024 Nov 1;15(11):e00765. [doi: 10.14309/ctg.0000000000000765] [Medline: 39212302]

XSL•FO
RenderX

38. Veronique I. Report on the condition of education 2024. : Institute of Education Sciences; 2024 URL: https://nces.ed.gov/pubs2024/2024144.pdf
39. Oei F, Putra IB, Jusuf NK. The relationship between skin color and keloid. Bali Med J 2021;10(2):835-838. [doi: 10.15562/bmj.v10i2.2619]
40. DeTemple DE, Meine TC. Comparison of the readability of ChatGPT and Bard in medical communication: a meta-analysis. BMC Med Inform Decis Mak 2025 Sep 1;25(1):325. [doi: 10.1186/s12911-025-03035-2] [Medline: 40890707]
41. Whittaker P, Sun M. Quality and readability of chatbot responses to patient questions: a systematic cross-sectional meta-synthesis. Health Informatics J 2025;31(4):14604582251388879. [doi: 10.1177/14604582251388879] [Medline: 41106853]
42. Li M, Zhang Y, Li Z, et al. From quantity to quality: boosting LLM performance with self-guided data selection for instruction tuning. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico p. 7602-7635. [doi: 10.18653/v1/2024.naacl-long.421]
43. Swanson K, He S, Calvano J, et al. Biomedical text readability after hypernym substitution with fine-tuned large language models. PLOS Digit Health 2024 Apr;3(4):e0000489. [doi: 10.1371/journal.pdig.0000489] [Medline: 38625843]
44. Daulat SR, Dholaria N, Burnet G, et al. Prompt engineering and follow-up questioning improves the readability of spine surgery questions in large language models. World Neurosurg 2025 Nov;203:124423. [doi: 10.1016/j.wneu.2025.124423] [Medline: 40889596]
45. Aljohani M, Hou J, Kommu S, Wang X. A comprehensive survey on the trustworthiness of large language models in healthcare. Presented at: Proceedings of the 2025 Findings of the Association for Computational Linguistics EMNLP'25; Nov 4-9, 2025; Suzhou, China p. 6720-6748. [doi: 10.18653/v1/2025.findings-emnlp.356]
46. Kim Y, Jeong H, Chen S, et al. Medical hallucination in foundation models and their impact on healthcare. Health Systems and Quality Improvement. Preprint posted online on 2025. [doi: 10.1101/2025.02.28.25323115]
47. Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics 2025;13(5):856. [doi: 10.3390/math13050856]
48. Bhattacharya R. Strategies to mitigate hallucinations in large language models. AMA 2024;10(1):62. [doi: 10.69554/NXXB8234]
49. Wołk K. Evaluating retrieval-augmented generation variants for clinical decision support: hallucination mitigation and secure on-premises deployment. Electronics (Basel) 2025;14(21):4227. [doi: 10.3390/electronics14214227]
50. Agrawal G, Kumarage T, Alghamdi Z, Liu H. Can knowledge graphs reduce hallucinations in LLMs?: a survey. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico p. 3947-3960. [doi: 10.18653/v1/2024.naacl-long.219]
51. Zhang H, Deng H, Ou J, Feng C. Mitigating spatial hallucination in large language models for path planning via prompt engineering. Sci Rep 2025;15(1):8881. [doi: 10.1038/s41598-025-93601-5]
52. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
53. Ferrag MA, Tihanyi N, Debbah M. From LLM reasoning to autonomous ai agents: a comprehensive review. arXiv. Preprint posted online on Apr 28, 2025. [doi: 10.48550/arXiv.2504.19678]
54. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024 Sep 25;26:e59505. [doi: 10.2196/59505] [Medline: 39321458]
55. Reddit user age, gender, & demographics. Exploding Topics. 2025. URL: https://explodingtopics.com/blog/reddit-users [accessed 2025-11-26]
56. Calloway C. Why do different LLMs give different answers to the same question? Model uncertainty and variability in LLM-based intrusion detection systems ranking: Norfolk State University; 2025. URL: https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1123&context=covacci-undergraduateresearch [accessed 2024-05-20]
57. Yona G, Aharoni R, Geva M. Can large language models faithfully express their intrinsic uncertainty in words? Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, FL p. 632-702. [doi: 10.18653/v1/2024.emnlp-main.443]
58. Fu T, Conde J, Martínez G, Grandury M, Reviriego P. Multiple choice questions: reasoning makes large language models (LLMS) more self-confident even when they are wrong. arXiv. Preprint posted online on Jan 16, 2025. [doi: 10.48550/arXiv.2501.09775]

Abbreviations

AI: artificial intelligence
GQS: Global Quality Scale
ICC: intraclass correlation coefficient
LLM: large language model
NLAT-AI: Natural Language Assessment Tool for Artificial Intelligence
PEMAT: Patient Education Materials Assessment Tool

XSL•FO
RenderX

SMOG: Simple Measure of Gobbledygook

Edited by A Benis; submitted 10.Jun.2025; peer-reviewed by BAF Assunção, F Fukuzawa, J Klewer, KS Rawat; accepted 29.Dec.2025; published 27.Feb.2026.

Please cite as:
Rao M, Xiujun T, Haoyu W
Evaluating GPT-4 Responses on Scars or Keloids for Patient Education: Large Language Model Evaluation Study
JMIR Med Inform 2026;14:e78838
URL: https://medinform.jmir.org/2026/1/e78838
doi:10.2196/78838

© Mingjun Rao, Tang Xiujun, Wang Haoyu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Models in Patient Health Communication for Atherosclerotic Cardiovascular Disease: Pilot Cross-Sectional Comparative Analysis

Pengfei Li[1*], MD; Yinfei Xu[1*], MD; Xiang Liu[2*], MD; Zhean Shen[1*], MD; Yi Wang[1], MD; Xinyi Lv[1], MD; Ziyi Lu[1], MD; Hui Wu[1], MD; Jiaqi Zhuang[1], MD; Yan Chen[1,3,4], MD

[1]Department of Emergency and Critical Care Medicine, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, Jiangsu, China

[2]Department of Rehabilitation, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China

[3]Department of Emergency Management, School of Health Policy & Management, Nanjing Medical University, Nanjing, Jiangsu, China

[4]Department of Emergency and Critical Care Medicine, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Yan Chen, MD
Department of Emergency and Critical Care Medicine
The Affiliated Suzhou Hospital of Nanjing Medical University
Suzhou Municipal Hospital, Gusu School, Nanjing Medical University
140 Hanzhong Rd
Suzhou, Jiangsu, 210029
China
Phone: 86 512 6515 53
Email: chenyandoc@njmu.edu.cn

Abstract

Background: Large language models (LLMs) have emerged as promising tools for enhancing public access to medical information, particularly for chronic diseases such as atherosclerotic cardiovascular disease (ASCVD). However, their effectiveness in patient-centered health communication remains underexplored, especially in multilingual contexts.

Objective: Our study aimed to conduct a comparative evaluation of 3 advanced LLMs—DeepSeek R1, ChatGPT-4o, and Gemini—in generating responses to ASCVD-related patient queries in both English and Chinese, assessing their performance across the domains of accuracy, completeness, and comprehensibility.

Methods: We conducted a cross-sectional evaluation based on 25 clinically validated ASCVD questions spanning 5 domains—definitions, diagnosis, treatment, prevention, and lifestyle. Each question was submitted 5 times to each of the 3 LLMs in both English and Chinese, yielding 750 responses in total, all generated under default settings to approximate real-world conditions. Three board-certified cardiologists blinded to model identity independently scored the responses using standardized Likert scales with predefined anchors. The assessment followed a rigorous multistage process that incorporated randomization, washout periods, and final consensus scoring.

Results: DeepSeek R1 achieved the highest "good response" rates (24/25, 96% in both English and Chinese), substantially outperforming ChatGPT-4o (21/25, 84%) and Gemini (12/25, 48% in English and 17/25, 68% in Chinese). DeepSeek R1 demonstrated superior median accuracy scores (6, IQR 6-6 in both languages) and completeness scores (3, IQR 2-3 in both languages) compared to the other models (P<.001). All models had a median comprehensibility score of 3; however, in English, DeepSeek R1 and ChatGPT-4o were rated significantly clearer than Gemini (P=.006 and P=.03, respectively), whereas no significant between-model differences were observed in Chinese (P=.08). Interrater reliability was moderate (Kendall W: accuracy=0.578; completeness=0.565; comprehensibility=0.486). Performance was consistently stronger for definitional and diagnostic questions than for treatment and prevention topics across all models. Specifically, none of the models consistently provided responses aligned with the latest clinical guidelines for the following key guideline-facing question "What is the standard treatment regimen for ASCVD?"

Conclusions: DeepSeek R1 exhibited promising and consistent performance in generating high-quality, patient-facing ASCVD information across both English and Chinese, highlighting the potential of open-source LLMs in promoting digital health literacy

XSL•FO
RenderX

and equitable access to chronic disease information. However, a clinically critical weakness was observed in guideline-sensitive treatment: the models did not reliably provide guideline-concordant standard treatment regimens, suggesting that LLM use should be limited to lower-risk informational subqueries (eg, definitions, diagnosis, and lifestyle education) unless augmented by expert oversight and safety controls.

(JMIR Med Inform 2026;14:e81422) doi:10.2196/81422

KEYWORDS

atherosclerotic cardiovascular disease; large language models; DeepSeek R1; ChatGPT-4o; Gemini; artificial intelligence; AI

Introduction

Atherosclerotic cardiovascular disease (ASCVD) is a major public health concern worldwide, significantly affecting morbidity and mortality rates and contributing to the global burden of disease [1]. Effective management and prevention strategies are paramount in reducing its impact on human health and longevity. In recent years, the development of advanced natural language processing technologies, particularly large language models (LLMs), has introduced a novel paradigm in health care communication [2]. LLMs are deep learning models trained on massive corpora of text data, enabling them to generate humanlike responses, summarize complex information, and interact with users in natural language. Models such as ChatGPT exemplify this technological advancement, offering scalable tools for disseminating medical information and supporting patient education, with the potential to enhance chronic disease prevention and self-management [3,4].

Despite their promise, the application of LLMs in health care—especially in clinical decision-making—remains constrained by several limitations [5,6]. First, LLMs do not inherently possess medical reasoning capabilities; their outputs are generated based on learned language patterns rather than grounded clinical judgment [7]. This introduces risks of producing plausible-sounding but clinically inappropriate advice. Second, the variability in training data sources and lack of domain-specific fine-tuning may result in inconsistent or outdated information [8]. Finally, while individual models may excel in certain tasks, performance can vary widely across different LLMs, especially in medical settings where accuracy and contextual appropriateness are vital [9]. Therefore, rigorous comparative evaluations are needed to identify the models that are most reliable, transparent, and suitable for patient-centered communication within health care and public health contexts.

To our knowledge, there have been no studies specifically focused on the performance of LLMs in providing information about ASCVD. Therefore, this study aimed to assess the performance of various advanced LLMs (ChatGPT-4o [OpenAI], Gemini [Google], and DeepSeek R1) in delivering accurate, comprehensive, and comprehensible information related to ASCVD. Using open-ended questions and simulated patient scenarios, we evaluated the quality and reliability of responses from a patient-centered perspective.

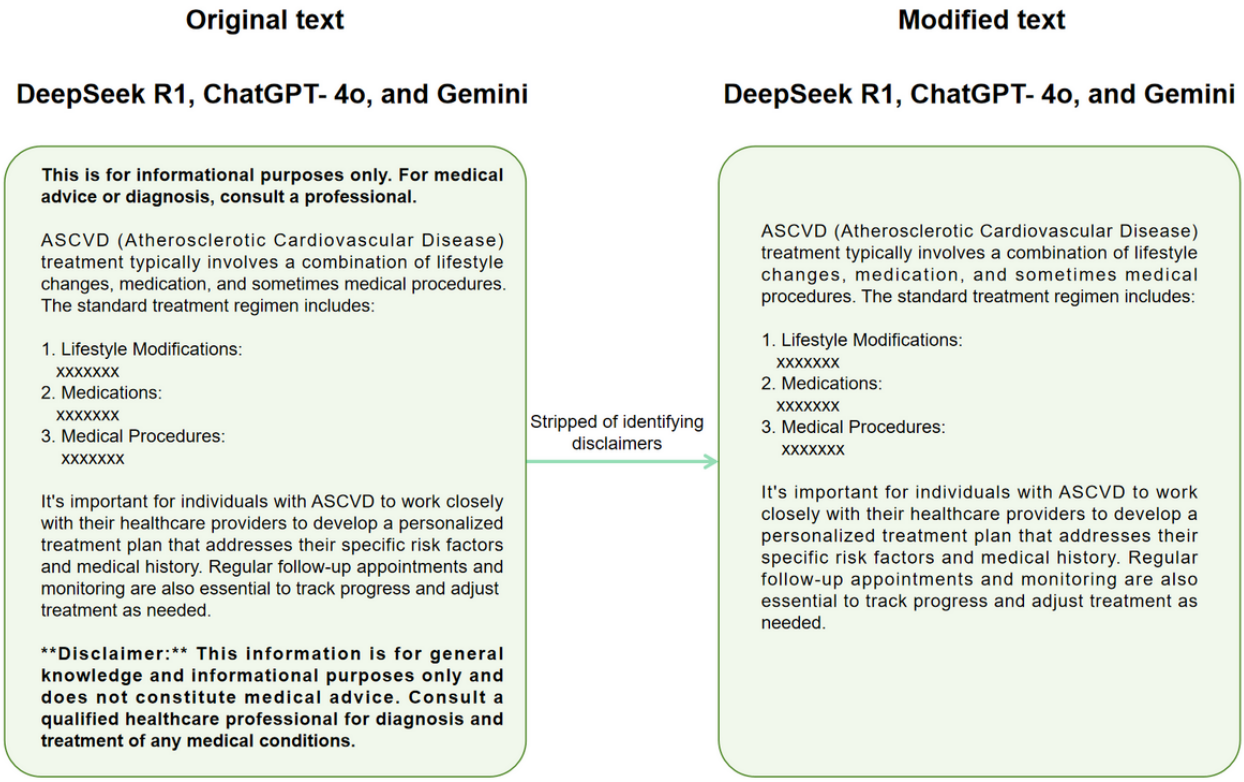
Methods

Question Design and Acquisition of Responses

Our study was conducted from May 15 to 30, 2025. Three qualified physicians (PL, YX, and YW) devised 25 comprehensive questions based on common concerns of patients regarding ASCVD, focusing on patient-centered care and disease management. These queries covered definitions, diagnosis, treatment, prevention, and lifestyle. To mitigate language bias and evaluate multilingual performance, each question was posed in both English and Chinese. To reduce randomness and assess response consistency, each question was submitted 5 separate times to 3 state-of-the-art LLMs—DeepSeek R1, ChatGPT-4o, and Gemini—in both languages, generating a total of 750 responses (25 questions × 3 models × 2 languages × 5 repetitions). To simulate real-world use scenarios, all models were assessed using their default, publicly available versions without any parameter adjustments (eg, temperature, maximum tokens, or top-p nucleus sampling). Crucially, the web-browsing or retrieval-augmented generation capabilities enabled by default in these public interfaces were allowed to function. This ensured that our findings reflected the typical performance accessible to end users who rely on these tools for real-time information retrieval. Each chat was conducted using the “new chat” function to avoid bias from correlation interference and ensure that each session was unaffected by previous prompts. All generated responses were exported as plain text. For DeepSeek R1, the reasoning content and blocks were programmatically removed before human review. Subsequently, all responses were stripped of identifying disclaimers (Figure 1). The complete set of 750 generated responses has been made publicly available in Multimedia Appendix 1 to ensure full transparency and facilitate further research.

XSL•FO
RenderX

Figure 1. Deidentification and preprocessing of model outputs before blinded scoring: raw large language model responses were exported as plain text and stripped of disclaimers and model-identifying metadata, yielding anonymized text for review by 3 cardiologists.



Assessment Methodology

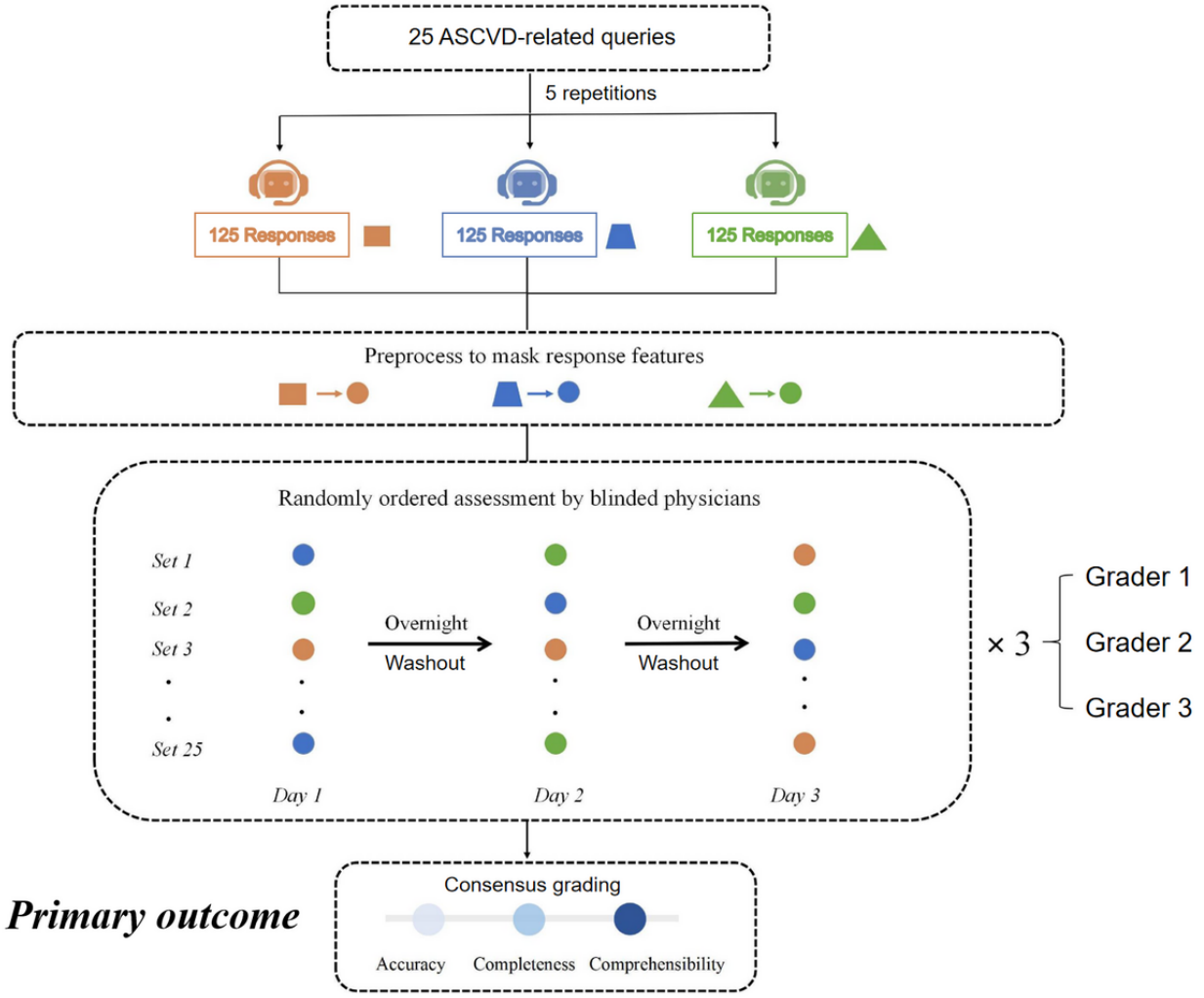
An evaluation panel of 3 board-certified cardiologists assessed the model outputs through a rigorous multistage process. For each of the 25 questions in both languages, 5 independent responses were generated per model. To avoid model selection bias, responses were randomly shuffled within their respective question sets, ensuring that the reviewers remained unaware of the specific language model that generated each response.

To minimize memory bias, 3 cardiologists independently assessed the responses in 3 separate rounds, each conducted on a distinct day with an overnight washout period. Assessments were benchmarked primarily to the 2021 European Society of Cardiology guidelines for cardiovascular disease prevention and the 2024 guidelines for managing chronic coronary syndromes, with cross-checks against the 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes [10-12]. We verified that no major guideline updates relevant to our end points were released during the evaluation window (May 15-30, 2025). Using predefined anchors detailed in Table S1 in Multimedia Appendix 2, each reviewer rated every response on 3 dimensions: accuracy (6-point Likert scale: 1="completely incorrect"; 6="completely correct"), completeness (3-point Likert scale: 1="incomplete"; 3="complete"), and comprehensibility (3-point Likert scale: 1="difficult to understand"; 3="easy to understand") [13]. The scoring protocol comprised 3 steps. First, for each question, each model generated 5 responses; the 3 reviewers independently rated every response on the 3 dimensions. Second, the reviewers convened to finalize one consensus score per response per dimension: identical ratings were carried forward, whereas discrepancies were resolved through discussion until unanimity was reached—or, if needed, with brief input from a senior cardiologist. Third, for each question, the arithmetic mean of the 5 consensus scores served as the final per-question score; because per-question scores were nonnormally distributed, between-platform comparisons used the median of the 25 per-question scores. Detailed scoring procedures for the blinded assessment are provided in Tables S2 to S4 in Multimedia Appendix 2.

The overall study design and process are illustrated in Figure 2. We established the following quality thresholds based on the arithmetic mean of the 5 repeated responses per question: an accuracy score of at least 4 was considered acceptable, whereas a score of at least 2 was required for both comprehensibility and completeness. Within the acceptable performance levels, we introduced a new concept called "good response" to facilitate the comparison of the LLMs' excellent responses. A "good response" was defined as having a mean accuracy score of ≥4, mean comprehensibility and completeness scores of ≥2, and a mean total score (sum of accuracy, completeness, and comprehensibility) score of ≥10. The total score threshold (10) was intentionally set higher than the sum of the minimum component scores (8) to ensure that a "good response" exceeded the baseline requirements in at least one dimension (eg, higher accuracy or completeness) rather than merely meeting the minimum acceptable standard in all categories.

XSL•FO
RenderX

Figure 2. Study design and data flow for the blinded comparison of large language models on atherosclerotic cardiovascular disease (ASCVD) questions. Outputs were generated under default settings; all model-identifying markers and disclaimers were removed; and responses were randomized and presented model blind to cardiologists for independent rating, consensus finalization, per-question aggregation, and cross-model comparison.



Data Analysis

All data analyses were performed using Prism (version 9.5.1; GraphPad Software) and SPSS (version 26.0; IBM Corp). Continuous variables were described using measures of central tendency and dispersion expressed as median and range. Graphical representations were used to enhance the clarity and interpretability of the data. When the data deviated from a normal distribution, differences between 2 groups were assessed using the Mann-Whitney *U* test, whereas differences among the 3 LLMs were evaluated using the Kruskal-Wallis rank sum test followed by the Dunn multiple-comparison test. The Fisher exact test was used for categorical variables. Interrater reliability was assessed using the Kendall *W* coefficient of concordance, with levels of agreement interpreted as follows: poor (0-0.2), fair (0.21-0.4), moderate (0.41-0.6), good (0.61-0.8), and excellent (0.81-1) [14]. A 2-sided *P* value of <.05 was considered statistically significant.

Ethical Considerations

The data analyzed in our study consisted of deidentified, synthetic outputs generated by LLMs in response to 25 guideline-informed ASCVD questions curated from publicly accessible sources; no patients or personal data were involved. Accordingly, participant recruitment and written informed consent were not applicable, and—per institutional policy—this work did not constitute human participant research and did not require ethics committee review.

Results

Performance

Overview

The "good response" rates for the 3 LLMs were as follows (Table 1): DeepSeek R1 achieved a "good response" rate of 96% (24/25) in both English and Chinese, outperforming ChatGPT-4o (21/25, 84%) and Gemini (12/25, 48% in English and 17/25, 68% in Chinese). Figure 3 illustrates the comparative distribution of scores across the 3 dimensions. In terms of accuracy, DeepSeek R1 maintained a dominant profile with high consistency, whereas Gemini exhibited the widest score variance. Regarding completeness, DeepSeek R1 frequently achieved maximum scores, whereas ChatGPT-4o often provided partially complete responses. Comprehensibility remained high for all models, with DeepSeek R1 and ChatGPT-4o showing slightly more stability than Gemini. Performance analysis revealed a critical pattern: model performance was consistently

XSL•FO
RenderX

and substantially weaker in the treatment domain than in all others, with "good response" rates of 80% (8/10) for DeepSeek R1, a total of 60% (6/10) for ChatGPT-4o, and 30% (3/10) for Gemini (Multimedia Appendix 3). For instance, none of the models consistently provided answers that aligned with the latest clinical guidelines in response to the question "What is the standard treatment regimen for ASCVD?" Similarly, the responses varied significantly in accuracy and detail for the question "How common is coronary artery bypass grafting (CABG) in ASCVD patients?" The following sections provide a detailed analysis of each model's performance across the domains of accuracy, completeness, and comprehensibility.

XSL•FO
RenderX

Table 1. Performance of the 3 large language models on atherosclerotic cardiovascular disease (ASCVD) questions in a blinded, bilingual evaluation. Questions spanned 5 domains (definitions, diagnosis, treatment, prevention, and lifestyle) and were scored by 3 board-certified cardiologists using predefined anchors. A "good response" was defined as an arithmetic mean score across the 5 repetitions of ≥4 for accuracy, ≥2 for completeness, and ≥2 for comprehensibility, with a mean total score (sum of accuracy, completeness, and comprehensibility) of ≥10 after consensus.

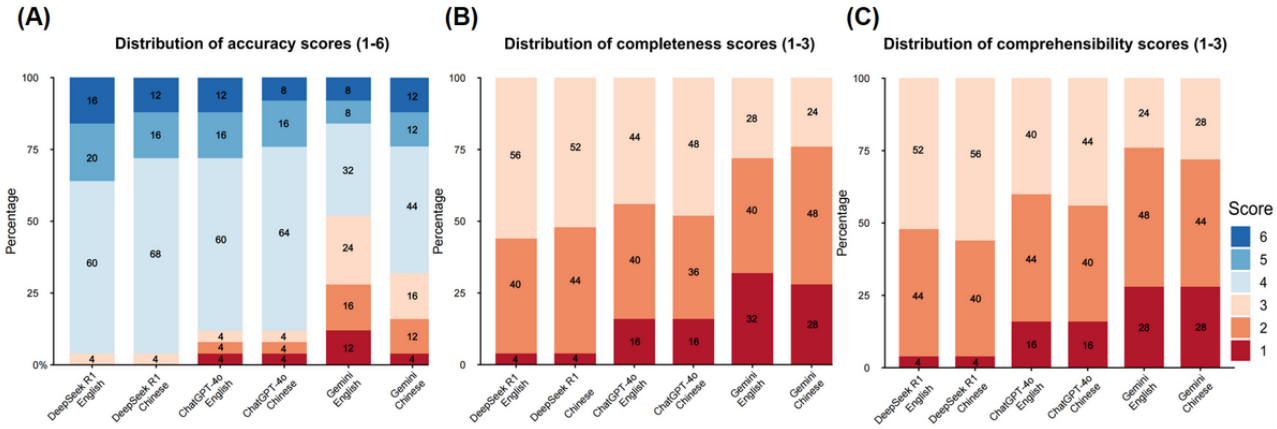
	DeepSeek English	DeepSeek Chinese	GPT-4o English	GPT-4o Chinese	Gemini English	Gemini Chinese
Good response (N=25), n (%)	24 (96)	24 (96)	21 (84)	21 (84)	12 (48)	17 (68)
Definitions						
"What is ASCVD?"	✓	✓	✓	✓		✓
"How does ASCVD develop?"	✓	✓	✓	✓		✓
"What diseases does ASCVD specifically include?"	✓	✓	✓	✓	✓	✓
"How does ASCVD affect the heart and blood vessels?"	✓	✓				
"What is the connection between ASCVD and heart disease and stroke?"	✓	✓	✓	✓		
Diagnosis						
"How can I determine my risk for ASCVD?"	✓	✓	✓	✓		
"What are the early symptoms of ASCVD?"	✓	✓	✓		✓	✓
"Which diagnostic tests are used to confirm ASCVD?"	✓	✓	✓	✓		
"Does a family history of cardiac diseases increase my risk of ASCVD?"	✓	✓	✓	✓	✓	✓
"Does my hypertension, diabetes, or hyperlipidemia increase my risk of ASCVD?"	✓	✓		✓		✓
Treatment						
"What is the standard treatment regimen for ASCVD?"						
"Is long-term use of lipid-lowering agents safe for individuals with ASCVD?"	✓	✓	✓	✓		
"Should individuals diagnosed with ASCVD undergo regular coronary angiography?"	✓	✓	✓	✓	✓	
"What are the nonpharmacological treatments for ASCVD?"	✓	✓	✓	✓	✓	✓
"How common is CABG[a] in ASCVD patients?"	✓	✓				
Prevention						
"Can I prevent ASCVD through diet?"	✓	✓	✓	✓		✓

XSL•FO
RenderX

	DeepSeek English	DeepSeek Chinese	GPT-4o English	GPT-4o Chinese	Gemini English	Gemini Chinese
"How important is physical activity in preventing ASCVD?"	✓	✓	✓	✓		✓
"Can antihypertensive and lipid-lowering medications prevent ASCVD?"	✓	✓	✓	✓		✓
"How effective is smoking cessation in preventing ASCVD?"	✓	✓	✓	✓	✓	✓
"How should I monitor my cardiovascular health?"	✓	✓	✓	✓	✓	✓
Lifestyle						
"With ASCVD, how should I adjust my dietary habits?"	✓	✓	✓	✓	✓	✓
"What types of physical activities are safe for someone with ASCVD? Are there any recommended forms of exercise?"	✓	✓	✓	✓	✓	✓
"What is the relationship between ASCVD and body weight?"	✓	✓	✓	✓	✓	✓
"How important is quitting smoking and reducing alcohol intake in managing ASCVD?"	✓	✓	✓	✓	✓	✓
"Can long-term psychological stress affect ASCVD?"	✓	✓	✓	✓	✓	✓

[a]CABG: coronary artery bypass grafting.

Figure 3. Distribution of accuracy, completeness, and comprehensibility ratings for large language model (LLM) responses to atherosclerotic cardiovascular disease–related questions. (A) Shows the distribution of accuracy scores across LLMs. (B) Shows the distribution of completeness scores across LLMs. (C) Shows the distribution of comprehensibility scores across LLMs.



Accuracy

Regarding accuracy, 97% (146/150) of the questions met the acceptable standard, with English-language responses reaching 99% (74/75) accuracy and Chinese-language responses reaching 96% (72/75) accuracy. For both English- and Chinese-language questions, the median accuracy score for DeepSeek R1 was consistently 6 (IQR 6-6 for both languages), whereas for ChatGPT-4o and Gemini, it was 5 (ChatGPT: IQR 5-6 for both languages; Gemini: IQR 4-5 for English and 5-6 for Chinese; Table 2). As detailed in Table 3, for all English-language queries, the mean accuracy score of DeepSeek R1 responses was 5.88 (SD 0.33), whereas the mean scores for ChatGPT-4o and Gemini were 5.40 (SD 0.64) and 4.84 (SD 0.68), respectively. For all Chinese-language queries, the mean

XSL•FO
RenderX

accuracy score for DeepSeek R1 responses was 5.80 (SD 0.40), whereas the mean ChatGPT-4o and Gemini scores were 5.20 (SD 0.76) and 5.00 (SD 0.86), respectively.

Table 2. Median scores for responses from the 3 large language models.

Assessment metric	DeepSeek R1, median (IQR)		ChatGPT-4o, median (IQR)		Gemini, median (IQR)	
	English	Chinese	English	Chinese	English	Chinese
Accuracy (1-6)	6 (6-6)	6 (6-6)	5 (5-6)	5 (5-6)	5 (4-5)	5 (5-6)
Completeness (1-3)	3 (2-3)	3 (2-3)	2 (2-2)	2 (2-2)	2 (2-2)	2 (2-3)
Comprehensibility (1-3)	3 (3-3)	3 (3-3)	3 (3-3)	3 (3-3)	3 (2-3)	3 (3-3)

Table 3. Mean scores for responses from the 3 large language models.

Assessment metric	DeepSeek R1, mean (SD)		ChatGPT-4o, mean (SD)		Gemini, mean (SD)	
	English	Chinese	English	Chinese	English	Chinese
Accuracy (1-6)	5.88 (0.33)	5.80 (0.40)	5.40 (0.64)	5.20 (0.76)	4.84 (0.68)	5.00 (0.86)
Completeness (1-3)	2.72 (0.45)	2.76 (0.43)	2.24 (0.43)	2.12 (0.43)	2.00 (0.57)	2.36 (0.63)
Comprehensibility (1-3)	2.96 (0.20)	2.92 (0.27)	2.88 (0.33)	2.84 (0.37)	2.64 (0.48)	2.92 (0.27)

Completeness

In terms of completeness, 95% (142/150) of responses were acceptable, including 95% (71/75) in both English and Chinese. Regardless of the language of the questions, the median completeness score for DeepSeek R1 was consistently 3 (IQR 2-3 for both languages), with ChatGPT-4o and Gemini both achieving median scores of 2 (ChatGPT: IQR 2-2 for both languages; Gemini: IQR 2-2 for English and 2-3 for Chinese; Table 2). For all English-language questions, the mean completeness score for DeepSeek R1's responses was 2.72 (SD 0.45), with ChatGPT-4o and Gemini achieving mean scores of 2.24 (SD 0.43) and 2.00 (SD 0.57), respectively. For all Chinese-language questions, the mean completeness score for responses from DeepSeek R1 was 2.76 (SD 0.43), with ChatGPT-4o and Gemini scoring a mean of 2.12 (SD 0.43) and 2.36 (SD 0.63; Table 3), respectively.

Comprehensibility

Regarding comprehensibility, all responses were at an acceptable level. Regardless of the language of the questions or the model used, the median comprehensibility score for all questions was consistently 3 (DeepSeek R1 and ChatGPT: IQR 3-3 for both languages; Gemini: IQR 2-3 for English and 3-3 for Chinese; Table 2). As shown in Table 3, for all English-language questions, the mean comprehensibility score for DeepSeek R1's responses was 2.96 (SD 0.20), whereas ChatGPT-4o and Gemini scored a mean of 2.88 (SD 0.33) and 2.64 (SD 0.48), respectively. For all Chinese-language questions, the mean comprehensibility score for DeepSeek R1 responses was 2.92 (SD 0.27), with ChatGPT-4o and Gemini scoring a mean of 2.84 (SD 0.37) and 2.92 (SD 0.27), respectively.

Interrater Reliability

For accuracy scores, the Kendall *W* coefficient of concordance ranged from 0.434 to 0.732, with a mean of 0.578 (SD 0.15). In terms of completeness scores, the coefficient varied from 0.360 to 0.782, averaging 0.565 (SD 0.21). For comprehensibility scores, the range was between 0.386 and 0.581, with a mean value of 0.486 (SD 0.10). This indicates that the scores for accuracy, completeness, and comprehensibility were at a moderate level of agreement.

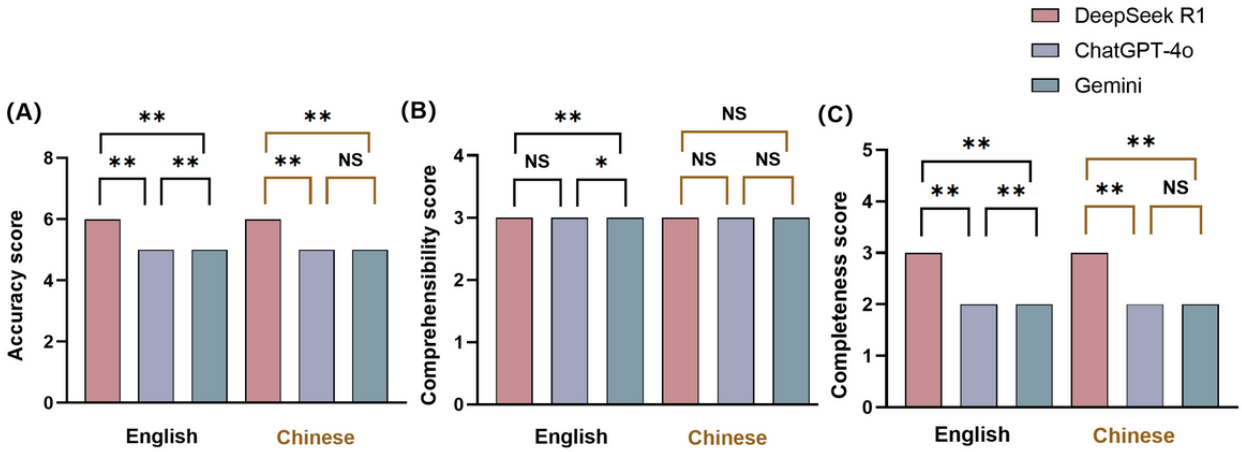
Model Comparison

In the overall comparison of English- and Chinese-language questions, DeepSeek R1 consistently outperformed the other models in terms of accuracy and completeness. For English-language questions, DeepSeek R1 achieved the highest median accuracy score of 6, significantly surpassing both ChatGPT-4o and Gemini (P<.001 for each pairwise comparison), whereas ChatGPT-4o also outperformed Gemini (P=.008). All 3 models received a median comprehensibility score of 3; the difference between DeepSeek R1 and ChatGPT-4o was not significant (P=.07), although DeepSeek R1 was significantly clearer than Gemini (P=.006), and ChatGPT-4o was moderately clearer than Gemini (P=.03; reflecting significant differences in mean ranks per the Mann-Whitney *U* test despite identical medians). Regarding completeness, DeepSeek R1 again led the models, scoring significantly higher than both ChatGPT-4o and Gemini (P=.009), with ChatGPT-4o also providing more complete responses than Gemini (P=.007). For Chinese-language questions, DeepSeek R1 maintained a significant lead in accuracy over ChatGPT-4o and Gemini (P<.001 in both cases), whereas the difference between ChatGPT-4o and Gemini was not statistically significant (P=.07). All 3 models achieved the same median comprehensibility score of 3, with no significant differences (P=.08). In terms of completeness, DeepSeek R1 again outperformed both ChatGPT-4o and Gemini (P<.001), whereas no significant difference was observed between ChatGPT-4o and Gemini (P=.06). Overall, DeepSeek R1 demonstrated consistently superior performance in both accuracy and completeness across languages. Differences in comprehensibility were minor, with Gemini showing slightly lower clarity in English. These findings affirm DeepSeek R1's leading capabilities and highlight subtle language-specific variations in the outputs of the other models (Figure 4). Specific

XSL•FO
RenderX

examples of LLM failures contrasted with superior responses are detailed in Table S5 in Multimedia Appendix 2.

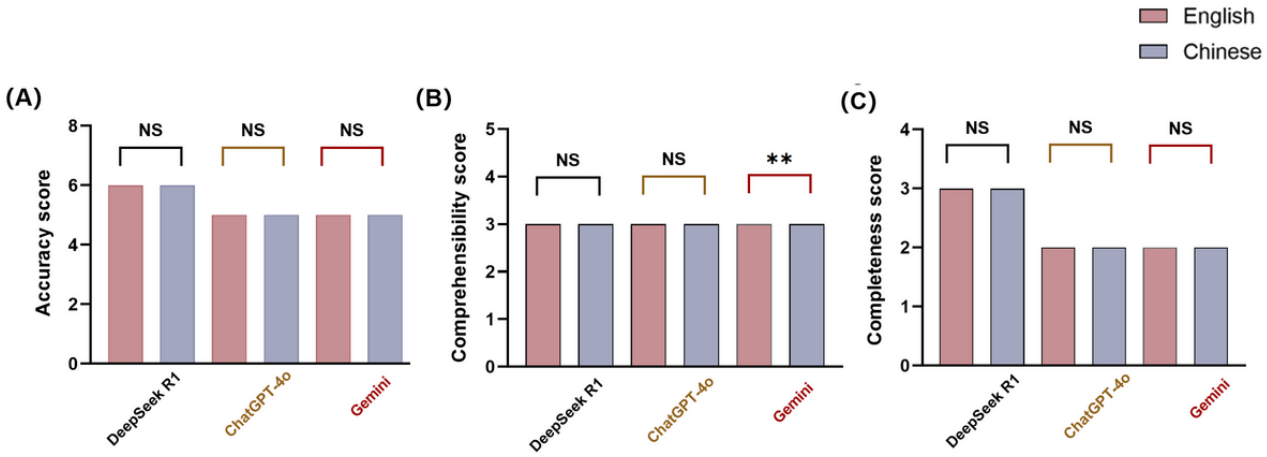
Figure 4. Comparison of large language model performance on atherosclerotic cardiovascular disease–related questions within languages. Scores are shown for accuracy, completeness, and comprehensibility. *P<.05; **P<.01; NS: not significant.



Language-Based Response Comparison

For DeepSeek R1, there were no significant differences between English and Chinese in terms of accuracy, completeness, and comprehensibility scores. Similarly, for ChatGPT-4o, there were no significant differences between English and Chinese across the same metrics, indicating a comparable performance in both languages. For Gemini, the accuracy and completeness scores showed no significant differences between English and Chinese. However, there was a significant difference in comprehensibility scores (P<.001) as the median comprehensibility score for English was 3 (IQR 2-3), whereas for Chinese, it was 3 (IQR 3-3). This indicated that the comprehensibility scores were more consistent in Chinese compared to having more variability in English (Figure 5).

Figure 5. Comparison of large language model performance on atherosclerotic cardiovascular disease–related questions across languages within models. For each model, scores for accuracy, completeness, and comprehensibility are shown for English and Chinese. **P<.01; NS: not significant.



Discussion

Principal Findings

Our study evaluated the performance of 3 advanced LLMs—DeepSeek R1, ChatGPT-4o, and Gemini—in providing accurate, comprehensive, and comprehensible information on ASCVD in both English and Chinese. This evaluation was conducted using a rigorous, blinded study design that incorporated randomization, washout periods, and consensus-based physician assessments. While the widespread emergence of LLMs has significantly expanded public access to medical information—particularly for underserved or non–English-speaking populations [15,16]—concerns about their reliability persist [17,18]. By systematically evaluating multiple LLMs across languages and using patient-centered ASCVD scenarios, our study offers a more holistic perspective on their capabilities and limitations. From a public health standpoint, this approach not only highlights the potential of LLMs to enhance digital health literacy and self-management of chronic diseases but also underscores the need for ongoing oversight and model validation to prevent the spread of misinformation [19,20].

The results demonstrated that DeepSeek R1 generally outperformed the other models in terms of accuracy and completeness and also achieved a slightly higher average score for understandability. Several factors may contribute to this superior performance. DeepSeek R1 is reported to rely heavily on reinforcement learning, which may contribute to improved

XSL•FO
RenderX

logical coherence and multistep reasoning. Additionally, the model is open source under the Massachusetts Institute of Technology license and may offer lower operating costs compared to proprietary models. This open-source nature confers several advantages—including reduced dependency on commercial application programming interfaces (APIs), lower implementation costs, and the potential for institution-specific fine-tuning—making it particularly attractive for health care systems with limited technical resources [21]. Critically, this open-source architecture offers a decisive advantage for clinical implementation: data sovereignty. Unlike proprietary cloud-based models (eg, ChatGPT-4o and Gemini) that necessitate the transmission of sensitive patient data to external commercial servers—raising significant compliance challenges with regulations such as HIPAA (Health Insurance Portability and Accountability Act; United States) and the General Data Protection Regulation (Europe)—DeepSeek R1 allows for local, on-premise deployment [22,23]. Health care systems can host the model within their secure firewalls, ensuring that patient data never leave the institution's control. This capability not only resolves privacy concerns but also creates a secure environment for the aforementioned fine-tuning on proprietary medical records. Furthermore, it is important to clarify that the high comprehensibility scores in this study were achieved without the reasoning chains being visible to the raters. This indicates that DeepSeek R1's internal chain-of-thought process effectively structures the final response for clarity independent of the visible reasoning trace [24]. Taken together, these characteristics suggest that DeepSeek R1 may be a promising candidate for disseminating reliable, cost-effective, and customizable medical information in both clinical and public health contexts.

Language-based comparisons revealed no significant performance differences between English and Chinese for both DeepSeek R1 and ChatGPT-4o. Although Gemini exhibited slightly higher scores in Chinese—particularly in comprehensibility—this difference was not statistically significant, indicating that the variation likely reflects incidental model behavior rather than a systematic language-based advantage. Achieving consistent performance across languages is essential for the global applicability of LLMs in health care communication. Our evaluation indicates a consistent and clinically relevant pattern: across all 3 LLMs, recommendations in the treatment and prevention domain were less reliable than responses to definitional and diagnostic queries. This observation is consistent with external reports that therapeutic, guideline-facing tasks show variable concordance and that performance can be highly sensitive to task framing and evaluation metrics beyond headline accuracy [25-27]. Recent benchmarks likewise describe mixed adherence across clinical scenarios, underscoring the difficulty of encoding and operationalizing rapidly evolving therapeutic knowledge within free-text generation [26-28]. Taken together, these findings suggest that, without explicit retrieval and provenance controls, current general-purpose architectures may not consistently reflect the most up-to-date evidence at inference time, which in turn can pose risks in high-stakes areas such as medication dosing and care timing [25-27]. They also help explain why standard offline evaluations may overestimate real-world reliability ("evaluation illusion") and why even low omission or error rates can have high clinical salience in treatment contexts [9,29].

In practical terms, our findings offer concrete guidance for implementing LLMs in clinical ASCVD care. The encouraging performance of models such as DeepSeek R1 supports adjunctive, nonautonomous use in patient-facing education platforms and, with clinician oversight, selective clinical decision support. For instance, models excelling in comprehensibility are best suited for generating patient education materials, whereas those with superior accuracy on guideline-specific content may better support the drafting of clinical summaries for specialist review. Successful integration will require attention to electronic health record interoperability; comprehensive staff training programs that cover the capabilities, limitations, and appropriate use of these artificial intelligence (AI) tools; and governance (privacy, auditability, and provenance). From a real-world clinical relevance standpoint, cost and deployment considerations (eg, licensing, API use, computing capacity and infrastructure, and local IT constraints) should inform model selection and scaling within existing health care workflows. In particular, open-source options (eg, DeepSeek R1) can lower licensing costs, enable local or hybrid deployment under strict data governance controls, and permit institution-specific tuning with auditable provenance—benefits that may improve feasibility in resource-constrained settings [30]. Given the observed fragility in guideline-sensitive domains, we recommend a clinician-in-the-loop workflow with tiered risk classification, automated guideline-concordance audits, and retrieval-augmented generation to anchor outputs in current evidence, with explicit human sign-off before the information enters care pathways [31,32]. Furthermore, transparent processes for obtaining patient consent for the use of AI-generated content in their care are essential. In addition, institutions should address regulatory and ethical implications (data protection, provenance, and accountability) and establish postdeployment monitoring and incident reporting processes to detect and mitigate potentially harmful responses [33].

Comparison to Prior Work

Unlike prior studies that have primarily examined single models within narrow clinical contexts [34,35], our investigation extends earlier designs by conducting a head-to-head, bilingual comparison of 3 advanced LLMs—DeepSeek R1, ChatGPT-4o, and Gemini—under a rigorous, blinded methodology incorporating randomization, washout, and consensus-based physician scoring. This design enabled more systematic cross-model comparisons than evaluations focused on 1 system or 1 language. The need for such rigor is underscored by reports from multidisciplinary areas such as cardio-oncology, where ChatGPT has answered only 68% of guideline-based queries correctly [36], highlighting the challenges that general-purpose models face when integrating complex clinical knowledge.

Focusing on ASCVD-related patient education and chronic disease self-management, we found that DeepSeek R1 achieved higher scores for accuracy and completeness and acceptable comprehensibility in both English and Chinese in our dataset,

XSL•FO
RenderX

suggesting suitability for patient-facing communication when used nonautonomously. These findings are concordant with recent *Nature Medicine* benchmarks showing that open-source systems can approach the performance of proprietary models in selected tasks [25,37]. Moreover, the open-source nature of DeepSeek R1 may enable privacy regulation–compliant local deployment and institution-specific fine-tuning—features that can support data-governed health systems—whereas its stable bilingual performance in our study aligns with goals of equitable information access [38]. At the same time, our domain-level analysis echoes and extends prior observations by showing that guideline-sensitive content remains comparatively fragile, reinforcing the need for clinician oversight and regular guideline concordance checks [39].

Limitations

Our study has several methodological limitations that should be considered when interpreting the findings. First, composite constraints in study design, item construction, and the evaluation perspective were present. The question set was clinician led, which may introduce selection bias; it used a fixed, evenly distributed, 5-domain preset (ie, definitions, diagnosis, treatment, prevention, and lifestyle) without direct patient involvement for external validation; model comparison was limited to 3 mainstream LLMs; and assessment relied entirely on expert review without patient-side measures such as comprehension or patient-reported outcomes. Collectively, these factors may affect clinical representativeness, diversity, and ecological validity [40].

Second, we acknowledge a moderate level of interrater agreement among the expert cardiologists as measured using the Kendall *W* (accuracy=0.578; completeness=0.565; comprehensibility=0.486). While such moderate reliability is not uncommon when evaluating nuanced, open-ended medical text where a degree of clinical subjectivity is inherent, it must be noted as a key limitation. To mitigate the impact of this disagreement, we used standardized anchors and a consensus adjudication step, and our statistical analysis (using medians and IQRs with nonparametric tests) was specifically chosen to limit the influence of rater heterogeneity. Nevertheless, this variability introduces uncertainty into the statistical significance of the differences found between models and suggests that expert opinions on AI-generated text quality can vary nontrivially. This is particularly relevant for clinical applications, where the use of such AI-generated content, especially for medical pathways or decision support, demands an even higher degree of caution. Consequently, the findings warrant a cautious interpretation. To enhance reliability in future work, we recommend developing more detailed scoring criteria with concrete examples and implementing a formal rater training and calibration process, including a pilot evaluation to establish baseline agreement.

Third, sample size and statistical power were not determined a priori. As an exploratory benchmarking study in a nascent area lacking established effect sizes, we did not perform an a priori power analysis. We used a coverage and stability-oriented design (25 items across 5 domains and 5 independent generations per model per language; 750 responses in total) to balance breadth and response stability. Although primary between-model contrasts reached statistical significance ($P<.001$)—suggesting adequate sensitivity post hoc—the absence of prespecified power planning limits generalizability; future confirmatory studies should use effect sizes and variance estimates from this work for formal power and sample size planning.

In addition, language and cultural generalizability are limited. Performance was evaluated only in English and Chinese. Differences in health beliefs, communication norms, and linguistic nuance across other languages and cultures may affect acceptability and applicability; broader multilingual, multicultural validation is needed.

Finally, geographic variation and reliance on the chosen reference guidelines (“benchmark binding”) may influence concordance judgments across health systems. While we benchmarked against specific international guidelines, ASCVD management varies by region and health system (eg, drug availability, targets, pathways, and reimbursement). Consequently, judgments of being concordant or nonconcordant are bound to the chosen reference standards, and local calibration is advisable for cross-region application.

Future Directions

First, longitudinal performance tracking should be prioritized. Standardized benchmarking with regular retesting across model updates, languages, and domains should be established to quantify performance drift, with early warning triggers for degradation; preregistered, multicenter confirmatory studies should be favored.

In addition, direct head-to-head comparisons with human materials should be conducted. LLM outputs should be benchmarked against human-authored patient education materials, assessing quantitative metrics (eg, accuracy and readability) and qualitative outcomes (eg, patient preference and trust) within a composite evaluation framework.

Moreover, pragmatic integration with clinical systems should be pursued. Low-coupling approaches to clinical systems under data minimization and privacy constraints (eg, standardized APIs, compliance and audit trails, and provenance and version management) should be explored to improve deployability.

Furthermore, communication should be tailored to patient literacy, language, culture, and context. Adaptive presentation methods based on health literacy, cultural background, preferred language, and clinical context should be developed and validated, maintaining clinical correctness while improving accessibility and understanding [41]. To ensure that these tailored communication strategies are effective, future work must include patient-centered validation, assessing the correlation between expert ratings and actual patient understanding and incorporating patient-reported outcome measures.

Finally, safety governance should be strengthened with structured risk management and monitoring. Future work should refine scoring anchors and operational definitions, broaden and diversify rater panels with structured calibration, conduct sensitivity and robustness analyses, implement routine guideline

XSL•FO
RenderX

concordance audits and a risk stratification framework (eg, dosing errors, contraindications, and care delay advice) [27], and establish continuous monitoring and incident reporting mechanisms to detect and mitigate potentially harmful or misleading outputs [26].

Conclusions

Our study evaluated the performance of several advanced LLMs in delivering information related to ASCVD, a major chronic disease burden in public health. In a blinded, bilingual comparison of DeepSeek R1, ChatGPT-4o, and Gemini, DeepSeek R1 demonstrated the highest accuracy and completeness, whereas all models showed comparable comprehensibility. Performance was consistently stronger for definitional and diagnostic queries than for treatment and prevention queries, indicating lower alignment with current therapeutic guidance in those domains. Notably, for the core question regarding the standard treatment regimen for ASCVD, none of the models consistently provided answers aligned with the latest clinical guidelines, representing a safety-relevant limitation of current general-purpose LLMs. These results delineate the present capability profile of LLMs for patient information on ASCVD and establish an empirical baseline for subsequent validation.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2023ZD0503900 and 2023ZD0503902). The authors also wish to acknowledge the developers of the large language models evaluated in this study, whose groundbreaking work made this research possible. While drafting this manuscript, the authors used DeepSeek with the sole purpose of checking the grammar of the text.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

Conceptualization: PL, YX, X Liu
Data curation: ZS, YW, X Lv
Formal analysis: ZL, HW
Funding acquisition: YC
Methodology: PL, YX
Visualization: PL, X Liu, JZ
Writing—original draft: PL, YX
Writing—review and editing: YC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full set of large language model responses (N=750; English and Chinese).
[XLSX File (Microsoft Excel File), 732 KB - medinform_v14i1e81422_app1.xlsx]

Multimedia Appendix 2
Detailed scoring rubric with anchors and examples.
[DOCX File , 30 KB - medinform_v14i1e81422_app2.docx]

Multimedia Appendix 3
Domain-specific “Good response” rates by model and language.
[DOCX File , 16 KB - medinform_v14i1e81422_app3.docx]

References

1. Roeters van Lennep JE, Tokgözoğlu LS, Badimon L, Dumanski SM, Gulati M, Hess CN, et al. Women, lipids, and atherosclerotic cardiovascular disease: a call to action from the European Atherosclerosis Society. Eur Heart J 2023 Oct 14;44(39):4157-4173 [FREE Full text] [doi: 10.1093/eurheartj/ehad472] [Medline: 37611089]
2. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res 2025 Jul 11;27:e71916 [FREE Full text] [doi: 10.2196/71916] [Medline: 40644686]
3. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]

XSL•FO
RenderX

4. Ferber D, El Nahhas OS, Wölflein G, Wiest IC, Clusmann J, Leßmann ME, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer 2025 Aug;6(8):1337-1349. [doi: 10.1038/s43018-025-00991-6] [Medline: 40481323]
5. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024 Sep 25;26:e59505 [FREE Full text] [doi: 10.2196/59505] [Medline: 39321458]
6. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 07;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
7. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
8. Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432 [FREE Full text] [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
9. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
10. Visseren FL, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, ESC Scientific Document Group. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur J Prev Cardiol 2022 Feb 19;29(1):5-115. [doi: 10.1093/eurjpc/zwab154] [Medline: 34558602]
11. Rao SV, O'Donoghue ML, Ruel M, Rab T, Tamis-Holland JE, Alexander JH, Peer Review Committee Members. 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the management of patients with acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2025 Jun 10;85(22):2135-2237 [FREE Full text] [doi: 10.1016/j.jacc.2024.11.009] [Medline: 40013746]
12. Vrints C, Andreotti F, Koskinas KC, Rossello X, Adamo M, Ainslie J, ESC Scientific Document Group. 2024 ESC Guidelines for the management of chronic coronary syndromes. Eur Heart J 2024 Sep 29;45(36):3415-3537. [doi: 10.1093/eurheartj/ehae177] [Medline: 39210710]
13. Pugliese N, Wai-Sun Wong V, Schattenberg JM, Romero-Gomez M, Sebastiani G, NAFLD Expert Chatbot Working Group, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2024 Apr;22(4):886-9.e5 [FREE Full text] [doi: 10.1016/j.cgh.2023.08.033] [Medline: 37716618]
14. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
15. Campbell DJ, Estephan LE, Sina EM, Mastrolonardo EV, Alapati R, Amin DR, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid 2024 Mar;34(3):371-377. [doi: 10.1089/thy.2023.0491] [Medline: 38010917]
16. Nielsen JP, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 2023 Sep;143(9):779-782. [doi: 10.1080/00016489.2023.2254809] [Medline: 37694729]
17. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 2023 Sep 04;13:1268915 [FREE Full text] [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]
18. Southwell BG, Otero Machuca J, Cherry ST, Burnside M, Barrett NJ. Health misinformation exposure and health disparities: observations and opportunities. Annu Rev Public Health 2023 Apr 03;44:113-130 [FREE Full text] [doi: 10.1146/annurev-publhealth-071321-031118] [Medline: 36207008]
19. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120 [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
20. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
21. Le Guellec B, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. Radiol Artif Intell 2024 Jul;6(4):e230364 [FREE Full text] [doi: 10.1148/ryai.230364] [Medline: 38717292]
22. Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023 Apr;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
23. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med 2024 Jul 08;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
24. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 15, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]

XSL•FO
RenderX

25. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 Aug;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
26. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023 Dec;29(12):2988. [doi: 10.1038/s41591-023-02656-2] [Medline: 37957381]
27. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
28. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. J Med Internet Res 2024 Apr 22;26:e55037 [FREE Full text] [doi: 10.2196/55037] [Medline: 38648098]
29. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
30. Dorfner FJ, Jürgensen L, Donle L, Al Mohamad F, Bodenmann TR, Cleveland MC, et al. Comparing commercial and open-source large language models for labeling chest radiograph reports. Radiology 2024 Oct;313(1):e241139. [doi: 10.1148/radiol.241139] [Medline: 39470431]
31. Beltramin D, Bousquet C, Tiffet T. Large language models could revolutionize health care, but technical hurdles may limit their applications. J Med Internet Res 2025 Jun 25;27:e71618 [FREE Full text] [doi: 10.2196/71618] [Medline: 40561496]
32. Forero DA, Abreu SE, Tovar BE, Oermann MH. Automated analyses of risk of bias and critical appraisal of systematic reviews (ROBIS and AMSTAR 2): a comparison of the performance of 4 large language models. J Am Med Inform Assoc 2025 Sep 01;32(9):1471-1476. [doi: 10.1093/jamia/ocaf117] [Medline: 40680299]
33. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the risk of bias in randomized clinical trials with large language models. JAMA Netw Open 2024 May 01;7(5):e2412687. [doi: 10.1001/jamanetworkopen.2024.12687] [Medline: 38776081]
34. Law S, Oldfield B, Yang W, Global Obesity Collaborative. ChatGPT/GPT-4 (large language models): opportunities and challenges of perspective in bariatric healthcare professionals. Obes Rev 2024 Jul;25(7):e13746. [doi: 10.1111/obr.13746] [Medline: 38613164]
35. Wang L, Mao Y, Wang L, Sun Y, Song J, Zhang Y. Suitability of GPT-4o as an evaluator of cardiopulmonary resuscitation skills examinations. Resuscitation 2024 Nov;204:110404 [FREE Full text] [doi: 10.1016/j.resuscitation.2024.110404] [Medline: 39343124]
36. Li P, Zhang X, Zhu E, Yu S, Sheng B, Tham YC, et al. Potential multidisciplinary use of large language models for addressing queries in cardio-oncology. J Am Heart Assoc 2024 Mar 19;13(6):e033584 [FREE Full text] [doi: 10.1161/JAHA.123.033584] [Medline: 38497458]
37. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025 Aug;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
38. Normile D. Chinese firm's large language model makes a splash. Science 2025 Jan 17;387(6731):238. [doi: 10.1126/science.adv9836] [Medline: 39818899]
39. Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the confidence of large language models in answering clinical questions: cross-sectional evaluation study. JMIR Med Inform 2025 May 16;13:e66917 [FREE Full text] [doi: 10.2196/66917] [Medline: 40378406]
40. Sumner J, Wang Y, Tan SY, Chew EH, Wenjun Yip A. Perspectives and experiences with large language models in health care: survey study. J Med Internet Res 2025 May 01;27:e67383 [FREE Full text] [doi: 10.2196/67383] [Medline: 40310666]
41. Ji H, Wang X, Sia CH, Yap J, Lim ST, Djohan AH, et al. Large language model comparisons between English and Chinese query performance for cardiovascular prevention. Commun Med (Lond) 2025 May 16;5(1):177 [FREE Full text] [doi: 10.1038/s43856-025-00802-0] [Medline: 40379850]

Abbreviations

AI: artificial intelligence
API: application programming interface
ASCVD: atherosclerotic cardiovascular disease
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model

XSL•FO
RenderX

Edited by A Coristine; submitted 28.Jul.2025; peer-reviewed by R Kocaaslan, H Pilabré, AA Arafat; comments to author 16.Oct.2025; accepted 17.Dec.2025; published 07.Jan.2026.

Please cite as:
Li P, Xu Y, Liu X, Shen Z, Wang Y, Lv X, Lu Z, Wu H, Zhuang J, Chen Y
Large Language Models in Patient Health Communication for Atherosclerotic Cardiovascular Disease: Pilot Cross-Sectional Comparative Analysis
JMIR Med Inform 2026;14:e81422
URL: https://medinform.jmir.org/2026/1/e81422
doi:10.2196/81422
PMID:

©Pengfei Li, Yinfei Xu, Xiang Liu, Zhean Shen, Yi Wang, Xinyi Lv, Ziyi Lu, Hui Wu, Jiaqi Zhuang, Yan Chen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multi-Evidence Clinical Reasoning With Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study

Hang Sheung Wong[1], MBBS; Tsz Kwan Wong[2], MBChB, MSc

[1]Department of Accident & Emergency, Princess Margaret Hospital, Hong Kong (SAR), China (Hong Kong)

[2]Department of Accident & Emergency, North Lantau Hospital, Hong Kong (SAR), China (Hong Kong)

Corresponding Author:
Hang Sheung Wong, MBBS
Department of Accident & Emergency
Princess Margaret Hospital
2-10, Princess Margaret Hospital Road, Lai Chi Kok, Kowloon
Hong Kong (SAR)
China (Hong Kong)
Phone: 852 68716917
Email: hs.wong3@ha.org.hk

Abstract

Background: Emergency triage accuracy is critical but varies with clinician experience, cognitive load, and case complexity. Mis-triage can delay care for high-risk patients and exacerbate crowding through unnecessary prioritization. Large language models (LLMs) show promise as triage decision-support tools but are vulnerable to hallucinations. Retrieval-augmented generation (RAG) may improve reliability by grounding LLM reasoning in authoritative guidelines and real clinical cases.

Objective: This study aimed to evaluate whether a dual-source RAG system that integrates guideline- and case-based evidence improves emergency triage performance versus a baseline LLM and to assess how closely its urgency assignments align with expert consensus and outcome-defined clinical severity.

Methods: We developed a dual-source RAG system—Multi-Evidence Clinical Reasoning RAG (MECR-RAG)—that retrieves sections from the Hong Kong Accident and Emergency Triage Guidelines (HKAETG) and cases from a database of 3000 emergency department triage encounters. In a retrospective single-center evaluation, MECR-RAG and a prompt-only baseline LLM (both DeepSeek-V3) were tested on 236 routine triage encounters to predict 5-level triage categories. Expert consensus reference labels were assigned by blinded senior triage nurses. Primary outcomes were quadratic weighted kappa (QWK) and accuracy versus consensus labels. Secondary analyses examined performance within 3 operationally and clinically relevant triage bands—immediate (categories 1 and 2), urgent (category 3), and nonurgent (categories 4 and 5). In 226 encounters with follow-up, we also assigned outcome-based severity tiers (R1-R3) using a published 3-level urgency reference standard and defined a disposition-safety composite.

Results: MECR-RAG achieved a mean QWK of 0.902 (SD 0.0021; 95% CI 0.901-0.904) and accuracy of 0.802 (SD 0.0082; 95% CI 0.795-0.808), outperforming the baseline LLM (QWK 0.801, SD 0.004; accuracy 0.542, SD 0.0073; both $P<.001$) and demonstrating expert-comparable agreement with triage nurses (interrater QWK 0.887). In 3-group analysis, MECR-RAG reduced overtriage from 68/236 (28.8%) with the baseline LLM to 30/236 (12.7%) and maintained low undertriage from 4/236 (1.7%) to 3/236 (1.3%), with the largest gains in the diagnostically ambiguous yet operationally important categories 3 and 4. In a secondary outcome-based analysis defining high-severity courses as R1+R2, MECR-RAG detected high-risk patients more sensitively than initial nurse triage (124/130, 95.4% vs 117/130, 90.0%; $P=.02$) while maintaining nurse-level specificity. MECR-RAG yielded the lowest weighted harm index (13.7, 19.5, and 20.3 per 100 patients for MECR-RAG, nurses, and the baseline LLM, respectively).

Conclusions: A dual-source RAG triage system that combines guideline-based rules with case-based reasoning achieved expert-comparable agreement, reduced overtriage, and better aligned urgency assignments than a prompt-only baseline LLM. Secondary outcome–based analyses in this cohort suggested more favorable triage patterns than initial nurse triage, supporting MECR-RAG as a concurrent decision-support layer that flags discordant or high-risk assignments; prospective multicenter implementation studies are needed to determine effects on emergency department crowding, delays, and patient outcomes.

(JMIR Med Inform 2026;14:e82026) doi:10.2196/82026

XSL•FO
RenderX

KEYWORDS

artificial intelligence; clinical decision support; clinical informatics; DeepSeek; digital health; emergency medicine; large language models; nurses; retrieval-augmented generation; triage

Introduction

Emergency triage is a critical step in emergency department (ED) operations, determining which patients are seen first when demand exceeds capacity. However, triage accuracy varies with clinicians' experience, cognitive load, and the complexity of presenting symptoms, and even trained nurses may misclassify a substantial minority of encounters [1,2]. Undertriage risks delayed recognition of life-threatening illness and preventable morbidity or mortality, whereas overtriage diverts scarce resources toward lower-acuity patients and can exacerbate ED crowding and boarding, both of which are associated with worse outcomes [2]. These challenges are particularly pronounced in mid-acuity categories, which represent most ED visits and where small changes in urgency assignment can translate into large differences in waiting time, crowding pressure, and exposure to harm.

Large language models (LLMs) have emerged as potential tools for clinical decision support because they can interpret free-text notes, integrate medical knowledge, and generate structured recommendations. In emergency care, LLMs have been tested for triage on simulated vignettes and curated cases, but reported performance has been highly variable, with agreement with human raters (Cohen κ) ranging from 0.125 to 0.899 across models, datasets, and evaluation frameworks [3-8]. A fundamental limitation of current "prompt-only" LLMs is their tendency to hallucinate, generating outputs that appear plausible but are factually incorrect or unsupported [9]. In emergency triage, such hallucinations may lead to urgency levels being assigned based on fabricated reasoning, outdated rules, or misinterpretation of presenting complaints, posing potential risks to patient safety and consistency.

Retrieval-augmented generation (RAG) has been proposed as a way to mitigate hallucinations by grounding LLM outputs in external, task-specific knowledge sources such as clinical guidelines, textbooks, electronic health records, and biomedical literature [10-12]. By retrieving relevant evidence at inference time and conditioning generation on that material, RAG can improve access to current, domain-specific knowledge and make model reasoning more transparent. Early work in emergency care has shown that adding retrieval can improve triage or referral accuracy compared with nonretrieval models. Yazaki et al [13] integrated RAG with a triage guideline database and tested GPT-3.5 on 100 structured, simulated scenarios derived from the Japanese National Examination for Emergency Medical Technicians, finding that the RAG-enhanced model achieved 70/100 (70%) accuracy and outperformed both emergency medical technicians and physicians. Gaber et al [14] used a curated dataset of 2000 real-world ED cases from the MIMIC-IV (Medical Information Mart for Intensive Care IV) database and retrieved evidence from a corpus of 30,000 PubMed abstracts, demonstrating superior triage-level prediction and exact-match accuracy compared with nonretrieval workflows.

Despite these advances, existing applications of RAG to emergency triage and referral have important gaps. One study relied largely on textbook-style, simulated examination cases [13], whereas another used routine labels from structured research datasets without clinician-adjudicated ground truth [14]. In both studies, RAG was applied to a single knowledge source—either structured triage guidelines or biomedical literature—rather than explicitly combining protocol-based rules with experiential knowledge from prior cases.

To our knowledge, no prior work has evaluated a dual-source RAG system that jointly retrieves local triage guidelines and real-world past triage cases, or directly compared such a system with a prompt-only LLM using raw, unstructured triage documentation paired with a blinded expert consensus reference standard and outcome-based validity checks. This design better captures the complexity and variability of real-world practice and strengthens the translational relevance of performance estimates.

There is also growing recognition that agreement with human labels alone is an incomplete measure of triage quality. Large cohort studies of systems such as the Emergency Severity Index (ESI) and the Manchester Triage System (MTS) have shown that nurse-assigned categories can underestimate or overestimate "true" acuity when judged against downstream outcomes such as hospitalization, intensive care unit (ICU) transfer, revisits, and short-term mortality [15-17]. In parallel, a *Journal of the American Medical Association* (*JAMA*) editorial on artificial intelligence (AI) in medicine has argued that evaluations of AI tools should prioritize clinically meaningful outcomes, patient-centered care, quality, and equity rather than focusing solely on narrow technical accuracy metrics [18]. Accordingly, evaluations of LLM-enabled triage should explicitly examine how assigned urgency levels align with downstream clinical severity and sentinel safety events, rather than treating concordance with nurse labels as sufficient.

In this context, we developed a dual-source RAG system—Multi-Evidence Clinical Reasoning RAG (MECR-RAG)—that retrieves both structured local triage guidelines (Hong Kong Accident and Emergency Triage Guidelines [HKAETG]) and analogous past triage cases and applies a structured Multi-Evidence Clinical Reasoning (MECR) framework to assign 5-level urgency categories. We retrospectively evaluated MECR-RAG on real ED triage notes, benchmarking it against a prompt-only baseline LLM and against expert nurse consensus labels. Our primary aim was to determine whether a dual-source RAG triage system (MECR-RAG) improves 5-level triage accuracy and agreement (quadratic weighted κ and accuracy) with expert nurse consensus compared with a prompt-only baseline LLM and to benchmark its performance against expert triage nurses. We hypothesized that MECR-RAG would achieve higher agreement than the baseline LLM and noninferior agreement relative to expert nurses. As a secondary aim, in a subset with follow-up data, we

XSL•FO
RenderX

assessed outcome-informed validity by examining how MECR-RAG, the baseline LLM, and initial nurse triage aligned with downstream clinical severity tiers and a composite of sentinel safety outcomes. This early-phase, single-center evaluation, reported in accordance with TRIPOD-LLM (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis - Large Language Models) [19], is intended to provide a methodologically rigorous foundation for future prospective and implementation studies of RAG-enhanced triage decision-support systems designed to augment, rather than replace, nurse-led triage.

Methods

Study Design and Setting

This study was a retrospective, proof-of-concept evaluation of a RAG-enhanced LLM for ED triage classification. The primary objective was to compare the performance of an RAG-based LLM that referenced both historical case data and the HKAETG against a prompt-only baseline LLM in assigning triage urgency levels.

The HKAETG is a 5-level triage system implemented across all 18 government-funded public EDs in Hong Kong. Its structure is conceptually aligned with other internationally recognized systems such as the MTS and ESI. Under the HKAETG, patients are assigned to 1 of 5 categories based on clinical urgency and time-to-treatment targets: category 1 (critical, 0 minutes), category 2 (emergency, 10 minutes), category 3 (urgent, 30 minutes), category 4 (semiurgent, 120 minutes), and category 5 (nonurgent, 180 minutes).

All study data were sourced from a single center, Princess Margaret Hospital (PMH), a major acute public hospital in Hong Kong and one of the city's largest tertiary care facilities. PMH operates a 24-hour ED and manages a high volume of complex and time-sensitive cases. In 2023 and 2024, the ED recorded 113,974 and 110,812 visits, respectively, highlighting its central role in emergency care delivery.

Pilot Study, Sample Size Calculation, and Model Selection

Before the main evaluation, we conducted a pilot study to inform model selection, prompting strategy, and sample size estimation. We randomly sampled 1000 ED triage cases from December 2024; 200 cases were used for pilot testing, and the remaining 800 cases were reserved to simulate a retrieval database. All pilot cases were excluded from the final study.

Three commercially available LLMs (DeepSeek-V3, GPT-4o, and Claude 3.7 Sonnet) were compared in both prompt-only and RAG configurations, with performance assessed against clinician-assigned triage labels using quadratic weighted kappa (QWK) and accuracy. Across models, performance was statistically similar in this pilot. DeepSeek-V3 was therefore selected as the backbone model for the main study primarily on the basis of its inference cost, application programming interface (API) accessibility, and comparable pilot performance.

In the pilot setting, the MECR-RAG configuration achieved an approximate QWK of 0.75 compared with 0.50 for the prompt-only baseline. Using the formula for comparing 2 kappa coefficients described by Fleiss, with α=.05 and power=.80 [20], the minimum required sample size to detect this difference was 48 cases. We nevertheless included 236 consensus-labeled triage cases in the final study to improve the precision of performance estimates and enhance generalizability.

Data Sources and Participants

The test set comprised ED triage cases collected at PMH between January 1 and December 31, 2023. All ED visits during this period were eligible regardless of age, sex, or presenting problem. To ensure temporal representativeness and reflect real-world variation in case mix, we used a stratified sampling strategy in which 1 calendar day was randomly selected from each month and 20 cases were sampled from that day. Within each sampled day, triage categories were sampled to approximate overall category distributions while modestly oversampling rare categories to improve precision for high-acuity strata. Four cases were excluded because case-number identification errors prevented data retrieval, yielding a final test set of 236 triage cases.

Reference labels were assigned using a blinded, multirater consensus process. Two advanced practice nurses with more than 10 years of emergency experience independently reviewed each case. Raters were blinded to the original nurse-assigned triage category and saw only information available at the time of triage: presenting complaint, condition on arrival, vital signs, and relevant past medical history when documented. Downstream data, including investigations, diagnoses, treatments, and ward progress, were withheld so that labels reflected only the initial triage decision context and matched the information provided to the LLMs. Disagreements were adjudicated by a third senior triage nurse to obtain a final consensus category for each encounter.

For retrieval, we constructed a separate database of 3000 anonymized ED visits to PMH between January 1 and December 31, 2024. All encounters in 2024 were eligible, with no clinical exclusion criteria, to preserve the natural diversity of documentation and presentations. Days were randomly sampled from each month, and cases were then drawn from those days to achieve broad temporal coverage; modest oversampling of rare triage categories (categories 1, 2, and 5) ensured sufficient representation while maintaining the real-world skew toward categories 3 and 4. The final database contained 94/3000 (3.1%) category 1, 120/3000 (4.0%) category 2, 1705/3000 (57.0%) category 3, 1035/3000 (34.5%) category 4, and 46/3000 (1.5%) category 5 encounters, with detailed monthly and category-specific sampling shown in Table S1 in Multimedia Appendix 1. To stress-test temporal robustness and reduce bias from contemporaneous documentation habits, 2023 test cases were evaluated against a 2024 retrieval database using the same triage guideline version; the full rationale is provided in Textbox S1 in Multimedia Appendix 1.

Preprocessing

Raw, unstructured triage documentation for both test cases and retrieval-database cases was exported from the electronic medical record, deidentified, and converted into a structured,

XSL•FO
RenderX

machine-readable text format. We retained only nurse-entered information available at the time of triage (eg, demographics, presenting complaint, condition on arrival, vital signs, and relevant past medical history) and excluded any downstream assessments, investigations, treatments, and outcomes. The latest version of the HKAETG (version 6, revised 2022) was similarly digitized and segmented into retrievable guideline sections. Both triage notes and guideline sections were then processed for semantic retrieval; further preprocessing and summarization details are provided in Textbox S2 in Multimedia Appendix 1.

LLM and RAG System Description

For all primary analyses, we used DeepSeek-V3 as the base LLM, accessed via a commercial API without any additional fine-tuning or architectural modification. In what follows, we refer to the DeepSeek-V3 model without retrieval as the "baseline LLM" and the dual-source retrieval-augmented configuration as MECR-RAG. Within each experimental configuration, the same LLM instance handled all stages of the pipeline (preprocessing, retrieval, and generation) with deterministic generation (temperature=0) to minimize stochastic variability. A separate embedding model was used to generate vector representations for semantic retrieval, and the same embedding configuration was applied during database construction and at query time to ensure consistency.

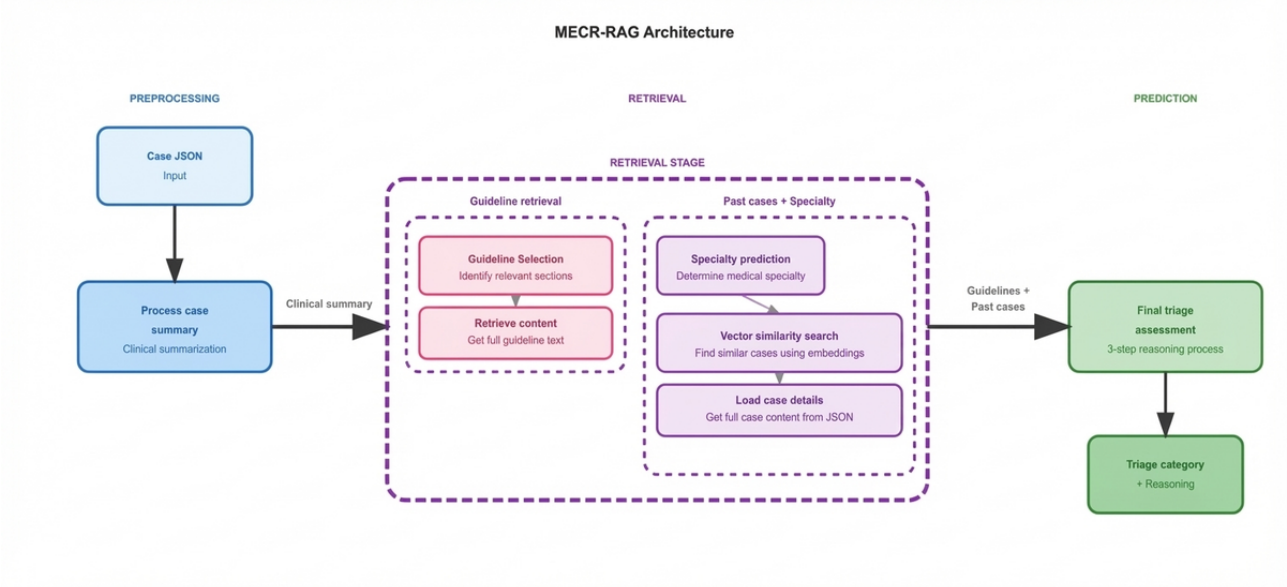
The pipeline was implemented in Python and executed on a standard desktop workstation without graphics processing unit acceleration. End-to-end processing time for the full MECR-RAG configuration was approximately 1 minute per case, including preprocessing, retrieval, and triage prediction, suggesting that near-real-time decision support would be technically feasible. Detailed model versions, libraries, and runtime benchmarks are provided in Section A1 in Multimedia Appendix 1.

MECR-RAG System Architecture

The MECR-RAG system follows a 5-stage pipeline designed to mirror how expert triage nurses combine protocol knowledge with experiential reasoning. First, the raw triage note is summarized into a concise, structured representation optimized for retrieval. Second, the model selects the most relevant sections of the HKAETG using an agentic reasoning step that prioritizes life-threatening conditions and documented physiologic abnormalities. Third, the system predicts the most likely attending specialty (eg, medicine, surgery, and orthopedics) to constrain subsequent case retrieval. Fourth, a hybrid retrieval module identifies similar past triage cases by combining metadata filters (age group and predicted specialty) with vector similarity search. Finally, the model applies a structured, MECR prompt that integrates guideline excerpts, retrieved cases, and the index presentation to assign a 5-level urgency category and provide a textual rationale.

The overall architecture is orchestrated using a modular, node-based workflow that preserves intermediate states, supports full audit trails, and enables systematic component ablation. A schematic overview of the MECR-RAG pipeline is shown in Figure 1, and additional implementation details are provided in Section A2 in Multimedia Appendix 1.

Figure 1. Overview of the Multi-Evidence Clinical Reasoning retrieval-augmented generation system architecture. MECR-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation.



The MECR-RAG pipeline consists of 5 sequential processing stages: clinical case summarization, agentic guideline section selection, specialty prediction for metadata filtering, hybrid past-case retrieval combining metadata and vector similarity search, and multievidence reasoning for the final triage assessment. LangGraph nodes manage state transitions and data flow throughout the pipeline.

Indexing and Database Construction

To support efficient retrieval, both historical triage cases and guideline content were converted into structured, machine-readable representations and indexed in a vector database. Historical cases were transformed into standardized triage-time summaries that captured presenting complaints, key vital signs, and brief clinical context, while preserving links to

XSL•FO
RenderX

metadata such as age group, triage category, and attending specialty. The same summarization template was applied to both database cases and incoming test cases to ensure that queries and indexed documents occupied a comparable embedding space.

The HKAETG document was similarly segmented into discrete sections and summarized to highlight the spectrum of presentations covered by each rule set. Each guideline section was stored with a short synopsis and structural metadata (eg, system, chief complaint, and red-flag criteria) to facilitate targeted retrieval during triage. The combined indexing strategy was designed to balance retrieval quality, computational efficiency, and scalability to larger case repositories. Details of prompt templates, sectioning rules, and indexing parameters are described in Section A3 in Multimedia Appendix 1.

Retrieval System Design

The retrieval layer implements a dual-source strategy that distinguishes authoritative protocol knowledge from experiential case patterns. Guideline retrieval is performed via an agentic selection step: given the summarized index case and a list of guideline section titles and synopses, the LLM selects up to 2 sections it deems most relevant, explicitly favoring conditions associated with high short-term risk or abnormal vital signs. This approach avoids purely lexical matching and aligns guideline selection more closely with clinical reasoning.

Past-case retrieval uses a multistage, hybrid approach. The model first predicts 1 or 2 likely specialties using triage-available information only, then filters the case database by age group and specialty to construct a clinically plausible candidate pool. Within this subset, vector similarity search identifies the closest matches based on the summarized clinical text. A small number of top-ranked cases are returned, along with similarity scores and metadata, for use in the final reasoning step. Similarity thresholds, top-k values, and fallback strategies were chosen empirically and are reported in Section A4 in Multimedia Appendix 1.

Generation and Prompting Strategy

The generation stage uses an MECR prompt that extends conventional chain-of-thought reasoning for the specific demands of triage. Rather than a single undifferentiated reasoning path, the prompt enforces a 3-tier structure: (1) a clinical risk assessment aligned with formal triage category definitions, (2) integration of condition-specific recommendations from the retrieved HKAETG sections, and (3) comparison with retrieved past cases to identify contextual factors that may support escalation or de-escalation (eg, frailty, social circumstances, and borderline physiology). Each evidence source produces an independent provisional triage recommendation.

The model is then instructed to reconcile these recommendations into a single final category, explicitly justifying the choice and indicating which sources carried the most weight. Outputs include the predicted triage category, a structured explanation, and a coarse confidence label (high, medium, or low), enabling clinicians to audit the reasoning process and gauge when additional scrutiny is warranted. The complete MECR prompt details are provided in Section A5 in Multimedia Appendix 1.

Ablation Configurations

To isolate the contribution of individual retrieval components, we implemented 3 reduced configurations alongside the full MECR-RAG system. The baseline configuration used the same triage definition table and reasoning structure but removed all retrieval, relying only on the index case description. The guideline-only configuration added agentic guideline selection but did not use past-case retrieval, whereas the case-only configuration retrieved similar past cases without guideline sections. In all ablation variants, preprocessing steps, LLM model, and evaluation procedures were held constant to ensure that performance differences could be attributed to the presence or absence of specific retrieval sources rather than to implementation artifacts. The details are provided in Section A6 in Multimedia Appendix 1.

Outcome Definition and Label-Based Evaluation

The primary outcome was the 5-level triage category (categories 1-5) predicted by each LLM configuration, compared with consensus labels assigned by expert triage nurses. Two configurations were evaluated: the MECR-RAG system, which incorporated both the HKAETG and retrieval of past triage cases, and a prompt-only baseline model with no retrieval component. Model outputs were assessed against the final adjudicated triage category for each case.

The primary performance metric was QWK, chosen because it accounts for the ordinal nature of triage categories and penalizes larger discrepancies more heavily, in line with their greater clinical consequence. Overall accuracy was reported as a complementary measure summarizing the proportion of exact matches between model predictions and consensus labels.

To provide a more clinically grounded view of performance, we also calculated precision, recall, and F_1-score for each triage category and for a post hoc grouping of triage levels into 3 urgency bands: immediate (categories 1 and 2), urgent (category 3), and nonurgent (categories 4 and 5). This mapping reflects local operational workflows, in which small changes in urgency assignment can translate into large differences in waiting time and risk exposure. These group-level metrics were used to characterize overtriage and undertriage patterns beyond aggregate QWK and accuracy estimates.

Because LLM outputs are probabilistic, all primary comparisons between MECR-RAG and the baseline LLM were repeated across 5 independent runs using identical test cases, prompts, retrieval databases, and API settings. For these analyses, we summarized performance as mean values with 95% CIs calculated across runs. For analyses based on a single representative run—for example, evaluations using the median-performing prediction set or post hoc subgroup comparisons—95% CIs were obtained via nonparametric case-level bootstrapping (1000 resamples). The specific CI method used is reported alongside each result, and full resampling details are provided in Textbox S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Secondary Outcome-Based Validity Analysis

Because there is no universally accepted gold standard for "true" acuity at ED presentation, triage tools are typically judged against surrogate outcomes such as hospitalization, ICU admission, revisits, and short-term mortality [15,16]. Large observational studies have shown that nurse-assigned categories can substantially underestimate or overestimate severity, with mis-triage affecting up to one-third of patients in some cohorts [17,21]. To complement our consensus label-based evaluation, we therefore derived outcome-based severity tiers and a disposition-safety composite (DSC) and re-examined how MECR-RAG, the baseline LLM, and initial nurse triage aligned with downstream clinical courses.

For all encounters with sufficient follow-up, we assigned an ordinal outcome-based severity tier (R1-R3) adapted from the multilevel reference standard used in MTS validation work, in which a multidisciplinary expert panel defined 3 urgency levels [15]. R1 represented high-acuity courses analogous to the "immediate or very urgent" reference category and was defined by the presence of any marker of life-threatening illness, such as markedly abnormal vital signs with modified early warning score ≥5 or depressed consciousness, emergency surgery or other high-intensity interventions shortly after ED arrival, or death in the ED, ICU, or during the index admission. R2 corresponded to an "urgent" course and included patients who did not meet R1 criteria but required substantial acute treatment (eg, intravenous medication, fluids, and nebulizers) or hospitalization. R3 comprised lower-acuity courses in which none of the R1 or R2 markers were present. Patients who left without being seen (LWBS) were coded separately and excluded from outcome-based binary analyses. Full operational definitions and mapping of specific outcome elements to R1-R3 are provided in Table S2 in Multimedia Appendix 1.

To capture sentinel deterioration events that may reflect underrecognition of severity at triage, we additionally defined a DSC based on components commonly used as ED quality and safety indicators. DSC was coded as positive if any of the following occurred after the index ED visit: (1) unplanned ICU transfer within 72 hours of ward admission, (2) ED revisit within 72 hours resulting in unplanned hospital admission (including ICU admission when present), or (3) all-cause death within 7 days [22-24]. A binary variable, DSC_any, was set to 1 if at least one component was present. Outcome adjudication for R1-R3 and DSC_any was performed using routinely collected outcome data and was fully blinded to the index triage category and all model outputs.

Using these constructs, we prespecified 4 binary analytic "lenses" for sensitivity analyses. A classic lens defined outcome+ as R1 only and outcome– as R2 or R3, with test+ as triage categories 1-2 versus 3-5, mirroring conventional MTS validation [15]. An operational lens defined outcome+ as R1 or R2 versus R3 and test+ as categories 1-3 versus 4-5, reflecting the separation between patients who should not wait in the lowest-priority stream and those who can safely do so. Two augmented lenses incorporated DSC_any: an augmented classic lens (outcome+=R1 or DSC_any; test+=categories 1 and 2) and an augmented operational lens (outcome+=R1 or R2 or DSC_any; test+=categories 1-3). These lenses were intended to test robustness across different, clinically motivated definitions of high severity and to align with prior recommendations on composite triage outcomes. Full definitions are provided in Textbox S4 in Multimedia Appendix 1.

For each lens and each triage method (nurse triage, baseline LLM, and MECR-RAG), we constructed 2×2 tables of test+ versus outcome+ and calculated sensitivity, specificity, and 95% CIs, along with positive and negative likelihood ratios (LR+ and LR–) and diagnostic odds ratios (DORs). To compare models on the same patients, we used paired McNemar tests to assess differences in sensitivity (within outcome+ cases) and specificity (within outcome– cases). As additional clinically interpretable summaries, we examined (1) the proportion of R1 cases assigned to immediate (categories 1 and 2), (2) the proportion of R2 cases assigned to nonurgent (categories 4 and 5), and (3) a weighted harm index per 100 patients that assigns higher penalties to more dangerous errors (R1 or R2 assigned to nonurgent) than to overtriage of R3 cases. Case-level harm scores were compared between models using paired signed tests. Detailed formulas for likelihood ratios, DORs, CIs, and harm index computation are provided in Textbox S5 in Multimedia Appendix 1.

Statistical Analysis

All analyses were conducted using R (version 4.5.0; R Core Team). Interrater agreement and kappa statistics were calculated using standard packages for reliability analysis. Classification metrics were obtained using established machine learning libraries, and nonparametric CIs were estimated using bootstrap resampling. Between-model comparisons used tests appropriate to the data structure, including Friedman tests with Nemenyi post hoc procedures for repeated-run summary metrics and Cochran Q and McNemar tests for paired binary outcomes. Exploratory subgroup, scaling, and cross-model generalizability analyses were treated as descriptive and were not formally powered for hypothesis testing. Full details of the statistical workflow, including package names, resampling schemes, handling of multiple runs, and exploratory analyses, are provided in Textbox S6 in Multimedia Appendix 1.

Ethical Considerations

This study was approved by the Hospital Authority Central Institutional Review Board (reference number CIRB-2024-561-4). Because this was a retrospective analysis of anonymized patient records with no impact on clinical care, the Central Institutional Review Board granted a waiver of informed consent. All data used in this study were fully anonymized prior to analysis, and no identifiable patient information was retained or accessible to the study team. No compensation was provided, as no participants were recruited or contacted for this study.

Results

Participants

The final test set comprised 236 ED triage encounters from 2023. After expert consensus adjudication, the distribution of triage categories was 23/236 (9.7%) category 1, 23/236 (9.7%)

XSL•FO
RenderX

category 2, 82/236 (34.7%) category 3, 98/236 (41.5%) category 4, and 10/236 (4.2%) category 5. The mean patient age was 55.1 (SD 26.2; range 0.3-98.0) years. Overall, 131/236 (55.5%) patients were male and 105/236 (44.5%) were female. The most common clinical specialties were medicine (115/236, 48.7%), orthopedics (36/236, 15.3%), surgery (33/236, 14.0%), and pediatrics (13/236, 5.5%), with the remaining 39/236 (16.5%) classified as other specialties.

The retrieval database used by the MECR-RAG system consisted of 3000 anonymized ED encounters from 2024. The triage category distribution was constructed to reflect real-world attendance patterns, which are typically skewed toward lower-acuity presentations, with modest upsampling of categories 1, 2, and 5 to ensure adequate representation of rare but clinically important groups. The mean age in the retrieval database was 56.7 (SD 25.6; range 0.2-104.0) years, with an equal sex distribution (1500/3000, 50.0% male and 1500/3000, 50.0% female). The specialty mix was similar to the test set, with medicine accounting for 1395/3000 (46.5%) encounters, followed by orthopedics (441/3000, 14.7%), surgery (435/3000, 14.5%), pediatrics (210/3000, 7.0%), and other specialties (519/3000, 17.3%). Baseline characteristics of both datasets are summarized in Table 1, and characteristics of the 1000- and 2000-case retrieval subsets used in the scaling analysis are provided in Table S3 in Multimedia Appendix 1.

Table 1. The demographic and triage category distribution of the test set and 3000-case retrieval-augmented generation database.

Characteristic	Test set (N=236)	3000-case database (N=3000)
Age (years), mean (SD; range)	55.1 (26.2; 0.3-98.0)	56.7 (25.6; 0.2-104.0)
Sex, n (%)		
Male	131 (55.5)	1500 (50.0)
Female	105 (44.5)	1500 (50.0)
Triage categories, n (%)		
Category 1	23 (9.7)	94 (3.1)
Category 2	23 (9.7)	120 (4.0)
Category 3	82 (34.7)	1705 (57.0)
Category 4	98 (41.5)	1035 (34.5)
Category 5	10 (4.2)	46 (1.5)
Specialty, n (%)		
Medicine	115 (48.7)	1395 (46.5)
Orthopedics	36 (15.3)	441 (14.7)
Surgery	33 (14.0)	435 (14.5)
Pediatrics	13 (5.5)	210 (7.0)
Others	39 (16.5)	519 (17.3)

Age is presented as mean (SD). Triage category and specialty values are reported as percentages of total cases within each dataset. Other specialties include dentistry; dermatology; ear, nose, and throat; neurosurgery; obstetrics and gynecology; oncology; ophthalmology; and psychiatry. Both datasets were constructed to reflect real-world triage distributions, with modest oversampling of rare categories (eg, categories 1, 2, and 5) to ensure sufficient representation for model evaluation. Full sampling rationale is provided in the “Methods” section. The test set and retrieval database exhibited broadly comparable distributions in age, sex, and specialty.

Comparative Performance of MECR-RAG and the Baseline LLM, Benchmarked Against Expert Rater Agreement

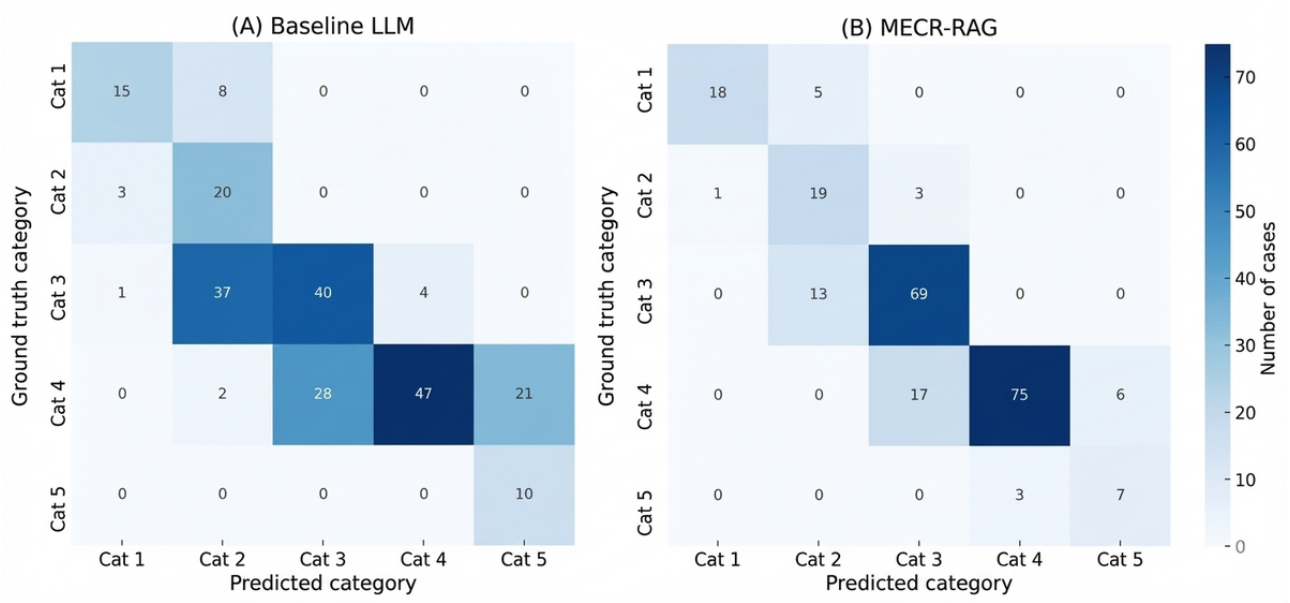
The MECR-RAG system showed substantially higher agreement with consensus triage labels than the baseline LLM. Across 5 independent runs, MECR-RAG achieved a mean QWK of 0.902 (SD 0.0021; 95% CI 0.901-0.904), compared with 0.801 (SD 0.004; 95% CI 0.798-0.804) for the baseline model (mean paired difference 0.101, SD 0.0058; 95% CI 0.097-0.106; P<.001). Classification accuracy followed the same pattern, with a mean accuracy of 0.802 (SD 0.0082; 95% CI 0.795-0.808) for MECR-RAG versus 0.542 (SD 0.0073; 95% CI 0.536-0.548) for the baseline LLM (mean paired difference 0.264, SD 0.0150; 95% CI 0.251-0.278; P<.001). Similar effect sizes were observed when analyses were repeated using single median-run prediction sets, indicating that MECR-RAG’s advantage was robust to run-to-run variability.

To benchmark model performance against human experts, we evaluated interrater reliability between the 2 professional triage nurses who generated the reference labels; their QWK was 0.887. By comparison, MECR-RAG achieved a mean QWK of 0.902 (SD 0.0021; 95% CI 0.901-0.904). Using a prespecified noninferiority margin of −0.05, MECR-RAG met criteria for noninferiority to expert agreement, and this conclusion was unchanged in sensitivity analyses restricted to single median-performing runs. Taken together, these findings suggest that MECR-RAG attains expert-comparable consistency in 5-level triage classification while clearly outperforming a

XSL•FO
RenderX

prompt-only LLM. Confusion matrices comparing triage classification performance between the baseline LLM and MECR-RAG are shown in Figure 2. Results are shown for the median-performing run across 5 stochastic trials in this figure.

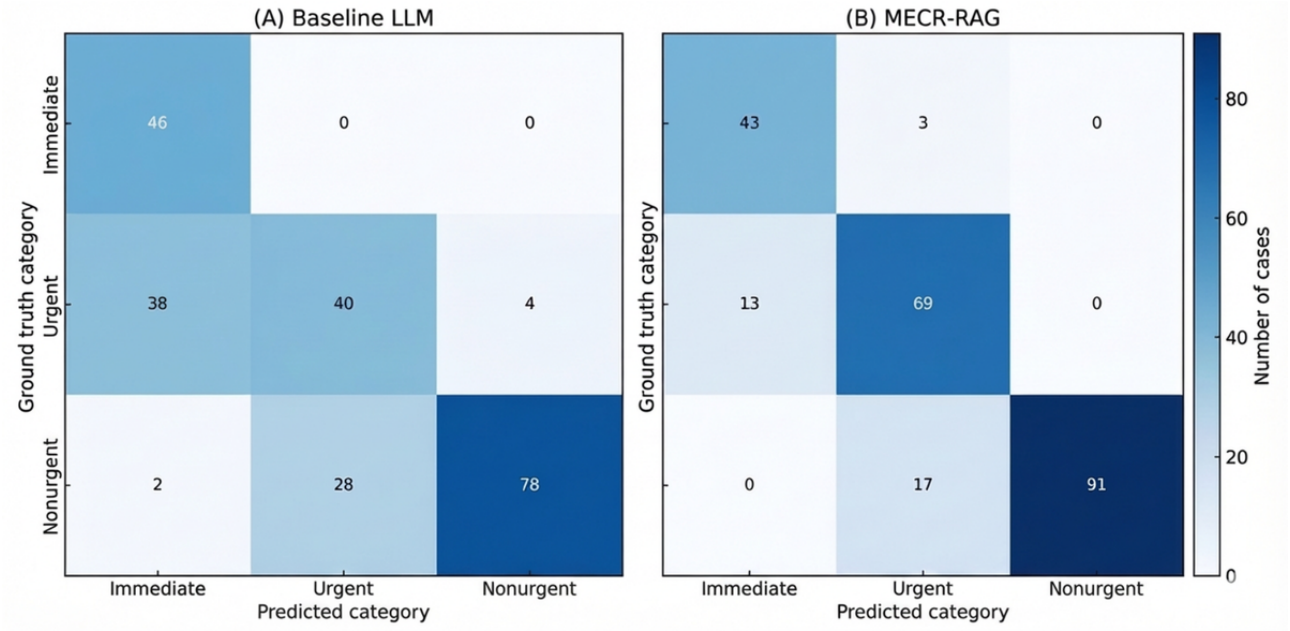
Figure 2. Confusion matrices comparing 5-level triage classification performance between the baseline prompt-only large language model (panel A) and the Multi-Evidence Clinical Reasoning retrieval-augmented generation system (panel B) on 236 emergency department triage encounters in the test set. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation.



Performance Using Clinically Grouped Triage Categories

To reflect operational priorities in the ED, we conducted a post hoc analysis collapsing the 5 triage categories into 3 clinically meaningful groups: immediate (categories 1 and 2), urgent (category 3), and nonurgent (categories 4 and 5). This schema mirrors local workflows, in which immediate cases are seen without delay, urgent cases typically wait approximately 32 minutes, whereas nonurgent cases typically wait approximately 240 minutes based on 2024 institutional data. All metrics in this analysis were derived from the median-performing model run to approximate single-run deployment. Confusion matrices for grouped predictions are shown in Figure 3. Results reflect the median-performing run across 5 stochastic trials in this figure.

Figure 3. Confusion matrices comparing triage group classification performance between the baseline prompt-only large language model (panel A) and the Multi-Evidence Clinical Reasoning retrieval-augmented generation system (panel B) under a 3-group schema. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation.



Under this 3-group schema, MECR-RAG maintained high overall performance, achieving a QWK of 0.882 (95% CI 0.840-0.920) compared with 0.759 (95% CI 0.700-0.813) for the baseline LLM (paired difference 0.125, 95% CI 0.077-0.174; P<.001). Accuracy similarly favored MECR-RAG (0.859 vs 0.695; paired difference 0.164, 95% CI 0.110-0.225; P<.001).

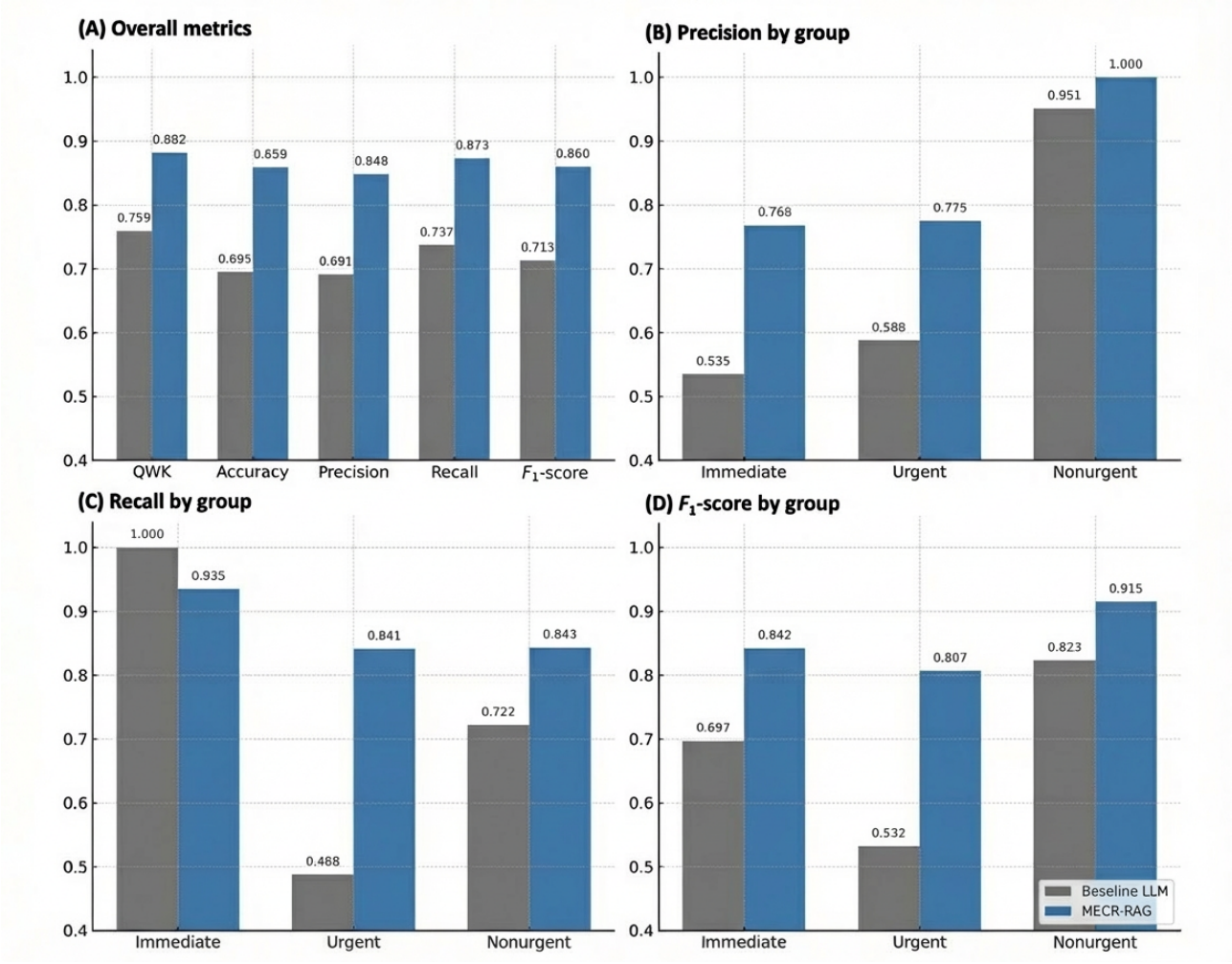
XSL•FO
RenderX

Macrolevel precision, recall, and F_1-score were also higher for MECR-RAG (precision 0.848 vs 0.691; recall 0.873 vs 0.737; F_1-score 0.860 vs 0.713).

Error-pattern analysis showed that MECR-RAG slightly reduced undertriage compared with the baseline LLM (3/236, 1.3% vs 4/236, 1.7%; McNemar P=.69), while substantially reducing overtriage. Overtriage decreased from 68/236 (28.8%) with the baseline LLM to 30/236 (12.7%) with MECR-RAG (McNemar P<.001). Thus, the model reduced unnecessary escalation without compromising detection of higher-acuity cases.

Performance advantages for MECR-RAG were consistent across triage groups. For immediate cases, MECR-RAG had slightly lower recall than the baseline LLM (0.935 vs 1.000) but substantially higher precision (0.768 vs 0.535), yielding a higher F_1-score (0.842 vs 0.697). Overall correctness did not differ significantly (McNemar P=.25), suggesting comparable sensitivity with fewer false positives. In the urgent group, MECR-RAG markedly outperformed the baseline model (recall 0.841 vs 0.488; precision 0.775 vs 0.588; F_1-score 0.807 vs 0.532; McNemar P<.001), indicating improved discrimination in this diagnostically challenging and operationally dominant band. For nonurgent cases, MECR-RAG again showed higher recall (0.843 vs 0.722) and F_1-score (0.915 vs 0.823), with perfect precision (1.000 vs 0.951; McNemar P=.01). These patterns, summarized in Figure 4, suggest that MECR-RAG delivers a more favorable balance between safety and operational efficiency, particularly in the mid- and low-acuity groups that account for most ED visits and drive waiting-time differentials.

Figure 4. Comparative model performance across overall and triage group-specific metrics. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation. QWK: quadratic weighted kappa.



Comparative model performance across overall and triage group-specific metrics. Performance is evaluated using a clinically relevant 3-level grouping: immediate (categories 1 and 2), urgent (category 3), and nonurgent (categories 4 and 5), corresponding to local time-to-physician targets and resource prioritization. Panel A summarizes overall performance across QWK, accuracy, precision, recall, and F_1-score for the baseline prompt-only LLM and the MECR-RAG system; MECR-RAG achieved higher QWK (0.882) and accuracy (0.859) than the baseline LLM (QWK 0.759; accuracy 0.695; both P<.001). Panels B-D show precision, recall, and F_1-score, respectively, by triage group, with MECR-RAG consistently outperforming the baseline LLM across all groups and particularly in the urgent group (F_1-score 0.807 vs 0.532). All results are based on the median-performing model run.

XSL•FO
RenderX

Outcome-Informed Validity Relative to Clinical Severity

Among the 236 triage encounters, 226 (95.8%) had sufficient follow-up information to assign an outcome-based severity tier (R1-R3); the remaining 10 (4.2%) patients were LWBS and were excluded from all outcome-based analyses but retained in label-based performance summaries. Within the 226 encounters with outcome labels, 13/226 (5.8%) met R1 criteria, 117/226 (51.8%) met R2 criteria, and 96/226 (42.5%) were classified as R3. When high-severity outcomes were defined as R1+R2 (n=130) and triage "positive" as categories 1-3 versus 4-5, MECR-RAG achieved the highest sensitivity for high-risk courses at 124/130 (95.4%), compared with 117/130 (90.0%) for initial nurse triage and 122/130 (93.8%) for the baseline LLM. The gain versus nurses was statistically significant (McNemar χ^2_1 =5.14; *P*=.02), whereas sensitivity did not differ from the baseline LLM (χ^2_1 =0.17; *P*=.68). Specificity for low-risk (R3) cases was 77/96 (80%) for MECR-RAG, 78/96 (81%) for nurses, and 66/96 (69%) for the baseline LLM. Thus, MECR-RAG maintained nurse-level specificity (χ^2_1 =0.00; *P*=1.00) while significantly improving specificity over the baseline LLM (χ^2_1 =5.26; *P*=.02). All 3 systems correctly avoided assigning nonurgent (categories 4 and 5) to any R1 cases (0/13 each). The DOR for distinguishing R1+R2 from R3 was 83.8 (95% CI 32.0-218.9) for MECR-RAG, 39.0 (95% CI 18.1-84.1) for nurses, and 33.6 (95% CI 14.6-77.4) for the baseline LLM, consistent with a more favorable overall trade-off between undertriage and overtriage for MECR-RAG, although CIs overlapped because of limited event counts.

To examine clinically important undertriage, we focused on R2 patients assigned to nonurgent triage (categories 4 and 5). Among 117 R2 encounters, 13/117 (11.1%) were initially labeled nonurgent by nurses, compared with 8/117 (6.8%) by the baseline LLM and 6/117 (5.1%) by MECR-RAG. In paired analysis, MECR-RAG approximately halved R2 nonurgent misclassification relative to nurses (13 to 6 cases; McNemar χ^2_1 =5.14; *P*=.02), with a smaller, nonsignificant reduction compared with the baseline LLM (8 to 6 cases; χ^2_1 =0.17; *P*=.68). Considering the entire triage distribution, MECR-RAG also produced the safest nonurgent queue in terms of hidden high-risk patients: among patients predicted as nonurgent, the proportion with R1+R2 outcomes was 6/83 (7%) for MECR-RAG, compared with 13/91 (14%) for nurses and 8/74 (11%) for the baseline LLM. immediate groups were correspondingly enriched for high-severity outcomes, with 47/47 (100%) R1+R2 for nurses, 56/56 (100%) for MECR-RAG, and 82/86 (95%) for the baseline LLM. These patterns suggest that MECR-RAG reduces undertriage of R2 patients into the lowest-priority stream while preserving a clear outcome-based risk gradient across immediate, urgent, and nonurgent groups.

As a complementary summary of mis-triage burden, we constructed a weighted harm index that assigns higher cost to more dangerous errors: R1 assigned to nonurgent (categories 4 and 5) was weighted as 5, R2 assigned to nonurgent as 2, and R3 assigned to immediate or urgent (categories 1-3) as 1. Across the 226 cases with outcome labels, initial nurse triage accumulated a weighted error sum of 44 (harm index 19.47 per 100 patients), the baseline LLM 46 (20.35 per 100 patients), and MECR-RAG 31 (13.72 per 100 patients). At the individual case level, the baseline LLM incurred higher harm scores than MECR-RAG in 19 discordant cases and lower scores in 6, with 201 ties (2-sided sign test *P*=.02), indicating a statistically significant net reduction in weighted harm with MECR-RAG. Compared with nurses, harm scores were lower for MECR-RAG in 14 cases and higher in 8 (204 ties; *P*=.29), suggesting a numerically favorable but underpowered difference. Overall, these outcome-based analyses indicate that MECR-RAG not only improves high-severity sensitivity and reduces R2 undertriage but also yields the lowest aggregate mis-triage burden when accounting for both the direction and severity of errors.

To test robustness to alternative definitions of clinical severity, we repeated binary analyses under classic and augmented outcome "lenses" that defined outcome+ using R1 alone, R1 combined with the DSC, and R1 or R2 combined with the composite (see Methods). Across all lenses, MECR-RAG consistently showed the most favorable trade-off between sensitivity and specificity, with the highest DORs and positive likelihood ratios compared with nurses and the baseline LLM. In a classic Manchester-style lens (outcome+=R1 only; test+=categories 1-2 vs 3-5), all 3 systems achieved perfect sensitivity for R1 cases, but MECR-RAG improved specificity relative to the baseline LLM and approached nurse performance. When the DSC (unplanned ICU transfer within 72 hours, 72-hour revisit with admission, or death within 7 days) was incorporated into augmented lenses, MECR-RAG again maintained the highest discrimination despite wide and overlapping confidence intervals reflecting the small number of sentinel events. Full 2×2 tables, effect estimates, CIs and details for each lens are provided in Tables S4-S7 and Section B in Multimedia Appendix 1. Complementary exploratory analyses, including subgroup performance stratified by age and sex and model behavior on adjudicated disagreement cases, are reported in Sections C and D in Multimedia Appendix 1.

Component-Wise Ablation of Retrieval Mechanisms

To assess the contribution of individual retrieval components, we conducted a component-wise ablation across four configurations: (1) a prompt-only baseline LLM without retrieval, (2) a guideline-only RAG model retrieving sections from the HKAETG, (3) a case-only RAG model retrieving similar past triage cases, and (4) the full MECR-RAG system combining both guideline and case retrieval. Each configuration used an identical preprocessing and generation pipeline and was evaluated over 5 independent runs. QWK was the primary outcome, with overall accuracy, macrolevel precision, and macrolevel recall as secondary metrics.

Across configurations, MECR-RAG achieved the highest agreement with expert consensus (mean QWK 0.902, SD 0.0021; 95% CI 0.901-0.904; accuracy 0.802, SD 0.0082), followed by the case-only model, then the guideline-only model, and finally the baseline LLM. Both guideline-based and case-based retrieval improved performance relative to the baseline, but their combined use in MECR-RAG produced the largest and most consistent gains. Friedman tests on QWK and accuracy across the 4 models were statistically significant, and

XSL•FO
RenderX

post hoc Nemenyi comparisons identified MECR-RAG as the only configuration with a statistically superior rank to the baseline LLM under conservative multiple-comparison control. Full numerical results are shown in Figure 5 (QWK) and Figure 6 (accuracy) and are detailed in Table S8 in Multimedia Appendix 1.

Figure 5. Quadratic weighted kappa performance of model variants and professional raters. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation; RAG: retrieval-augmented generation.

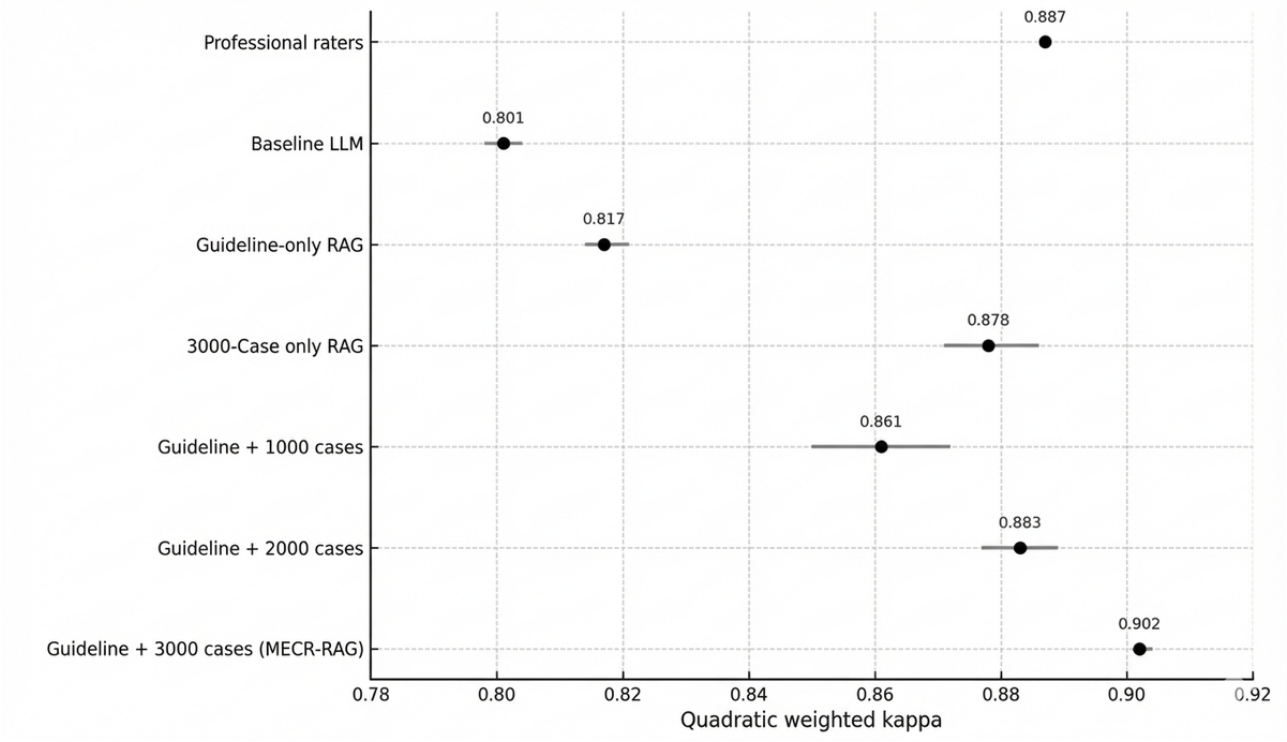
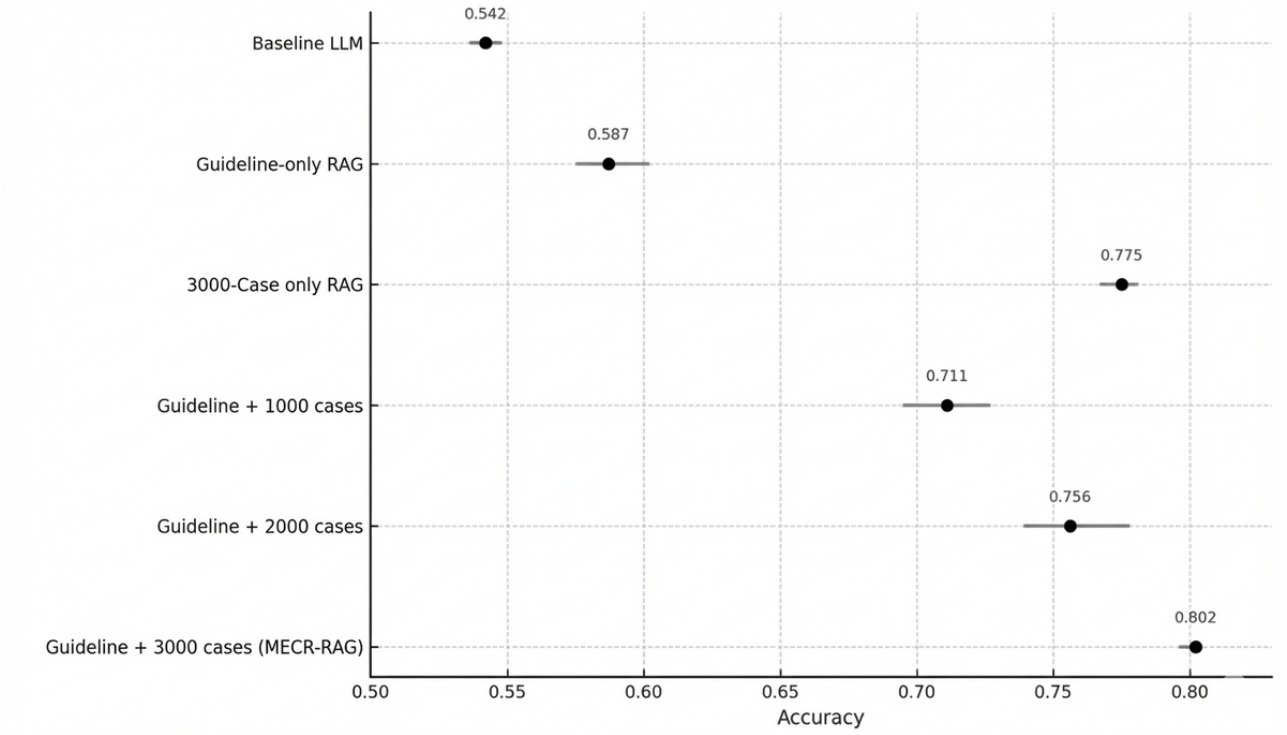


Figure 6. Accuracy performance across model variants. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation; RAG: retrieval-augmented generation.

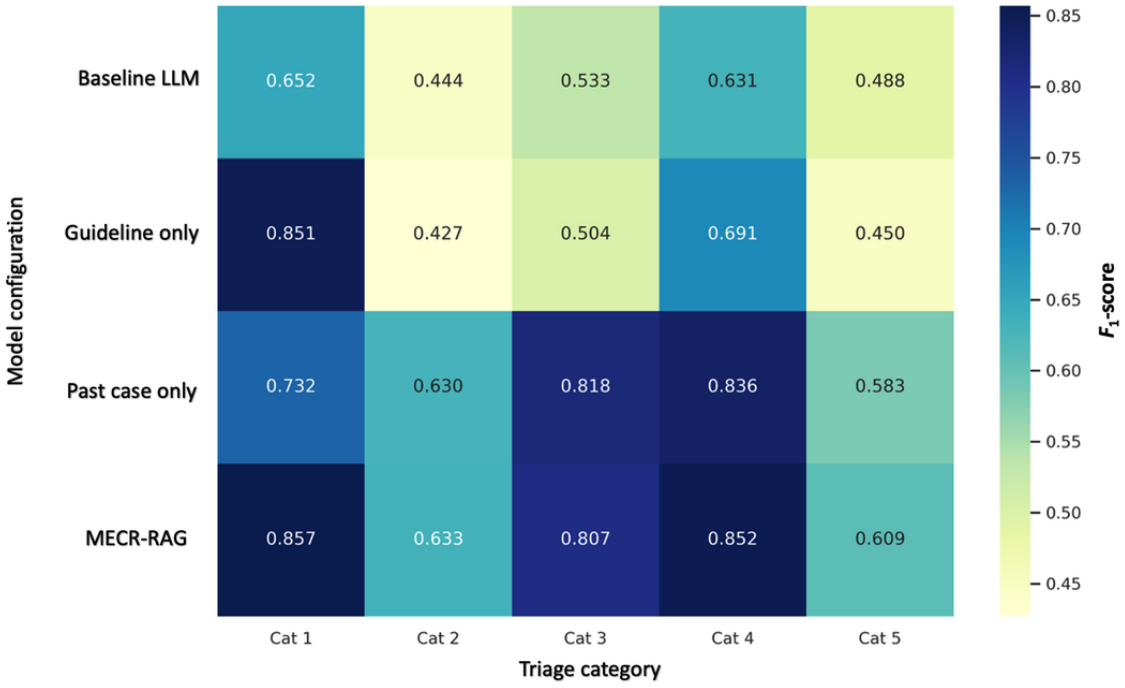


To understand how retrieval affected different acuity levels, we examined per-category accuracy and F_1-scores for triage categories 1-5. Guideline retrieval contributed most strongly to category 1 (critical) performance, where the guideline-only model achieved the highest accuracy and F_1-score, consistent with the value of structured, rule-based protocols for recognizing

XSL•FO
RenderX

life-threatening presentations. In contrast, case retrieval had the largest impact on categories 3 and 4, where the case-only and MECR-RAG models substantially outperformed both the baseline and guideline-only configurations, reflecting the importance of experiential pattern recognition in diagnostically ambiguous, midacuity presentations. For categories 2 and 5, model differences were smaller and not statistically robust, although MECR-RAG generally retained the highest F_1-scores. Overall, MECR-RAG delivered the best balance of precision and recall across all categories, supporting the complementary value of combining guideline-based and case-based retrieval (Figure 7; detailed estimates in Table S9 in Multimedia Appendix 1).

Figure 7. Component-wise ablation analysis of model performance across triage categories. LLM: large language model; MERC-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation.



As an exploratory scaling analysis, we also varied the size of the past-case retrieval database from 1000 to 3000 cases and observed monotonic gains in QWK and accuracy with reduced between-run variability at larger retrieval pools (Tables S3 and S10 in Multimedia Appendix 1).

In Figure 5, the mean QWK values with 95% CIs are shown for the baseline and retrieval-augmented model variants. Points represent mean QWK across 5 stochastic runs, and horizontal error bars indicate 95% CIs across these runs. The professional benchmark (QWK 0.887) reflects interrater agreement between 2 expert triage nurses. The figure includes both ablation variants—a guideline-only RAG model and a 3000-case-only RAG model—and scaling variants that combine guideline and past case retrieval with databases containing 1000, 2000, and 3000 past triage cases. The MECR-RAG model (guideline plus 3000 past cases) achieved the highest mean QWK and was statistically superior to the baseline prompt-only large language model (LLM; mean paired difference 0.101, SD 0.0058; 95% CI 0.097-0.106; P<.001), while demonstrating noninferior agreement compared with professional raters.

In Figure 6, the mean accuracy values with 95% CIs are shown for the baseline prompt-only LLM, RAG variants, and combined retrieval configurations. The x-axis shows accuracy, with points representing mean values across 5 stochastic runs and horizontal error bars indicating 95% CIs. The figure includes ablation variants (guideline-only and 3000-case-only RAG) and scaling variants that combine guideline retrieval with 1000, 2000, or 3000 past cases. The MECR-RAG configuration (guideline plus 3000 past cases) achieved the highest accuracy and demonstrated a statistically significant improvement compared with the baseline LLM.

F_1-scores for each triage category (categories 1-5) across 4 model configurations. Rows represent the baseline prompt-only LLM, a guideline-only RAG model, a 3000-case-only RAG model, and the MECR-RAG system, and columns represent triage categories. Values reflect outputs from the median-performing evaluation run. Guideline retrieval contributed most prominently to performance in category 1, while case retrieval improved performance in categories 3 and 4, where patient presentations are more heterogeneous and require contextual interpretation. The MECR-RAG system generally achieved the highest or near-highest F_1-scores across triage categories, illustrating the additive value of combining structured guideline knowledge with experiential case-based retrieval.

XSL•FO
RenderX

Generalizability Across LLM Architectures

The primary model used for all prior evaluations was DeepSeek-V3. To evaluate whether the MECR-RAG framework generalizes beyond this development model, we conducted exploratory assessments using 2 additional foundation models: Claude 3.7 and GPT-4o. Each model was tested on the same 236-case triage dataset with the same 3000-case retrieval database under 2 configurations: the MECR-RAG system and a baseline, prompt-only LLM without retrieval augmentation.

In these single-run evaluations, MECR-RAG consistently outperformed the corresponding baseline configuration. For Claude 3.7, QWK improved from 0.840 under the baseline to 0.890 with MECR-RAG. Similarly, GPT-4o achieved a QWK of 0.890 with MECR-RAG, compared with 0.778 without retrieval. Although these results are based on single inference runs and are not statistically powered for formal comparison, they provide preliminary evidence that the dual-source retrieval framework may confer similar relative gains across distinct LLM back ends. A detailed breakdown of model predictions versus consensus triage labels for Claude 3.7 and GPT-4o, under both MECR-RAG and baseline conditions, is provided in Figures S1 and S2 in Multimedia Appendix 1, respectively.

Taken together, these preliminary findings suggest that the benefits of dual-source retrieval generalize beyond DeepSeek-V3 and may confer cross-model robustness of MECR-RAG across LLMs from different developers, supporting its broader applicability in diverse clinical settings. Further benchmarking across a wider range of model families, input formats, and deployment environments is warranted to validate robustness and assess trade-offs in latency, interpretability, and resource efficiency for real-world clinical deployment.

Discussion

Principal Findings

This retrospective evaluation suggests that a dual-source, retrieval-augmented LLM can achieve expert-comparable agreement with consensus nurse triage labels while substantially outperforming a prompt-only baseline model. MECR-RAG improved QWK and accuracy over the baseline LLM under both the 5-level and 3-group schemas and met prespecified noninferiority criteria compared with professional triage nurses. The largest gains occurred in Categories 3 and 4 and in the urgent versus nonurgent grouping, which together accounted for most ED visits and are operationally dominant, with large differences in waiting times. Ablation analyses indicated that both guideline retrieval and case retrieval contributed to performance, with the full dual-source MECR-RAG configuration consistently outperforming guideline-only and case-only variants. Performance also improved as the retrieval database increased from 1000 to 3000 cases, suggesting that retrieval quality benefits from a richer case pool. Exploratory cross-model experiments with Claude 3.7 Sonnet and GPT-4o further suggested that these benefits generalize across different LLM back ends.

Ground Truth Validity and Construct Interpretation

In this study, the primary reference standard was expert nurse consensus at the time of presentation, which reflects routine clinical decision-making but is an imperfect proxy for underlying acuity. Large US data on ESI version 4 report mis-triage in 32.2% of encounters (3.3% undertriage and 28.9% overtriage), underscoring that nurse-level labels are imperfect proxies for “true” acuity [17]. Even with trained emergency staff, approximately 17% of triage assignments may be inaccurate [21]. To address these limitations, we complemented consensus labels with outcome-informed constructs, ordinal severity tiers (R1-R3) and a DSC, adapted from MTS validation work and recommendations on triage outcome selection [15,22,24].

Within these outcome-based frameworks, MECR-RAG showed higher sensitivity for R1+R2 courses than initial nurse triage at comparable specificity, reduced misclassification of R2 patients into the nonurgent stream, and yielded the lowest weighted harm index, suggesting closer alignment between assigned urgency and downstream clinical course. At the same time, event counts for the most severe outcomes were modest, and both nurse consensus and the composite outcomes remain pragmatic proxies rather than definitive measures of “true” severity. We therefore interpret these findings as supportive evidence that MECR-RAG performs well against clinically grounded, outcome-informed constructs of acuity, rather than as proof that it fully captures an underlying severity state.

Comparison With Prior Work

To our knowledge, this is the first study to evaluate a dual-source, RAG-enhanced LLM for emergency triage using raw triage documentation paired with expert-consensus reference labels and complementary outcome-informed severity constructs. Prior RAG studies in emergency care have shown that retrieval can improve triage or referral performance on simulated exam-style cases or curated ED datasets but have relied on structured scenarios or routine triage labels without independent adjudication [13,14]. Our design addresses the broader evidence gap noted in a recent systematic review, in which only 5% of 519 published evaluations of health care LLMs used real patient care data for testing [25]. This provides a more realistic and stringent test of RAG-enhanced triage than has been reported previously.

Masanneck et al [4] established an early benchmark by evaluating LLM triage performance against expert assessments, but their dataset consisted of physician-curated vignettes from a single day, with selected patient attributes adjusted according to predefined standard operating procedures. In contrast, our study applies dual-source RAG to a year-long sample of anonymized triage notes without manual editing or reinterpretation, preserving the fidelity of frontline documentation and temporal variation. Zaboli et al [3] and Paslı et al [7] have likewise emphasized the importance of handling unstructured narratives and integrating local triage frameworks; our dual-retrieval design directly addresses these needs by combining structured guideline content with real-world case patterns.

XSL•FO
RenderX

Our findings also align with broader work on nurse triage systems, showing that midacuity presentations are both the most common and the most difficult to classify. In a large cohort of over 5.3 million adult ED encounters using ESI version 4, Sax et al [17] found that approximately 29% of visits were overtriaged and that the sensitivity of ESI version 4 for identifying low-acuity, low-resource patients (correctly assigning ESI 4-5 among patients using <2 resources and with no critical interventions) was only 50%. A majority of encounters were assigned to the midlevel category ESI 3, highlighting poor differentiation between lower- and higher-acuity patients. Validation of the HKAETG has similarly shown the lowest interrater agreement for categories 3 and 4, reflecting the intrinsic difficulty of distinguishing "urgent" from "semiurgent" presentations [26]. In our cohort, categories 3 and 4 comprised about three-quarters of all cases, and the prompt-only LLM reproduced the overtriage patterns documented in both human and LLM triage studies. In contrast, MECR-RAG substantially reduced overtriage and improved precision and recall specifically in these mid- and low-acuity strata, suggesting that dual-source retrieval may help address a long-standing bottleneck in emergency triage performance.

Clinical and Operational Implications

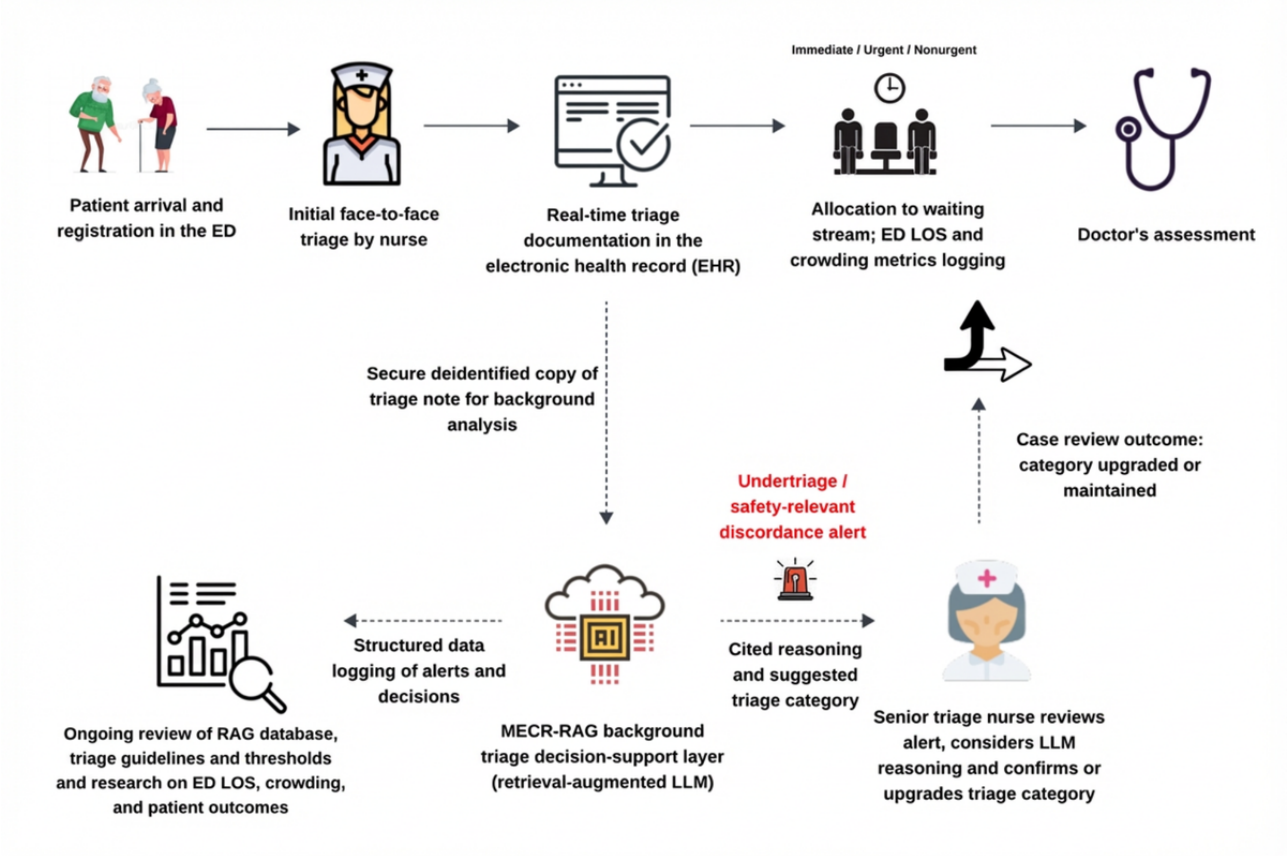
Although we did not directly measure ED length of stay (LOS), boarding, or mortality, several findings suggest how MECR-RAG could plausibly influence safety and flow if implemented as decision support. Under a 3-group schema (immediate, urgent, and nonurgent), MECR-RAG achieved higher agreement and accuracy than the prompt-only LLM, with markedly less overtriage and similarly low undertriage. Gains were concentrated in the urgent and nonurgent tiers, which dominate the case mix and are associated with substantially longer waits for lower-priority patients at our center. In outcome-based analyses, MECR-RAG showed higher sensitivity for R1+R2 courses than initial nurse triage at nurse-level specificity, reduced misclassification of R2 patients into the nonurgent stream, and yielded the lowest severity-weighted harm index. Together, these patterns indicate fewer high-risk patients "hidden" in the longest-wait queue and fewer low-risk patients unnecessarily occupying high-acuity resources, aligning urgency assignments more closely with downstream clinical course.

These redistributions of risk are clinically relevant because prolonged ED LOS, crowding, and boarding have been repeatedly linked to delayed treatment, longer inpatient stays, and higher short-term mortality. More efficient separation between urgent and nonurgent queues could, in principle, help decompress monitored areas, reduce avoidable time in high-acuity spaces for low-risk patients, and support more timely care for those with high-severity courses. However, our evaluation was retrospective and did not include LOS, crowding, or mortality as prespecified end points. Any potential contribution of MECR-RAG therefore remains hypothesis-generating and must be tested in prospective, outcome-focused studies.

As illustrated in the proposed human-in-the-loop workflow (Figure 8), MECR-RAG is intended to function as a background triage decision-support layer: triage nurses retain primary responsibility for assigning categories, while a secure, deidentified copy of the triage note is analyzed in parallel, with the system remaining silent for most encounters and surfacing alerts only for predefined safety-relevant discordances (eg, potential undertriage of urgent or immediate cases). Alerts and explanations can then be reviewed by a senior nurse, who may confirm or upgrade the triage category, while all alerts, decisions, and associated operational metrics are logged to support ongoing quality assurance and research on ED LOS, crowding, and downstream outcomes. A staged implementation strategy, beginning with shadow-mode deployment and progressing to focused pilots of discordance alerts with monitoring for automation bias, alert fatigue, throughput, and equity, will be essential before wider rollout.

XSL•FO
RenderX

Figure 8. Conceptual human-in-the-loop workflow for Multi-Evidence Clinical Reasoning retrieval-augmented generation as a background triage decision-support layer in the emergency department. ED: emergency department; LLM: large language model; LOS: length of stay; MECR-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation; RAG: retrieval-augmented generation.



Patients follow the routine care pathway shown in the top row: arrival and registration in the ED, initial face-to-face triage by a nurse, real-time triage documentation in the electronic health record, allocation to an urgency-based waiting stream (immediate, urgent, or nonurgent) with logging of ED LOS and crowding metrics, and subsequent doctor's assessment. In parallel, a secure deidentified copy of the triage note is sent to the MECR-RAG system, which operates as a background triage decision-support layer. MECR-RAG generates its own triage recommendation with cited reasoning and remains silent for most encounters, but triggers an undertriage or other safety-relevant discordance alert when its recommendation conflicts with the nurse-assigned category. Alerts and explanations are routed to a senior triage nurse, who may confirm or upgrade the triage category; all alerts, decisions, and outcomes are logged to support ongoing quality assurance and research on ED LOS, crowding, and patient-centered outcomes.

Limitations and Future Directions

This study has several important limitations. First, within the staged evaluation pathway outlined by DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence), this retrospective, offline prediction study represents preclinical (in silico) validation and should not be interpreted as evidence of real-world clinical effectiveness or safety [27]. Moreover, strong offline discrimination does not reliably translate into clinical benefit after implementation; prior early warning and sepsis tools (including the Rothman Index and the Epic Sepsis Model) have shown materially weaker performance and limited impact in independent, postdeployment evaluations, partly due to differences in data quality, missingness, case mix, alert logic, and workflow across sites [28-32]. Retrospective deep learning models are also vulnerable to dataset shift, class imbalance, and missing data, further limiting transportability across settings and time [33]. Randomized evaluations of AI-enabled interventions similarly suggest that even high-performing models may fail to improve primary clinical end points without careful workflow integration and attention to system constraints [34,35]. Accordingly, our retrospective, single-center results for MECR-RAG should be viewed as predeployment evidence and require prospective, multicenter evaluation before inferring equivalent real-time performance or improved patient outcomes.

Second**,** the evaluation was conducted in a single ED using the HKAETG; performance may differ under other triage frameworks (eg, MTS and ESI), languages, documentation practices, workflows, and patient case mix. Although the retrieval-augmented design is modular and could, in principle, be adapted by reindexing local guidelines and locally representative triage notes without retraining the underlying LLM, portability remains an empirical question and requires external validation and prospective multicenter evaluation.

Third, the retrieval database and text summaries were constructed automatically from routine triage notes without manual adjudication, and prompts were generated without clinician editing. This reflects realistic deployment conditions and improves scalability, but some retrieved exemplars may be

XSL•FO
RenderX

mislabeled or incomplete, and clinically relevant nuances may be lost during summarization. In addition, we did not directly evaluate usability, interpretability, automation bias, or workflow impact, which are increasingly recognized as critical determinants of the safety and effectiveness of AI-enabled clinical decision-support systems [32,36].

These limitations point to clear priorities for future work. Prospective, multicenter studies across institutions using different triage frameworks (HKAETG, MTS, and ESI) are needed to assess generalizability and to determine how MECR-RAG performs in real-time workflows. Consistent with DECIDE-AI and recent guidance from major journals and professional societies, future evaluations should move beyond agreement metrics to prespecify patient-centered and operational end points, including ED and hospital LOS, crowding and boarding times, unexpected ICU transfer, time-critical treatment delays, and short-term mortality [18,37-39]. Implementation-focused studies should also examine how MECR-RAG is integrated into practice (eg, as a second reader or discordance alert), its effects on clinician behavior, workload, and equity, and unintended consequences such as alert fatigue or overreliance on model outputs, ideally using pragmatic randomized or quasi-experimental designs where feasible [34,35]. Framing this work as an early-phase evaluation and explicitly outlining the need for prospective, outcome-focused trials in line with current AI evaluation guidance positions MECR-RAG as a promising candidate for further study rather than a tool ready for immediate widespread deployment.

Conclusions

In summary, this retrospective, single-center study provides one of the first evaluations of a dual-source, retrieval-augmented LLM for emergency triage on real triage notes with expert consensus labels, extending prior work beyond simulated scenarios and routine labels. MECR-RAG achieved agreement comparable to expert nurses and clearly outperformed a prompt-only LLM, particularly in the high-volume category 3-4 bands, while approximately halving overtriage and keeping undertriage very low. In exploratory outcome-based analyses within this cohort, MECR-RAG also showed higher sensitivity for high-severity courses, a safer nonurgent queue, and the lowest severity-weighted harm index compared with initial nurse triage. These findings support the potential of retrieval-augmented LLMs as clinically coherent decision-support tools that function as background second readers—surfacing guidelines and analogous cases, flagging discordant or high-risk assignments for senior review—and thereby potentially improving allocation of high-acuity resources in ways that could, if replicated prospectively, help reduce ED crowding, LOS, and related adverse outcomes. However, this remains an early-phase, retrospective, single-center evaluation based on proxy outcomes, and rigorous prospective multicenter studies across different triage frameworks, with prespecified patient-centered and operational end points and careful attention to workflow, usability, and equity, are essential before such systems can be recommended for routine clinical deployment.

Acknowledgments

We would like to thank KH Leung for organizing and facilitating the consensus triage labeling process. We are also grateful to the emergency department nursing staff at Princess Margaret Hospital for their triage assessments, which formed the basis of the clinical documentation used in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional and regulatory restrictions imposed by the Hospital Authority of Hong Kong, but are available from the corresponding author on reasonable request. The underlying code supporting the findings of this study is available at MECR-RAG-Clinical-Research [40].

Authors' Contributions

HS conceptualized the study design, developed the methodology, collected and curated the data, performed the analysis, created the figures, and drafted the initial manuscript. TK developed the computational architecture, configured the model framework, implemented and executed the MECR-RAG system.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed Multi-Evidence Clinical Reasoning retrieval-augmented generation implementation, prompts, and additional analyses.
[DOCX File , 1124 KB - medinform_v14i1e82026_app1.docx]

XSL•FO
RenderX

References

1. Suamchaiyaphum K, Jones AR, Markaki A. Triage accuracy of emergency nurses: an evidence-based review. J Emerg Nurs 2024;50(1):44-54. [doi: 10.1016/j.jen.2023.10.001] [Medline: 37930287]
2. Hinson JS, Martinez DA, Cabral S, George K, Whalen M, Hansoti B, et al. Triage performance in emergency medicine: a systematic review. Ann Emerg Med 2019;74(1):140-152. [doi: 10.1016/j.annemergmed.2018.09.022] [Medline: 30470513]
3. Zaboli A, Brigo F, Brigiari G, Massar M, Parodi M, Pfeifer N, et al. Chat-GPT in triage: still far from surpassing human expertise - an observational study. Am J Emerg Med 2025;92:165-171. [doi: 10.1016/j.ajem.2025.03.028] [Medline: 40120387]
4. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: comparative study. J Med Internet Res 2024;26:e53297 [FREE Full text] [doi: 10.2196/53297] [Medline: 38875696]
5. Arslan B, Nuhoglu C, Satici M, Altinbilek E. Evaluating LLM-based generative AI tools in emergency triage: a comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. Am J Emerg Med 2025;89:174-181. [doi: 10.1016/j.ajem.2024.12.024] [Medline: 39731895]
6. Pasli S, Yadigaroğlu M, Kirimli EN, Beşer MF, Unutmaz İ, Ayhan A, et al. ChatGPT-supported patient triage with voice commands in the emergency department: a prospective multicenter study. Am J Emerg Med 2025;94:63-70. [doi: 10.1016/j.ajem.2025.04.040] [Medline: 40273640]
7. Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: insights from a study with ChatGPT. Am J Emerg Med 2024;78:170-175. [doi: 10.1016/j.ajem.2024.01.037] [Medline: 38295466]
8. Kaboudi N, Firouzbakht S, Shahir Eftekhar M, Fayazbakhsh F, Joharivarnoosfaderani N, Ghaderi S, et al. Diagnostic accuracy of chatgpt for patients' triage; a systematic review and meta-analysis. Arch Acad Emerg Med 2024;12(1):e60. [doi: 10.22037/aaem.v12i1.2384] [Medline: 39290765]
9. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel) 2025;13(6):603 [FREE Full text] [doi: 10.3390/healthcare13060603] [Medline: 40150453]
10. Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc 2025;32(6):1032-1039. [doi: 10.1093/jamia/ocaf059] [Medline: 40220286]
11. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. In: Ku LW, Srikumar V, Martins A, editors. Findings of the Association for Computational Linguistics ACL. Bangkok, Thailand: Association for Computational Linguistics; 2024:6233-6251.
12. Zhang G, Xu Z, Jin Q, Chen F, Fang Y, Liu Y, et al. Leveraging long context in retrieval augmented language models for medical question answering. NPJ Digit Med 2025;8(1):239 [FREE Full text] [doi: 10.1038/s41746-025-01651-w] [Medline: 40316710]
13. Yazaki M, Maki S, Furuya T, Inoue K, Nagai K, Nagashima Y, et al. Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model. Prehosp Emerg Care 2025;29(3):203-209. [doi: 10.1080/10903127.2024.2374400] [Medline: 38950135]
14. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med 2025;8(1):263 [FREE Full text] [doi: 10.1038/s41746-025-01684-1] [Medline: 40346344]
15. Zachariasse JM, Seiger N, Rood PPM, Alves CF, Freitas P, Smit FJ, et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. PLoS One 2017;12(2):e0170811 [FREE Full text] [doi: 10.1371/journal.pone.0170811] [Medline: 28151987]
16. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
17. Sax DR, Warton EM, Mark DG, Vinson DR, Kene MV, Ballard DW, et al. Evaluation of the emergency severity index in us emergency departments for the rate of mistriage. JAMA Netw Open 2023;6(3):e233404 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3404] [Medline: 36930151]
18. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, Park H, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
19. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
20. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980;36(2):343-346. [Medline: 26625475]

XSL•FO
RenderX

21. Hinson JS, Martinez DA, Schmitz PSK, Toerper M, Radu D, Scheulen J, et al. Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. Int J Emerg Med 2018;11(1):3 [FREE Full text] [doi: 10.1186/s12245-017-0161-8] [Medline: 29335793]
22. Delgado MK, Liu V, Pines JM, Kipnis P, Gardner MN, Escobar GJ. Risk factors for unplanned transfer to intensive care within 24 hours of admission from the emergency department in an integrated healthcare system. J Hosp Med 2013;8(1):13-19. [doi: 10.1002/jhm.1979] [Medline: 23024040]
23. Gabayan GZ, Derose SF, Asch SM, Yiu S, Lancaster EM, Poon KT, et al. Patterns and predictors of short-term death after emergency department discharge. Ann Emerg Med 2011;58(6):551-558.e2 [FREE Full text] [doi: 10.1016/j.annemergmed.2011.07.001] [Medline: 21802775]
24. Ling D, Sung C, Fang C, Ko C, Chou EH, Herrala J, et al. High-risk return visits to United States emergency departments, 2010-2018. West J Emerg Med 2022;23(6):832-840 [FREE Full text] [doi: 10.5811/westjem.2022.7.57028] [Medline: 36409935]
25. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
26. Fan M, Leung L. Validation of the Hong Kong accident and emergency triage guidelines. The HKU Scholars Hub. 2013. URL: https://hub.hku.hk/handle/10722/210102 [accessed 2026-01-02]
27. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
28. Rothman MJ, Rothman SI, Beals J. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. J Biomed Inform 2013;46(5):837-848 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.011] [Medline: 23831554]
29. Arnold J, Davis A, Fischhoff B, Yecies E, Grace J, Klobuka A, et al. Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: a prospective observational study. BMJ Open 2019;9(10):e032187 [FREE Full text] [doi: 10.1136/bmjopen-2019-032187] [Medline: 31601602]
30. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(8):1065-1070 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
31. Habib AR, Lin AL, Grant RW. The epic sepsis model falls short-the importance of external validation. JAMA Intern Med 2021;181(8):1040-1041. [doi: 10.1001/jamainternmed.2021.3333] [Medline: 34152360]
32. Downing N, Rolnick J, Poole S, Hall E, Wessels A, Heidenreich P, et al. Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation. BMJ Qual Saf 2019;28(9):762-768 [FREE Full text] [doi: 10.1136/bmjqs-2018-008765] [Medline: 30872387]
33. Mann KD, Good NM, Fatehi F, Khanna S, Campbell V, Conway R, et al. Predicting patient deterioration: a review of tools in the digital hospital setting. J Med Internet Res 2021;23(9):e28209 [FREE Full text] [doi: 10.2196/28209] [Medline: 34591017]
34. Mazor T, Farhat KS, Trukhanov P, Lindsay J, Galvin M, Mallaber E, et al. Clinical trial notifications triggered by artificial intelligence-detected cancer progression: a randomized trial. JAMA Netw Open 2025;8(4):e252013 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.2013] [Medline: 40257799]
35. Mandair D, Elia MV, Hong JC. Considerations in translating AI to improve care. JAMA Netw Open 2025;8(4):e252023 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.2023] [Medline: 40257804]
36. Schinkel M, van der Poll T, Wiersinga WJ. Artificial intelligence for early sepsis detection: a word of caution. Am J Respir Crit Care Med 2023;207(7):853-854 [FREE Full text] [doi: 10.1164/rccm.202212-2284VP] [Medline: 36724366]
37. Chiu I, Lin Y, Syue Y, Kung C, Wu K, Li C. The influence of crowding on clinical practice in the emergency department. Am J Emerg Med 2018;36(1):56-60. [doi: 10.1016/j.ajem.2017.07.011] [Medline: 28705743]
38. Sartini M, Carbone A, Demartini A, Giribone L, Oliva M, Spagnolo AM, et al. Overcrowding in emergency department: causes, consequences, and solutions-a narrative review. Healthcare (Basel) 2022;10(9):1625 [FREE Full text] [doi: 10.3390/healthcare10091625] [Medline: 36141237]
39. Armoundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, American Heart Association Institute for Precision Cardiovascular Medicine; Council on CardiovascularStroke Nursing; Council on Lifelong Congenital Heart DiseaseHeart Health in the Young; Council on Cardiovascular RadiologyIntervention; Council on Hypertension; Council on the Kidney in Cardiovascular Disease;Stroke Council. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the american heart association. Circulation 2024;149(14):e1028-e1050 [FREE Full text] [doi: 10.1161/CIR.0000000000001201] [Medline: 38415358]
40. Multi-Evidence Clinical Reasoning RAG for Emergency Medicine Triage - Journal Submission Supporting Code. GitHub. URL: https://github.com/tkwr565/MECR-RAG-Clinical-Research [accessed 2026-01-05]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
API: application programming interface
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence
DOR: diagnostic odds ratio
DSC: disposition-safety composite
ED: emergency department
ESI: Emergency Severity Index
HKAETG: Hong Kong Accident and Emergency Triage Guidelines
ICU: intensive care unit
JAMA: Journal of the American Medical Association
LLM: large language model
LOS: length of stay
LWBS: left without being seen
MECR: Multi-Evidence Clinical Reasoning
MECR-RAG: Multi-Evidence Clinical Reasoning retrieval-augmented generation
MIMIC-IV: Medical Information Mart for Intensive Care IV
MTS: Manchester Triage System
PMH: Princess Margaret Hospital
QWK: quadratic weighted kappa
RAG: retrieval-augmented generation
TRIPOD-LLM: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis - Large Language Models

Edited by A Benis; submitted 07.Aug.2025; peer-reviewed by S Mohanadas, D Yoo; comments to author 05.Sep.2025; revised version received 05.Dec.2025; accepted 29.Dec.2025; published 26.Jan.2026.

Please cite as:
Wong HS, Wong TK
Multi-Evidence Clinical Reasoning With Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study
JMIR Med Inform 2026;14:e82026
URL: https://medinform.jmir.org/2026/1/e82026
doi:10.2196/82026
PMID:

©Hang Sheung Wong, Tsz Kwan Wong. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning Algorithms to Predict Venous Thromboembolism in Patients With Sepsis in the Intensive Care Unit: Multicenter Retrospective Study

Yan Zhang[1*], MD; Xia Ren[2*], MS; Luojie Liu[2], MD; Junjie Zha[1], MS; Yijie Gu[1], BS; Hongwei Ye[1], MD

[1]Department of Emergency and Critical Care Medicine, Changshu No.1 People's Hospital, Changshu, Jiangsu, China

[2]Department of Gastroenterology, Changshu No.1 People's Hospital, Changshu, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Hongwei Ye, MD
Department of Emergency and Critical Care Medicine
Changshu No.1 People's Hospital
No. 1 Shuyuan Street
Changshu, Jiangsu, 215500
China
Phone: 86 13915657982
Email: zy_rocky@163.com

Abstract

Background: Venous thromboembolism (VTE) is a common and severe complication in intensive care unit (ICU) patients with sepsis. Conventional risk stratification tools lack sepsis-specific features and may inadequately capture complex, nonlinear interactions among clinical variables.

Objective: This study aimed to develop and validate an interpretable machine learning (ML) model for the early prediction of VTE in ICU patients with sepsis.

Methods: This multicenter retrospective study used data from the Medical Information Mart for Intensive Care IV database for model development and internal validation, and an independent cohort from Changshu Hospital for external validation. Candidate predictors were selected through univariate analysis, followed by least absolute shrinkage and selection operator regression. Retained variables were used in multivariable logistic regression to identify independent predictors, which were then used to develop 9 ML models, including categorical boosting, decision tree, k-nearest neighbor, light gradient boosting machine, logistic regression, multilayer perceptron, naive Bayes, random forest, and support vector machine. Performance was evaluated by discrimination (area under the curve [AUC]), calibration, and clinical use (decision curve analysis). A subgroup analysis stratified by the Sequential Organ Failure Assessment score was conducted in the external cohort to assess model stability across sepsis severity levels. Model interpretability was assessed using Shapley Additive Explanations (SHAP) to quantify the contribution of features to the predicted risk.

Results: A total of 25,197 patients from the Medical Information Mart for Intensive Care IV cohort and 328 patients from the external cohort were included, with VTE incidences of 844 out of 25,197 (3.4%) and 30 out of 328 (9.2%), respectively. The light gradient boosting machine model performed best, achieving an AUC of 0.956 in internal validation. Despite the higher VTE incidence and clinical severity in the external validation, the model maintained robust generalization with an AUC of 0.786. Notably, the model achieved enhanced discriminative ability in the severe sepsis subgroup (Sequential Organ Failure Assessment score >6) with an AUC of 0.816, compared with 0.769 in the mild to moderate sepsis subgroup. Calibration curves indicated strong agreement between predicted and observed outcomes, and decision curve analysis showed superior net benefit across clinically relevant thresholds. SHAP analysis identified central venous catheterization, serum chloride and bicarbonate levels, arterial catheterization, and prolonged partial thromboplastin time as the most influential predictors. Partial dependence plots revealed both linear and nonlinear associations between these variables and VTE risk. Individual-level force plots further enhanced interpretability by visualizing personalized risk profiles.

Conclusions: We developed a high-performing and interpretable ML model for predicting VTE in ICU patients with sepsis. The model demonstrated robustness across cohorts and enhanced performance in the severe sepsis population. By integrating

XSL•FO
RenderX

diverse clinical data and leveraging SHAP for transparent explanations, this tool may support personalized prophylaxis and early diagnostic strategies.

(JMIR Med Inform 2026;14:e80969) doi:10.2196/80969

KEYWORDS

venous thromboembolism; sepsis; machine learning; SHAP; MIMIC-IV

Introduction

Sepsis is a life-threatening condition characterized by acute organ dysfunction resulting from a dysregulated host response to infection and remains a leading cause of mortality in the intensive care unit (ICU) [1]. According to the Global Burden of Disease Study, approximately 48.9 million cases of sepsis occur annually worldwide, resulting in an estimated 11.0 million deaths, which accounts for 19.7% of all global deaths [2]. Venous thromboembolism (VTE), comprising deep vein thrombosis and pulmonary embolism (PE), is a common and severe complication of sepsis that significantly worsens patient outcomes. Despite adherence to guideline-recommended thromboprophylaxis, the incidence of VTE in patients with sepsis remains high at 37.2%, up to 10-fold greater than in ICU populations with no sepsis [3]. Previous studies have shown that the presence of VTE is associated with a 28-day mortality rate of 28.6% among patients with sepsis, and the occurrence of PE nearly doubles the risk of in-hospital death (odds ratio [OR] 1.94) [4,5]. These findings underscore the urgent need for early and accurate VTE risk stratification in sepsis to guide timely clinical interventions and reduce morbidity and mortality in this high-risk population.

Current clinical methods for predicting VTE in patients with sepsis have notable limitations. Traditional biomarkers such as D-dimer, although widely used, lack specificity in the context of sepsis, as the condition itself activates coagulation pathways and elevates fibrin degradation products [6,7]. Furthermore, commonly used scoring systems such as the Wells and Autar scores were primarily developed and validated in trauma or postoperative populations. These tools do not account for sepsis-specific indicators, such as the Sequential Organ Failure Assessment (SOFA) score or lactate levels, thereby limiting their predictive use in patients with sepsis [8-10]. Although multivariable logistic regression (LR) models that incorporate clinical and laboratory variables have shown promise, with reported area under the curve (AUC) values reaching up to 0.87, these models are inherently linear and may fail to capture the complex and nonlinear interactions among sepsis-related risk factors [11,12].

Machine learning (ML) techniques have emerged as powerful tools for handling high-dimensional data, uncovering nonlinear associations, and automating feature selection. In recent years, ML models have demonstrated superior performance in predicting complications such as septic shock and acute kidney injury (AKI) in ICU patients, with reported AUC values reaching up to 0.90 [13-15]. However, earlier generations of ML models often functioned as “black boxes,” which limited their adoption in clinical settings due to the lack of interpretability. Explainable ML approaches, such as Shapley Additive Explanations (SHAP), have been developed to address this challenge. By quantifying the contribution of each feature, SHAP enhances model transparency and facilitates clinical understanding [16]. For instance, Liu et al [17] developed an interpretable ML model for septic shock prediction, whose explanatory outputs aligned closely with hemodynamic mechanisms described in the Sepsis-3 guidelines, thereby reinforcing its clinical relevance [18].

Therefore, this study aimed to develop an interpretable ML model that integrates clinical features and laboratory indicators to predict the risk of VTE in ICU patients with sepsis. By applying SHAP to interpret model outputs and identify key predictors, we sought to create a decision support tool that combines high predictive accuracy with strong clinical interpretability, ultimately facilitating personalized prevention strategies.

Methods

Data Sources

The training and internal validation datasets were extracted from the Medical Information Mart for Intensive Care IV version 3.0 (MIMIC-IV v3.0) database, which contains comprehensive and high-quality data from 65,366 patients admitted to the ICU at Beth Israel Deaconess Medical Center between 2008 and 2022. One of the authors (YZ) completed the Collaborative Institutional Training Initiative examination (record ID: 60227322) and was granted access to the database for research purposes.

Ethical Considerations

Ethical approval for the use of MIMIC-IV data was granted by the institutional review boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, with a waiver of informed consent due to the use of deidentified data. The data were handled in accordance with the Health Insurance Portability and Accountability Act standards to ensure patient privacy and confidentiality. In addition, ICU patients admitted to Changshu Hospital affiliated to Soochow University between January 2019 and August 2024 were included as an external validation cohort. Ethical approval for this component of the study was obtained from the ethics review committee of Changshu Hospital (number L2024055). As the study was observational in nature and used routinely collected clinical data, the requirement for written informed consent was waived. All electronic data were anonymized before analysis, and access to the dataset was restricted to the primary research team to maintain strict confidentiality. All procedures conformed to the ethical standards set forth in the Declaration of Helsinki. No compensation was provided to any participants involved in this study.

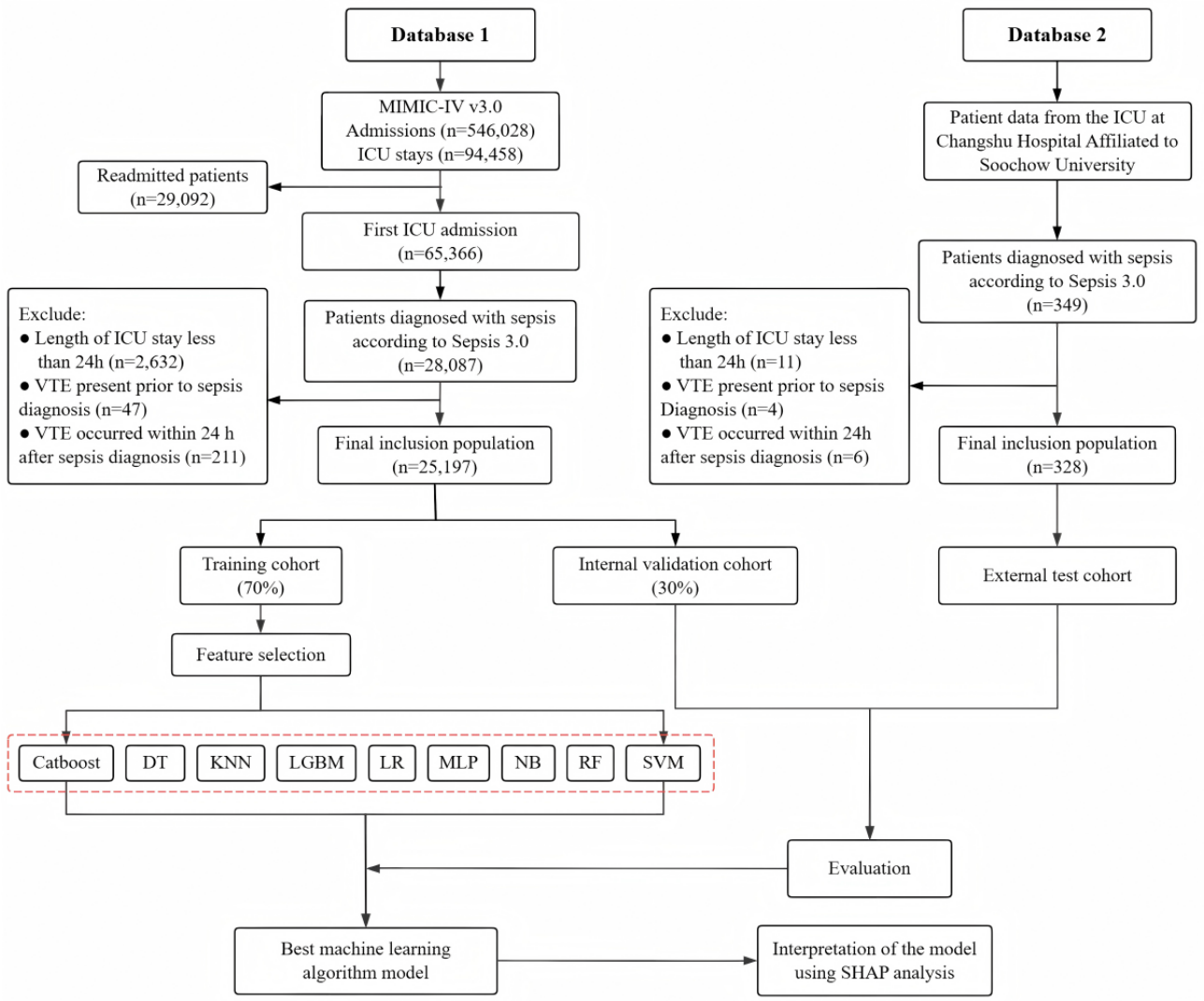
XSL•FO
RenderX

Study Population

This study included adult ICU patients (aged 18 years and older) with sepsis. For patients with multiple ICU admissions, only the first admission was analyzed. Sepsis was defined according to the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), which require a suspected or confirmed infection accompanied by an increase in the SOFA score of 2 points or more [18]. VTE was defined as an acute episode of deep vein thrombosis (either upper or lower extremities), superficial vein thrombosis (either upper or lower extremities), or PE. VTE diagnoses were identified based on either the *ICD-9* (*International Classification of Diseases, Ninth Revision*; codes 45119, 4512, 45181, 45182, 45183, 45184, 45189, 4519, 4532, 4538, 45381, 45382, 45383, 45384, 45385, 45386, 45387, 45389, 4539, 4150, 41511, 41512, 41513, 41519, 45340, 45341, 45342, 4510, 452, 4530, 4531, and 4533) or the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*; codes I808, I809, I8290, I82890, I2699, I2692, I2690, I2602, I2609, I8000, I8001, I8002, I81, I820, and I821) in the MIMIC-IV database, or through diagnostic imaging reports from Changshu Hospital, including duplex venous ultrasonography and contrast-enhanced computed tomography. In the Changshu Hospital cohort, the ICU implemented a routine lower limb venous ultrasound screening protocol for all patients with sepsis meeting high-risk criteria. Patients were excluded meeting any of the following criteria: (1) ICU stay of less than 24 hours, (2) a documented VTE diagnosis prior to sepsis onset, and (3) VTE events occurring within 24 hours following the sepsis diagnosis.

Patients from the MIMIC-IV cohort were randomly split into a training set and an internal validation set in a 7:3 ratio. The training set was used for variable selection and model development, while the internal and external validation sets were used to assess model performance. The overall study workflow is illustrated in Figure 1.

Figure 1. Flowchart of patient enrollment and cohort selection. CatBoost: categorical boosting; DT: decision tree; ICU: intensive care unit; KNN: k-nearest neighbor; LGBM: light gradient boosting machine; LR: logistic regression; MIMIC: Medical Information Mart for Intensive Care Unit; MLP: multilayer perceptron; NB: naive Bayes; RF: random forest; SHAP: SHapley Additive exPlanations; SVM: support vector machine; VTE: venous thromboembolism.



XSL•FO
RenderX

Data Extraction and Processing

Data from the MIMIC-IV database were extracted using Structured Query Language via pgAdmin 4 (version 6.21; pgAdmin Development Team). All clinical and laboratory variables were collected within the first 24 hours of ICU admission, with only the initial measurement included for variables with multiple readings. A set of variables was extracted for each patient, encompassing (1) demographics: age (years), sex, race (Asian, Black, White, Hispanic, and other), height (cm), weight (kg), BMI (kg/m^2), and insurance (Medicare or other); (2) vital signs: heart rate (HR, times per minute), respiratory rate (RR, times per minute), systolic blood pressure (mm Hg), diastolic blood pressure (mm Hg), temperature (°C), and oxygen saturation (SpO_2, %); (3) severity scores: Glasgow Coma Scale, SOFA score, and simplified acute physiology score II (SAPS II); (4) laboratory data: white blood cell count (K/μL), hemoglobin (g/dL), platelet (PLT, K/μL), C-reactive protein (mg/dL), alanine aminotransferase (U/L), aspartate aminotransferase (U/L), albumin (g/dL), creatinine (mg/dL), sodium (mEq/L), potassium (mEq/L), chloride (mEq/L), calcium (mg/dL), blood urea nitrogen (mg/dL), hematocrit (%), glucose (mg/dL), international normalized ratio, prothrombin time (PT, seconds), partial thromboplastin time (PTT, seconds), pH, partial pressure of oxygen (PaO_2, mm Hg), partial pressure of carbon dioxide ($PaCO_2$, mm Hg), base excess (mEq/L), lactic acid (mEq/L), and bicarbonate (HCO_3^-, mEq/L); (5) ventilation parameters: positive end-expiratory pressure (PEEP, cmH_2O), fraction of inspired oxygen (FiO_2, %), and PEEP/FiO_2 ratio; (6) comorbidities: hypertension, diabetes, myocardial infarction (MI), chronic obstructive pulmonary disease (COPD), asthma, pancreatitis, acute respiratory failure (ARF), AKI, acute respiratory distress syndrome (ARDS), and malignant cancer; and (7) therapeutic interventions: arterial catheterization, central venous catheterization, cardiopulmonary bypass, continuous renal replacement therapy (CRRT), ventilation, heparin, aspirin, and vasopressors use. The same set of variables was extracted from the external validation cohort at Changshu Hospital affiliated with Soochow University, ensuring consistency between the datasets.

For missing data, variables with less than 5% missingness were imputed using mean substitution. For those with 5%-30% missing values, multiple imputation methods were used. Variables with more than 30% missing data were excluded from analysis to minimize bias and preserve model robustness. Outliers were identified and treated as missing values and subsequently handled using the same imputation strategies.

Statistical Analysis

All statistical analyses were performed using Stata software (version 16.0; StataCorp LLC), R software (version 4.1.2; R Foundation for Statistical Computing), DecisionLinnc (version 1.0; DecisionLinnc Corp), and Python (version 3.9.12; Python Software Foundation). The Kolmogorov-Smirnov test was used to assess the normality of continuous variables. Variables with a normal distribution were expressed as means and SDs, and group comparisons were conducted using the 2-tailed independent samples *t* test. Non–normally distributed variables were presented as median and IQR and analyzed using the Mann-Whitney *U* test. Categorical variables were summarized as counts (percentages) and compared using the chi-square test. Two-sided *P* values of <.05 were considered statistically significant.

Variable Selection and Model Development

Variables that showed statistically significant differences in the univariate analysis were considered candidate predictors. These variables were further refined using the least absolute shrinkage and selection operator regression to minimize overfitting and reduce multicollinearity. Predictors with nonzero coefficients were subsequently entered into a multivariate LR model. Independent predictors identified through this process were incorporated into the ML models.

Following variable selection, the Synthetic Minority Oversampling Technique (SMOTE) preprocessing algorithm was applied to address class imbalance between VTE and non-VTE cases in the training cohort. In the training cohort, VTE cases (576/17,637, 3.27%) were highly unbalanced compared with non-VTE cases (17,061/17,637, 96.73%). After SMOTE application, the VTE class was balanced to match the non-VTE class, resulting in a 50:50 distribution with a total of 34,122 samples used for model training. This oversampling method helped ensure that the models were trained on a balanced dataset, thereby improving their ability to detect minority class events. A total of 9 ML algorithms were trained on the development cohort, including categorical boosting (CatBoost), decision tree, k-nearest neighbor, light gradient boosting machine (LGBM), LR, multilayer perceptron, naive Bayes, random forest (RF), and support vector machine. For all ML models, we used a 5-fold cross-validation strategy, combined with Grid Search, for hyperparameter optimization and model selection. Notably, SMOTE was applied exclusively to the training folds of the cross-validation process to prevent data leakage into the validation folds or the internal validation set.

Model Evaluation and Interpretation

Model performance was assessed using the internal and external validation datasets. Discrimination was evaluated by plotting receiver operating characteristic curves and calculating the AUC. Calibration curves were used to compare predicted probabilities with observed outcomes. Clinical use was assessed using decision curve analysis, which estimated the net benefit across a range of probability thresholds. In addition, several evaluation metrics were computed, including accuracy, sensitivity, specificity, and the Youden index, to offer a comprehensive assessment of model performance.

The best performing model was selected based on comprehensive evaluation of these metrics. To assess the model's stability and performance in different risk populations, a stratified analysis was conducted on the external validation cohort based on the SOFA score. Patients were divided into 2 subgroups: mild to moderate sepsis (SOFA score ≤6) and severe sepsis (SOFA score >6). To enhance model interpretability, SHAP analysis was conducted. SHAP values were computed to quantify the contribution of individual features to the model's predictions and to explore feature interactions. Visualization

XSL•FO
RenderX

tools from the SHAP package in Python, including feature importance plots, bee swarm plots, force plots, and partial dependence plots (PDPs), were used to provide an intuitive understanding of the model's decision-making process.

Results

Baseline Characteristics

A total of 25,197 patients from the MIMIC-IV cohort and 328 patients from the external validation cohort were included in the study. Among them, 844 out of 25,197 (3.4%) patients in the MIMIC-IV cohort and 30 out of 328 (9.2%) patients in the external cohort developed VTE. The baseline characteristics of all patients are summarized in Table 1. In the MIMIC-IV cohort, patients who developed VTE were younger and exhibited significantly higher HRs and RRs than those with no VTE (all $P<.001$). Laboratory parameters revealed higher PLT counts, serum chloride levels, and PTT in the VTE group (all $P<.05$). In addition, patients with VTE had higher illness severity scores, as indicated by elevated SOFA and SAPS II scores (both $P<.05$). Comorbidities such as ARF, AKI, and ARDS were significantly more common in the VTE group (all $P<.001$). Furthermore, these patients were more likely to have undergone invasive procedures, including arterial and central venous catheterization (both $P<.001$). In the external validation cohort, similar patterns were observed. Patients with VTE demonstrated higher rates of ARDS and more frequent use of mechanical ventilation than patients with no VTE (both $P<.001$). These findings highlight significant differences in demographics, clinical severity, comorbidities, and therapeutic interventions between patients with and with no VTE across both cohorts, emphasizing the relevance of these factors in VTE risk stratification.

XSL•FO
RenderX

Table 1. Baseline characteristics of the MIMIC-IV[a] and external validation cohorts.

Variables	MIMIC-IV cohort (N=25,197)			External Validation cohort (N=328)		
	Non-VTE[b] (n=24,353)	VTE (n=844)	*P* value	Non-VTE (n=298)	VTE (n=30)	*P* value
Demographic data						
Age (years), mean (SD)	66.61 (16.28)	63.83 (16.75)	<.001	73.33 (12.60)	75.90 (9.88)	.19
Sex (male), n (%)	14173 (58.2)	502 (59.5)	.46	210 (70.5)	22 (73)	.74
BMI, mean (SD)	27.71 (6.20)	28.06 (6.15)	.11	22.78 (2.59)	21.93 (3.45)	.20
Race, n (%)	N/A[c]	N/A	.14	N/A	N/A	N/A
Asian	1994 (8.2)	80 (9.5)	N/A	N/A	N/A	N/A
Black	15812 (64.9)	568 (67.3)	N/A	N/A	N/A	N/A
White	806 (3.3)	27 (3.2)	N/A	N/A	N/A	N/A
Hispanic	5038 (20.7)	147 (17.4)	N/A	N/A	N/A	N/A
Other	703 (2.9)	22 (2.6)	N/A	N/A	N/A	N/A
Insurance, n (%)	N/A	N/A	.02	N/A	N/A	.09
Other	4,027 (16.5)	166 (19.7)	N/A	83 (27.9)	4 (13)	N/A
Medicare	20326 (83.5)	678 (80.3)	N/A	215 (72.2)	26 (87)	N/A
Vital signs						
HR[d], mean (SD)	90.00 (20.58)	94.38 (21.12)	<.001	102.86 (24.64)	104.73 (23.56)	.68
RR[e], mean (SD)	19.53 (6.44)	20.38 (6.42)	<.001	22.81 (5.97)	23.97 (6.39)	.35
SBP[f], mean (SD)	121.21 (24.40)	122.17 (25.61)	.29	119.34 (28.29)	119.00 (29.08)	.95
DBP[g], mean (SD)	67.30 (18.12)	68.01 (19.08)	.29	66.14 (14.18)	66.00 (14.80)	.96
Temperature, mean (SD)	36.81 (0.77)	36.81 (0.84)	.88	37.45 (1.71)	37.40 (2.31)	.91
SpO_2[h], median (Q1-Q3)	98 (95-100)	98 (95-100)	.005	92 (88-97)	91 (87-98)	.18
Severity scoring system						
GCS[i], median (Q1-Q3)	15 (13-15)	15 (13-15)	.38	15 (12-15)	14 (11-15)	.32
SOFA[j], median (Q1-Q3)	5 (3-8)	5 (3-8)	.04	6 (5-8)	6 (4-8)	.87
SAPS[k] II, mean (SD)	40.31 (14.22)	41.70 (15.31)	.01	39.69 (14.19)	42.43 (15.58)	.36
Laboratory data						
WBC[l], mean (SD)	12.62 (6.81)	13.05 (7.38)	.10	14.13 (6.88)	13.26 (5.50)	.43
Hemoglobin, mean (SD)	10.51 (2.24)	10.62 (2.13)	.13	10.87 (2.65)	9.71 (2.64)	.03
PLT[m], mean (SD)	195.42 (100.25)	206.41 (111.67)	.005	189.43 (128.43)	191.00 (103.37)	.94
ALT[n], median (Q1-Q3)	26 (16-52)	29 (17-63)	.03	51 (25-81)	27 (19-53)	.11
Creatinine, mean (SD)	1.48 (1.46)	1.41 (1.33)	.12	2.23 (1.82)	2.30 (2.43)	.87
Sodium, mean (SD)	137.89 (5.59)	138.43 (5.64)	.007	137.45 (5.17)	138.38 (5.86)	.41
Potassium, mean (SD)	4.24 (0.77)	4.14 (0.76)	<.001	4.17 (0.72)	4.28 (0.84)	.49
Chloride, mean (SD)	104.04 (6.69)	105.40 (6.80)	<.001	104.44 (7.00)	105.38 (6.74)	.47
BUN[o], median (Q1-Q3)	20 (14-34)	21 (14-35)	.85	14 (9-21)	13 (6-21)	.90
Hematocrit, mean (SD)	31.99 (6.67)	32.03 (6.41)	.85	32.29 (9.06)	29.36 (7.93)	.07
Glucose, median (Q1-Q3)	130 (106-165)	127 (105-162)	.14	8 (6-11)	8 (6-10)	.90
INR[p], median (Q1-Q3)	1.3 (1.2-1.6)	1.3 (1.2-1.5)	.81	1.3 (1.2-1.5)	1.2 (1.1-1.6)	.42
PT[q], median (Q1-Q3)	14.6 (12.8-17.0)	14.6 (13.0-17.1)	.95	15.7 (14.4-17.6)	14.9 (14.2-16.8)	1.00

XSL•FO
RenderX

Variables	MIMIC-IV cohort (N=25,197)			External Validation cohort (N=328)		
	Non-VTE[b] (n=24,353)	VTE (n=844)	*P* value	Non-VTE (n=298)	VTE (n=30)	*P* value
PTT[r], median (Q1-Q3)	31.1 (27.3-37.6)	32.2 (27.7-41.6)	<.001	31.9 (27.7-38.2)	39.9 (35.1-46.8)	.17
pH[s], mean (SD)	7.36 (0.10)	7.37 (0.10)	.15	7.40 (0.10)	7.37 (0.11)	.33
PaO_2[t], median (Q1-Q3)	140.0 (87.0-273.0)	129.0 (84.0-239.5)	.006	93.5 (74.3-121.0)	85.5 (67.0-112.0)	.17
$PaCO_2$[u], mean (SD)	42.35 (12.06)	41.04 (12.52)	.003	38.51 (13.15)	39.59 (17.02)	.74
BE[v], mean (SD)	–1.43 (5.24)	–1.47 (5.44)	.82	–2.05 (5.50)	–2.76 (5.31)	.49
Lac[w], median (Q1-Q3)	1.8 (1.2-2.7)	1.8 (1.2-2.6)	.55	2.3 (1.5-3.5)	2.0 (1.5-3.0)	.82
HCO_3^-[x], mean (SD)	22.62 (4.76)	23.12 (4.81)	.003	22.76 (5.11)	22.27 (5.11)	.62
Ventilation Variables, median (Q1-Q3)						
PEEP[y]	5 (5-5)	5 (5-5)	.14	3 (2-5)	5 (5-5)	.001
FiO_2[z]	60 (50-100)	60 (50-100)	.09	40 (33-500)	40 (33-50)	.82
PEEP[y]/FiO_2[z]	247 (153-352)	245 (146-345)	.52	230 (168-307)	203 (152-295)	.24
Comorbidities, n (%)						
Hypertension	10083 (41.4)	374 (44.3)	.09	214 (71.8)	21 (70)	.83
Diabetes	3521 (14.5)	0 (0.0)	<.001	85 (28.5)	5 (17)	.17
Myocardial infarction	1085 (4.5)	11 (1.3)	<.001	43 (14.4)	4 (13)	.87
COPD[aa]	1647 (6.8)	0 (0.0)	<.001	40 (13.4)	3 (10)	.60
Asthma	1912 (7.9)	78 (9.2)	.14	19 (6.4)	0 (0)	.15
Pancreatitis	600 (2.5)	38 (4.5)	<.001	10 (3.4)	4 (13)	.01
ARF[bb]	8202 (33.7)	339 (40.2)	<.001	180 (60.4)	24 (80)	.04
AKI[cc]	16710 (68.6)	652 (77.3)	<.001	187 (62.8)	16 (53)	.31
ARDS[dd]	5604 (23.0)	260 (30.8)	<.001	54 (18.1)	16 (53)	<.001
Malignant cancer	3829 (15.7)	152 (18.0)	.07	64 (21.5)	8 (27)	.51
Therapeutic interventions, n (%)						
Arterial catheterization	1595 (6.6)	166 (19.7)	<.001	55 (18.5)	30 (100)	<.001
Central venous catheterization	2028 (8.3)	218 (25.8)	<.001	45 (15.1)	27 (90)	<.001
CPB[ee]	2888 (11.9)	77 (9.1)	.02	30 (10.1)	1 (3)	.23
CRRT[ff]	2175 (8.9)	96 (11.4)	.02	30 (10.1)	10 (33)	<.001
Ventilation	12004 (49.3)	416 (49.3)	1.00	148 (49.7)	25 (83)	<.001
Heparin	21304 (87.5)	786 (93.1)	<.001	254 (85.2)	29 (97)	.08
Aspirin	12522 (51.4)	384 (45.5)	<.001	135 (45.3)	7 (23)	.02

XSL•FO
RenderX

Variables	MIMIC-IV cohort (N=25,197)			External Validation cohort (N=328)		
	Non-VTE[b] (n=24,353)	VTE (n=844)	*P* value	Non-VTE (n=298)	VTE (n=30)	*P* value
Vasopressors	7538 (31.0)	290 (34.4)	.04	107 (35.9)	25 (83)	<.001

[a]MIMIC: Medical Information Mart for Intensive Care Unit.
[b]VTE: venous thromboembolism.
[c]N/A: not applicable.
[d]HR: heart rate.
[e]RR: respiratory rate.
[f]SBP: systolic blood pressure.
[g]DBP: diastolic blood pressure.
[h]SpO_2: oxygen saturation.
[i]GCS: Glasgow Coma Scale.
[j]SOFA: Sequential Organ Failure Assessment.
[k]SAPS II: simplified acute physiology score II.
[l]WBC: white blood cell count.
[m]PLT: platelet.
[n]ALT: alanine aminotransferase.
[o]BUN: blood urea nitrogen.
[p]INR: international normalized ratio.
[q]PT: prothrombin time.
[r]PTT: partial thromboplastin time.
[s]pH: potential of hydrogen.
[t]PaO_2: partial pressure of oxygen.
[u]$PaCO_2$: partial pressure of carbon dioxide.
[v]BE: base excess.
[w]Lac: lactic acid.
[x]HCO_3^{--}: bicarbonate.
[y]PEEP: positive end-expiratory pressure.
[z]FiO_2: fraction of inspired oxygen.
[aa]COPD: chronic obstructive pulmonary disease.
[bb]ARF: acute respiratory failure.
[cc]AKI: acute kidney injury.
[dd]ARDS: acute respiratory distress syndrome.
[ee]CPB: cardiopulmonary bypass.
[ff]CRRT: continuous renal replacement therapy.

Selection of Predictors

Patients in the training set were divided into VTE and non-VTE groups. Univariate analysis identified several variables significantly associated with VTE, including admission age, insurance, HR, RR, SAPS II, alanine aminotransferase, potassium, chloride, glucose, PTT, PaO_2, $PaCO_2$, PEEP, and comorbidities such as diabetes, MI, COPD, pancreatitis, ARF, AKI, and ARDS (all *P*<.05). In addition, therapeutic interventions including arterial catheterization, central venous catheterization, cardiopulmonary bypass, CRRT, heparin use, and aspirin use were significantly associated with VTE (all *P*<.05; Table 2). Least absolute shrinkage and selection operator regression identified 18 variables with nonzero coefficients at the optimal lambda value of 0.0029. These included age, insurance, HR, potassium, chloride, PTT, diabetes, MI, COPD, pancreatitis, ARF, AKI, ARDS, arterial catheterization, central venous catheterization, CRRT, heparin use, and aspirin use (Figures 2A and 2B). These variables were subsequently entered into a multivariate LR model, which revealed that potassium (OR 0.8711, 95% CI 0.7541-0.9756; *P*=.02), chloride (OR 1.0206, 95% CI 1.0081-1.0326; *P*=.001), PTT (OR 1.0086, 95% CI 1.0054-1.0115; *P*<.001), MI (OR 0.2851, 95% CI 0.1101-0.5999; *P*=.003), AKI (OR 1.5274, 95% CI 1.2166-1.9369; *P*<.001), arterial catheterization (OR 1.8985, 95% CI 1.4102-2.5648; *P*<.001), central venous catheterization (OR 2.0805, 95% CI 1.5213-2.8382; *P*<.001), and heparin use (OR 1.5509, 95% CI 1.1203-2.0851; *P*=.01) were independently associated with VTE risk (Table 3). We noted instances of complete separation in the training set for comorbidities such as diabetes and COPD, where the VTE subgroup contained zero cases, leading to extreme ORs in the multivariable LR.

XSL•FO
RenderX

Table 2. Univariate analysis of the clinical features in the training set.

Variables	Total (N=17,637)	Non-VTE[a] (n=17,061)	VTE (n=576)	*P* value
Demographic data				
Age (years), mean (SD)	66.49 (16.30)	66.57 (16.27)	64.10 (16.93)	<.001
Sex (male), n (%)	10268 (58.2)	9934 (58.2)	334 (58.0)	.91
BMI, mean (SD)	27.75 (6.20)	27.74 (6.20)	27.90 (6.04)	.53
Race, n (%)				.42
Asian	1415 (8.0)	1360 (8.0)	55 (9.6)	
Black	11463 (65.0)	11,085 (65.0)	378 (65.6)	
White	604 (3.4)	583 (3.4)	21 (3.7)	
Hispanic	3642 (20.7)	3538 (20.7)	104 (18.1)	
Other	513 (2.9)	495 (2.9)	18 (3.1)	
Insurance, n (%)				.004
Other	2937 (16.7)	2816 (16.5)	121 (21.0)	
Medicare	14700 (83.4)	14245 (83.5)	455 (79.0)	
Vital signs				
HR[b], mean (SD)	90.17 (20.61)	90.06 (20.57)	93.50 (21.61)	<.001
RR[c], mean (SD)	19.52 (6.47)	19.50 (6.46)	20.18 (6.59)	.02
SBP[d], mean (SD)	121.14 (24.40)	121.13 (24.38)	121.45 (25.14)	.77
DBP[e], mean (SD)	67.35 (18.17)	67.34 (18.14)	67.48 (19.10)	.86
Temperature, mean (SD)	36.81 (0.77)	36.81 (0.77)	36.74 (0.84)	.06
SpO_2[f], median (Q1-Q3)	98 (95-100)	98 (95-100)	98 (95-100)	.07
Severity scoring system				
GCS[g], median (Q1-Q3)	15 (13-15)	15 (13-15)	15 (13-15)	.67
SOFA[h], median (Q1-Q3)	5 (3-8)	5 (3-8)	5 (4-8)	.12
SAPS[i] II, mean (SD)	40 (14)	40 (14)	42 (16)	.01
Laboratory data				
WBC[j], mean (SD)	12.65 (6.84)	12.64 (6.82)	13.04 (7.35)	.19
Hemoglobin, mean (SD)	10.51 (2.25)	10.50 (2.25)	10.58 (2.11)	.42
PLT[k], mean (SD)	195.34 (100.17)	195.06 (99.70)	203.51 (113.05)	.08
ALT[l], median (Q1-Q3)	26 (16-53)	26 (16-53)	30 (17-62)	<.001
Creatinine, mean (SD)	1.48 (1.45)	1.48 (1.46)	1.43 (1.28)	.39
Sodium, mean (SD)	137.90 (5.62)	137.89 (5.62)	138.25 (5.65)	.13
Potassium, mean (SD)	4.24 (0.77)	4.24 (0.77)	4.12 (0.73)	<.001
Chloride, mean (SD)	104.10 (6.68)	104.06 (6.68)	105.29 (6.73)	<.001
BUN[m], median (Q1-Q3)	20 (14-34)	20 (14-34)	22 (14-35)	.20
Hematocrit, mean (SD)	31.98 (6.71)	31.98 (6.72)	31.87 (6.37)	.67
Glucose, median (Q1-Q3)	130 (106-165)	130 (106-165)	125 (103-159)	.03
INR[n], median (Q1-Q3)	1.3 (1.2-1.6)	1.3 (1.2-1.6)	1.3 (1.2-1.5)	.54
PT[o], median (Q1-Q3)	14.6 (12.8-17.0)	14.6 (12.8-16.9)	14.6 (13.1-17.1)	.35
PTT[p], median (Q1-Q3)	31.2 (27.3-37.6)	31.2 (27.3-37.6)	32.6 (27.9-42.2)	<.001

XSL•FO
RenderX

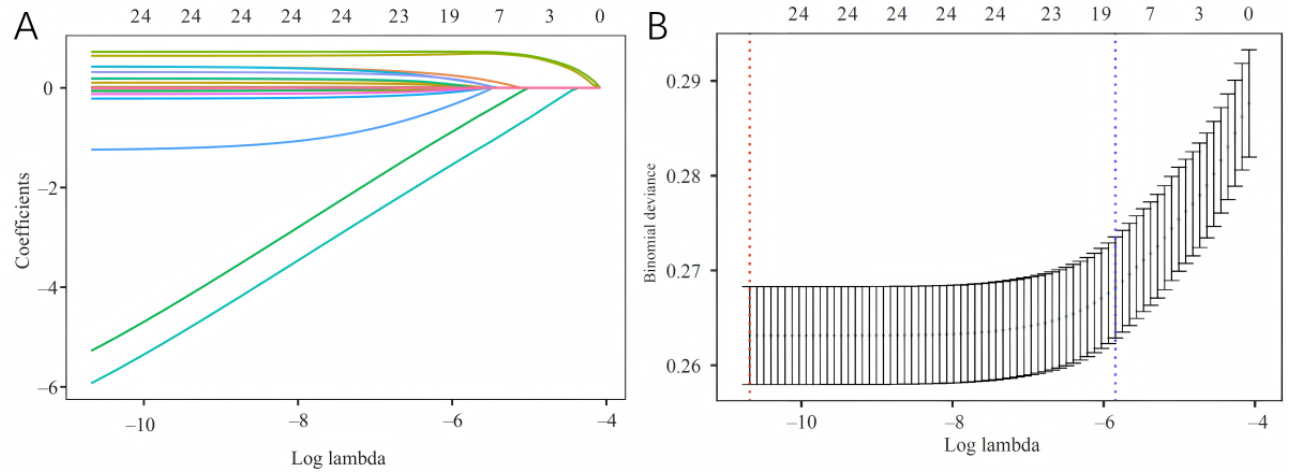
Variables	Total (N=17,637)	Non-VTE[a] (n=17,061)	VTE (n=576)	*P* value
pH[q], mean (SD)	7.36 (0.10)	7.36 (0.10)	7.36 (0.10)	.37
PaO_2[r], median (Q1-Q3)	140 (87-273)	141 (87-275)	129 (86-243)	.04
$PaCO_2$[s], mean (SD)	42.32 (12.07)	42.36 (12.04)	41.14 (13.06)	.03
BE[t], mean (SD)	–1.47 (5.26)	–1.47 (5.25)	–1.51 (5.58)	.87
Lac[u], median (Q1-Q3)	1.8 (1.2-2.7)	1.8 (1.2-2.7)	1.8 (1.2-2.7)	.63
HCO_3^-[v], mean (SD)	22.58 (4.76)	22.57 (4.75)	22.87 (4.82)	.14
Ventilation Variables, median (Q1-Q3)				
PEEP[w]	5 (5-5)	5 (5-5)	5 (5-5)	.045
FiO_2[x]	60 (50-100)	60 (50-100)	60 (50-100)	.13
PEEP[w]/FiO_2[x]	246 (152-353)	246 (152-353)	254 (149-357)	.84
Comorbidities, n (%)				
Hypertension	7333 (41.6)	7075 (41.5)	258 (44.8)	.11
Diabetes	2463 (14.0)	2463 (14.4)	0 (0.0)	<.001
Myocardial infarction	793 (4.5)	787 (4.6)	6 (1.0)	<.001
COPD[y]	1142 (6.5)	1142 (6.7)	0 (0.0)	<.001
Asthma	1381 (7.8)	1327 (7.8)	54 (9.4)	.16
Pancreatitis	449 (2.6)	421 (2.5)	28 (4.9)	<.001
ARF[z]	5965 (33.8)	5727 (33.6)	238 (41.3)	<.001
AKI[aa]	12163 (69.0)	11,712 (68.7)	451 (78.3)	<.001
ARDS[bb]	4117 (23.3)	3943 (23.1)	174 (30.2)	<.001
Malignant cancer	2790 (15.8)	2685 (15.7)	105 (18.2)	.11
Therapeutic interventions, n (%)				
Arterial catheterization	1265 (7.2)	1150 (6.7)	115 (20.0)	<.001
Central venous catheterization	1580 (9.0)	1443 (8.5)	137 (23.8)	<.001
CPB[cc]	2090 (11.9)	2038 (12.0)	52 (9.0)	.03
CRRT[dd]	1605 (9.1)	1533 (9.0)	72 (12.5)	.004
Ventilation	8718 (49.4)	8442 (49.5)	276 (47.9)	.46
Heparin	15479 (87.8)	14943 (87.6)	536 (93.1)	<.001
Aspirin	9026 (51.2)	8771 (51.4)	255 (44.3)	<.001

XSL•FO
RenderX

Variables	Total (N=17,637)	Non-VTE[a] (n=17,061)	VTE (n=576)	*P* value
Vasopressors	5492 (31.1)	5294 (31.0)	198 (34.4)	.09

[a]VTE: venous thromboembolism.
[b]HR: heart rate.
[c]RR: respiratory rate.
[d]SBP: systolic blood pressure.
[e]DBP: diastolic blood pressure.
[f]SpO_2: oxygen saturation.
[g]GCS: Glasgow Coma Scale.
[h]SOFA: Sequential Organ Failure Assessment.
[i]SAPS II: simplified acute physiology score II.
[j]WBC: white blood cell count.
[k]PLT: platelet.
[l]ALT: alanine aminotransferase.
[m]BUN: blood urea nitrogen.
[n]INR: international normalized ratio.
[o]PT: prothrombin time.
[p]PTT: partial thromboplastin time.
[q]pH: potential of hydrogen.
[r]PaO_2: partial pressure of oxygen.
[s]$PaCO_2$: partial pressure of carbon dioxide.
[t]BE: base excess.
[u]Lac: lactic acid.
[v]HCO_3^-: bicarbonate.
[w]PEEP: positive end-expiratory pressure.
[x]FiO_2: fraction of inspired oxygen.
[y]COPD: chronic obstructive pulmonary disease.
[z]ARF: acute respiratory failure.
[aa]AKI: acute kidney injury.
[bb]ARDS: acute respiratory distress syndrome.
[cc]CPB: cardiopulmonary bypass.
[dd]CRRT: continuous renal replacement therapy.

Figure 2. Screening predictors of venous thromboembolism using least absolute shrinkage and selection operator regression. (A) Coefficient profile plotted against the logarithm of the lambda sequence. Each colored line represents the coefficient path of an individual candidate predictor. (B) Cross-validation plot for determining the optimal penalty term. The green dots represent the mean binomial deviance with error bars. The left red vertical dashed line indicates the log lambda value that minimizes the deviance (log lambda minimum = –10.683), and the right blue vertical dashed line indicates the log lambda value within 1 standard error of the minimum (log lambda 1se = –5.845).



XSL•FO
RenderX

Table 3. Multivariate logistic regression analysis of the selected clinical features in the training set.

Variables	β	*Z* value	Odds ratio	95% CI	*P* value
Admission age (years)	–.0018	–0.6384	0.9982	–0.0072 to 0.0037	.52
Insurance	–.2206	–1.9143	0.8020	–0.4436 to 0.0084	.06
HR[a]	.0035	1.6880	1.0035	–0.0006 to 0.0076	.09
Potassium	–.1380	–2.3558	0.8711	–0.2541 to –0.0244	.02
Chloride	.0204	3.2561	1.0206	0.0081 to 0.0326	.001
PTT[b]	.0085	5.5201	1.0086	0.0054 to 0.0115	<.001
Diabetes	–15.8672	–0.0766	0.0000	–162.9076 to –157.5869	.94
Myocardial infarction	–1.2549	–2.9781	0.2851	–2.1976 to –0.5169	.003
COPD[c]	–15.5638	–0.0530	0.0000	–222.8419 to –215.7841	.96
Pancreatitis	.3493	1.6805	1.4181	–0.0786 to 0.7391	.09
ARF[d]	.1127	1.2010	1.1192	–0.0721 to 0.2957	.23
AKI[e]	.4236	3.9539	1.5274	0.2166 to 0.6369	<.001
ARDS[f]	.1806	1.8616	1.1979	–0.0116 to 0.3689	.06
Arterial catheterization	.6411	5.5322	1.8985	0.4102 to 0.8648	<.001
Central venous catheterization	.7326	6.8935	2.0805	0.5213 to 0.9382	<.001
CRRT[g]	.2571	1.8355	1.2931	–0.0240 to 0.5257	.07
Heparin	.4388	2.5944	1.5509	0.1203 to 0.7851	.01
Aspirin	–.0882	–0.9590	0.9156	–0.2687 to 0.0917	.34

[a]HR: heart rate.

[b]PTT: partial thromboplastin time.

[c]COPD: chronic obstructive pulmonary disease.

[d]ARF: acute respiratory failure.

[e]AKI: acute kidney injury.

[f]ARDS: acute respiratory distress syndrome.

[g]CRRT: continuous renal replacement therapy.

Development and Validation of Models

All selected variables were incorporated into the development of 9 ML models to predict the risk of VTE in ICU patients with sepsis. The 9 ML models exhibited varying degrees of predictive performance in the training set (Figure 3A). Internal validation based on receiver operating characteristic curve analysis demonstrated that the LGBM model achieved the highest predictive performance (AUC=0.956), followed by CatBoost (AUC=0.924), RF (AUC=0.794), decision tree (AUC=0.704), k-nearest neighbors (AUC=0.673), multilayer perceptron (AUC=0.660), support vector machine (AUC=0.621), LR (AUC=0.621), and naïve Bayes (AUC=0.588; Figure 3B). Detailed performance metrics, including sensitivity, specificity, accuracy, precision, F_1-score, and AUC for each model, are shown in Table 4. Notably, the LGBM model exhibited the highest specificity (0.942), accuracy (0.910), precision (0.956), and F_1-score (0.921), consistently outperforming the other algorithms across these key evaluation indicators.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic (ROC) curves of the test and validation sets of 9 machine learning models. (A) ROC curves of the training set. (B) ROC curves of the internal validation set. (C) ROC curves of the external validation set. AUC: area under the curve; CatBoost: categorical boosting; DT: decision tree; KNN: k-nearest neighbor; LGBM: light gradient boosting machine; LR: logistic regression; MLP: multilayer perceptron; NB: naive Bayes; RF: random forest; SVM: support vector machine.

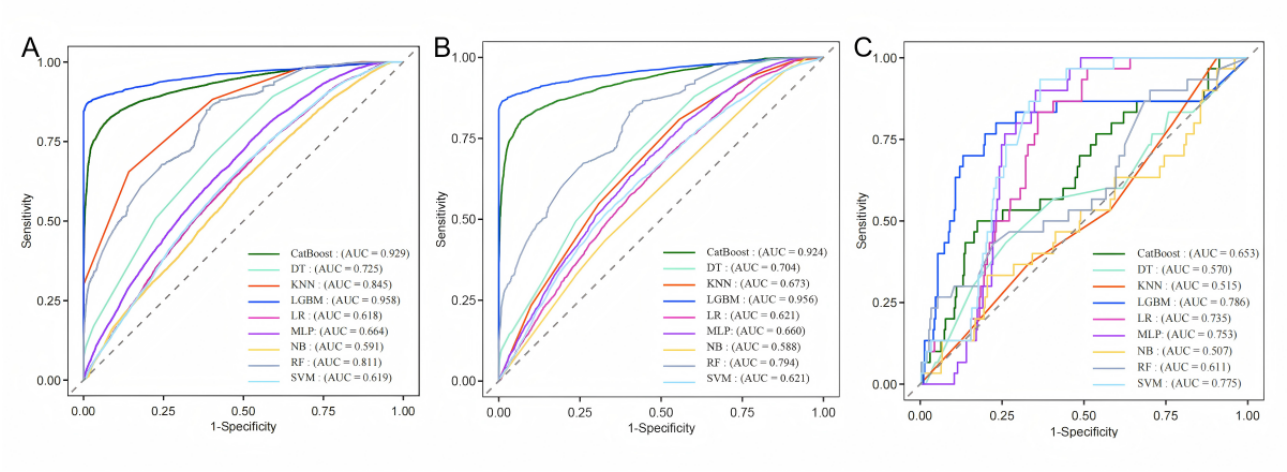


Table 4. Comprehensive evaluation of machine learning model performance.

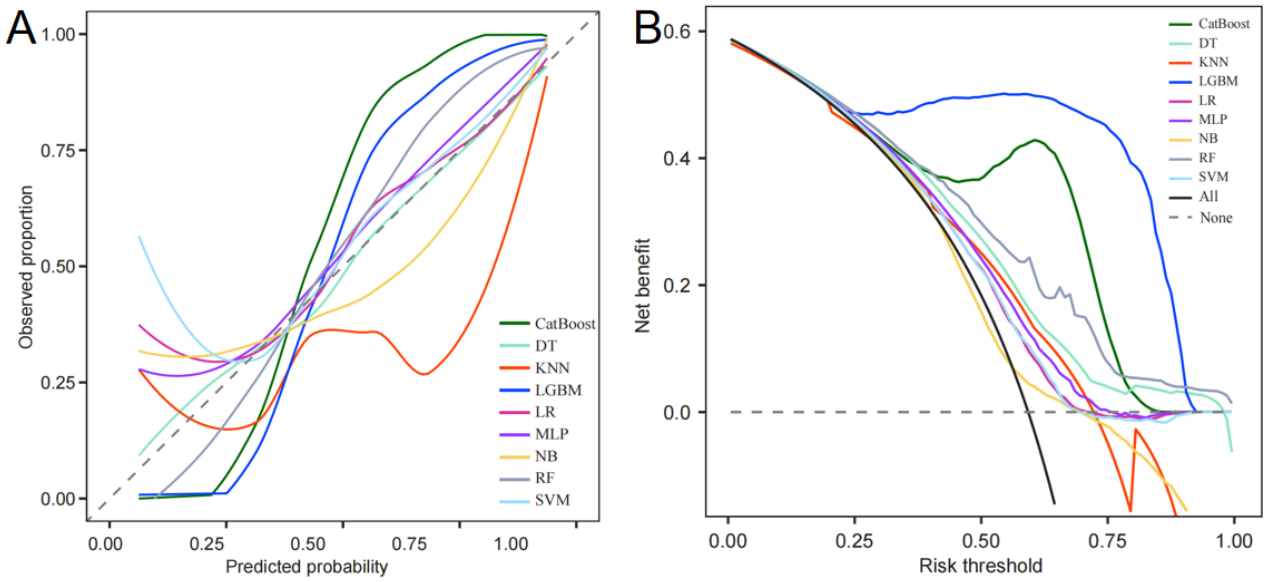
Methods	Sensitivity	Specificity	Accuracy	Precision	F_1	AUC[a]
CatBoost[b]	0.926	0.598	0.791	0.766	0.839	0.924
DT[c]	0.881	0.394	0.680	0.675	0.764	0.704
KNN[d]	0.809	0.444	0.658	0.674	0.735	0.673
LGBM[e]	0.888	0.942	0.910	0.956	0.921	0.956
LR[f]	0.938	0.199	0.633	0.625	0.750	0.621
MLP[g]	0.803	0.504	0.680	0.698	0.747	0.660
NB[h]	0.308	0.769	0.498	0.655	0.419	0.588
RF[i]	0.893	0.438	0.705	0.694	0.781	0.794
SVM[j]	0.991	0.027	0.594	0.592	0.741	0.621

[a]AUC: area under the curve.
[b]CatBoost: categorical boosting.
[c]DT: decision tree.
[d]KNN: k-nearest neighbor.
[e]LGBM: light gradient boosting machine.
[f]LR: logistic regression.
[g]MLP: multilayer perceptron.
[h]NB: naive Bayes.
[i]RF: random forest.
[j]SVM: support vector machine.

Calibration analysis revealed that the LGBM model demonstrated strong alignment between predicted probabilities and observed outcomes, reflecting robust reliability (Figure 4A). In the decision curve analysis, the LGBM model consistently provided the highest net benefit across a broad range of clinically relevant threshold probabilities, further supporting its clinical applicability in decision-making scenarios (Figure 4B).

XSL•FO
RenderX

Figure 4. Calibration curve of the LGBM model in the internal validation set (A). Decision curve analysis of the LGBM model in the internal validation set (B). CatBoost: categorical boosting; DT: decision tree; KNN: k-nearest neighbor; LGBM: light gradient boosting machine; LR: logistic regression; MLP: multilayer perceptron; NB: naive Bayes; RF: random forest; SVM: support vector machine.



External validation using an independent test cohort further confirmed the robustness and generalizability of the LGBM model, which achieved an AUC of 0.786, outperforming all other ML models in this cohort (Figure 3C). The consistent performance observed across both internal and external validation cohorts underscores the strong generalizability and clinical potential of the LGBM model for predicting VTE in ICU patients with sepsis.

Following stratification based on SOFA score, the LGBM model demonstrated enhanced discriminative ability in the high-risk group. The AUC for the severe sepsis group was 0.816, which was higher than the AUC of 0.769 observed in the mild to moderate sepsis group. The detailed performance metrics are shown in Table 5.

Table 5. Subgroup analysis of LGBM[a] model performance stratified by SOFA[b] score.

Subgroup	SOFA Criteria	Sensitivity	Specificity	AUC[c]
Mild to moderate sepsis	2 ≤ SOFA ≤ 6	0.722	0.839	0.769
Severe sepsis	SOFA > 6	0.667	0.856	0.816

[a]LGBM: light gradient boosting machine.

[b]SOFA: Sequential Organ Failure Assessment.

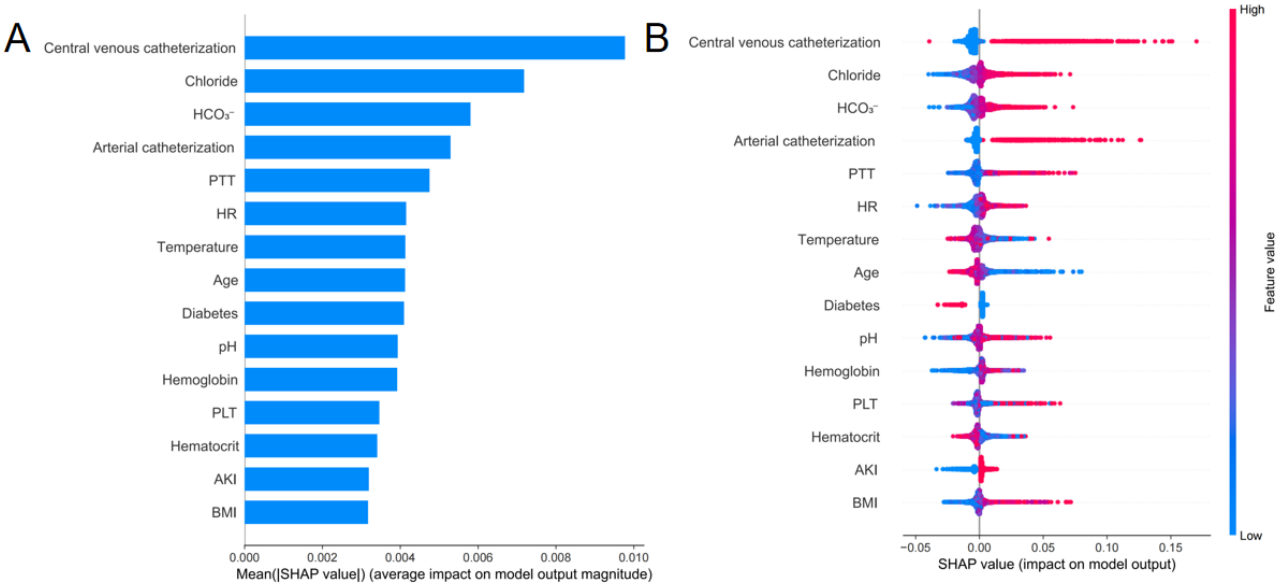
[c]AUC: area under the curve.

Explanation of the Model

Figure 5 displays the top 15 predictive features for VTE incidence as identified by the LGBM model using SHAP analysis. The feature importance plot, ranked by the mean absolute SHAP values, highlights the variables with the greatest overall impact on model predictions (Figure 5A). The 5 most influential features were central venous catheterization, chloride, bicarbonate, arterial catheterization, and PTT. The corresponding SHAP beeswarm plot (Figure 5B) illustrates each feature's influence on individual predictions, where SHAP values to the right of zero reflect increased predicted risk. It revealed that the presence of central venous or arterial catheters, elevated chloride and bicarbonate levels, and prolonged PTT consistently contributed to a higher VTE risk.

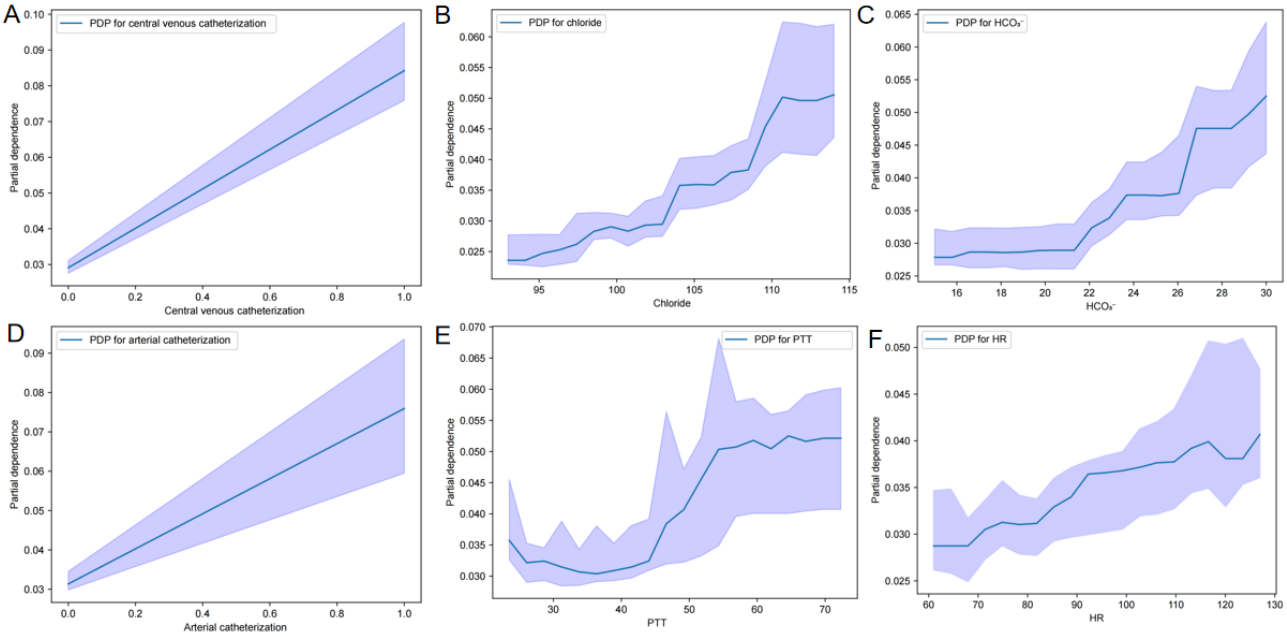
XSL•FO
RenderX

Figure 5. SHAP summary plot for clinical variables contributing to the light gradient boosting machine (LGBM) model. (A) Feature importance ranking plot based on the LGBM model. (B) Scatter plot of variables for SHAP analysis based on the LGBM model. AKI: acute kidney injury; HCO3-: bicarbonate; HR: heart rate; pH: potential of hydrogen; PLT: platelet; PTT: partial thromboplastin time; SHAP: SHapley Additive exPlanations.



To further examine the marginal effects of key predictors on model output, PDPs were generated for the top-ranked features (Figures 6A-6F). These plots depict how the predicted risk of VTE changes on average with varying values of a single feature, while holding others constant. For dichotomous variables such as central venous (Figure 6A) and arterial (Figure 6D) catheterization, the plots showed a marked increase in the average predicted risk when catheters were present compared with their absence. Continuous variables including serum chloride (Figure 6B), bicarbonate (Figure 6C), and HR (Figure 6F) exhibited monotonic relationships, with risk increasing steadily as the values rose. In contrast, PTT (Figure 6E) exhibited a distinct threshold effect. The predicted risk remained low and stable within the normal range and then increased sharply for values beyond approximately 45 seconds before plateauing at highly prolonged levels, confirming its nonlinear risk association identified by SHAP.

Figure 6. SHapley Additive exPlanations feature PDP of the light gradient boosting machine model. (A) Partial dependence plot of central venous catheter. (B) Partial dependence plot of chloride. (C) Partial dependence plot of bicarbonate. (D) Partial dependence plot of arterial catheter. (E) Partial dependence plot of partial thromboplastin time. (F) Partial dependence plot of heart rate. HCO3-: bicarbonate; HR: heart rate; PDP: partial dependence plot; PTT: partial thromboplastin time.

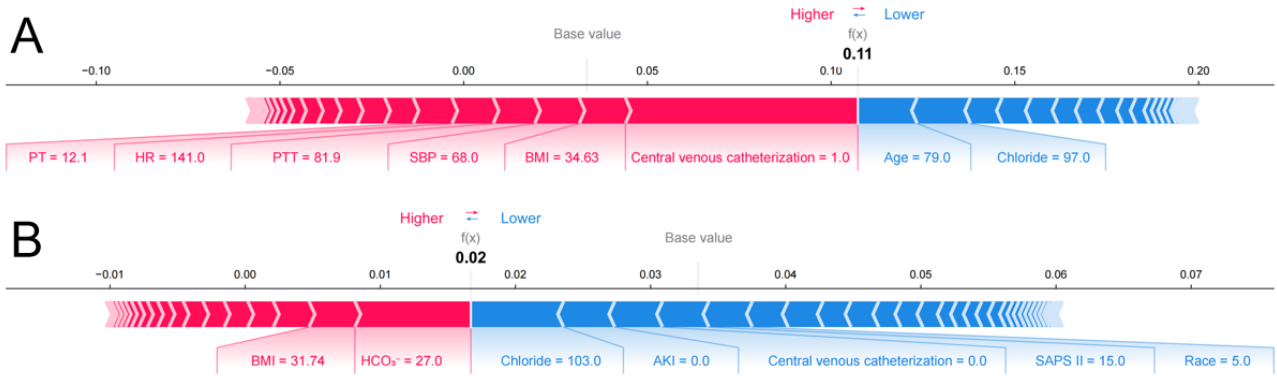


XSL•FO
RenderX

Interpretation of Individual Prediction

To enhance case-level interpretability, SHAP force plots were used to visualize the contribution of individual features to specific predictions (Figure 7). In these plots, each feature was represented as a colored bar that moves the predicted value away from a baseline, which corresponds to the average prediction across the dataset. Features that increase the predicted risk are shown in red, while those that reduce the risk are shown in blue. The length of each bar reflects the extent of the feature's influence. A longer bar indicates a stronger impact on the final predicted probability.

Figure 7. SHapley Additive exPlanations (SHAP) force plots for explaining of individual's prediction results according to the light gradient boosting machine model. (A) SHAP force plot of a high-risk example. (B) SHAP force plot of a low-risk example. AKI: acute kidney injury; HCO3-: bicarbonate; HR: heart rate; PT: prothrombin time; PTT: partial thromboplastin time; SAPS: simplified acute physiology score II; SBP: systolic blood pressure.



For example, in Figure 7A, the predicted VTE probability for a specific patient was 0.11, which was higher than the baseline value. This elevated risk was primarily attributed to the presence of a central venous catheter, a markedly prolonged PTT of 81.9 seconds, and a high HR of 141 min^{-1}. In contrast, Figure 7B showed a patient with a predicted probability of 0.02, which was lower than the baseline. In this case, protective features including the absence of a central venous catheter, absence of AKI, and a low SAPS II score collectively outweighed the influence of risk factors such as elevated BMI and bicarbonate level. These examples demonstrated that the model was capable of producing clear and clinically relevant explanations for individual predictions.

Discussion

Principal Findings

In this multicenter retrospective study, we developed and validated several ML models to predict the risk of VTE in critically ill patients with sepsis. Among the models tested, the LGBM algorithm demonstrated the highest predictive accuracy with an AUC of 0.956 in the internal validation set and maintained strong generalizability with an AUC of 0.786 in an external cohort. The observed VTE incidence in our external validation cohort (30/328, 9.15%) was substantially higher than the 3.35% (844/25,197) incidence recorded in the MIMIC-IV development cohort. While this difference contributed to the drop in the AUC during external validation, this discrepancy is likely a consequence of multiple factors, including divergent diagnostic practices, disparities in clinical risk profiles, and racial composition differences between the 2 cohorts.

First, variations in VTE surveillance and diagnostic criteria likely served as a primary driver. The MIMIC-IV cohort relied on retrospective identification via *ICD* (*International Classification of Diseases*) codes, a method known to carry a risk of misclassification bias and potential underascertainment, particularly of asymptomatic events. Conversely, the external validation cohort used explicit diagnostic imaging reports and implemented a routine lower limb venous ultrasound screening program for high-risk patients with sepsis. This systematic, active surveillance methodology inherently tends to yield a higher VTE capture rate compared with passive, coding-based case ascertainment. Second, the external validation group exhibited a substantially higher baseline severity and utilization of aggressive life support interventions. Patients with VTE in this cohort showed markedly higher rates of established risk factors such as mechanical ventilation (83.33% vs 49.29% in MIMIC-IV VTE group), ARDS (53.33% vs 30.81%), and CRRT (33.33% vs 11.37%) [19-21]. This heightened clinical severity and iatrogenic risk profile provide a sound clinical explanation for the elevated VTE incidence. Third, the inherent demographic differences between the cohorts represent a major challenge to model generalizability. The MIMIC-IV cohort is composed of a racially diverse population, while the external validation cohort is predominantly Asian, representing a geographically and ethnically distinct cohort. Given that VTE risk is known to vary significantly across different racial groups, this change in demographic distribution may be a contributing factor to the observed performance decrease [22].

Despite these significant differences in cohort characteristics and VTE incidence, the LGBM model maintained the highest predictive performance in both internal and external validation, underscoring its inherent robustness and generalization capacity. These performance advantages align with recent findings that gradient boosting methods often outperform logistic models in predictive accuracy when applied to large clinical datasets [23,24]. Crucially, we applied SHAP to interpret the model's predictions. This analysis identified a set of key risk factors driving VTE predictions, including the presence of central and arterial catheters, serum chloride and bicarbonate levels, and the PTT. By quantifying each feature's contribution to individual risk scores, SHAP facilitated insight into why certain patients

XSL•FO
RenderX

were classified as high-risk. This combination of robust performance and interpretability suggests that our model may assist clinicians in identifying ICU patients with sepsis at an increased risk of VTE, thereby supporting timely prophylaxis or diagnostic evaluation. Conventional instruments such as the Caprini, Padua, or Wells scores were primarily developed for general medical or surgical populations and relied on a linear, additive combination of risk factors [10,25]. These tools lack the specificity required to capture the distinct pathophysiology of sepsis which involves a complex interplay among inflammation, endothelial dysfunction, and coagulopathy collectively referred to as sepsis-induced coagulopathy [26,27]. The pathophysiology of sepsis-induced VTE is inherently nonlinear. Complex interactions among organ failure, systemic inflammation, and iatrogenic factors such as the use of catheters and vasopressors jointly modulate VTE risk in a manner that traditional linear models cannot fully capture [28,29]. In contrast, our LGBM model is consistent with previous ML applications in critical care and automatically captures complex relationships and interactions among predictors while integrating heterogeneous ICU data to generate a more accurate and nuanced risk estimate [30]. Our study found that the LGBM model significantly outperformed the LR model, achieving an AUC value of 0.956 compared with 0.621, which underscores the advantage of advanced ML algorithms in modeling the complex relationships among multiple risk factors. We recognize that prior studies have used advanced ML models for VTE prediction in critically ill patients. For instance, Jin et al [31] developed an RF model for VTE prediction in ICU patients that achieved an AUC of 0.788 in their validation dataset. Our model demonstrates equivalent predictive power in a challenging, sepsis-specific cohort. This robust performance likely stems from 2 key strengths of our approach: first, the inherent efficiency and predictive power of the LGBM algorithm compared with other boosting techniques; second, and more critically, our model's unique inclusion of metabolic and acid-base markers, such as serum chloride and bicarbonate, may offer a more nuanced reflection of the prothrombotic state induced by sepsis-associated organ dysfunction, thereby enhancing predictive accuracy.

The comparative analysis between the LR findings and the SHAP feature ranking reveals the core limitation of traditional statistical approaches in septic VTE prediction. The discrepancy observed for bicarbonate is illustrative. While bicarbonate was not retained in the final LR model, suggesting that its linear signal lacked statistical significance for VTE risk, the LGBM model ranked it as the third most influential predictor overall. This strongly suggests that bicarbonate's predictive power does not stem from a simple linear effect but from its complex nonlinear relationship or crucial interaction with other factors, reflecting the severity of systemic acid-base disturbances. Similarly, MI appeared as a protective factor in the final multivariable regression (OR 0.2851), but its marginal contribution was deemed less vital by SHAP than core metabolic and iatrogenic factors. This protective association in the linear model was likely influenced by confounding or collinearity among chronic conditions, potentially because patients with a history of MI often receive stricter chronic anticoagulation management, which may suppress their VTE risk and mask the true underlying risk in the linear model. The issue of complete separation for variables such as diabetes and COPD further highlights the inherent defect of LR when processing sparse data, as it led to unreliable, extreme coefficients. These findings collectively validate the choice of the nonlinear LGBM model and underscore the necessity of relying on its transparent SHAP feature analysis to evaluate the true, complex VTE risk drivers that are often obscured by the restrictive linearity of traditional statistical methods. Furthermore, the finding that the LGBM model achieved a higher AUC (0.816 vs 0.769) in the severe sepsis subgroup is significant. This enhanced discriminative power in the highest risk population confirms the model's clinical use, suggesting that it is particularly effective for targeted VTE prophylaxis and early intervention in patients with multiorgan dysfunction.

The interpretability enabled by SHAP analysis is essential for clinical translation and offers valuable pathophysiological insights. The top predictors identified by our SHAP analysis are closely related to the pathophysiology of sepsis-associated VTE and reinforce the model's clinical plausibility. The paramount importance of central venous catheterization aligns with Virchow's triad as the catheter can directly cause endothelial injury and lead to local blood flow stasis, both of which are potent triggers for thrombosis [32]. This finding is consistent with numerous studies highlighting indwelling catheters as a major iatrogenic risk factor for VTE in ICU patients [33,34]. The contribution of arterial catheterization, although less pronounced, further underscores the role of endothelial disruption from invasive devices.

Beyond these established risk factors, the model identified the predictive value of metabolic parameters such as serum chloride and bicarbonate. Elevated serum chloride is frequently associated with metabolic acidosis, which commonly develops in severe sepsis and shock [35]. Acidosis has been shown to enhance PLT aggregation and impair fibrinolysis, thereby promoting a prothrombotic state [36]. The association with bicarbonate likely reflects the severity of systemic acid-base disturbances commonly observed in septic shock. These findings suggest that fluid resuscitation strategies and acid-base management may represent modifiable components of VTE risk. The role of PTT is particularly noteworthy. Although a prolonged PTT is traditionally associated with bleeding risk, in sepsis it may paradoxically indicate consumptive coagulopathy in which widespread microthrombosis rapidly consumes clotting factors and leads to their depletion, a hallmark of disseminated intravascular coagulation [37]. The PDP revealed a nonlinear risk profile, where VTE risk increased markedly once PTT exceeded approximately 45 seconds. This nonlinear relationship would likely be overlooked in traditional linear analyses and highlights the model's capacity to identify clinically meaningful physiological thresholds. The contribution of HR appears more direct as its PDP showed a monotonic increase in predicted VTE risk with rising HR. Tachycardia is a cardinal manifestation of the systemic inflammatory response and catecholamine surge in sepsis, reflecting physiological stress and a hyperdynamic state that may promote endothelial dysfunction and a prothrombotic environment [38,39].

XSL•FO
RenderX

Strengths and Limitations

A major strength of our study lies in the integration of rigorous model development with interpretability, which enhances its potential clinical applicability. The model was developed using the large, high-quality public MIMIC-IV database and externally validated it in a geographically and ethnically distinct cohort from our own center. This 2-tiered validation strategy is essential for evaluating model generalizability, which remains a common limitation in single-center ML studies [40]. The decline in AUC from 0.956 to 0.786 during external validation, while still maintaining the best performance among all models, underscores the model's robust generalizability. This performance gap likely reflects intrinsic differences between the 2 cohorts, such as patient genetics, local clinical practices, VTE screening protocols, and data-recording standards. These findings underscore the critical need for local calibration or model retraining before deploying any predictive model in a new clinical setting. In addition, our use of SHAP enhances model transparency and interpretability rather than producing a purely opaque prediction tool. By generating individualized explanations through force plots, our model provides clinicians with transparent and actionable insights that support clinical decision-making and facilitate integration into routine workflows. Such interpretability is essential for translating ML tools from research settings to real-world clinical practice.

Despite these strengths, our study has several limitations. First, its retrospective design allows for the identification of associations but not the establishment of causality. The reliance on *ICD* codes for VTE diagnosis in the MIMIC-IV database may lead to some misclassification, as coding practices can vary and may not capture all clinically recognized events [41]. Second, we included data from only the first 24 hours of ICU admission. While this time frame is critical for early risk stratification, VTE is a dynamic process, and models incorporating time series data may provide greater predictive accuracy [42]. Third, despite comprehensive data extraction, some potential confounders, such as the specific type and dosage of thromboprophylaxis, patient mobility status, or the specific subtypes of malignancy, were not available. This includes the lack of detailed or consistently documented data on heparin dosage, such as prophylactic versus therapeutic, or duration, which forced us to treat heparin use as a binary variable. Finally, although the use of imputation for missing data is standard practice, it may still introduce a small degree of bias.

Future research should aim to prospectively validate our LGBM model in multicenter interventional trials to confirm its clinical use and impact on patient outcomes. Further refinement of the model may involve incorporating dynamic variables throughout the ICU stay and integrating emerging biomarkers related to coagulopathy and endothelial dysfunction. In addition, exploring federated learning approaches could enable the development of more robust and generalizable models across institutions while preserving patient data privacy [43].

Conclusions

This study presents a high-performing and interpretable LGBM model for predicting VTE in ICU patients with sepsis by effectively integrating a broad spectrum of clinical and laboratory data. By leveraging SHAP to enhance transparency, the model extends beyond a simple predictive tool to function as a decision support system capable of elucidating complex risk profiles and enabling clinicians to implement more personalized and effective VTE prevention strategies. This work lays the foundation for the adoption of sophisticated yet interpretable artificial intelligence–driven tools aimed at reducing the burden of VTE in this high-risk patient population.

Acknowledgments

This work was supported by the Suzhou Special Project for Clinical Key Disease Diagnosis and Treatment Technologies (LCZX202334), Key Projects of the Changshu Science and Technology Development Program (CSWS202209), and the Special Research Fund Projects of the China International Medical Foundation (Z-2014-08-2309-1).

Data Availability

The datasets and code used in this study are available from the corresponding author on reasonable request.

Authors' Contributions

YZ and XR designed the study and wrote the manuscript. YZ, XR, and LL collected, analyzed, and interpreted the data. JZ and YG contributed to literature search. XR, and HY critically reviewed and improved the drafts of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Pandey S. Nanotheranostics 2024;8(3):270-284 [FREE Full text] [doi: 10.7150/ntno.94071] [Medline: 38577320]
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study. Lancet 2020;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]

XSL•FO
RenderX

3. Kaplan D, Casper TC, Elliott CG, Men S, Pendleton RC, Kraiss LW, et al. VTE incidence and risk factors in patients with severe sepsis and septic shock. Chest 2015;148(5):1224-1230 [FREE Full text] [doi: 10.1378/chest.15-0287] [Medline: 26111103]
4. Hasegawa D, Sato R, Lee YI, Wang HY, Nishida K, Steiger D. The prevalence, risk factors, and outcomes of acute pulmonary embolism complicating sepsis and septic shock: a national inpatient sample analysis. Sci Rep 2024;14(1):16049 [FREE Full text] [doi: 10.1038/s41598-024-67105-7] [Medline: 38992133]
5. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001;29(7):1303-1310. [doi: 10.1097/00003246-200107000-00002] [Medline: 11445675]
6. Innocenti F, Lazzari C, Ricci F, Paolucci E, Agishev I, Pini R. D-dimer tests in the emergency department: current insights. Open Access Emerg Med 2021;13:465-479 [FREE Full text] [doi: 10.2147/OAEM.S238696] [Medline: 34795538]
7. Franchini M, Focosi D, Pezzo MP, Mannucci PM. How we manage a high D-dimer. Haematologica 2024;109(4):1035-1045 [FREE Full text] [doi: 10.3324/haematol.2023.283966] [Medline: 37881856]
8. Shen R, Gao M, Tao Y, Chen Q, Wu G, Guo X, et al. Prognostic nomogram for 30-day mortality of deep vein thrombosis patients in intensive care unit. BMC Cardiovasc Disord 2021;21(1):11 [FREE Full text] [doi: 10.1186/s12872-020-01823-4] [Medline: 33407152]
9. Chopard R, Albertsen IE, Piazza G. Diagnosis and treatment of lower extremity venous thromboembolism: a review. JAMA 2020;324(17):1765-1776. [doi: 10.1001/jama.2020.17272] [Medline: 33141212]
10. Ashrafi Z, Ameri M, Khosravi A, Mirzaei M, Shaker M, Ebrahimi H. Comparison of Autar scale and Wells criteria in DVT risk assessment by nurses in patients with lower extremity trauma. J Vasc Nurs 2022;40(3):148-152. [doi: 10.1016/j.jvn.2022.05.001] [Medline: 36414370]
11. Benfield M, He J, Arnall J, Kaizen W, Jandrisevits E, Eboli-Lopes K, et al. Venous thromboembolism incidence and risk factors in patients undergoing hematopoietic stem cell transplantation. Transplant Cell Ther 2025;31(2):111.e1-111.e10. [doi: 10.1016/j.jtct.2024.10.016] [Medline: 39505211]
12. Viarasilpa T, Panyavachiraporn N, Marashi SM, Van Harn M, Kowalski RG, Mayer SA. Prediction of symptomatic venous thromboembolism in critically ill patients: the ICU-venous thromboembolism score. Crit Care Med 2020;48(6):e470-e479. [doi: 10.1097/CCM.0000000000004306] [Medline: 32187076]
13. Yue S, Li S, Huang X, Liu J, Hou X, Zhao Y, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. J Transl Med 2022;20(1):215 [FREE Full text] [doi: 10.1186/s12967-022-03364-0] [Medline: 35562803]
14. Fan Z, Jiang J, Xiao C, Chen Y, Xia Q, Wang J, et al. Construction and validation of prognostic models in critically ill patients with sepsis-associated acute kidney injury: interpretable machine learning approach. J Transl Med 2023;21(1):406 [FREE Full text] [doi: 10.1186/s12967-023-04205-4] [Medline: 37349774]
15. Chen Y, Jiang H, Wei Y, Qiu Y, Su L, Chen J, China National Critical Care Quality Control Center Group (China-NCCQC group). Blood pressure response index and clinical outcomes in patients with septic shock: a multicenter cohort study. EBioMedicine 2024;106:105257 [FREE Full text] [doi: 10.1016/j.ebiom.2024.105257] [Medline: 39059317]
16. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025;79:103470 [FREE Full text] [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
17. Liu G, Zheng S, He J, Zhang Z, Wu R, Yu Y, et al. An easy and quick risk-stratified early forewarning model for septic shock in the intensive care unit: development, validation, and interpretation study. J Med Internet Res 2025;27:e58779 [FREE Full text] [doi: 10.2196/58779] [Medline: 39913913]
18. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
19. Havlicek EE, Palumbo J, Soto-Campos G, Goldenberg NA, Sochet AA. Invasive mechanical ventilation and risk of hospital-acquired venous thromboembolism. Respir Care 2024;69(11):1392-1399. [doi: 10.4187/respcare.11698] [Medline: 38889926]
20. Theochari CA, Theochari NA, Mylonas KS, Papaconstantinou D, Giannakodimos I, Spartalis E, et al. Venous thromboembolism following major abdominal surgery for cancer: a guide for the surgical intern. Curr Pharm Des 2022;28(10):787-797. [doi: 10.2174/1381612828666220217140639] [Medline: 35176975]
21. Chen X, Huang J, Liu J, Deng H, Pan L. Venous thromboembolism risk factors and prophylaxis of elderly intensive care unit patients in a Chinese general hospital. Ann Palliat Med 2021;10(4):4453-4462 [FREE Full text] [doi: 10.21037/apm-21-464] [Medline: 33894730]
22. Chen YW, Fong ZV, Qadan M, Kunitake H, Mullen JT, Chang DC. Should all patients receive the same prophylaxis? Racial variation in the risk of venous thromboembolism after major abdominal operations. Am J Surg 2021;222(5):884-889. [doi: 10.1016/j.amjsurg.2021.05.020] [Medline: 34144805]
23. Chen R, Petrazzini BO, Malick WA, Rosenson RS, Do R. Prediction of venous thromboembolism in diverse populations using machine learning and structured electronic health records. Arterioscler Thromb Vasc Biol 2024;44(2):491-504 [FREE Full text] [doi: 10.1161/ATVBAHA.123.320331] [Medline: 38095106]

XSL•FO
RenderX

24. Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep 2022;12(1):15889 [FREE Full text] [doi: 10.1038/s41598-022-20149-z] [Medline: 36220875]
25. Hayssen H, Sahoo S, Nguyen P, Mayorga-Carlin M, Siddiqui T, Englum B, et al. Ability of Caprini and Padua risk-assessment models to predict venous thromboembolism in a nationwide study. medRxiv 2023 [FREE Full text] [doi: 10.1101/2023.03.20.23287506] [Medline: 36993603]
26. Maneta E, Aivalioti E, Tual-Chalot S, Emini Veseli B, Gatsiou A, Stamatelopoulos K, et al. Endothelial dysfunction and immunothrombosis in sepsis. Front Immunol 2023;14:1144229 [FREE Full text] [doi: 10.3389/fimmu.2023.1144229] [Medline: 37081895]
27. Iba T, Levy JH, Warkentin TE, Thachil J, van der Poll T, Levi M, Scientific Standardization Committee on DIC ,the Scientific Standardization Committee on Perioperative Critical Care of the International Society on Thrombosis Haemostasis. Diagnosis and management of sepsis-induced coagulopathy and disseminated intravascular coagulation. J Thromb Haemost 2019;17(11):1989-1994 [FREE Full text] [doi: 10.1111/jth.14578] [Medline: 31410983]
28. Su J, Tie X, Zhou R, Zou T, Hong M, Yang L, et al. Risk factors and a nomogram model for deep vein thrombosis in critically ill patients with sepsis: a retrospective analysis. Sci Rep 2025;15(1):16641 [FREE Full text] [doi: 10.1038/s41598-025-01660-5] [Medline: 40360645]
29. van der Poll T, van de Veerdonk FL, Scicluna BP, Netea MG. The immunopathology of sepsis and potential therapeutic targets. Nat Rev Immunol 2017;17(7):407-420. [doi: 10.1038/nri.2017.36] [Medline: 28436424]
30. Li M, Han S, Liang F, Hu C, Zhang B, Hou Q, et al. Machine learning for predicting risk and prognosis of acute kidney disease in critically ill elderly patients during hospitalization: internet-based and interpretable model study. J Med Internet Res 2024;26:e51354 [FREE Full text] [doi: 10.2196/51354] [Medline: 38691403]
31. Jin J, Lu J, Su X, Xiong Y, Ma S, Kong Y, et al. Development and validation of an ICU-venous thromboembolism prediction model using machine learning approaches: a multicenter study. Int J Gen Med 2024;17:3279-3292 [FREE Full text] [doi: 10.2147/IJGM.S467374] [Medline: 39070227]
32. Parienti J, Mongardon N, Mégarbane B, Mira J, Kalfon P, Gros A, 3SITES Study Group. Intravascular complications of central venous catheterization by insertion site. N Engl J Med 2015;373(13):1220-1229. [doi: 10.1056/NEJMoa1500964] [Medline: 26398070]
33. Faustino EVS. Central venous catheter-associated deep venous thrombosis in critically ill children. Semin Thromb Hemost 2018;44(1):52-56. [doi: 10.1055/s-0037-1603938] [Medline: 28778103]
34. Krishnamoorthy A, Hansdak SG, Peter JV, Pichamuthu K, Rajan SJ, Sudarsan TI, et al. Incidence and risk factors for deep venous thrombosis and its impact on outcome in patients admitted to medical critical care. Indian J Crit Care Med 2024;28(6):607-613. [doi: 10.5005/jp-journals-10071-24723] [Medline: 39130389]
35. Filis C, Vasileiadis I, Koutsoukou A. Hyperchloraemia in sepsis. Ann Intensive Care 2018;8(1):43 [FREE Full text] [doi: 10.1186/s13613-018-0388-4] [Medline: 29589205]
36. de Souza JR, Yokoyama AP, Magnus MM, Boin I, de Ataide EC, Munhoz DC, et al. Association of acidosis with coagulopathy and transfusion requirements in liver transplantation. J Thromb Thrombolysis 2022;53(4):887-897. [doi: 10.1007/s11239-021-02609-x] [Medline: 34800258]
37. Levi M, van der Poll T. Inflammation and coagulation. Crit Care Med 2010;38(2 Suppl):S26-S34. [doi: 10.1097/CCM.0b013e3181c98d21] [Medline: 20083910]
38. Vallejos A, Olivares P, Gatica S, Villegas V, Echeverria C, Cabello-Verrugio C, et al. Endotoxemia-induced endothelial fibrosis inhibition improves hypotension, tachycardia, multiple organ dysfunction syndrome, cytokine response, oxidative stress, and survival. Lab Invest 2019;99(8):1173-1192 [FREE Full text] [doi: 10.1038/s41374-019-0237-6] [Medline: 30911151]
39. Sethi P, Peiris CD. A review of catecholamine associated cardiomyopathies and channelopathies. Cureus 2020;12(2):e6957 [FREE Full text] [doi: 10.7759/cureus.6957] [Medline: 32195067]
40. Dung-Hung C, Cong T, Zeyu J, Yu-Shan O, Yung-Yan L. External validation of a machine learning model to predict hemodynamic instability in intensive care unit. Crit Care 2022;26(1):215 [FREE Full text] [doi: 10.1186/s13054-022-04088-9] [Medline: 35836294]
41. Walraven CV. A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. Int J Epidemiol 2018;47(2):605-616. [doi: 10.1093/ije/dyx253] [Medline: 29253160]
42. Cong Y, Wang B, Fei C, Zhang H, Li Z, Zhu Y, et al. Dynamic observation and risk factors analysis of deep vein thrombosis after hip fracture. PLoS One 2024;19(6):e0304629 [FREE Full text] [doi: 10.1371/journal.pone.0304629] [Medline: 38829867]
43. Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. Patterns (N Y) 2024;5(7):100974 [FREE Full text] [doi: 10.1016/j.patter.2024.100974] [Medline: 39081567]

Abbreviations

AKI: acute kidney injury

XSL•FO
RenderX

ARDS: acute respiratory distress syndrome
ARF: acute respiratory failure
AUC: area under the curve
CatBoost: categorical boosting
COPD: chronic obstructive pulmonary disease
CRRT: continuous renal replacement therapy
FiO_2: fraction of inspired oxygen
HR: heart rate
ICD: International Classification of Diseases
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICU: intensive care unit
LGBM: light gradient boosting machine
LR: logistic regression
MI: myocardial infarction
MIMIC-IV: Medical Information Mart for Intensive Care IV
ML: machine learning
OR: odds ratio
$PaCO_2$: partial pressure of carbon dioxide
PaO_2: partial pressure of oxygen
PDP: partial dependence plot
PE: pulmonary embolism
PEEP: positive end-expiratory pressure
PLT: platelet
PT: prothrombin time
PTT: partial thromboplastin time
RF: random forest
RR: respiratory rate
SAPS II: simplified acute physiology score II
SHAP: Shapley Additive Explanations
SMOTE: Synthetic Minority Oversampling Technique
SOFA: Sequential Organ Failure Assessment
SpO_2: oxygen saturation
VTE: venous thromboembolism

Edited by A Benis; submitted 20.Jul.2025; peer-reviewed by G Wang, S Wan; comments to author 07.Oct.2025; accepted 29.Dec.2025; published 30.Jan.2026.

Please cite as:
Zhang Y, Ren X, Liu L, Zha J, Gu Y, Ye H
Machine Learning Algorithms to Predict Venous Thromboembolism in Patients With Sepsis in the Intensive Care Unit: Multicenter Retrospective Study
JMIR Med Inform 2026;14:e80969
URL: https://medinform.jmir.org/2026/1/e80969
doi:10.2196/80969
PMID:

©Yan Zhang, Xia Ren, Luojie Liu, Junjie Zha, Yijie Gu, Hongwei Ye. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of Large Language Models for Radiologists' Support in Multidisciplinary Breast Cancer Teams: Comparative Study

Hong Jiang[1,2*], MD, PhD; Chun Yang[3*], BM; Wenbin Zhou[4], MD, PhD; Cheng-liang Yin[5], MD, PhD; Shan Zhou[6,7], PhD; Rui He[8], PhD; Guanghui Ran[9], BM; Wujie Wang[9], BM; Meixian Wu[2], MS; Juan Yu[9], MD, PhD

[1]Faculty of Medicine, Macau University of Science and Technology, Macao, China

[2]Department of Statistics, Zhuhai Clinical Medical College of Jinan University, Zhuhai, China

[3]Qiandongnan Prefecture Hospital of Traditional Chinese Medicine, Kai Li, China

[4]Guangdong Provincial Key Laboratory of Tumor Interventional Diagnosis and Treatment, Zhuhai, China

[5]Department of Medical Innovation Research, Chinese PLA General Hospital, Beijing, China

[6]Cancer Virology Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

[7]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

[8]Grammar and Cognition Lab (GraC), Department of Translation & Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

[9]Department of Radiology, The First Affiliated Hospital of Shenzhen University, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Juan Yu, MD, PhD
Department of Radiology
The First Affiliated Hospital of Shenzhen University
Department of Radiology, Shenzhen Second People's Hospital, The First Affiliated of Shenzhen University, Sungang Road 3002
Shenzhen, 518003
China
Phone: 86 18576685849
Email: yujuan0072@qq.com

Abstract

Background: Artificial intelligence tools, particularly large language models (LLMs), have shown considerable potential across various domains. However, their performance in the diagnosis and treatment of breast cancer remains unknown.

Objective: This study aimed to evaluate the performance of LLMs in supporting radiologists within multidisciplinary breast cancer teams, with a focus on their roles in facilitating informed clinical decisions and enhancing patient care.

Methods: A set of 50 questions covering radiological and breast cancer guidelines was developed to assess breast cancer. These questions were posed to 9 popular LLMs and clinical physicians, with the expectation of receiving direct "Yes" or "No" answers along with supporting analysis. The performances of the 9 models, including ChatGPT-4.0, ChatGPT-4o, ChatGPT-4o mini, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Tongyi Qianwen 2.5, ChatGLM, and Ernie Bot 3.5, were evaluated against that of radiologists with varying experience levels (resident physicians, fellow physicians, and attending physicians). Responses were assessed for accuracy, confidence, and consistency based on alignment with the 2024 National Comprehensive Cancer Network Breast Cancer Guidelines and the 2013 American College of Radiology Breast Imaging-Reporting and Data System recommendations.

Results: Claude 3 Opus and ChatGPT-4 achieved the highest confidence scores of 2.78 and 2.74, respectively, while ChatGPT-4o led in accuracy with a score of 2.92. In terms of response consistency, Claude 3 Opus and Claude 3.5 Sonnet led the pack with scores of 3.0, closely followed by ChatGPT-4o, Gemini 1.5 Pro, and ChatGPT-4o mini, all recording impressive scores exceeding 2.9. ChatGPT-4o mini excelled in clinical diagnostics with a top score of 3.0 among all LLMs, and this score was also higher than all physician groups; however, no statistically significant differences were observed between it and any physician group (all $P>.05$). ChatGPT-4 also had a higher score than the physician groups but showed comparable statistical performance to them ($P>.05$). Across radiological diagnostics, clinical diagnosis, and overall performance, ChatGPT-4o mini and the Claude models achieved higher mean scores than all physician groups. However, these differences were statistically significant only when compared to fellow physicians ($P<.05$). However, ChatGLM and Ernie Bot 3.5 underperformed across diagnostic areas, with lower scores than all physician groups but no statistically significant differences (all $P>.05$). Among physician groups, attending

XSL•FO
RenderX

physicians and resident physicians exhibited comparable high scores in radiological diagnostic performance, whereas fellow physicians scored somewhat lower, though the difference was not statistically significant (P>.05).

Conclusions: LLMs such as ChatGPT-4o and Claude 3 Opus showed potential in supporting multidisciplinary teams for breast cancer diagnostics and therapy. However, they cannot fully replicate the intricate decision-making processes honed through clinical experience, particularly in complex cases. This highlights the need for ongoing artificial intelligence refinement to ensure robust clinical applicability.

(JMIR Med Inform 2026;14:e68182) doi:10.2196/68182

KEYWORDS

breast cancer; large language models; LLMs; radiology assistance; clinical decision-making; ACR BI-RADS; NCCN guidelines; radiologist; National Comprehensive Cancer Network; American College of Radiology Breast Imaging-Reporting and Data System

Introduction

Breast cancer is one of the most common malignancies in women worldwide, with over 2 million diagnoses each year, and it remains the leading cause of cancer-related deaths in women [1,2]. According to a World Health Organization departmental news report in 2021, breast cancer mortality and incidence rates are higher in low- and middle-income countries than in high-income nations, largely due to disparities in early detection and treatment. However, the biological characteristics of breast cancer vary significantly, including imaging features, pathological traits, and lymph node assessment, making early screening and personalized treatment challenging yet crucial.

A mature early screening system relies on well-established professional teams and standardized, specialized guidelines. Although there have been recent advancements, developing countries still face significant challenges in breast cancer care in terms of multidisciplinary coordination among specialist teams (surgeons, oncologists, pathologists, and radiologists), well-balanced medical resources, and adequately trained physicians [3,4]. Furthermore, many physicians rely on empiricism and practice medicine rather than evidence-based guidelines [5,6]. Therefore, there is an urgent need for effective artificial intelligence (AI) tools to support health care professionals in resource-limited settings, improve diagnostic accuracy, and enhance patient outcomes [7,8].

In recent years, several prominent large language models (LLMs) have emerged, including ChatGPT (OpenAI), Gemini Advanced (Google LLC), Claude (Anthropic), Tongyi Qianwen (Alibaba Corporation), ChatGLM (Zhipu AI), and Ernie Bot (Baidu). These models, trained on extensive datasets, offer significant potential in health care by generating complex text through deep learning. ChatGPT, for instance, gained rapid popularity, attracting over 1 million users within days of its release. (knowledge base updated as of October 2023). Studies have shown its potential in medicine, performing well on the United States Medical Licensing Examination and the Board of Radiology-style examination [9,10]. ChatGPT can help radiologists simplify cumbersome diagnostic imaging descriptions [11,12]. Notably, the recommendations provided by ChatGPT for the clinical management of early-diagnosed breast cancer closely align with decisions made by multidisciplinary teams (MDTs) [13]. ChatGPT has indicated a level of congruence with the National Comprehensive Cancer Network (NCCN) guidelines in identifying a wide range of therapeutic agents for the treatment of advanced metastases or advanced primary tumors [14]. However, recent studies highlighted significant limitations in ChatGPT-4.0, revealing a “curse reversal,” where a model trained on “A is B” may struggle to infer “B is A” [15].

The rapid advancement of AI has ushered in a new era of innovation, with the emergence and continuous refinement of various LLMs profoundly transforming the health care sector. This study evaluates the performance of various LLMs in addressing breast cancer guideline-related questions, comparing them with radiologists of varying expertise levels to assess their potential in enhancing radiologists’ diagnostic capabilities.

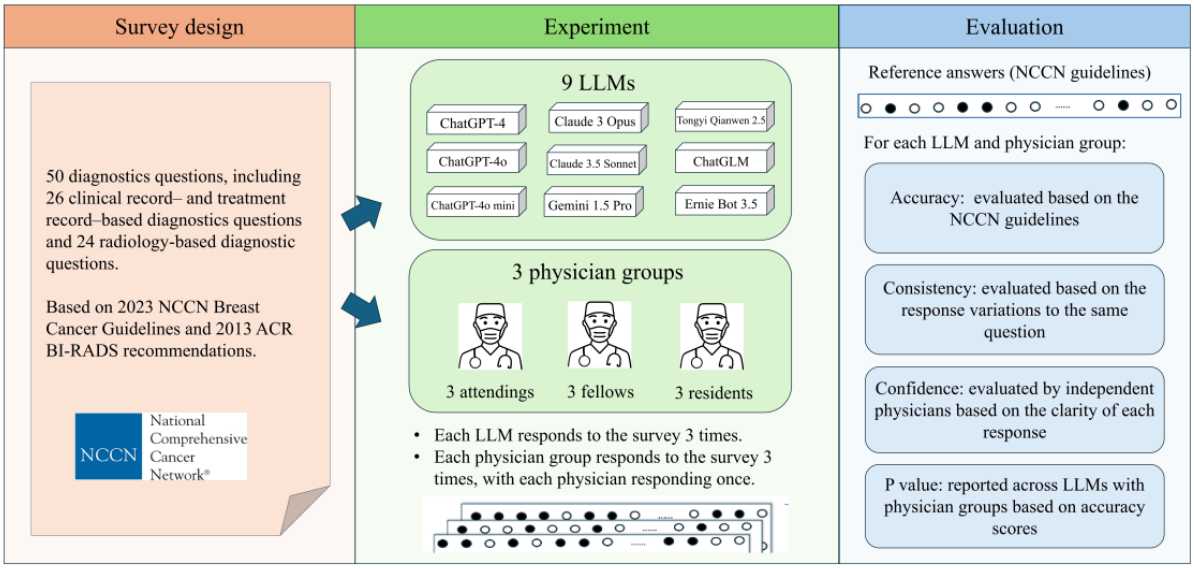
Methods

Study Design

The study aimed to evaluate the performance of 9 LLMs in breast cancer imaging diagnosis and to compare their performance with that of breast imaging specialists. We developed a set of 50 questions focused on breast cancer diagnosis and treatment based on the 2013 version of the American College of Radiology Breast Imaging-Reporting and Data System (ACR BI-RADS) classification and the 2024 NCCN Breast Cancer Guidelines. The first 24 questions focused on radiological diagnostics, while the remaining 26 questions pertained to clinical diagnosis and treatment. Questions were then presented to breast radiologists and the 9 LLMs for diagnostic responses. The accuracy, confidence, and consistency of each model in responding to the diagnostic queries were assessed. Furthermore, we conducted a comparative analysis of each model’s performance against radiologists possessing varying levels of clinical experience. The study design is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Graph in the full-text abstract: the study design, including question categorization, experiment, and evaluation workflow. ACR BI-RADS: American College of Radiology Breast Imaging-Reporting and Data System; LLM: Large Language Model; NCCN: National Comprehensive Cancer Network.



Participants and Models

The 9 LLMs included ChatGPT-4.0, ChatGPT-4o, ChatGPT-4o mini, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Tongyi Qianwen 2.5, ChatGLM, and Ernie Bot 3.5. In addition, the study recruited a total of nine board-certified radiologists, stratified into three distinct experience levels: (1) resident physicians (n=3; 1-5 years of postlicensure experience), undergoing subspecialty training to develop foundational clinical expertise; (2) fellow physicians (n=3; 5-10 years of experience), having completed residency and pursuing advanced fellowship training (1-3 years) to specialize in breast imaging diagnostics; and (3) attending physicians (n=3; >10 years of experience), senior specialists with independent diagnostic authority, responsible for clinical decision-making and trainee supervision. This hierarchical cohort enabled comparative benchmarking of LLM performance against progressive stages of radiologist expertise. Each physician received the same set of 50 questions.

Questions Design

The set of 50 questions was developed based on guidelines extracted from the 2024 NCCN Breast Cancer Guidelines and the 2013 ACR BI-RADS diagnostic criteria, which are widely regarded as the gold standards in clinical breast cancer diagnosis. These criteria are commonly adhered to by physicians in clinical breast cancer diagnosis. In this study, we used these same standards to evaluate the responses provided by the LLMs and physicians to ensure validity and fair comparison. The questions covered a range of common clinical diagnostic decision-making scenarios, providing radiological evidence, such as mammography, ultrasound, and magnetic resonance imaging, as well as clinical information indicators such as medical history, physical examination findings, and laboratory test results, across preoperative, intraoperative, and postoperative stages.

Procedure

Nine physicians were divided into 3 groups based on their experience levels, with each physician answering all 50 questions once. They were prohibited from consulting guidelines or reference materials and were required to complete the questions within 30 minutes. Each physician independently completed the assessment to minimize recall bias. The study aimed to simulate real-world conditions by controlling time and eliminating external aids.

All LLMs were queried via the Google application programming interface on a single date (August 21, 2024) without any additional context (such as task background or medical knowledge priming) and generated responses accordingly. Each LLM independently answered every question 3 times. Standard responses were defined as “yes” or “no.” A response was classified as “yes” if it demonstrated a strong affirmative inclination, supported by compelling theoretical evidence and logical reasoning. Conversely, a response was classified as “no” if it showed a clear negative inclination, backed by equally robust theoretical evidence and sound logic. If an LLM produced an ambiguous response, lacked sufficient evidence, or contained flawed reasoning, it was prompted to reanswer. If the reanswered response remained unsatisfactory, 2 senior physicians with over 10 years of experience were recruited to classify the response based on their professional judgment, thereby evaluating the AI model’s accuracy. In cases of disagreement, the evaluators reevaluated the response based on the 2024 NCCN Guidelines and the 2013 ACR BI-RADS criteria to reach a consensus through discussion.

All LLM-generated responses were anonymized, randomized, and then assessed by evaluators who were blinded to the source (LLM identity) of each response. The accuracy of each response was assessed against the diagnostic NCCN guidelines.

XSL•FO
RenderX

Responses consistent with the guidelines were assigned a score of 1, while inconsistent responses received a score of 0. The accuracy score for each question was calculated by summing the scores of the 3 responses. The overall accuracy performance of both the LLMs and physician groups was determined by averaging the accuracy scores across the 50 questions, resulting in a score ranging from 0 to 3. Subsequently, the accuracy performance of the LLMs was compared to that of the physician groups through rigorous statistical analysis.

Confidence and Consistency

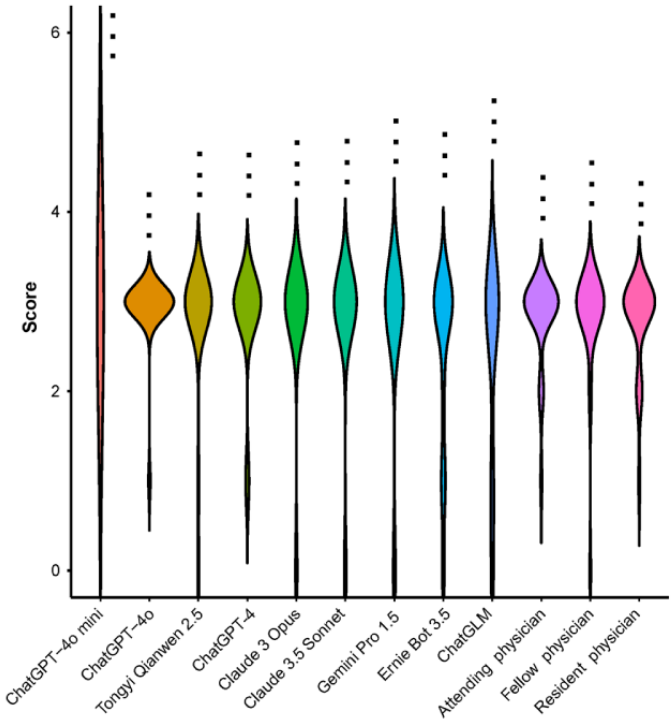
Confidence refers to the clarity of responses, categorized as either "confident" or "nonconfident." A confident answer provides an explicit "yes" or "no" to questions, while a nonconfident answer lacks a direct "yes or no" statement, despite potentially containing detailed explanations, requiring additional prompting to elicit a definitive response. Confident answers received are scored 1, while nonconfident answers are scored 0. The confidence score for each question is calculated by summing the scores from 3 independent response attempts. Each LLM's overall confidence performance is assessed by averaging these scores across 50 questions, resulting in a score ranging from 0 to 3. The confidence ratio represents each LLM's average confidence score divided by the maximum possible score (ie, 3).

Consistency evaluates whether an LLM provides uniform responses across 3 independent attempts to answer the same question. If all 3 answers are identical, the responses are considered consistent, regardless of their alignment with guideline criteria. For consistency assessment, an LLM receives 3 points if all 3 answers to a question are identical, and 2 points if one answer differs from the others. Each LLM's consistency performance is evaluated by averaging these scores across 50 questions, with possible scores ranging from 2 to 3. The consistency ratio is defined as each LLM's average consistency score divided by the maximum possible score (ie, 3).

Statistical Analysis

R software (version 4.4.2; R Foundation for Statistical Computing) was used for data management and statistical analysis. Continuous variables were described using means, and specific group instances were expressed as absolute values. The Wilcoxon rank-sum test was used for score comparisons between 2 different groups. We opted not to perform a Kruskal-Wallis test (for overall multigroup comparison) before pairwise analyses because our research focus was on hypothesis-driven, targeted comparisons between 2 specific groups: each LLM vs each physician subgroup (attendings, fellows, and residents). This design directly aligned with our aim—to evaluate whether single LLMs differ in performance from radiologists at distinct experience levels—rather than testing for overall differences across all groups. Differences between groups were considered significant at $P<.05$. P value correction for multiple comparisons was applied using the Benjamini-Hochberg method. Adjusted P values are provided in Multimedia Appendix 1, Tables 2-4, and corresponding Figures 2-4 to present direct pairwise comparison results; raw P values for multiple comparisons are provided exclusively in Multimedia Appendix 2.

Figure 2. Accuracy scores of large language models (LLMs) compared to physicians in clinical diagnosis and treatment questions of breast cancer. 3-star or dot symbols on each violin from top to bottom represent P values compared to each LLM to attending, fellow, and resident physicians, using the Wilcoxon rank-sum test. Symbols are as follows: "***" represents $P<.001$, "**" represents $.001\leq P<.01$, "*" represents $.01\leq P<.05$, and "black dots" represents $P\geq.05$.



XSL•FO
RenderX

Figure 3. Accuracy scores of large language models (LLMs) compared to those of physicians in radiological diagnosis questions of breast cancer.3-star or dot symbols on each violin from top to bottom represent *P* values compared to each LLM to attending, fellow, and resident physicians, using the Wilcoxon rank-sum test. Symbols are as follows: "***" represents *P*<.001, "**" represents .001≤*P*<.01, "*" represents .01≤*P*<.05, and "black dots" represents P≥.05.

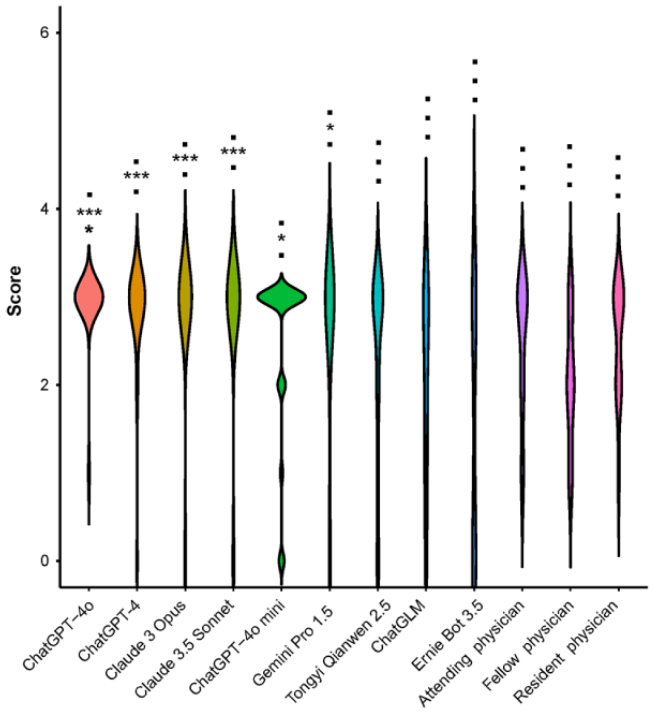
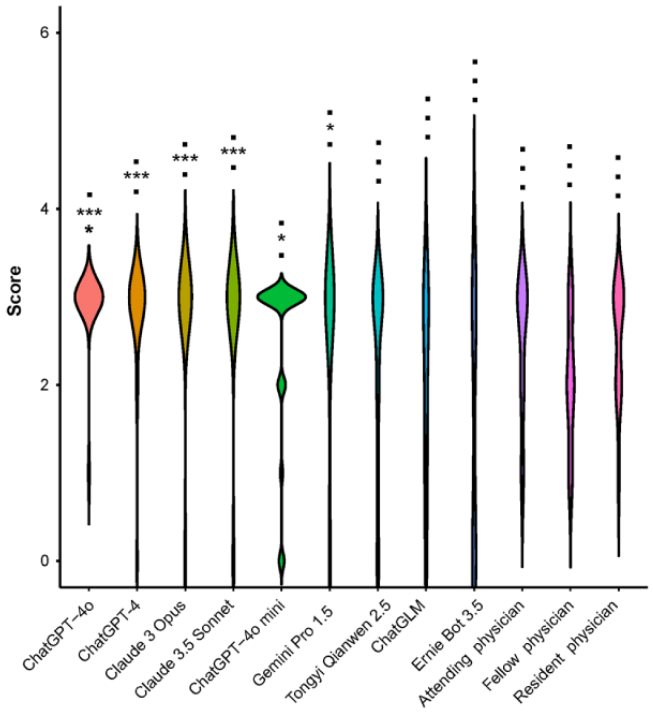


Figure 4. Accuracy scores of large language models (LLMs) compared to those of physicians in totalfor all 50 questions of breast cancer questions.3-star or dot symbols on each violin from top to bottom represent *P* values compared to each LLM to attending, fellow, and resident physicians, using the Wilcoxon rank-sum test. Symbols are as follows: "***" represents *P*<.001, "**" represents .001≤*P*<.01, "*" represents .01≤*P*<.05, and "black dots" represents P≥.05.



Ethical Considerations

Upon review by the Ethics Committee of Zhuhai People's Hospital, the information and content provided by the researcher did not involve the collection, storage, or use of human samples, and, therefore, met the requirements for exemption from ethical review.

XSL•FO
RenderX

Results

Representative Examples

Textbox 1 provides a representative selection of question-answer examples from the 9 LLMs, while a more comprehensive collection of cases is provided in the Multimedia Appendix 3.

Textbox 1. Representative examples of questions and answers from the 9 large language models (LLMs).

Sample question:

- Can an oval, hypoechoic lesion, growing nonparallel or vertically, with a heterogeneous echo pattern and obscured margins on breast ultrasound, be classified as Breast Imaging Reporting and Data System (BI-RADS 4b)?

ChatGPT-4 response:

- Yes, an oval, hypoechoic lesion growing nonparallel (vertically) with a heterogeneous echo pattern and obscured margins observed on breast ultrasound could be classified as BI-RADS 4b. This classification indicates a moderate suspicion of malignancy, meaning the lesion exhibits characteristics that are more suspicious than those typically found in BI-RADS 4a but not as definitively malignant as those in BI-RADS 4c.
- The characteristics you described—nonparallel orientation (often termed “taller than wide”), heterogeneous echo pattern, and obscured margins—suggest a higher level of concern. These features are often associated with a greater likelihood of malignancy:
  - Nonparallel or vertical orientation: this orientation is more typical of malignant lesions, as benign masses often grow in parallel with the skin.
  - Heterogeneous echo pattern: this can indicate variability within the lesion, which is more common in malignant tumors than in benign growths.
  - Obscured margins: this suggests that the lesion’s borders are not clearly defined, which can be an indication of invasive growth, a characteristic more typical of malignant processes.
- Conclusion: a lesion under 5 mm without a lipid core showing a washout pattern on magnetic resonance imaging should be approached with a high degree of suspicion for malignancy. This situation requires proactive management, including possible biopsy and close imaging follow-up, to ensure timely diagnosis and treatment.

ChatGPT-4o response:

- Yes, an oval, hypoechoic lesion that is growing nonparallel (vertically oriented), with a heterogeneous echo pattern and obscured margins on breast ultrasound can be classified as BI-RADS 4b. BI-RADS 4b is used to describe lesions with a moderate suspicion of malignancy.
- Key features of BI-RADS 4b:
  - Oval shape: while an oval shape can suggest a benign lesion, other features, such as orientation, margin appearance, and internal echo pattern, significantly influence the level of suspicion.
  - Nonparallel growth (vertically oriented): a lesion that grows nonparallel, or “taller than wide,” raises concern for malignancy. This orientation is often associated with invasive tumors because malignant lesions tend to invade tissue in a vertical direction.
  - Heterogeneous echo pattern: a heterogeneous echo pattern, where the lesion appears uneven or varied in echogenicity, is suspicious because it can indicate the presence of both solid and cystic components, or areas of differing cellular density, which is common in malignant tumors.
  - Obscured margins: obscured or indistinct margins are another red flag for malignancy. Clear, well-defined margins are typically seen in benign lesions, while indistinct margins suggest invasion into surrounding tissues.
- In conclusion, the combination of nonparallel growth, heterogeneous echo pattern, and obscured margins justifies a BI-RADS 4b classification, meaning the lesion has a moderate risk of malignancy and requires biopsy for a definitive diagnosis.

ChatGPT-4o mini response:

- Yes, an oval, hypoechoic lesion with a nonparallel (vertically oriented) growth pattern, heterogeneous echo pattern, and obscured margins on breast ultrasound can be classified as BI-RADS 4b.
- Here’s why this classification might be appropriate:
  - Oval shape: while typically less suspicious than irregular shapes, an oval shape alone does not rule out malignancy.
  - Hypoechoic: a darker appearance on ultrasound is often associated with both benign and malignant lesions.
  - Nonparallel growth pattern: a vertical or irregular growth pattern can be more concerning, as it may suggest infiltration into surrounding tissues.
  - Heterogeneous echo pattern: variability in echo patterns within a lesion can be indicative of more complex internal structures, which may be suggestive of malignancy.
  - Obscured margins: margins that are not well-defined or are irregularly shaped can be a sign of malignancy.
- BI-RADS 4b indicates a moderate suspicion of malignancy, with a likelihood of 10%-50%. This classification suggests that while there is a notable suspicion of cancer, the probability is not as high as in BI-RADS 4c. A biopsy is recommended to obtain a definitive diagnosis.

XSL•FO
RenderX

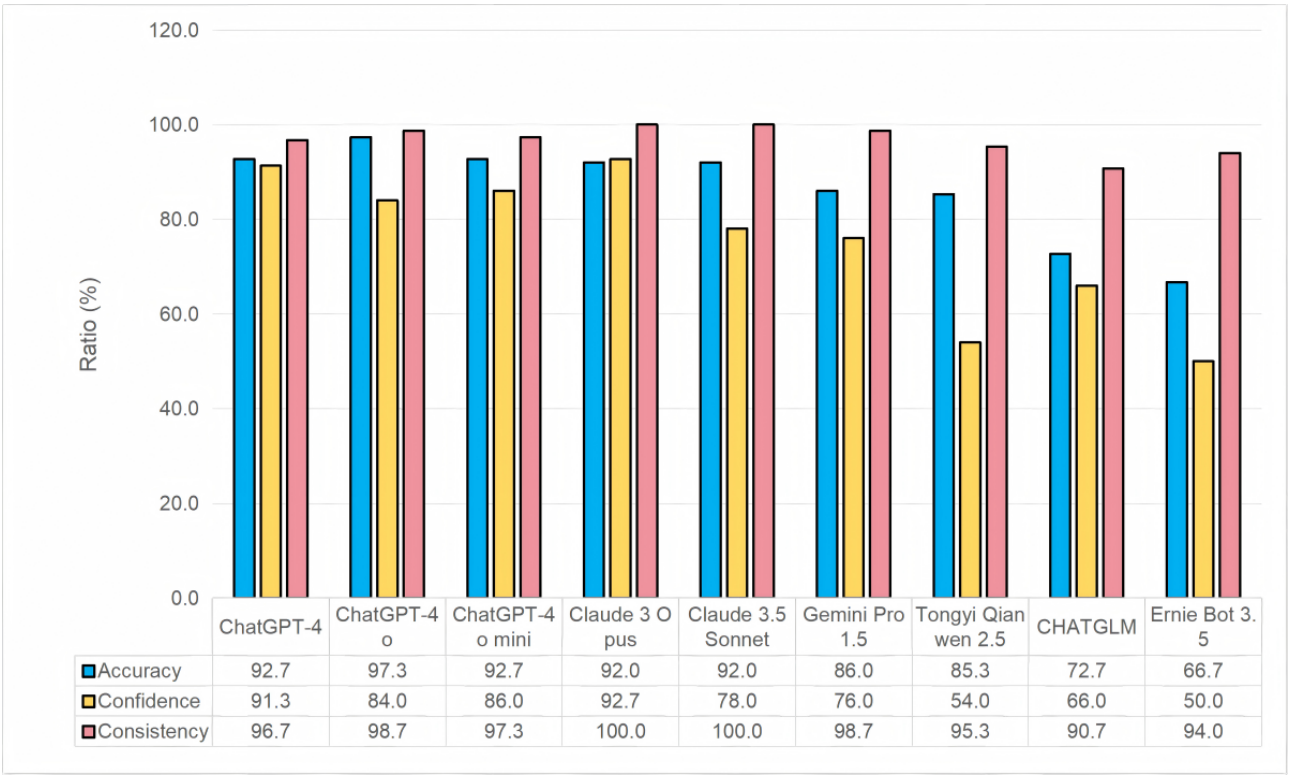
Confidence Level

Claude 3 Opus achieved the highest average score of 2.78, followed by ChatGPT-4 with a score of 2.74. The remaining models scored as follows: ChatGPT-4o 2.52, ChatGPT-4o mini 2.58, Claude 3.5, Sonnet 2.34, Gemini 1.5 Pro 2.28, Tongyi Qianwen 2.5 1.62, ChatGLM 1.98, and Ernie Bot 3.5 1.50. Claude 3 Opus and ChatGPT-4 demonstrated significantly higher confidence compared to the other models (Table 1 and Figure 5).

Table 1. Confidence and consistency levels of the large language models (LLMs).

Methods	Confidence level	Consistency level
ChatGPT-4	2.74	2.90
ChatGPT-4o	2.52	2.96
ChatGPT-4o mini	2.58	2.92
Claude 3 Opus	2.78	3.00
Claude 3.5 Sonnet	2.34	3.00
Gemini 1.5 Pro	2.28	2.96
Tongyi Qianwen 2.5	1.62	2.86
ChatGLM	1.98	2.72
Ernie Bot 3.5	1.50	2.82

Figure 5. Percentage of accuracy, confidence, and consistency levels of large language models (LLMs) across all 50 questions.



Consistency

Claude 3 Opus and Claude 3.5 Sonnet achieved the highest score of 3.0, demonstrating their superiority. The remaining models achieved the following consistency scores: ChatGPT-4o: 2.96, Gemini 1.5 Pro: 2.96, ChatGPT-4o mini: 2.92, ChatGPT4: 2.90, Tongyi Qianwen 2.5: 2.86, Ernie Bot 3.5: 2.82, and ChatGLM: 2.71 (Table 1 and Figure 5).

Performances on Clinical Diagnostics and Treatment Questions

In the comparison of clinical diagnostic problems, the mean accuracy scores are 3 (SD 0), 2.92 (SD 0.39), 2.81 (SD 0.69), 2.81 (0.49), 2.81 (SD 0.63), 2.77 (SD 0.65), 2.77 (SD 0.82), 2.77 (SD 0.82), 2.65 (SD 0.98), 2.38 (SD 1.02), and 2.35 (SD 1.13) for ChatGPT-4o mini, ChatGPT-4o, Tongyi Qianwen 2.5, attending physicians, fellow physicians, ChatGPT-4, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Ernie Bot 3.5, and ChatGLM, respectively. No significant differences were observed between any language modeling group and the various

XSL•FO
RenderX

physician groups. Regarding comparisons among physician groups, no significant differences were detected between different physician subgroups (Table 2, Figure 2, and Multimedia Appendices 1 and 4).

Table 2. Accuracy score comparisons of large language models (LLMs) with physician groups for clinical diagnostic and treatment questions.

Methods	Value, mean (SD)	FDR[a]-adjusted *P* value		
		Attending physician	Fellow physician	Resident physician
ChatGPT-4	2.77 (0.65)	.84	.97	.66
ChatGPT-4o	2.92 (0.39)	.44	.81	.49
ChatGPT-4o mini	3 (0)	.37	.42	.23
Claude 3 Opus	2.77 (0.82)	.62	.90	.49
Claude 3.5 Sonnet	2.77 (0.82)	.62	.90	.49
Gemini 1.5 Pro	2.65 (0.98)	.84	.97	.66
Tongyi Qianwen 2.5	2.81 (0.69)	.62	.90	.49
ChatGLM	2.35 (1.13)	.44	.44	.49
Ernie Bot 3.5	2.38 (1.02)	.44	.42	.49
Attending physician	2.81 (0.49)	—[b]	.90	.74
Fellow physician	2.81 (0.63)	—	—	.66
Resident physician	2.77 (0.51)	—	—	—

[a]FDR: false discovery rate.

[b]Not applicable.

In Figure 2, 3-star or dot symbols on each violin from top to bottom represent *P* values compared to each LLM to attending, fellow, and resident physicians, using the Wilcoxon rank-sum test. Symbols are as follows: "***" represents $P<.001$, "**" represents $.001\leq P<.01$, "*" represents $.01\leq P<.05$, and "." represents $P\geq.05$.

Performances on Radiological Diagnostics Problems

In the comparison of radiological diagnostic problems, the scores were as follows: ChatGPT-4 2.79, ChatGPT-4o 2.92, ChatGPT-4o mini 2.54, Claude 3 Opus 2.75, Claude 3.5 Sonnet 2.75, Gemini 1.5 Pro 2.5, Tongyi Qianwen 2.29, ChatGLM 2.0, and Ernie Bot 3.5 1.58. The attending physicians scored 2.46, the fellow physicians scored 2.04, and the resident physicians scored 2.5. The differences between ChatGPT-4, ChatGPT-4o, ChatGPT-4o mini, Claude 3 Opus, Claude 3.5 Sonnet, and Gemini 1.5 Pro and the fellow physicians were statistically significant ($P<.05$). There were statistically significant differences between ChatGPT-4o and the resident physicians ($P=.04$). In contrast, the differences between Tongyi Qianwen 2.5, ChatGLM, and the physician groups were not statistically significant ($P>.05$). In comparisons among physician groups, the differences were not statistically significant ($P>.05$; Table 3, Multimedia Appendices 1 and 4, and Figure 5).

XSL•FO
RenderX

Table 3. Accuracy score comparisons of large language models (LLMs) with physician groups for radiological diagnostic questions.

Methods	Value, mean (SD)	FDR[a]-adjusted *P* value		
		Attending physician	Fellow physician	Resident physician
ChatGPT-4	2.79 (0.66)	.10	<.001	.07
ChatGPT-4o	2.92 (0.41)	.06	<.001	.04
ChatGPT-4o mini	2.54 (0.93)	.51	.01	.45
Claude 3 Opus	2.75 (0.85)	.09	<.001	.07
Claude 3.5 Sonnet	2.75 (0.85)	.09	<.001	.07
Gemini 1.5 Pro	2.5 (1.06)	.50	.01	.41
Tongyi Qianwen 2.5	2.29 (1.08)	.81	.15	.97
ChatGLM	2 (1.1)	.22	.84	.22
Ernie Bot 3.5	1.58 (1.44)	.09	.49	.09
Attending physician	2.46 (0.78)	—[b]	.07	.99
Fellow physician	2.04 (0.75)	—	—	.07
Resident physician	2.5 (0.66)	—	—	—

[a]FDR: false discovery rate.

[b]Not applicable.

In Figure 3, 3-star or dot symbols on each violin from top to bottom represent *P* values compared to each LLM to attending, fellow, and resident physicians, using the Wilcoxon rank-sum test. Symbols are as follows: "***" represents $P<.001$, "**" represents $.001\leq P<.01$, "*" represents $.01\leq P<.05$, and "." represents $P\geq.05$.

Performances on Full Set Questions

In the comparison of overall performance, the scores were as follows: ChatGPT-4o 2.92, ChatGPT-4 2.78, ChatGPT-4o mini 2.78, Claude 3 Opus 2.76, Claude 3.5 Sonnet 2.76, Gemini 1.5 Pro 2.58, Tongyi Qianwen 2.56, ChatGLM 2.18, and Ernie Bot 3.5 2.0. The attending physicians scored 2.64, the fellow physicians scored 2.44, and the resident physicians scored 2.64. The differences between ChatGPT-4, ChatGPT-4o, ChatGPT-4o mini, Claude 3 Opus, Claude 3.5 Sonnet, and the fellow physicians were statistically significant ($P<.05$). Significant differences were observed between ChatGPT-4o and resident physicians ($P=.01$) and between ChatGPT-4o and attending physicians ($P=.03$). In the comparisons among physician groups, none of the differences were statistically significant ($P>.05$; Table 4, Multimedia Appendices 1 and 4, and Figure 4).

XSL•FO
RenderX

Table 4. Accuracy score comparisons of large language model (LLM) models with physician groups for all 50 questions.

Methods	Value, mean (SD)	FDR[a]-adjusted *P* value		
		Attending physician	Fellow physician	Resident physician
ChatGPT-4	2.78 (0.65)	.13	.006	.09
ChatGPT-4o	2.92 (0.4)	.03	<.001	.01
ChatGPT-4o mini	2.78 (0.68)	.13	.006	.09
Claude 3 Opus	2.76 (0.82)	.08	.003	.05
Claude 3.5 Sonnet	2.76 (0.82)	.08	.003	.05
Gemini 1.5 Pro	2.58 (1.01)	.47	.06	.30
Tongyi Qianwen 2.5	2.56 (0.93)	.84	.18	.67
ChatGLM	2.18 (1.12)	.09	.47	.12
Ernie Bot 3.5	2 (1.29)	.08	.26	.09
Attending physician	2.64 (0.66)	—[b]	.19	.78
Fellow physician	2.44 (0.79)	—	—	.30
Resident physician	2.64 (0.6)	—	—	—

[a]FDR: false discovery rate.

[b]Not applicable.

In Figure 4, 3-star or dot symbols on each violin from top to bottom represented *P* values compared to each LLM model to attending, fellow, and resident physicians, with the Wilcoxon rank-sum test. Symbols are as follows: "***" represents *P*<.001, "**" represents .001≤*P*<.01, "*" represents .01≤*P*<.05, and "." represents *P*≥.05.

Discussion

Principal Findings

With the global population aging, the rising incidence of cancer, and the increasing complexity of treatment options, there is a growing demand for MDT discussions. However, implementing high-level MDTs in underdeveloped countries remains a significant challenge [16], largely because of the knowledge barriers between specialist doctors. An important question is how specialist doctors in less developed regions can achieve the same medical standards as those in developed regions, easily access authoritative and professional medical knowledge, and overcome barriers between specialties. The recently developed chatbot can theoretically provide instant, evidence-based responses, demonstrating significant potential as an ideal tool for enhancing high-quality health care in underserved regions. In our comparison of 9 LLMs, Claude 3 Opus demonstrated the highest confidence score at 2.78, followed by ChatGPT-4 (2.74) and ChatGPT-4o mini (2.58). These scores suggest that these models exhibit a higher degree of certainty when managing complex medical issues. In contrast, Ernie Bot 3.5's confidence score of 1.50 was significantly lower, indicating a reduced ability to comprehend and process medical problems. These findings underscore the varying levels of proficiency among language models in accurately addressing the same problem. Interestingly, unlike human practitioners, not all chatbots confidently respond to closed-ended questions with a definitive "yes" or "no," suggesting that training data diversity and model architecture significantly impact diagnostic accuracy. This uncertainty may lead to reduced consistency and accuracy of their outputs, thus limiting their independent use in practical clinical applications, making them inseparable from close collaboration with physicians during decision-making.

In addition, in the consistency assessment, Claude 3 Opus and Claude 3.5 Sonnet achieved the highest scores of 3, followed by ChatGPT-4o and Gemini 1.5 Pro, both scoring 2.96, with ChatGPT-4o mini at 2.92. In contrast, ChatGLM and Ernie Bot 3.5 exhibited lower consistency scores of 2.72 and 2.82, respectively. These results highlight the differences in response stability across the LLMs, a factor that is especially critical for clinical decision support. In clinical applications, particularly in dynamic and complex environments, models lacking sufficient consistency may undermine confidence in their outputs. Although Claude 3 Opus and Claude 3.5 Sonnet surpassed the ChatGPT models in response consistency (scoring 3), the ChatGPT models demonstrated superior performance in diagnostic accuracy. However, none of the models provided correct answers to all question sets. Therefore, while AI introduces new possibilities in fields such as medicine, it also presents challenges that require careful expert scrutiny to avoid imposing additional burdens on patients and health care professionals [17,18].

By comparing the nine LLMs with the responses of nine radiologists on the overall 50 guideline questions, ChatGPT-4o achieved the highest score of 2.92, significantly outperforming all physician groups. The difference between ChatGPT-4o and all physician groups (*P*<.05) was statistically significant. ChatGPT-4, ChatGPT-4o mini, Claude 3 Opus, and Claude 3.5 Sonnet all scored higher on average than the physician groups. However, the difference was statistically significant only for the comparison with fellow physicians (*P*<.05). In contrast, Gemini 1.5 Pro and Tongyi Qianwen 2.5 had lower scores, with no statistically significant differences compared to each group

XSL•FO
RenderX

of doctors (P>.05). ChatGLM and Ernie Bot 3.5 showed lower scores than all physician groups, but the differences were not statistically significant. No statistically significant differences (P>.05) were found in the comparisons between each group of doctors, suggesting that attending physicians, fellow physicians, and resident physicians performed comparably in answering the entire question set. Attending physicians and resident physicians both scored 2.64, which may reflect a more consistent ability to answer certain routine questions across all years of experience. Fellow physicians, on the other hand, scored slightly lower, but the difference was not significant, suggesting that their performance in overall diagnostic and therapeutic tasks remains competitive.

In the evaluation of radiological diagnostic performance, ChatGPT-4o scored 2.92. A significant difference was observed between ChatGPT-4o and resident physicians (P=.04), and a significant difference was also detected between ChatGPT-4o and fellow physicians (P<.05). ChatGPT-4 achieved a score of 2.79, surpassing all physician groups on average. However, this difference reached statistical significance solely in the comparison with fellow physicians (P<.05). The difference between ChatGPT-4o mini and the fellow physicians was statistically significant (P=.01). Claude 3 Opus and Claude 3.5 Sonnet both scored 2.75, higher than all physician groups; in particular, the difference between these LLMs and fellow physicians was statistically significant (P<.05), suggesting superior performance in radiological diagnostics. Gemini 1.5 Pro scored 2.50, showing small differences compared with attending physicians and resident physicians but significantly higher than fellow physicians (P=.01). Tongyi Qianwen 2.5 and ChatGLM scored lower, with no statistically significant differences compared to physician groups (P>.05). Ernie Bot 3.5 had the lowest score (1.58) among all LLMs, with no statistical significance compared to any physician group, indicating relatively poorer performance. We conclude that the ChatGPT and Claude series outperform junior medical doctors in addressing complex medical issues, highlighting their potential clinical utility. However, ChatGLM and Ernie Bot 3.5 may require further optimization to improve their performance in medical diagnostics. In the diagnostic radiology domain, the average performance of attending physicians and resident physicians was comparable, whereas the small group of fellow physicians scored lower on average. It is noteworthy that resident physicians, who have recently completed standardized residency training, may be more familiar with the specific guidelines on which the questions were based. Attending physicians, benefiting from extensive clinical experience, would be expected to exhibit an advantage in diagnosing complex imaging findings. The performance variation across the small physician subgroups underscores the preliminary nature of these comparisons.

In the assessment of clinical diagnosis and treatment performance, ChatGPT-4o mini achieved the highest score of 3 among all LLMs, although no statistically significant differences were found when compared with any physician group. ChatGPT-4o scored 2.92 and showed no significant difference compared with physician groups (P>.05); however, it outperformed all physician groups on average. Other LLMs, including ChatGPT-4, Claude 3 Opus, and Claude 3.5 Sonnet, exhibited similar performance to physician groups. ChatGLM and Ernie Bot 3.5 showed weaker performance. Overall, the ChatGPT and Claude series of LLMs outperformed some physician groups with lower seniority in complex health care challenges, indicating potential for clinical applications. Further optimization may be necessary for ChatGLM and Ernie Bot 3.5 to enhance their diagnostic performance. Scores between attending physicians and fellow physicians were identical, while resident physicians scored slightly lower; however, these differences did not reach statistical significance. Radiologists of varying experience levels may possess adequate knowledge of clinical guidelines in diagnosis and treatment, particularly when dealing with standardized procedures or common diseases, resulting in minimal differences. Despite differences in experience, attending physicians and fellow physicians demonstrated similar proficiency in applying clinical guidelines.

It is noteworthy that Tongyi Qianwen 2.5 was the only model unable to generate a response to a specific question, instead providing the feedback, “I’m very sorry, but I don’t think I fully understand what you mean. Let’s change the subject first, shall we?” This issue may be attributable to the complexity and specialized nature of the question, or potential biases and gaps in the model’s training data, leading to insufficient knowledge in specific domains. Our data indicate that when Tongyi Qianwen 2.5 fails to respond, the model’s accuracy may be overstated, and its actual performance is significantly below expectations.

Comparison With Previous Work

In our analysis of the 9 language models, Ernie Bot 3.5, Tongyi Qianwen 2.5, and ChatGLM were more likely to recommend referring the questions to medical experts when faced with specialized imaging terminology and clinical scenarios, rather than providing a definitive answer. In contrast, Claude 3 Opus and the ChatGPT series provided more detailed diagnostic imaging feedback, including recommendations for additional biopsies to support diagnosis and an emphasis on multidisciplinary collaboration. Regarding clinical considerations for surgery, the ChatGPT series excelled by offering surgical recommendations for breast cancer patients, while also demonstrating conventional surgical approaches and discussing recent therapeutic advancements. Furthermore, the ChatGPT family outperformed other language models in breast cancer diagnosis and treatment, with each model excelling in its respective area of expertise. This finding highlights the critical need to select the most appropriate language model based on task-specific performance, thereby enhancing problem-solving across various scenarios. These interactive language models provided feedback significantly faster than breast cancer radiologists, reducing time costs and increasing productivity. With continued iteration and improvement, ChatGPT has demonstrated considerable potential in radiology, including concise report generation, support for medical education, clinical decision aids, patient communication optimization, and data analysis [19,20]. Analyzing the reasons for comparable performance between leading language models reveals several key advantages: (1) large-scale, high-quality training data: both the developers of Claude 3 Opus and Chat

XSL•FO
RenderX

GPT have invested substantial resources in acquiring and curating authoritative medical literature, textbooks, and specialized professional databases; (2) comparable model scale and architecture: as state-of-the-art LLMs, both likely use similar parameter scales and analogous underlying architectural designs; (3) equivalent alignment methodologies: both organizations have likely implemented similar reinforcement learning from human feedback techniques, using extensive human evaluator input to enhance the accuracy of model outputs; (4) competition-driven enhancement: market competition has compelled both companies to continuously improve model performance, particularly in high-value vertical domains such as health care. This competitive environment has resulted in both products demonstrating enhanced logical reasoning capabilities; (5) domain-specific knowledge processing: both models have likely incorporated specialized training methodologies for processing technical terminology and complex semantic relationships within specialized conceptual frameworks; (6) parallel safety control mechanisms: in high-risk domains such as medicine, both models have presumably implemented stringent output controls and prudence measures to ensure responsible performance. Given the continuous evolution of LLMs, their capabilities in complex domains, particularly medicine, are rapidly advancing. Future research should further investigate the potential of these models to maximize their value in clinical diagnosis and treatment [21].

However, LLMs were not specifically trained for the medical field and rely on large datasets collected from public sources, such as web pages, books, and websites. Chinese LLMs have reported average accuracy in general medical question-answering of around 82%; yet, performance drops significantly in complex clinical reasoning tasks, such as breast cancer diagnosis (below 50%) [22]. Advanced models, such as ChatGPT-3.5, ChatGPT-4.0, and Claude-2, achieve slightly higher performance (approximately 60%) in breast cancer clinical assessments, though this still highlights critical limitations in high-stakes medical applications [23]. Despite the vast amount of training data, these models lack comprehensive common sense, which limits their reasoning ability. The literature suggests that LLM performance can be significantly enhanced by guiding reasoning steps incrementally, particularly in tasks involving longer reasoning chains [24,25]. In our study, however, the 9 LLMs were unable to accurately diagnose all rigorous and logical medical problems, and a "curse of reversal" phenomenon was observed. Furthermore, unlike real-time search engines such as Google and Bing, LLMs can only generate responses based on their training data and do not have real-time access to current information [26,27]. In addition, LLMs cannot reliably differentiate between factual and fictional statements, posing a potential risk as sources of misinformation in certain cases [28-30]. Moreover, the training data for these models are not publicly available, and the sources of their information are not disclosed, making it difficult to verify the accuracy of their outputs [31]. Therefore, radiologists must continue to rely on traditional evidence-based education and approach LLMs critically, cross-referencing their outputs with credible medical sources [32].

Limitations

This study has several limitations. First, our findings are subject to the inherent limitations of the author-crafted, closed-ended question format. Although designed for objective assessment, this format may not fully capture the reasoning capabilities of LLMs and could potentially underestimate their performance in real-world, open-ended clinical scenarios. Second, the binary definition of a "confident" response as a direct "yes" or "no" may not adequately reflect clinical decision-making, where acknowledging uncertainty is a sign of expertise. This could have introduced bias by favoring LLMs that generate definitive answers over physicians who appropriately consider diagnostic nuances. Third, the physician sample size, while sufficient for initial comparisons, remains relatively small and may not be fully representative of the broader population of radiologists, which could affect the generalizability of our pairwise comparisons. Fourth, the study's generalizability is constrained by its specific design: the assessment is limited to 50 selected questions that do not encompass all aspects of the guidelines, and the performance of the LLMs was evaluated on a single, specific task (answering structured concordance questions). Consequently, the results may not be directly extrapolated to other clinical datasets or different natural language processing tasks. Fifth, the study conditions, including a strict 30-minute time limit and a "no-consultation" rule for the participating physicians, may not reflect real-world clinical practice, where complex cases often benefit from more time and collaborative discussion. These methodological choices, while necessary for standardization, may have disadvantaged the physicians and thus the conclusions of this study must be heavily qualified by this context. In addition, it has been reported that LLMs may generate different responses to identical prompts at different times, due to model updates [31,33]. Future studies should use larger and more diverse physician cohorts, incorporate more nuanced assessments of confidence, explore the impact of different time constraints and collaborative settings, and use a broader range of question types and clinical tasks to validate and extend our findings.

Future Directions

In medical AI applications, LLMs participating in clinical decision-making face critical responsibility challenges, including decision output liability, misinformation risks, and privacy breach accountability [30]. Typically, LLMs do not independently generate decision outputs but function as tools for physicians. Clinicians must integrate guidelines, relevant literature, and their clinical expertise to formulate final clinical decisions. When using LLMs, it is essential to conceal patients' identifying information to safeguard privacy.

This necessitates establishing explicit legal frameworks that delineate responsibility boundaries among developers, health care institutions, and physicians, while constructing comprehensive accountability mechanisms [34,35]. Concurrently, LLMs present a risk of generating inaccurate medical information that could potentially mislead diagnostic and therapeutic decisions. To address this concern, model developers must implement rigorous content review protocols

XSL•FO
RenderX

and ensure professional, authoritative verification of training corpora.

Conclusions

In conclusion, within the context of our study, Claude 3 Opus and the ChatGPT series, particularly ChatGPT-4o, excelled in addressing breast cancer–related guideline questions and demonstrated significant potential for clinical application. In contrast, the less consistent performance of other LLMs, especially Ernie Bot 3.5 and ChatGLM, suggests that both confidence and consistency should be carefully considered when selecting and applying LLMs to ensure efficacy and stability in clinical practice. This preliminary evaluation indicates that LLMs can provide radiologists with extensive cross-disciplinary knowledge, potentially enhancing contributions within MDTs. However, they cannot fully replace human expertise, particularly in complex diagnostic scenarios that require nuanced decision-making honed by clinical experience.

Acknowledgments

We also sincerely thank Dr Siping Luo, Dr Xiao Luo, Huimei Lü, Zhichunmei Yuan, and Qiuxinmei Chen for their invaluable suggestions in designing the 50 breast cancer questions.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Funding

This research was funded by the Department of Science and Technology of Guangdong Province (grant no 112175605105) and the research project “Preoperative Localization and Navigation of Non-Mass Breast Cancer with DCE-MRI and Ultrasound Fusion Imaging” (grant no 2023yjlcyj019), and the Shenzhen Basic Research Project (Natural Science Foundation; grant no JCYJ20220530150416036). The funding period was 3 years, starting in June 2023, with a total budget of CNY $35,615 (US $5000).

Authors' Contributions

CY was responsible for conceptualization, developing the conceptual framework, and formulating the ideas for this study. HJ and CY prepared the original draft, including substantive translation, and contributed to manuscript creation. HJ, CY, JY, and CY participated in the review and editing of the manuscript, covering both prepublication and postpublication stages. Supervision of the study was provided by HJ, CY, and JY, who oversaw and took leadership responsibility for the planning and execution of research activities. Resource support, including the provision of breast cancer–related guideline materials, was provided by JY, HJ, and CY. Formal analysis was conducted by WZ, who applied statistical techniques to analyze the research data, while data curation, including management of LLM models and essential software code, was performed by WZ and SZ. Investigation, including data collection during the study, was carried out by GR, WW, and MW. Project administration was coordinated by JY, who ensured the smooth planning and execution of research activities. Funding acquisition was secured by HJ and JY.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Scores of large language models (LLMs) and physician groups on 50 questions, with comparison analysis.
[XLSX File (Microsoft Excel File), 14 KB - medinform_v14i1e68182_app1.xlsx]

Multimedia Appendix 2
Comparison of accuracy scores between large language models (LLMs) and physicians across 2 domains.
[DOCX File , 23 KB - medinform_v14i1e68182_app2.docx]

Multimedia Appendix 3
Scores of large language models (LLMs) and physician groups on 50 questions, including raw *P* values for comparison.
[DOCX File , 20 KB - medinform_v14i1e68182_app3.docx]

Multimedia Appendix 4
Specific responses from 9 large language models (LLMs) to the 50 questions (each question asked 3 times) and the corresponding physicians’ responses.
[XLSX File (Microsoft Excel File), 854 KB - medinform_v14i1e68182_app4.xlsx]

XSL•FO
RenderX

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. New global breast cancer initiative highlights renewed commitment to improve survival. World Health Organization (WHO). 2021 Mar 8. URL: https://www.who.int/news/item/08-03-2021-new-global-breast-cancer-initiative-highlights-renewed-commitment-to-improve-survival [accessed 2025-01-01]
3. Trapani D, Ginsburg O, Fadelu T, Lin NU, Hassett M, Ilbawi AM, et al. Global challenges and policy solutions in breast cancer control. Cancer Treat Rev 2022;104:102339. [doi: 10.1016/j.ctrv.2022.102339] [Medline: 35074727]
4. Azadnajafabad S, Saeedi Moghaddam S, Mohammadi E, Delazar S, Rashedi S, Baradaran HR, et al. Patterns of better breast cancer care in countries with higher human development index and healthcare expenditure: insights from GLOBOCAN 2020. Front Public Health 2023;11:1137286 [FREE Full text] [doi: 10.3389/fpubh.2023.1137286] [Medline: 37124828]
5. Hisham R, Ng CJ, Liew SM, Hamzah N, Ho GJ. Why is there variation in the practice of evidence-based medicine in primary care? A qualitative study. BMJ Open 2016;6(3):e010565 [FREE Full text] [doi: 10.1136/bmjopen-2015-010565] [Medline: 26962037]
6. Yuan L, Cao J, Wang D, Yu D, Liu G, Qian Z. Regional disparities and influencing factors of high quality medical resources distribution in China. Int J Equity Health 2023;22(1):8 [FREE Full text] [doi: 10.1186/s12939-023-01825-6] [Medline: 36627636]
7. Zuhair V, Babar A, Ali R, Oduoye MO, Noor Z, Chris K, et al. Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. J Prim Care Community Health 2024;15:21501319241245847 [FREE Full text] [doi: 10.1177/21501319241245847] [Medline: 38605668]
8. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. Radiology 2019;293(2):246-259 [FREE Full text] [doi: 10.1148/radiol.2019182627] [Medline: 31549948]
9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
11. Maroncelli R, Rizzo V, Pasculli M, Cicciarelli F, Macera M, Galati F, et al. Probing clarity: AI-generated simplified breast imaging reports for enhanced patient comprehension powered by ChatGPT-4o. Eur Radiol Exp 2024;8(1):124. [doi: 10.1186/s41747-024-00526-1] [Medline: 39477904]
12. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(5):2817-2825 [FREE Full text] [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
13. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023;308(6):1831-1844 [FREE Full text] [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
14. Schulte B. Capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. Cureus 2023;15(4):e37938 [FREE Full text] [doi: 10.7759/cureus.37938] [Medline: 37220429]
15. Berglund L, Tong M, Kaufmann M, Balesni M, Stickland A, Korbak T. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". arXiv. Preprint posted online on Sep 21, 2023 2023. [doi: 10.48550/arXiv.2309.12288]
16. Eaton V, Zambrano A, Sanabria V, Lopez R, Kyei I, Mra R, et al. Innovative methodology for strengthening a multidisciplinary team approach in cities in low- and middle-income countries. JCO Glob Oncol 2022;8:e2200149 [FREE Full text] [doi: 10.1200/GO.22.00149] [Medline: 36252159]
17. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019;69(2):127-157 [FREE Full text] [doi: 10.3322/caac.21552] [Medline: 30720861]
18. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. PLOS Digit Health 2024;3(4):e0000341 [FREE Full text] [doi: 10.1371/journal.pdig.0000341] [Medline: 38630683]
19. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
20. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023;104(6):269-274 [FREE Full text] [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]
21. Hadi M, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh M. A survey on large language models: applications, challenges, limitations, and practical usage. arXiv. Preprint posted online on Jul 19, 2023 2023 Unreviewed arXiv preprint (arXiv:2307.10169); not formally peer-reviewed or published [FREE Full text] [doi: 10.36227/techrxiv.23589741.v1]

XSL•FO
RenderX

22. Shanghai Artificial Intelligence Laboratory, China Software Testing Center (CSTC). Analysis Report on the Development of Large Models in the Healthcare Field (2024). WeChat Official Account (Shanghai Artificial Intelligence Laboratory).: Shanghai Artificial Intelligence Laboratory; 2024. URL: https://mp.weixin.qq.com/s/XCd5FiW2DacnfYXn1XdkSQ
23. Deng L, Wang T, Yangzhang, Zhai Z, Tao W, Li J, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg 2024;110(4):1941-1950 [FREE Full text] [doi: 10.1097/JS9.0000000000001066] [Medline: 38668655]
24. Jin M, Yu Q, Zhao H, Hua W, Meng Y, Zhang Y. The impact of reasoning step length on large language models. arXiv. Preprit posted online on Jan 10, 2024 2024. [doi: 10.48550/arXiv.2401.04925]
25. Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Duan N, et al. CRITIC: large language models can self-correct with tool-interactive critiquing. In: Proceedings of the International Conference on Learning Representations (ICLR). 2024 Presented at: 12th International Conference on Learning Representations (ICLR) 2024; May 7–11, 2024; Vienna, Austria URL: https://openreview.net/forum?id=Sx038qxjek
26. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023;25:e48659 [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
27. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
28. Simon FM, Altay S, Mercier H. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. HKS Misinfo Review 2023 5;4(5):N/A-N/A. [doi: 10.37016/mr-2023-054]
29. Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell 2024;6(8):852-863 [FREE Full text] [doi: 10.1038/s42256-024-00872-2]
30. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
31. Bernal J, Mazo C. Transparency of artificial intelligence in healthcare: insights from professionals in computing and healthcare worldwide. Appl Sci 2022 Oct 11;12(20):10228. [doi: 10.3390/app122010228]
32. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum 2023;4(5):e231938 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1938] [Medline: 37200013]
33. Kahng M, Tenney I, Pushkarna M, Liu M, Wexler J, Reif E, et al. LLM comparator: visual analytics for side-by-side evaluation of large language models. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24). 2024 Presented at: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24); May 11-16, 2024; Honolulu, HI p. 1-7. [doi: 10.1145/3613905.3650755]
34. Ferdaus MM, Abdelguerfi M, Loup E, N. Niles K, Pathak K, Sloan S. Towards trustworthy AI: a review of ethical and robust large language models. arXiv. Preprint posted online on Jun 1, 2024 2024 [FREE Full text]
35. Bouhouita-Guermech S, Haidar H. Scoping review shows the dynamics and complexities inherent to the notion of "Responsibility" in artificial intelligence within the healthcare context. Asian Bioeth Rev 2024;16(3):315-344. [doi: 10.1007/s41649-024-00292-7] [Medline: 39022380]

Abbreviations

AI: artificial intelligence
ACR BI-RADS: American College of Radiology Breast Imaging-Reporting and Data System
LLM: large language model
MDT: multidisciplinary team
NCCN: National Comprehensive Cancer Network

Edited by A Coristine; submitted 16.Nov.2024; peer-reviewed by S Matsuda, L Wreyford; comments to author 11.Feb.2025; revised version received 02.Dec.2025; accepted 02.Dec.2025; published 02.Feb.2026.

Please cite as:
Jiang H, Yang C, Zhou W, Yin CL, Zhou S, He R, Ran G, Wang W, Wu M, Yu J
Evaluation of Large Language Models for Radiologists' Support in Multidisciplinary Breast Cancer Teams: Comparative Study
JMIR Med Inform 2026;14:e68182
URL: https://medinform.jmir.org/2026/1/e68182
doi:10.2196/68182
PMID:

XSL•FO
RenderX

©Hong Jiang, Chun Yang, Wenbin Zhou, Cheng-liang Yin, Shan Zhou, Rui He, Guanghui Ran, Wujie Wang, Meixian Wu, Juan Yu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence Models for Predicting Triage in Emergency Departments: Seven-Month Retrospective Comparative Study of Natural Language Processing, Large Language Model, and Joint Embedding Predictive Architectures

Edouard Lansiaux[1], MD, MSc; Ramy Azzouz[2,3], MD; Emmanuel Chazard[3,4], MD, PhD; Amélie Vromant[5], MD, MSc; Eric Wiel[1,3], MD, PhD

[1]Emergency Department, Lille University Hospital, Lille, France

[2]Centre Antipoison, Lille University Hospital, Lille, France

[3]ULR 2694-METRICS, Lille University, Lille, France

[4]Department of Public Health, EA 2694, Lille University, Lille, France

[5]Emergency Department, Hôpital Pitié-Salpêtrière, Assistance Publique des Hôpitaux de Paris, Paris, France

Corresponding Author:
Edouard Lansiaux, MD, MSc
Emergency Department
Lille University Hospital
2 Avenue Oscar Lambret
Lille, 59800
France
Phone: 33 789222432
Email: edouard.lansiaux@orange.fr

Abstract

Background: Triage errors in emergency departments (EDs), including undertriage and overtriage, pose significant risks to patient safety and resource allocation. With increasing patient volumes and staffing challenges, artificial intelligence (AI) integration into triage protocols has gained attention as a potential solution.

Objective: This study aims to develop and compare 3 AI models—natural language processing (NLP), large language model (LLM), and Joint Embedding Predictive Architecture (JEPA)—for predicting triage outcomes according to the French Emergency Nurses Classification in Hospital (FRENCH) scale and to assess their performance relative to nurse triage and clinical expert consensus.

Methods: We conducted a retrospective analysis of prospectively collected data from adult patients triaged at Roger Salengro Hospital ED (Lille, France) over 7 months (June-December 2024). Three AI models were developed: TRIAGEMASTER (NLP with Doc2Vec + MLP), URGENTIAPARSE (LLM with FlauBERT + Extreme Gradient Boosting [XGBoost]), and EMERGINET (JEPA with variance-invariance-covariance regularization). Of 73,236 ED visits, 657 (0.90%) had complete audio recordings and structured data. Data were split 80:20 into training and validation sets with stratification. Gold-standard labels were established by senior clinician consensus (minimum 5 years of ED experience). The primary outcome was concordance with the gold-standard FRENCH triage level, assessed using weighted κ, Spearman correlation, F1-score, area under the receiver operating characteristic (AUC-ROC) curve, mean absolute error (MAE), and root mean square error (RMSE). Secondary analyses evaluated Groupes d'Etude Multicentrique des Services d'Accueil (GEMSA) prediction and performance by input data type.

Results: URGENTIAPARSE demonstrated superior performance, with a composite z score of 2.514 compared with EMERGINET (0.438), TRIAGEMASTER (–3.511), and nurse triage (–4.343). URGENTIAPARSE achieved an F1-score of 0.900 (95% CI 0.876-0.924), an AUC-ROC of 0.879 (95% CI 0.851-0.907), a weighted κ of 0.800 (P<.001), a Spearman correlation of 0.802 (P<.001), an MAE of 0.228, and an RMSE of 0.790. Exact agreement was 90.0%, with near-agreement (+1 or –1 level) of 92.8%. However, training showed perfect accuracy (1.0) with poor validation performance (~0.5), indicating overfitting. EMERGINET achieved moderate performance (F1-score=0.731, AUC 0.686), while TRIAGEMASTER and nurse triage performed poorly (F1-score=0.618 and 0.303, respectively). For GEMSA prediction, URGENTIAPARSE maintained superiority (κ=0.863, Spearman=0.864, P<.001). Class 1 (highest acuity) was underrepresented (4/657, 0.61%), limiting undertriage risk assessment.

XSL•FO
RenderX

Conclusions: The LLM-based architecture (URGENTIAPARSE) demonstrated the highest accuracy for ED triage prediction among the tested models, outperforming traditional NLP, JEPA, and current nurse triage practices. However, severe overfitting, extreme selection bias (657/73,236, 0.90%, inclusion), a monocentric design, and sparse high-acuity representation limit clinical applicability. Before deployment, the model requires regularization, external validation across diverse EDs, prospective testing, and comprehensive safety evaluation, particularly for undertriage detection. Integration of AI triage support systems shows promise but demands rigorous validation, bias mitigation, and transparent uncertainty quantification to ensure patient safety.

(JMIR Med Inform 2026;14:e83318) doi:10.2196/83318

KEYWORDS

triage; emergency department; artificial intelligence; natural language processing; large language model; Joint Embedding Predictive Architecture; clinical decision support; FRENCH scale

Introduction

Triage in emergency departments (EDs) is vital for prioritizing patients by severity, directly affecting outcomes, efficiency, and resource use. In busy or understaffed settings, accurate triage is essential to ensure safe and equitable care.

In France, the French Emergency Nurses Classification in Hospital (FRENCH) triage scale has become a standardized reference tool for emergency medical services [1]. This 6-level urgency stratification system (1=immediate life-threatening to 6=nonurgent) was designed to promote consistency and reproducibility. However, in practice, its application often varies [2], shaped by disparities in nurse experience, time pressure, interpretive inconsistencies, and fatigue. These factors lead to 2 key errors: undertriage, in which serious cases are underestimated, risking delays in care and potentially preventable mortality; and overtriage, in which minor cases are overprioritized, leading to crowding and inefficient use of resources [3].

Against this backdrop, artificial intelligence (AI) has emerged as a promising complement to traditional triage [4]. AI systems in emergency medicine can process diverse data, operate continuously, and provide consistent and transparent decisions; however, their success over the past decade has been mixed.

Early AI triage systems relied on natural language processing (NLP) to extract clinical meaning from text, using hand-crafted rules or statistical models [5]. Although these early systems were effective, they lacked the ability to understand context, nuance, or rare cases. Advances emerged with large language models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and its French version, FlauBERT [6]. These transformer-based models provided deep contextual understanding. When combined with boosting algorithms such as Extreme Gradient Boosting (XGBoost), LLMs formed hybrid architectures that balanced high performance with interpretability [7], 2 pillars of clinical deployment. More recently, Joint Embedding Predictive Architectures (JEPA) have introduced a third paradigm. JEPA models do not predict outcomes directly; instead, they embed both inputs and targets (eg, triage categories) into a shared latent space. Using techniques such as energy minimization and variance-invariance-covariance regularization (VICReg) [8,9], these models can uncover subtle cross-modal patterns. However, their complexity poses ongoing challenges for interpretability, an essential requirement in medical contexts.

Few studies have compared AI models under consistent conditions [10,11]. Most focus on narrow tasks, rely on limited or noisy data, and are often developed outside the French health care system, reducing their relevance for users of the FRENCH scale. Furthermore, recent work has highlighted concerns regarding judgment biases in triage [11,12], hallucinations in LLMs [13], and the need for robust calibration and uncertainty quantification in clinical AI systems [14].

This study aims to address this gap. We conducted a retrospective analysis of triage encounters at the Centre Hospitalier Universitaire (CHU) Lille ED, one of France's major academic hospitals [15]. Using rich clinical data and nurse-patient dialogues, 3 AI models—TRIAGEMASTER, URGENTIAPARSE, and EMERGINET—were developed to predict FRENCH triage levels. These models were evaluated against expert consensus to identify the best-performing system and to assess their clinical suitability and limitations.

Methods

Study Design

This study was designed as a retrospective, observational, monocentric analysis conducted in the Adult Emergency Department of the CHU Lille, France. The ED is a high-volume academic tertiary care facility serving a diverse patient population, with approximately 73,000 adult visits annually. The investigation covered a continuous 7-month period from June 1, 2024, to December 31, 2024. This time frame was selected to capture seasonal variations in ED activity and to ensure a sufficiently large and heterogeneous dataset.

Ethical Compliance

The study adhered to strict French and European ethical standards, with approval and registration from the Comité d'Éthique de la Recherche en Santé, Health Data Hub filing (project number 18605502), and Commission Nationale de l'Informatique et des Libertés (CNIL) Méthodologie de Référence-004 compliance [16]. Audio recordings were collected with patient information at triage. Personal data were present before anonymization and underwent rigorous deidentification, including removal of names, dates, identifiers, and geographic information smaller than the department level, with transcript verification performed by 2 independent

XSL•FO
RenderX

reviewers. As this was a retrospective analysis of deidentified data, individual informed consent was waived in accordance with French regulations. The study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)-AI guidelines (see Multimedia Appendix 1). Owing to patient confidentiality, the full dataset and code are not publicly available; only the code used to define the gold standard is shared (see Multimedia Appendix 2).

Population and Inclusion Criteria

The study included adult patients (≥18 years) evaluated in the Lille CHU ED with complete structured clinical data and recorded nurse-patient audio interviews. Exclusion criteria were minors (<18 years), legally protected adults, patients who declined data use (n=2), and cases with incomplete recordings, corrupted files, missing structured data, or failed transcription quality checks (n=22).

Data Acquisition and Preprocessing

The study relied on 2 primary data sources. First, structured clinical data were extracted from the ResUrgences ED management software (version 2024.1.148; Berger-Levrault). Variables collected included age, sex, vital signs (systolic and diastolic blood pressure, heart rate, temperature, oxygen saturation, and supplemental oxygen requirement), pain score (visual analog scale 0-10), time of presentation, comorbidities, chief complaint category, the initial FRENCH triage decision, and the Groupes d'Etude Multicentrique des Services d'Accueil (GEMSA) scale (corresponding to patient outcome after diagnostic evaluation and treatment in the ED) [17].

Unstructured data from nurse-patient triage interviews were audio-recorded using standardized digital recorders and manually transcribed by trained personnel, with accuracy verified by 2 independent reviewers. Transcripts were linked using unique patient identifiers and time stamps, anonymized in accordance with CNIL guidelines, and securely stored on Health Data Hub–approved servers for modeling. Recording participation was limited to volunteer nurses (n<10) during their regular shifts, introducing potential selection bias in conversation style and patient characteristics.

Comparison of Included Versus Excluded Cohorts

To assess selection bias, we compared demographic and clinical characteristics between included patients (n=657) and the broader ED population (n=72,579) during the study period. The included cohort had a similar mean age (42.6 vs 43.2 years, *P*=.68), sex distribution (330/657, 50.2%, male, *P*=.67), and vital sign distributions. However, included patients had slightly lower acuity (mean nurse FRENCH 3.4 vs 3.1, *P*=.02) and higher admission rates (15.5% vs 12.3%, *P*=.03), likely reflecting volunteer nurse participation patterns and recording feasibility during less chaotic triage encounters (Table 1).

XSL•FO
RenderX

Table 1. Patient demographics, admission characteristics, vital parameters, and triage distributions.

Characteristic	Overall	Training	Validation
Age (years), mean (SD)	42.6 (19.7)	42.5 (19.8)	43 (19.5)
Sex, n (%)			
Male	330 (50.2)	264 (50.2)	66 (50.4)
Female	327 (49.8)	262 (49.8)	65 (49.6)
Admission time, n (%)			
Night (9 PM-6 AM)	211 (32.1)	169 (32.1)	42 (32.1)
Morning (6 AM-2 PM)	241 (36.7)	193 (36.7)	48 (36.6)
Afternoon (2 PM-9 PM)	205 (31.2)	164 (31.2)	41 (31.3)
Chief complaint, n (%)			
Trauma, n (%)	127 (19.3)	101 (19.2)	26 (19.8)
Abdominal, n (%)	102 (15.5)	81 (15.4)	21 (16.0)
Cardiovascular	86 (13.1)	68 (12.9)	18 (13.7)
Neurological	69 (10.5)	55 (10.5)	14 (10.7)
ENT (ear-nose-throat)/stomatology	62 (9.4)	49 (9.3)	13 (9.9)
Other[a]	211 (32.1)	172 (32.7)	39 (29.8)
Comorbidities, n (%)			
None	496 (75.5)	398 (75.7)	98 (74.8)
≥1 comorbidity	161 (24.5)	128 (24.3)	33 (25.2)
Vital signs, mean (SD)			
Systolic blood pressure (mmHg)	143 (24)	143 (24)	143 (23)
Diastolic blood pressure (mmHg)	83 (16)	83 (16)	83 (17)
Heart rate (bpm)	89 (18)	89 (18)	89 (17)
Temperature (°C)	37.5 (0.6)	37.5 (0.6)	37.5 (0.6)
Pain score (0-10)	4 (3.6)	4 (3.6)	4.1 (3.7)
Peripheral capillary oxygen saturation (SpO_2, %)	97 (2)	97 (2)	97 (2)
Supplemental O2 (L/min), mean (SD)	0.04 (0.4)	0.04 (0.4)	0.03 (0.3)
Nurse triage (FRENCH[b]), n (%)			
1 (immediate), n (%)	4 (0.6)	3 (0.6)	1 (0.8)
2 (very urgent), n (%)	86 (13.1)	68 (12.9)	18 (13.7)
3A (urgent, invasive)	0 (0)	0 (0)	0 (0)
3B (urgent, noninvasive)	354 (53.9)	283 (53.8)	71 (54.2)
4 (less urgent)	118 (18.0)	94 (17.9)	24 (18.3)
5 (nonurgent)	95 (14.5)	78 (14.8)	17 (13.0)
Disposition (GEMSA[c]), n (%)			
1	0 (0)	0 (0)	0 (0)
2 (discharged)	509 (77.5)	407 (77.4)	102 (77.9)
3 (short stay)	18 (2.7)	14 (2.7)	4 (3.1)
4 (ward admission)	102 (15.5)	82 (15.6)	20 (15.3)
5 (intensive care unit admission)	10 (1.5)	9 (1.7)	1 (0.8)
Not specified	18 (2.7)	14 (2.7)	4 (3.1)

XSL•FO
RenderX

[a]Other complaints are genitourinary, gynecology/obstetrics, infectious, toxicology, ophthalmology, dermatology, psychiatry, pulmonary, and rheumatology presentations.

[b]FRENCH: French Emergency Nurses Classification in Hospital.

[c]GEMSA: Groupes d'Etude Multicentrique des Services d'Accueil.

Gold-Standard Construction

A gold-standard triage label was established through independent review by 3 senior emergency physicians (minimum 5 years of postresidency ED experience, board-certified in emergency medicine) who were strictly blinded to the original nurse triage assignment, all AI model outputs, and patient outcomes (laboratory results, imaging, disposition, and diagnoses established after the index triage encounter). Each adjudicator independently reviewed the complete triage transcript, vital signs, and demographic data, and then assigned a FRENCH level based solely on information available at the time of triage.

Initial interrater agreement was substantial (Fleiss κ=0.72, 95% CI 0.68-0.76). For the cases with disagreement (155/657, 23.6%), adjudicators met for a structured consensus discussion without knowledge of which assignments were discordant. Final consensus was reached in all cases through majority voting after discussion. This gold standard served as the reference for model training and evaluation. To assess potential incorporation bias, we conducted a sensitivity analysis excluding cases in which the gold standard differed from nurse triage by more than 1 level (n=47), and found similar model performance rankings.

Model Architectures and Training Pipelines

Three distinct AI models were developed, each representing a major class of current machine learning techniques. All models were implemented in Python 3.9 (Python Foundation) with PyTorch 1.12 and scikit-learn 1.1, and trained on NVIDIA A100 graphics processing units (40 GB VRAM), with training times of 2-4 hours per model.

Leakage Safeguards and Hyperparameter Tuning

Overview

Critical measures were implemented to prevent data leakage. All preprocessing steps (Doc2Vec fitting, FlauBERT fine-tuning, and feature scaling) were confined strictly to the training folds, with transformations subsequently applied to validation data. Hyperparameter tuning was performed using 5-fold stratified cross-validation nested within the training set only, without accessing validation data. Random seeds were fixed (seed=42) to ensure reproducibility. Early stopping monitored validation loss with a patience of 10 epochs to prevent overfitting.

TRIAGEMASTER (NLP)

This model used paragraph vector embeddings (Doc2Vec, distributed memory model; vector dimension=100, window=5, min_count=2, epochs=40) to represent free-text input [18], concatenated with *z* score–normalized numerical features derived from structured data (age, sex encoded as binary, and all vital signs). The architecture consisted of a 3-layer feedforward neural network (input→128→64→6 output classes) with Rectified Linear Unit activations [19], dropout regularization (P=.32, applied after each hidden layer), batch normalization, and an L2 weight penalty ($\lambda=1\times10^{-5}$). Optimization used the Adam optimizer with an initial learning rate of 0.001, batch size of 32, and learning rate decay (factor=0.5, patience=5 epochs). The maximum sequence length was 512 tokens. Missing data (<2% of vital signs) were imputed using median values from the training set.

Classification probabilities were derived via softmax activation. CIs for the area under the curve (AUC) were calculated using the DeLong method [20]. Model performance was evaluated across the full range of predicted probabilities. Optimal probability thresholds were defined as the predicted probability that maximized the F_1-score [21]. Inference latency averaged 47 ms per patient on production CPU hardware (Intel Xeon Gold 6248R).

URGENTIAPARSE (LLM)

Built on the FlauBERT language model (flaubert-base-cased, 137 million parameters) [6], this architecture fine-tuned contextual embeddings from patient complaints over 5 epochs with a learning rate of 2×10^{-5}, and then extracted the [CLS (classification)] token embedding (768 dimensions). These embeddings were concatenated with normalized structured variables and fed into an XGBoost classifier (n_estimators=200, max_depth=6, learning_rate=0.1, min_child_weight=3, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, reg_lambda=1.0, scale_pos_weight adjusted for class imbalance). This hybrid design enabled both interpretability and high-dimensional pattern recognition [22]. Feature importance was derived using Shapley Additive Explanations (SHAP) values [23], providing transparent model behavior traceable to individual tokens and vital signs. The maximum sequence length was 512 tokens with truncation. Class weights were inversely proportional to class frequencies to address imbalance. Inference latency averaged 312 ms per patient.

EMERGINET (JEPA)

This JEPA encoded textual and structured inputs into a shared latent space (dimension=256) through dual encoders (text encoder: FlauBERT base + 2-layer bidirectional Long Short-Term Memory with hidden size 128 [24]; structured encoder: 3-layer MLP with hidden sizes (64, 128, 256)). The loss function minimized contrastive energy between predicted and target class embeddings using cosine similarity, regularized with VICReg (variance weight=1.0, invariance weight=1.0, covariance weight=0.04) to enforce variance >1.0, invariance via mean squared error, and a near-identity covariance matrix [8]. Training used the Adam optimizer (learning rate=1×10^{-4}, batch size=16, 100 epochs). Final classification was performed using nearest-neighbor matching in embedding space with temperature-scaled softmax (temperature=0.07). Inference latency averaged 189 ms per patient.

XSL•FO
RenderX

Training and Validation Procedures

Each model was trained on 80% of the dataset (n=526) and validated on the remaining 20% (n=131), stratified by gold-standard triage level to maintain class balance (class proportions: 1=4/657, 0.61%; 2=86/657, 13.1%; 3B=354/657, 53.9%; 4=118/657, 18.0%; and 5=95/657, 14.5%; classes 3A and 6 had 0 cases). Hyperparameter grids were searched using 5-fold cross-validation on the training data only. All models were trained 3 times with different random initializations to assess stability; reported metrics are means across runs, with SDs <0.03 for all primary metrics. Training and validation loss curves were monitored to detect overfitting.

Evaluation Metrics

Performance Metrics for Ordinal Outcomes

Model performance was assessed using both absolute error measures and classification agreement metrics. Unless otherwise specified, all *P* values were 2-tailed, with a significance threshold of α=.05. Given the ordered nature of the FRENCH scale, metrics that account for ordinality (weighted κ, Spearman correlation, and mean absolute error [MAE]) were prioritized.

MAE

This measures the average deviation between predicted and actual triage levels.

where is the predicted value, y_i is the true value, and n is the total number of cases.

Root Mean Square Error

This emphasizes larger misclassification errors:

Weighted Cohen κ

This accounts for agreement beyond chance and penalizes larger discrepancies using quadratic weights:

where $fo_{i,j}$ are observed frequencies, $fe_{i,j}$ are expected frequencies under independence, and $w_{i,j}=(i-j)^2/(k-1)^2$ with k classes. Significance was tested via bootstrap (10,000 iterations).

Spearman Correlation Coefficient

This measures the monotonic association between predicted and true ordinal ranks:

$r_s=[\text{cov}(rg_X,rg_Y)]/[\sigma rg_X \sigma rg_Y]$ **(4)**

F1-scores

Macro F_1-score reflects the unweighted average across all triage classes (treating each class equally), whereas micro F_1-score is weighted by class frequency:

F_1-score = (TP)/[TP+1/2(FP+FN)] **(5)**

where TP denotes true positive, FP false positive, and FN false negative. Class-specific F_1-scores were calculated to assess performance in rare high-acuity classes.

Exact Agreement Percentage

This represents the proportion of predictions that exactly match the gold standard.

Near-Agreement Percentage

This represents the proportion of predictions within the +1 or –1 triage level of the gold standard.

Area Under the Receiver Operating Characteristic Curve

This summarizes discrimination across all classification thresholds, calculated using a one-versus-rest strategy for multiclass classification, with 95% CIs estimated via the DeLong method [20].

Composite z Score

This is an integrated performance indicator combining multiple metrics for ranking.

Score = Z(–MAE) + Z(–RMSE) + $Z(\kappa_w)$ + $Z(r_s)$ **(6)**

where $Z(X) = (X - \mu_X)/\sigma_X$ standardizes each metric to mean 0 and SD 1 across all models. Negative MAE and RMSE were used, so higher z scores indicate better performance for all components.

Multiclass Brier Score

This is the mean squared difference between predicted probabilities and one-hot encoded outcomes, providing global probabilistic accurac

where is the predicted probability for class j of case i, y_{ij} is a binary indicator (1 if true class, 0 otherwise), and k is the number of classes.

Asymmetric Error Analysis

Given the clinical importance of undertriage (predicting lower acuity than the true acuity, potentially delaying critical care), we analyzed error asymmetry by calculating the undertriage rate (proportion of predictions <true class) and the overtriage rate (proportion of predictions >true class). For classes 1-2 (highest acuity), we report sensitivity, specificity, positive predictive value, and negative predictive value separately, as undertriage in these categories poses the greatest patient safety risk.

Calibration Analysis

Model calibration was assessed by grouping predicted probabilities into quantile-based bins (10 bins per class) and comparing mean predicted probabilities with observed frequencies. Reliability diagrams plotted mean predicted probability versus observed frequency for each bin, with perfect calibration represented by the diagonal. Expected calibration error and maximum calibration error were computed as measures of calibration quality. Additionally, risk distributions were

XSL•FO
RenderX

visualized using density ridge plots stratified by true triage class and faceted by predicted class.

Secondary Analyses

A secondary analysis compared actual versus AI-predicted GEMSA scores (patient disposition: 2=discharged home, 3=short stay unit, 4=admitted to ward, and 5=intensive care unit [ICU]) using models trained separately on nurse-recorded structured histories and full interview transcripts. Another analysis assessed model performance according to input data type (structured data only vs unstructured transcripts only). Both analyses used evaluation methods consistent with the primary analysis. All statistical analyses were performed in R version 4.2 (R Foundation) and Python 3.9.

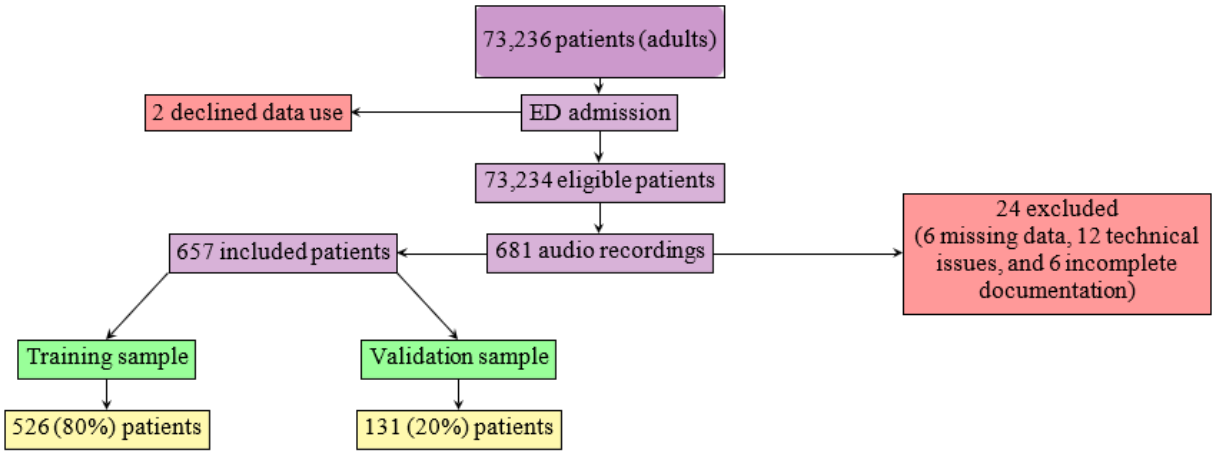
Results

Population Characteristics

During the inclusion period, 73,236 patients visited the ED of Roger Salengro Hospital. Two patients declined data use. Triage nurse interviews were recorded and transcribed for 681 (0.93%) patients. After excluding 24 cases (6 with missing data and 18 with transcription quality issues), 657 (0.90%) patients met the inclusion criteria (Figure 1).

Included patients had a mean age of 42.6 (SD 19.7) years and a median age of 38 years. The sample was gender-balanced (330/657, 50.2%, male). Most admissions occurred during morning hours (241/657, 36.7%, 6 AM-2 PM). The primary chief complaints were trauma (127/657, 19.3%), abdominal (102/657, 15.5%), and cardiovascular (81/657, 13.1%). One-quarter of patients (161/657, 24.5%) had 1 or more comorbidities, predominantly cardiovascular. Most patients were triaged as FRENCH 3B (354/657, 53.9%), with sparse representation in the highest acuity categories (class 1: 4/657, 0.61%).

Figure 1. Patient flow diagram showing inclusion, exclusion, and dataset splitting. ED: emergency department.



Training Performance and Overfitting Assessment

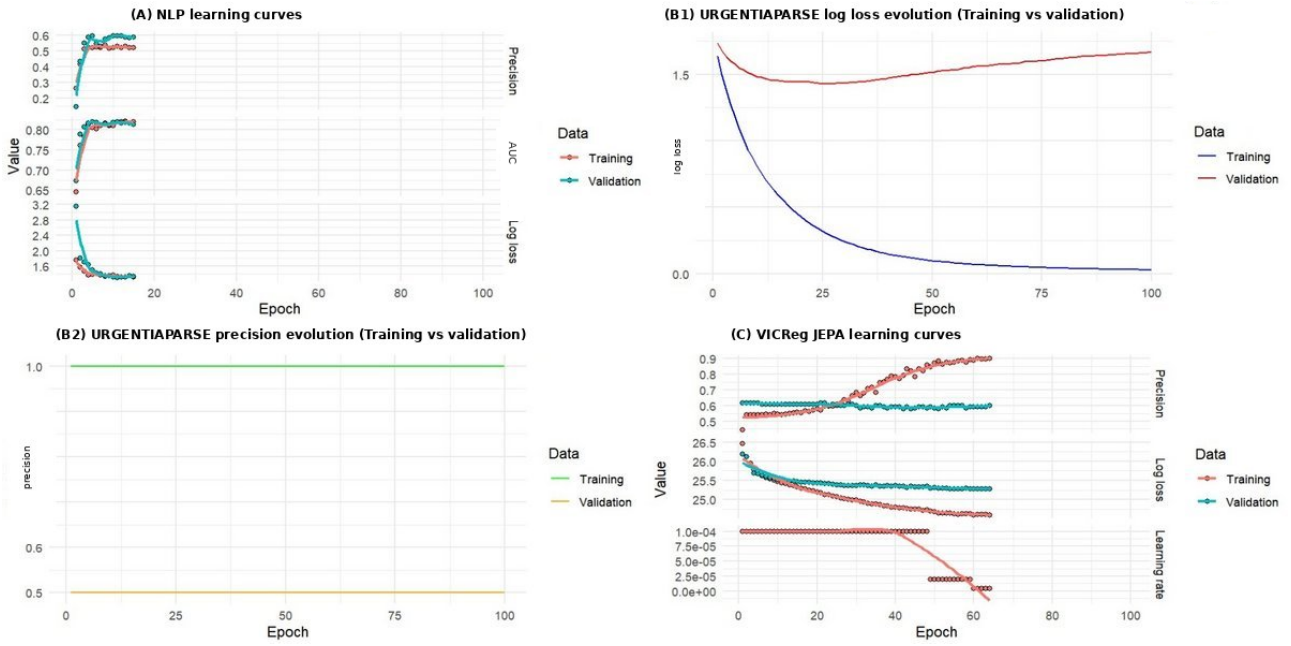
Over 100 training epochs, TRIAGEMASTER achieved a training accuracy plateau of approximately 75%, with validation accuracy similarly stable at 72%, while log loss decreased and stabilized for both sets, indicating reasonable generalization (Figure 2A).

URGENTIAPARSE exhibited severe overfitting, achieving perfect training accuracy (1.0) by epoch 20 but persistently poor validation accuracy (~0.5 throughout training), indicating memorization rather than learning generalizable patterns (Figure 2B.2). Training log loss approached 0, whereas validation loss remained high and flat at approximately 1.5, with an AUC plateau of 0.69 (Figure 2B.1). This discrepancy indicates that, despite strong performance on the training data, the model fails to generalize to unseen cases—a critical limitation for clinical deployment.

EMERGINET reached a training accuracy of 0.85, but validation accuracy remained at 0.60, indicating moderate overfitting (Figure 2C). Training loss declined steadily, whereas validation loss remained higher and stable, reflecting the performance gap.

XSL•FO
RenderX

Figure 2. Learning curves for 3 artificial intelligence models. (A) TRIAGEMASTER shows stable learning with consistent training and validation performance. (B1) URGENTIAPARSE log-loss evolution demonstrates severe overfitting, with training loss approaching 0 while validation loss remains high. (B2) URGENTIAPARSE accuracy evolution confirms perfect training accuracy (1.0) but poor validation accuracy (0.5). (C) EMERGINET VICReg JEPA learning curves show moderate overfitting, with a gap between training and validation performance. JEPA: Joint Embedding Predictive Architecture; NLP: natural language processing; VICReg: variance-invariance-covariance regularization.



Overall Model Performance

URGENTIAPARSE (LLM) achieved the highest composite score (2.514), outperforming EMERGINET (JEPA, 0.438), TRIAGEMASTER (NLP, –3.511), and nurse triage (–4.343; Table 2). URGENTIAPARSE demonstrated strong discrimination, with a micro F_1-score of 0.900 (95% CI 0.876-0.924), macro F_1-score of 0.894 (95% CI 0.869-0.919), and AUC of 0.879 (95% CI 0.851-0.907). Agreement metrics were substantial: weighted κ of 0.800 (P<.001, 95% CI 0.761-0.839) and Spearman correlation of 0.802 (P<.001, 95% CI 0.768-0.832). Error metrics showed an MAE of 0.228 classes (approximately 1 triage level error every 4-5 patients) and an RMSE of 0.790. Exact agreement was 90.0%, with near agreement (+1 or –1 level) of 92.8%.

EMERGINET achieved moderate performance (micro F_1-score=0.731, macro F_1-score=0.747, AUC 0.686, κ=0.560, P<.001), whereas TRIAGEMASTER performed poorly (micro F_1-score=0.618, AUC 0.642, κ=0.370, P<.001; Spearman=0.005, P=.95). Current nurse triage showed the lowest performance (micro F_1-score=0.303, AUC 0.776, κ=0.080, P=.08; exact agreement=30.3%), likely reflecting the difference between real-time clinical judgment under time pressure and the gold standard established by senior physicians following comprehensive information review.

Receiver operating characteristic (ROC) curve analysis (Figure 3) confirmed the superiority of URGENTIAPARSE across most triage classes, with particularly strong performance for classes 2, 4, and 5 (AUC >0.85). However, class 1 performance was perfect (AUC 1.0) because of extreme sparsity (n=4), limiting the reliability of this estimate. Performance for class 3 was lower (AUC 0.748), likely reflecting the combined representation of 3A (invasive procedures) and 3B (noninvasive), with no 3A cases in the dataset.

Table 2. Primary analysis comparing triage prediction performance across artificial intelligence models and nurse triage.

Model	Mean absolute error	Root mean square error	κ[a]	Spearman[b]	Micro F_1-score	Macro F_1-score	Exact	Near[c]	z score[d]	Area under the curve
URGENTIAPARSE	0.228	0.790	0.800	0.802	0.900	0.894	0.900	0.928	2.514	0.879
EMERGINET	0.401	0.979	0.560	0.602	0.731	0.747	0.820	0.860	0.438	0.686
TRIAGEMASTER	0.637	1.180	0.370	0.005	0.618	0.613	0.554	0.696	-3.511	0.642
Nurse triage	1.393	1.834	0.080	0.024	0.303	0.275	0.303	0.498	-4.343	0.776

[a]Weighted Cohen κ; for all κ values, P<.001 except nurse triage (P=.08).

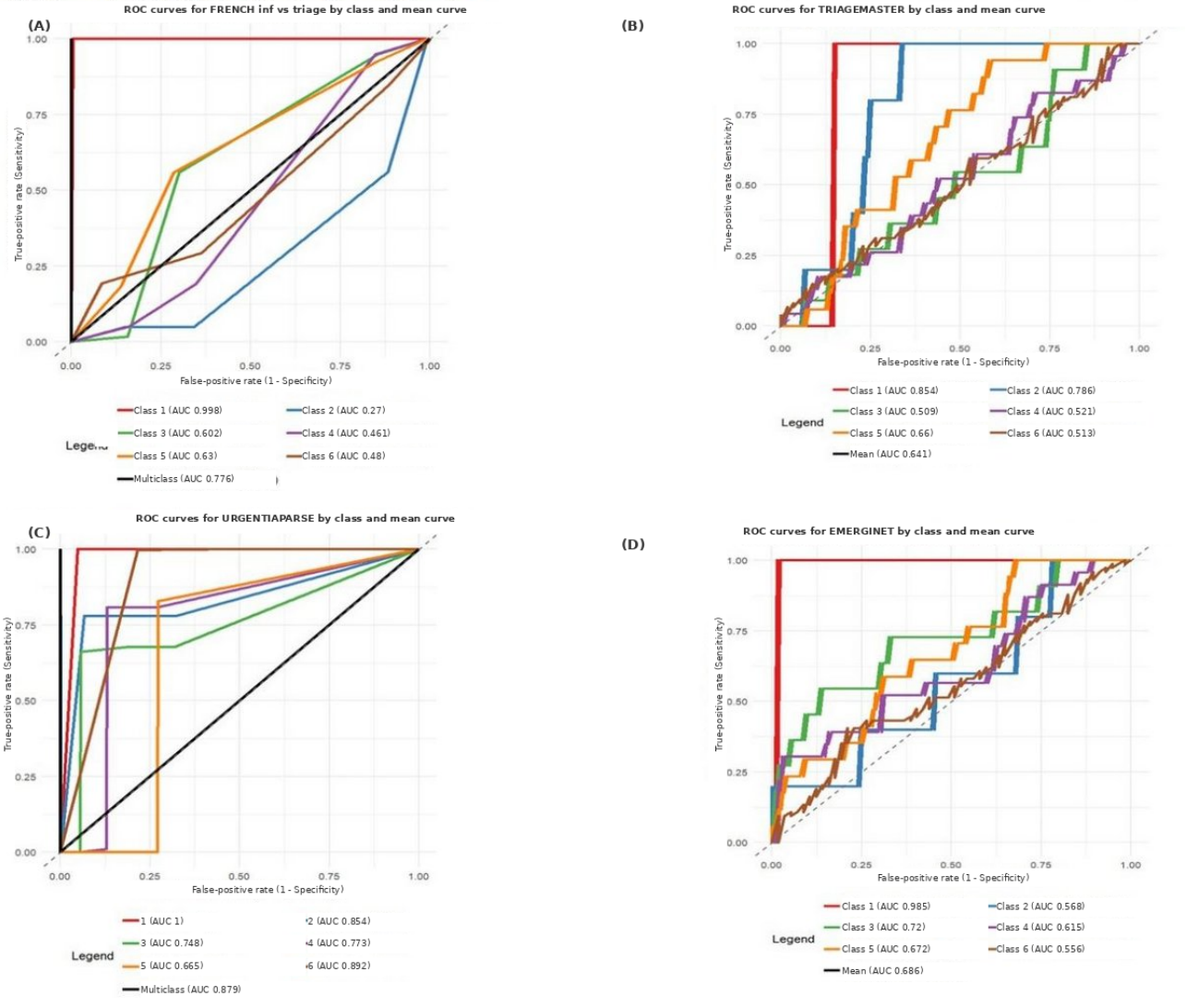
[b]Spearman rank correlation; all P<.001 except TRIAGEMASTER (P=.95) and nurse triage (P=.76).

[c]Near agreement: proportion within the +1 or –1 triage level.

[d]Composite z score combining standardized mean absolute error, root mean square error, κ, and Spearman metrics (higher is better).

XSL•FO
RenderX

Figure 3. Receiver operating characteristic (ROC) curves for each triage class and the mean curve across all models. URGENTIA-PARSE (large language model) demonstrates superior discrimination for most classes. The results for class 1 (area under the curve [AUC] 1.0 for multiple models) should be interpreted cautiously due to extreme sparsity (n=4). (A) FRENCH inf vs triage — This panel shows ROC curves for a rule-based French triage reference system, where Class 1 achieves near-perfect discrimination (AUC 0.998) but most other classes perform poorly (AUC 0.27–0.63), yielding a modest multiclass AUC of 0.776. (B) TRIAGEMASTER — This panel displays the per-class ROC performance of TRIAGEMASTER, a structured NLP triage model, with moderate overall discrimination (mean AUC 0.641) and relatively homogeneous class-level AUCs ranging from 0.509 to 0.854, suggesting balanced but limited performance across triage levels. (C) URGENTIAPARSE — This panel presents URGENTIAPARSE's ROC curves, the best-performing model overall (multiclass AUC 0.879), with Class 1 achieving perfect separation (AUC 1.0) and strong performance across most classes, reflecting its superior ability to discriminate triage severity levels from free-text clinical notes. (D) EMERGINET — This panel illustrates EMERGINET's ROC curves, showing strong Class 1 discrimination (AUC 0.985) but more heterogeneous performance across the remaining classes (AUC 0.556–0.72), resulting in an intermediate mean AUC of 0.686.



Error Pattern Analysis

Confusion matrices revealed distinct error patterns (Figure 4, top row). URGENTIAPARSE and EMERGINET slightly overestimated acuity for lower triage levels (predicting higher urgency than true), indicating a clinically conservative bias. TRIAGEMASTER systematically underestimated acuity across most classes. Nurse triage showed scattered and inconsistent patterns, with frequent misclassification across nonadjacent categories.

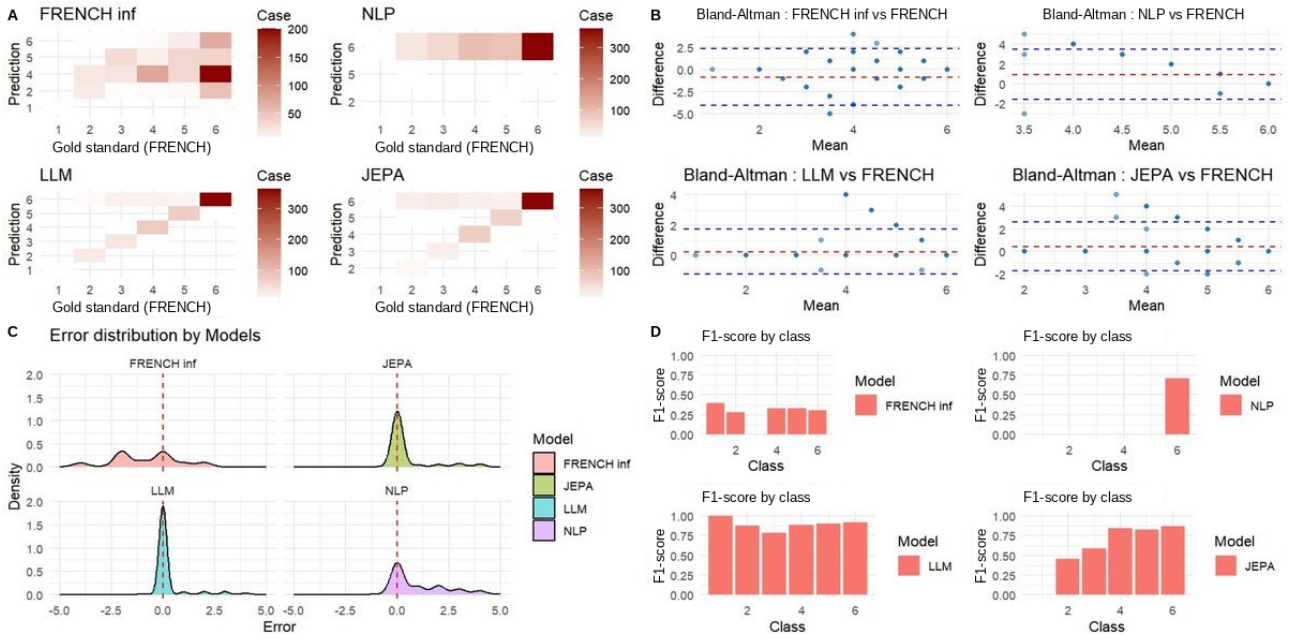
Bland-Altman analysis (Figure 4, middle row) confirmed low, well-centered errors for URGENTIAPARSE (mean difference –0.02, 95% limits of agreement –1.58 to 1.54) and EMERGINET (mean difference 0.08, limits –2.05 to 2.21). TRIAGEMASTER showed larger deviations (limits –3.02 to 2.98), and nurse triage demonstrated extreme disagreement (limits –4.21 to 3.87).

Error distribution analysis (Figure 4, bottom left) showed that URGENTIAPARSE and EMERGINET had narrow, well-centered error distributions (SD <1.0), whereas TRIAGEMASTER and nurse triage exhibited broad, dispersed distributions, indicating less reliable predictions.

Class-specific F_1-scores (Figure 4, bottom right) demonstrated that URGENTIAPARSE maintained consistently high performance across all represented classes (F_1-score >0.75 for classes 2-5), whereas the other approaches showed variable and often poor class-specific performance, particularly for minority classes.

XSL•FO
RenderX

Figure 4. Comprehensive evaluation of four AI-based triage models against the FRENCH gold standard. (A) Confusion matrices comparing predicted versus gold standard triage levels (1–6) for FRENCH inf, NLP, LLM, and JEPA models. (B) Bland-Altman plots assessing systematic bias and limits of agreement between each model and the FRENCH gold standard. (C) Kernel density distributions of prediction errors illustrating the bias and precision of each model. (D) Per-class F1-scores quantifying the discriminative performance of each model across the six triage severity levels.



Asymmetric Error and Undertriage Risk

Given the clinical importance of undertriage, we analyzed error directionality. URGENTIAPARSE had an undertriage rate of 4.6% (predicting lower acuity than true) and an overtriage rate of 5.4% (predicting higher acuity), reflecting a slight conservative bias favoring patient safety. For high-acuity classes (1-2, n=90), URGENTIAPARSE achieved a sensitivity of 87.8% (95% CI 79.2%-93.7%) and a specificity of 95.1% (95% CI 91.3%-97.6%), with a positive predictive value of 86.8% and a negative predictive value of 95.6%. However, the limited representation of class 1 (n=4) constrains the assessment of undertriage risk among the most critical patients. Future deployment would require prospective monitoring, with particular attention to undertriage events and associated patient outcomes.

Calibration and Reliability

Multiclass Brier scores indicated superior probabilistic calibration for URGENTIAPARSE (0.142), compared with EMERGINET (0.385), TRIAGEMASTER (0.524), and nurse triage (0.693). However, detailed calibration analysis revealed heterogeneous reliability across classes (Figure 5).

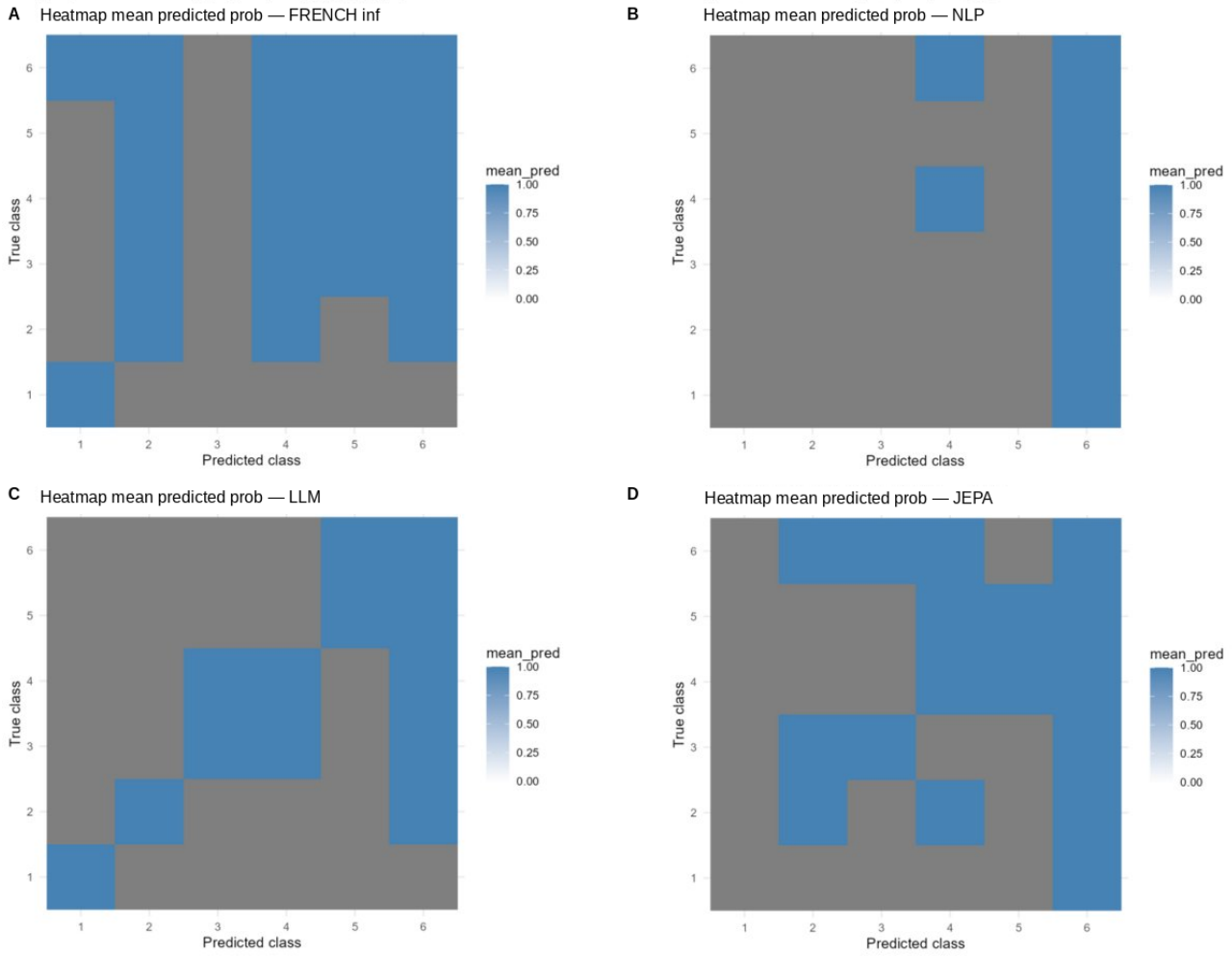
Class-specific calibration analysis showed the following patterns:

- Class 1: Predictions were overconfident, with limited probability range coverage because of extreme sparsity.
- Class 2: Predictions were dominated by high-probability values (>0.8), indicating limited use of intermediate confidence levels.
- Class 3: Displayed a partially calibrated diagonal trend, but with inconsistencies across probability bins.
- Class 4: Demonstrated the most balanced calibration, with broader probability coverage and closer agreement between predicted and observed frequencies.

These findings highlight substantial class-specific differences and underscore the need for post hoc calibration methods (eg, temperature scaling or isotonic regression) before clinical deployment to ensure reliable probability estimates for risk stratification.

XSL•FO
RenderX

Figure 5. Mean predicted probability heatmaps for the four triage models. Each panel displays, for each true triage class (y-axis, 1–6), the mean predicted probability assigned to each predicted class (x-axis, 1–6), with blue intensity encoding probability magnitude (0–1): (A) FRENCH inf, (B) NLP, (C) LLM, and (D) JEPA — blue blocks concentrated along or near the diagonal indicate well-calibrated class-specific predictions. Perfect calibration would be indicated by high intensity only along the diagonal.



Risk Distribution Analysis

Ridge plot analysis (Figure 6) revealed distinct behavioral patterns across approaches. Nurse triage demonstrated relatively consistent distributional patterns, with moderate overlap between adjacent classes and clear modal peaks, suggesting reasonable discrimination but substantial uncertainty.

TRIAGEMASTER exhibited variable performance, with notable gaps in certain class predictions (particularly panels 1 and 3), indicating an inability to consistently predict across all triage categories. Distributions were concentrated around specific probability values, suggesting either high confidence in limited scenarios or potential overfitting.

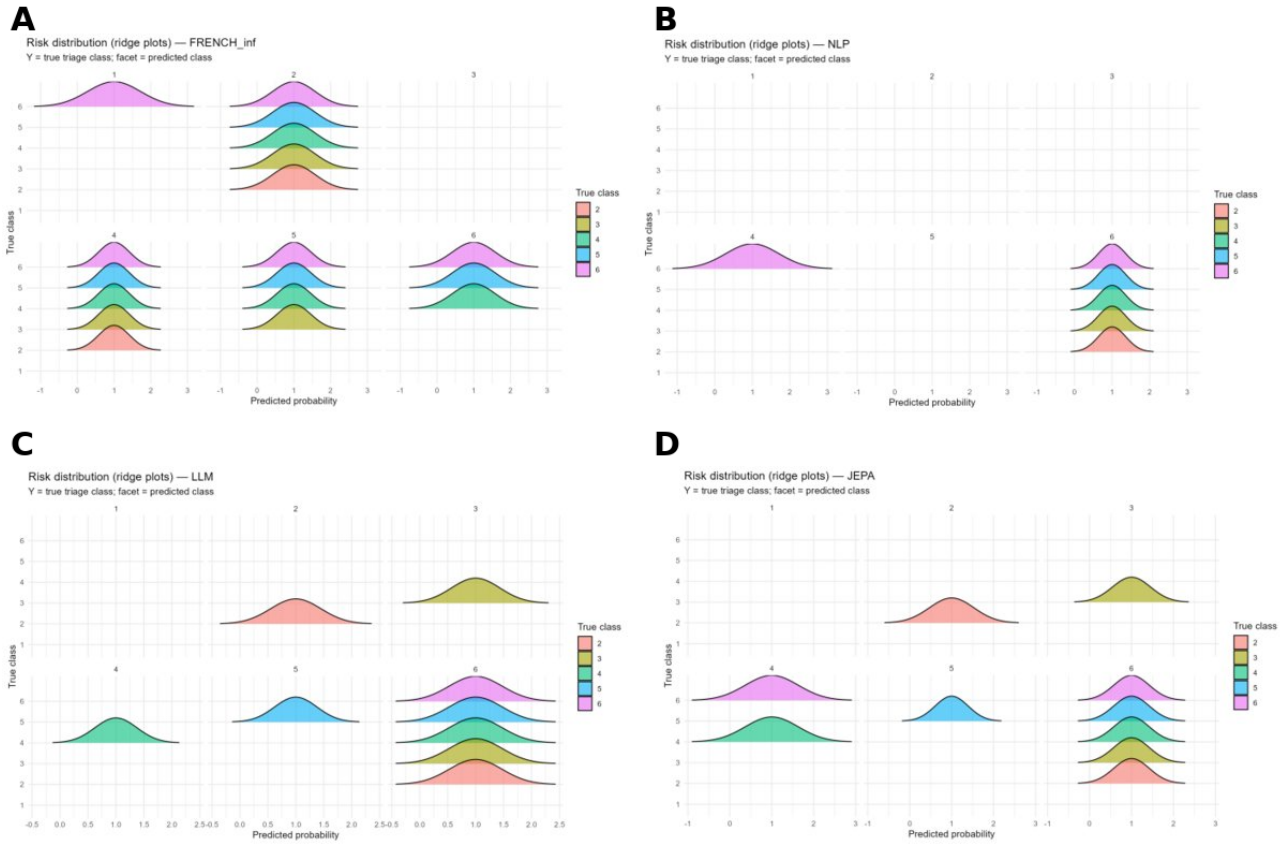
URGENTIAPARSE displayed markedly selective characteristics, with substantial gaps in class representation across prediction scenarios. Several triage classes appeared absent in certain combinations, suggesting that the LLM produced more selective predictions. When predictions were generated, distributions were tightly concentrated with distinct modal peaks, indicating high confidence. This sparse prediction pattern may reflect superior discrimination in clear-cut cases but raises concerns regarding coverage across diverse patient presentations. It may also suggest that the model effectively refrains from uncertain predictions—a behavior that requires explicit characterization and validation.

EMERGINET showed heterogeneous patterns, with some classes displaying broad, diffuse distributions and others exhibiting sharp, concentrated peaks, indicating variable predictive confidence across triage categories.

The distinct sparse pattern observed with URGENTIAPARSE warrants careful consideration for clinical implementation. Although it may indicate superior precision in unambiguous cases, the absence of predictions for certain class combinations presents challenges for comprehensive triage coverage. This pattern requires further validation to ensure adequate performance across all patient presentations, including edge cases and ambiguous scenarios.

XSL•FO
RenderX

Figure 6. Risk score distributions by predicted triage class across models (ridge plots). For each model, predicted probability distributions are shown separately per predicted class (facets 1–6, columns) and colored by true triage class (2–6, see legend). (A) FRENCH_inf model (French clinical rule-based inference). (B) NLP model (natural language processing). (C) LLM model (large language model). (D) JEPA model (Joint Embedding Predictive Architecture). The x-axis represents the predicted probability; the y-axis represents the true triage class. Substantial overlap between distributions within a facet indicates high inter-class confusion for that predicted class.



Secondary Analysis: GEMSA Prediction

URGENTIAPARSE substantially outperformed the other models in predicting patient disposition (GEMSA), achieving a composite *z* score of 2.473 versus –2.056 for TRIAGEMASTER and –4.399 for EMERGINET (Table 3). Performance metrics included an MAE of 0.082, RMSE of 0.402, weighted κ of 0.863 (P<.001), Spearman correlation of 0.864 (P<.001), micro F_1-score of 0.957, and macro F_1-score of 0.628. High exact agreement (95.7%) and near agreement (96.0%) further confirmed its accuracy and reliability.

Confusion matrices (Figure 7, top) showed that URGENTIAPARSE produced highly concentrated predictions along the diagonal (accurate), EMERGINET demonstrated moderate accuracy, and TRIAGEMASTER displayed scattered predictions. Bland-Altman analysis (Figure 7, middle) confirmed strong agreement for URGENTIAPARSE (mean difference 0.04, limits –0.83 to 0.91), moderate agreement for EMERGINET (limits –2.12 to 1.96), and large deviations for TRIAGEMASTER (limits –2.51 to 2.25).

Error distributions (Figure 7, bottom left) showed that EMERGINET had small, tightly clustered errors; URGENTIAPARSE exhibited slightly broader but well-centered errors; and TRIAGEMASTER demonstrated wide, poorly centered distributions. Class-specific F_1-scores (Figure 7, bottom right) revealed that URGENTIAPARSE achieved consistently high scores across the represented GEMSA categories, whereas EMERGINET and TRIAGEMASTER showed poorer and more uneven performance.

XSL•FO
RenderX

Table 3. Secondary analysis comparing artificial intelligence models for GEMSA[a] (disposition) prediction.

Model	Mean absolute error	Root mean square error	κ[b]	Spearman[c]	Micro F_1-score	Macro F_1-score	Exact	Near[d]	z score[e]
URGENTIAPARSE	0.082	0.402	0.863	0.864	0.957	0.628	0.957	0.960	2.473
TRIAGEMASTER	0.364	0.863	0.185	0.240	0.811	0.198	0.811	0.840	–2.056
EMERGINET	0.988	1.009	0	0	0.027	0.009	0.027	0.985	–4.399

[a]GEMSA: Groupes d'Etude Multicentrique des Services d'Accueil.

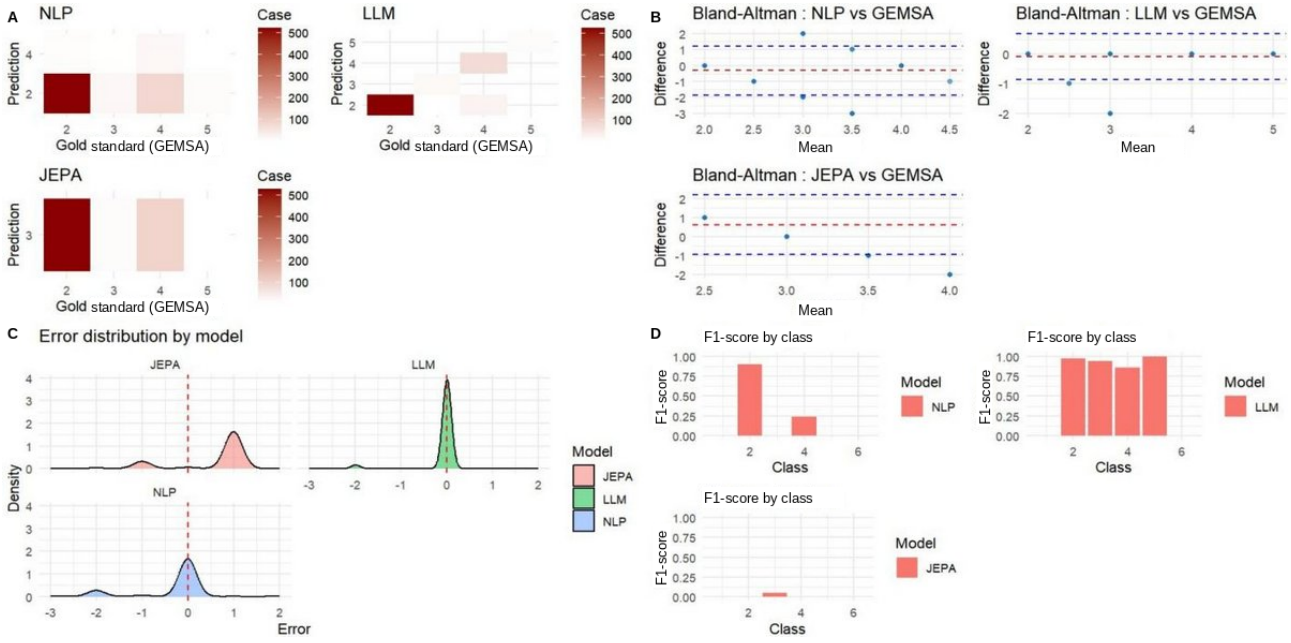
[b]Weighted Cohen κ; URGENTIAPARSE and TRIAGEMASTER, *P*<.001; EMERGINET, *P*=.99.

[c]Spearman rank correlation; URGENTIAPARSE, *P*<.001; TRIAGEMASTER, *P*=.006; EMERGINET, *P*=.99.

[d]Near agreement: proportion within the +1 or –1 GEMSA level.

[e]Composite *z* score (higher is better).

Figure 7. Evaluation of three AI triage models against the GEMSA gold standard. (A) Confusion matrices comparing predicted versus gold standard triage levels for NLP, LLM, and JEPA. (B) Bland-Altman plots showing bias and limits of agreement between each model and GEMSA. (C) Kernel density distributions of prediction errors illustrating model precision and systematic bias. (D) Per-class F1-scores quantifying discriminative performance across triage severity levels.



Impact of Data Modality: Structured Versus Unstructured Inputs

Analysis of input data type impact (Table 4) revealed that structured data (vital signs, demographics, and chief complaint codes) yielded slightly better performance than unstructured transcripts for URGENTIAPARSE (*z* score 1.806 vs 1.628) and TRIAGEMASTER (–4.790 vs –4.949), whereas EMERGINET showed minimal difference (–1.357 vs –0.993). These findings suggest that (1) vital signs and demographic data contain substantial triage-predictive information; (2) LLMs can extract meaningful patterns from both modalities; and (3) optimal performance likely requires multimodal integration. The relatively small performance gap between modalities for URGENTIAPARSE indicates that the transcript-based LLM captures much of the clinically relevant information, supporting the feasibility of voice-based triage assistance systems.

XSL•FO
RenderX

Table 4. Performance comparison across data modalities (structured vs unstructured inputs).

Model	Data type	Mean absolute error	Root mean square error	κ[a]	Spearman[b]	Micro F_1-score	Macro F_1-score	Exact	Near[c]	*z* score[d]
URGENTIAPARSE	Structured	0.213	0.748	0.821	0.826	0.903	0.898	0.903	0.928	1.806
URGENTIAPARSE	Unstructured	0.240	0.807	0.790	0.790	0.893	0.891	0.893	0.924	1.628
EMERGINET	Structured	0.510	1.247	0.420	0.559	0.807	0.551	0.807	0.836	–1.357
EMERGINET	Unstructured	0.516	1.240	0.430	0.560	0.791	0.539	0.791	0.843	–0.993
TRIAGEMASTER	Structured	0.916	1.557	0.057	0.160	0.572	0.190	0.572	0.712	–4.790
TRIAGEMASTER	Unstructured	0.991	1.629	-0.003	0.001	0.545	0.118	0.545	0.691	–4.949

[a]Weighted Cohen κ; $P<.001$ for all except TRIAGEMASTER unstructured ($P=.97$).

[b]Spearman rank correlation; $P<.001$ for all except TRIAGEMASTER ($P>.20$).

[c]Near agreement: proportion within the +1 or –1 triage level.

[d]Composite *z* score (higher is better).

Discussion

Principal Findings

This retrospective proof-of-concept study developed and compared 3 AI architectures for ED triage prediction using real nurse-patient interaction data and national triage standards. Of 73,236 ED visits over a 7-month period, 657 (0.90%) patients with complete audio recordings and structured data were included in the analysis. The LLM-based model (URGENTIAPARSE) demonstrated superior performance across multiple metrics (composite *z* score=2.514, F_1-score=0.900, AUC 0.879, weighted κ=0.800, $P<.001$) compared with the JEPA-based EMERGINET (*z* score=0.438), the NLP-based TRIAGEMASTER (*z* score=–3.511), and current nurse triage practices (*z* score=–4.343). Secondary analyses confirmed the superiority of URGENTIAPARSE for disposition prediction (GEMSA) and its robustness across structured and unstructured data modalities. However, severe methodological limitations—particularly extreme selection bias, overfitting, sparse representation of high-acuity cases, and the monocentric design—preclude immediate clinical deployment and underscore the need for substantial additional validation.

Comparison With Prior Work

Our findings align with and extend previous research on AI-assisted triage. The URGENTIAPARSE F_1-score (0.900) substantially exceeds that of prior French ED triage models: Clinical BELGPT-2 achieved an F_1-score of 0.62 [12], while external models such as Mistral 7B and Llama 3 8B demonstrated moderate performance, with larger models (Mixtral 8x7B) underperforming [11]. Our results also surpass those reported in the original FRENCH scale validation studies (AUC 0.83-0.86 for hospitalization prediction [1,25]) when applied by nurses, suggesting that AI has the potential to enhance triage accuracy.

However, unlike previous studies, we observed severe overfitting despite the use of regularization techniques. This likely reflects the extremely small effective sample size relative to model complexity (657 cases for the 137-million-parameter FlauBERT base), the high-dimensional input space (768-dimensional embeddings plus structured features), and substantial class imbalance with sparse representation of high-acuity cases. Recent studies on LLM hallucinations [13] and time-series forecasting [26] suggest potential mitigation strategies, including semantic entropy–based uncertainty quantification, prompt engineering, and backbone reprogramming—approaches that warrant investigation in future iterations.

Our calibration analysis revealed class-specific heterogeneity, consistent with recent findings on proper scoring rules for multiclass classification [14]. The observation that URGENTIAPARSE generates sparse, high-confidence predictions rather than well-distributed probability estimates suggests that the model may benefit from postprocessing approaches such as temperature scaling or Platt scaling to improve probability calibration before clinical implementation.

Prior studies have identified biases in triage decisions related to patient sex, nurse experience, and pain reporting [11,12]. Although our study did not explicitly assess these biases, the gold-standard consensus process and AI model evaluation framework established here could facilitate future bias detection studies. For example, systematic perturbation analyses, such as modifying patient sex references in transcripts while holding all other factors constant, could be conducted to evaluate potential model sensitivity to demographic attributes.

Clinical Implications and Deployment Considerations

If validated prospectively across diverse settings, LLM-based triage support could enhance patient safety and operational efficiency through real-time decision support that highlights high-risk cases for immediate review; quality assurance by flagging discrepancies between nurse and AI triage for senior clinician verification; training tools that provide immediate feedback to novice nurses; and workload redistribution during surge periods by prescreening low-acuity cases. However, multiple critical barriers must be addressed before deployment.

The perfect training accuracy (1.0) with poor validation performance (~0.5) observed for URGENTIAPARSE indicates

XSL•FO
RenderX

severe memorization rather than learning generalizable patterns. Potential solutions include substantial data augmentation, such as paraphrasing, synonym replacement, and back-translation for transcripts, combined with controlled perturbations of vital signs. More aggressive regularization strategies, including higher dropout rates, stronger L2 penalties, and techniques such as mixup or cutmix, may also help. Model simplification through adapter layers or low-rank adaptation (LoRA) fine-tuning of FlauBERT, rather than full model fine-tuning, could further reduce overfitting. In addition, ensemble methods that average predictions from multiple diverse architectures, transfer learning from related domains (eg, ambulance call triage or primary care urgency assessment), and active learning to selectively acquire cases with the highest model uncertainty all warrant investigation.

With only 4 (N=657) class 1 cases (0.61%) and 86 (13.1%) class 2 cases, we cannot reliably evaluate model performance for the most critical patients—those for whom undertriage could be fatal. Prospective deployment would therefore require conservative threshold selection that prioritizes overtriage over undertriage; mandatory senior clinician override for any AI-suggested downgrade relative to nurse triage; and real-time monitoring dashboards tracking undertriage events and associated patient outcomes, including ED length of stay, ICU admission, in-hospital mortality, and 72-hour return visits. In addition, automatic model deactivation should be triggered if the undertriage rate exceeds predefined safety thresholds.

Current probability estimates also require post hoc calibration—using methods such as temperature scaling, Platt scaling, or isotonic regression—to ensure reliable confidence estimates. Furthermore, implementing abstention mechanisms, whereby the model declines to issue a prediction when uncertainty exceeds a predefined threshold, may improve safety by deferring ambiguous cases to human judgment. Techniques from conformal prediction could additionally provide statistically rigorous uncertainty quantification with formal coverage guarantees.

Although SHAP values provide token- and feature-level explanations for URGENTIAPARSE predictions, clinicians require intuitive, actionable insights that align with real-world decision-making. Future work should therefore focus on generating natural language explanations—for example: “Triage level 2 suggested due to elevated heart rate (124 bpm), chest pain complaint, and description of respiratory distress.” Contrastive explanations could further enhance interpretability and trust, such as “If the heart rate were within the normal range, the predicted triage level would be 3.” Such counterfactual framing helps clinicians understand which features are most influential and how changes in patient presentation might alter the recommendation.

Real-time deployment requires inference latency under 5 seconds, a requirement met by the current 312-ms latency. However, seamless electronic health record integration without disrupting nurse workflow remains essential. An intuitive user interface that displays AI recommendations with confidence levels and key supporting features, combined with easy override mechanisms requiring mandatory documentation of the rationale, would facilitate adoption. Establishing feedback loops in which clinician corrections are logged for continuous model improvement completes the integration requirements.

AI triage support raises profound liability questions, with unclear responsibility among nurses, AI developers, and hospitals when undertriage occurs. Algorithmic bias is another concern, as model performance may vary by patient demographics, insurance status, or language. Privacy considerations regarding the secure handling of audio recordings and transcripts must be addressed, and equity requires ensuring that all patient subgroups receive accurate predictions. Comprehensive fairness audits, bias mitigation strategies, and clear governance frameworks must precede any deployment.

Limitations

This study has multiple substantial limitations that constrain interpretation and generalizability across several domains.

The severe selection bias represents perhaps the most concerning methodological limitation. Only 657 of 73,236 (0.90%) ED visits had usable recordings, representing <1% of eligible encounters. Recordings were limited to volunteer nurses (fewer than 10) during feasible moments, likely excluding the highest-acuity cases requiring immediate intervention without time for a structured interview; chaotic shifts with multiple simultaneous arrivals; patients with altered mental status, language barriers, or communication difficulties; nurses uncomfortable with recording technology; and specific periods, such as overnight and weekends. The included cohort had slightly lower acuity than the broader ED population (P=.02), introducing systematic bias that limits generalizability to the real-world case mix. Future studies must implement systematic recording protocols that capture consecutive cases across all shifts, nurses, and patient types to obtain representative samples.

The sparse representation of the highest-acuity classes constitutes the most critical limitation for clinical deployment. Out of 657 cases, class 1 included only 4 (0.61%), class 2 included 86 (13.1%), and no class 3A or class 6 cases were observed. This extreme imbalance severely limits model training on high-risk cases, where undertriage has the greatest consequences. Reliable performance estimation for rare but critical categories becomes impossible, as CIs for class 1 metrics are extremely wide. The ability to detect undertriage patterns that could lead to preventable morbidity or mortality therefore remains fundamentally constrained. Generalization to settings with different acuity distributions also cannot be assessed. Class-specific performance metrics, particularly sensitivity for classes 1-2, have limited reliability and require validation in datasets with adequate representation across the full triage spectrum. Techniques such as the synthetic minority oversampling technique or focal loss to address extreme class imbalance should be explored in future work.

The severe overfitting observed for URGENTIAPARSE fundamentally limits real-world applicability. The model achieved perfect training accuracy (1.0), whereas validation accuracy remained approximately 0.5, clearly indicating that the model memorized training examples rather than learning generalizable triage patterns. This overfitting persisted despite

XSL•FO
RenderX

dropout (P=0.3), L2 regularization ($\lambda=1\times10^{-5}$), and early stopping, suggesting that the effective sample size (n=657) is insufficient for the model complexity (137 million parameters in FlauBERT base). Although we report validation set performance rather than training set performance, the marked training-validation discrepancy signals a high risk of catastrophic failure on truly external data with different patient populations, nurse communication styles, or institutional protocols. Prospective external validation across multiple, diverse EDs is therefore essential before any consideration of clinical use. Potential solutions include more aggressive regularization, model simplification through adapter layers or LoRA fine-tuning, ensemble methods, or the collection of substantially larger training datasets, with a target of over 10,000 cases for stable LLM fine-tuning.

The monocentric design severely constrains external validity. Conducted at a single French academic tertiary care hospital, the results may not generalize to community hospitals with different patient demographics, resource levels, and triage processes. Rural or low-volume EDs, with different case mix and staffing patterns, represent an entirely different operational context. Non-French health care systems using alternative triage scales, such as the Emergency Severity Index, Canadian Triage and Acuity Scale, Manchester Triage System, or Australasian Triage Scale, would require complete model retraining and revalidation. Non-French language contexts would likewise necessitate retraining language models on appropriate corpora. EDs serving culturally, linguistically, or socioeconomically distinct patient populations may exhibit fundamentally different triage patterns. Multicenter validation across diverse settings—including academic and community hospitals, urban and rural locations, and multiple countries—is therefore required to assess transportability and identify factors that affect model performance.

The exclusion of repeat ED visits creates an important gap in our understanding of model performance. By including only index visits, we cannot assess model behavior in patients with frequent ED utilization, complex chronic conditions requiring a longitudinal perspective, or evolving clinical presentations across multiple encounters. Repeat visitors often present atypically and may benefit most from AI-assisted triage, yet they were systematically excluded from model development. This exclusion may bias performance estimates upward if repeat visitors are more difficult to triage accurately.

Several concerns regarding gold-standard construction could introduce incorporation bias. Adjudicators reviewed complete transcripts and vital signs with unlimited time for deliberation, in sharp contrast to real-time nurse triage conducted under time pressure and with incomplete information. This fundamental difference in decision-making context may overestimate AI-human agreement, as both the AI and gold-standard adjudicators had access to complete information unavailable to real-time nurses. Although blinded to outcomes, adjudicators were aware that they were establishing a research gold standard rather than making operational decisions with immediate consequences, which may have influenced risk tolerance and the willingness to assign higher acuity levels. The initial interrater agreement (κ=0.72) indicates substantial but not perfect consensus, with a 23.5% disagreement rate suggesting meaningful variability in expert interpretation even with complete information. The gold standard therefore reflects retrospective expert judgment, which may differ from optimal real-time triage, particularly when information available only after initial presentation is considered. A sensitivity analysis excluding cases with nurse-gold-standard discrepancies over 1 level yielded similar results, partially mitigating this concern but not eliminating it.

Limited independent validation of the FRENCH scale itself represents an often overlooked limitation. Although the FRENCH scale serves as the French national standard, independent validation studies remain limited beyond the original 2009 and 2019 publications [1,25]. The original validation reported an AUC of 0.858 for hospitalization prediction and moderate correlations with resource utilization. A 2019 validation study including 29,423 patients reported similar findings, with an AUC of 0.83 and moderate but significant correlations between triage level and resource utilization or testing [25]. Thus, even compared with FRENCH validation studies, the URGENTIAPARSE model demonstrates superior AUC performance relative to nurses who conducted triage in those studies. However, our gold-standard consensus process effectively establishes an “ideal FRENCH application” and does not validate whether FRENCH itself optimally predicts patient outcomes. Future research should therefore evaluate AI predictions against objective clinical end points, such as ICU admission, in-hospital mortality, 72-hour return visits, and length of stay, independent of triage scale assignment, to assess true clinical utility rather than agreement with an imperfect reference standard.

The lack of prospective validation and real-world testing fundamentally limits our ability to make clinical recommendations. This retrospective analysis cannot assess real-time inference performance or the integration challenges that inevitably arise during deployment. The impact on actual triage decisions and nurse acceptance remains unknown and could substantially affect real-world utility. Effects on patient outcomes, ED workflow, and resource utilization also cannot be determined from retrospective data. Model degradation over time, as patient populations, disease patterns, or institutional practices evolve through concept drift, represents a realistic threat that requires prospective monitoring. Rare failure modes and edge cases typically become apparent only during large-scale deployment involving thousands or tens of thousands of predictions. Unintended consequences, such as deskilling of nurses who become overly reliant on AI recommendations, or gaming of the system if performance metrics become targets, could further undermine potential benefits. Prospective stepped-wedge cluster randomized trials comparing AI-assisted versus standard triage across multiple sites, therefore, represent essential next steps before any clinical implementation can be responsibly considered.

The absence of cost-effectiveness and implementation analyses limits informed adoption decisions. We did not evaluate the computational infrastructure costs required for deployment, including graphics processing unit servers, maintenance, and

XSL•FO
RenderX

updates, which may be substantial for resource-constrained hospitals. Training costs for nurses to use the system effectively also remain unknown and could be prohibitive if extensive instruction is required. The impact on triage time and ED throughput could be positive or negative, depending on implementation details. Cost per quality-adjusted life year gained represents the gold standard for health care economic evaluation, but requires long-term outcome data that are currently unavailable. Scalability challenges related to broad dissemination across diverse health care settings also remain unaddressed. Economic evaluation from a health system perspective is therefore essential for adoption decisions in the context of finite health care resources.

The absence of fairness and bias assessment represents a significant ethical limitation. We did not systematically evaluate whether model performance varies by patient demographics—including age, sex, race or ethnicity, insurance status, language, or socioeconomic status—potentially exacerbating health disparities. Prior work has identified sex-based and other biases in human triage decisions [11,12]; AI systems may perpetuate or amplify these biases if they are not explicitly addressed during development. If the model performs worse for already disadvantaged populations, deployment could worsen existing health inequities. Comprehensive fairness audits stratified by protected characteristics are therefore required before deployment, along with bias mitigation strategies if disparities are identified. Partnerships with ethicists, patients, and advocacy groups to define fairness metrics aligned with societal values should precede any clinical implementation.

Technological and data quality constraints affected the study throughout. Audio recording quality varied substantially, leading to transcription errors that may have affected model input quality in unpredictable ways. Equipment loss and technical failures resulted in 24 exclusions (3.5% of recorded cases), suggesting that real-world deployment would encounter similar challenges. Missing data for certain vital signs (<2%) required imputation using training set medians, which may have introduced bias. Real-world deployment would face these and additional challenges, necessitating robust handling of noisy, incomplete, or corrupted inputs with graceful degradation rather than catastrophic failure. The model's behavior when confronted with poor-quality audio, incomplete vital signs, or technical failures remains unknown.

The lack of uncertainty quantification and explicit abstention mechanisms precludes safe clinical deployment. The model currently provides point predictions without well-calibrated CIs or explicit abstention when uncertainty is high. Clinical deployment requires principled uncertainty quantification—using methods such as conformal prediction, Bayesian approaches, or ensemble disagreement—to determine when deferral to human judgment is warranted. Without such mechanisms, the model may generate highly confident yet incorrect predictions in novel situations outside its training distribution. The sparse prediction patterns observed for URGENTIAPARSE may reflect implicit abstention behavior that should be explicitly characterized; however, we have not systematically examined when or why the model declines to issue predictions.

Limited interpretability persists despite SHAP analysis. Although SHAP values provide mathematical attributions indicating which input features most influenced each prediction, they may not align with clinical reasoning patterns or offer actionable insights for clinicians. A nurse or physician may not find numerical feature importance scores intuitive or helpful in understanding why the AI generated a specific recommendation. Natural language explanation generation that translates SHAP values into clinically meaningful statements remains underdeveloped. User studies assessing interpretability from the clinician perspective are therefore needed to determine whether these explanations facilitate, rather than hinder, clinical decision-making.

Regulatory and legal uncertainties cloud the path to clinical implementation. This research prototype has not undergone regulatory review, such as Conformité Européenne (CE) marking in Europe or Food and Drug Administration (FDA) clearance in the United States, which would likely be required for medical device classification of clinical decision support software. The regulatory pathway for AI-based triage support systems remains unclear and continues to evolve. Legal liability frameworks for AI-assisted triage are also poorly defined. When undertriage occurs with AI assistance, responsibility among nurses, AI developers, and hospitals remains ambiguous. These legal and regulatory uncertainties must be resolved before widespread deployment can proceed.

Finally, the absence of a continual learning framework limits long-term sustainability. Static models trained on historical data inevitably degrade as patient populations, diseases, and clinical practices evolve. Deployment would therefore require active monitoring of model performance in production; detection of performance degradation or concept drift; triggered retraining protocols when performance declines; and potentially online learning mechanisms to adapt to changing patterns while maintaining safety. We have not developed or validated any such infrastructure. Without continual learning capabilities, the model would become progressively obsolete, even if initially successful.

Despite these numerous and substantial limitations, this study provides important proof-of-concept evidence that LLM architectures can achieve high accuracy for ED triage prediction when evaluated using carefully collected retrospective data. The study also establishes a rigorous evaluation methodology that incorporates multiple complementary metrics and comprehensive error analysis, upon which future research can build. The limitations outlined above provide a detailed road map for the substantial additional work required before clinical translation can be responsibly considered, including the specific methodological improvements, additional studies, and infrastructure development necessary to support safe and effective deployment.

Future Research Directions

Building on these findings and addressing the identified limitations, future research should prioritize several key areas.

XSL•FO
RenderX

Trials across more than 10 diverse EDs—including academic and community hospitals, urban and rural settings, and varied patient demographics—with systematic consecutive case enrollment targeting more than 10,000 patients would enable assessment of external validity, detection of site-specific effects, and adequately powered subgroup analyses. A stepped-wedge cluster randomized design would allow causal inference regarding the impact on patient outcomes, including mortality, length of stay, ICU admission rate, and 72-hour return visits, while controlling for temporal trends.

Oversampling the highest-acuity encounters in classes 1-2 through deliberate recording protocols during trauma activations, stroke alerts, cardiac arrests, and other critical presentations would enable reliable detection of undertriage and robust safety evaluation. This targeted data collection would address one of the most important current limitations.

Systematic comparison of advanced overfitting mitigation strategies is also warranted. Such evaluation should include parameter-efficient fine-tuning approaches (eg, LoRA, adapters, and prompt tuning); aggressive data augmentation (eg, back-translation, paraphrasing, synonym replacement, and controlled vital sign perturbations); semisupervised learning leveraging unlabeled transcripts; multitask learning jointly predicting triage, disposition, resource utilization, and outcomes; transfer learning from related domains such as ambulance call triage or primary care urgency assessment; and ensemble methods combining diverse architectures with different inductive biases.

Systematic assessment of model behavior under a range of perturbations would strengthen confidence in its robustness. These should include adversarial attacks representing intentional gaming; distribution shifts due to seasonal variation, pandemics, or novel diseases; missing data resulting from incomplete vital signs, silent patients, or technical failures; transcription errors arising from speech recognition failures; and edge cases, such as rare presentations, complex comorbidities, or psychiatric crises. Developing certified robustness guarantees represents an important objective.

Stratified performance evaluation by protected characteristics—including age, sex, race or ethnicity, insurance status, language, and socioeconomic status—combined with detection of disparate impact, root cause analysis of identified biases, and rigorous testing of mitigation strategies (eg, reweighting, fairness constraints, or adversarial debiasing) would address equity concerns. Partnerships with ethicists, patients, and advocacy groups to define fairness metrics aligned with societal values are essential.

Development of natural language explanation generation, contrastive explanations that articulate how predictions would change if specific features differed, visualization tools highlighting salient features, and user interface optimization informed by clinician feedback would enhance interpretability. Human factors studies evaluating how AI recommendations influence decision-making—including risks of automation bias (overreliance) or dismissiveness (underreliance)—would further inform safe implementation.

Rigorous cost-effectiveness analysis from a health system perspective; budget impact modeling; implementation barrier assessment; stakeholder engagement with nurses, physicians, administrators, and patients; workflow redesign; training program development; and sustainment planning would support adoption. Qualitative studies exploring nurse and physician attitudes, concerns, and acceptance would help identify potential obstacles early.

Infrastructure for continual model performance monitoring, concept drift detection, triggered retraining protocols, online learning mechanisms, and transparent reporting to end users and regulators would maintain performance over time. Development of living systematic reviews tracking field performance as deployment expands would build the evidence base.

Exploration of synergies with AI-assisted diagnosis supporting downstream care decisions, predictive models for ED crowding and resource planning, automated documentation reducing nurse burden, and early warning systems for deterioration after initial triage could create an integrated ecosystem. Emerging work on AI transforming health care from acute care to early prevention [27,28] suggests potential for broader impact across the care continuum.

Collaboration with regulatory agencies, including the European Medicines Agency and FDA, to establish appropriate oversight frameworks balancing innovation with patient safety, defining clinical validation requirements, and clarifying liability and postmarket surveillance obligations for AI medical devices would facilitate responsible translation.

Conclusions

This study demonstrates that LLM-based AI (URGENTIAPARSE) can achieve high accuracy for ED triage prediction, outperforming traditional NLP (TRIAGEMASTER), JEPA (EMERGINET), and current nurse triage practices when validated against a senior clinician consensus gold standard. With an F_1-score of 0.900, AUC of 0.879, weighted κ of 0.800, and 90% exact agreement, the LLM model shows promise for enhancing triage consistency and potentially supporting clinical decision-making.

However, severe methodological limitations mandate extreme caution. Critical barriers to clinical deployment include severe overfitting, with perfect training accuracy but poor validation performance indicating an inability to generalize beyond the training data; extreme selection bias, with only a 0.9% (657/73,236) inclusion rate limiting representativeness; sparse high-acuity case representation, with only 4 class 1 cases precluding reliable undertriage risk assessment where stakes are highest; a monocentric design constraining external validity; retrospective analysis without prospective real-world validation; and absent economic, workflow, fairness, and regulatory analyses.

Before any consideration of clinical deployment, this AI triage support system requires large-scale prospective multicenter validation targeting more than 10,000 cases across diverse settings; comprehensive overfitting mitigation and model robustness enhancement; targeted high-acuity case collection

XSL•FO
RenderX

enabling undertriage safety evaluation; fairness audits and bias mitigation strategies; post hoc calibration and uncertainty quantification with abstention mechanisms; interpretability enhancement with natural language explanations; human factors studies and workflow integration testing; economic evaluation and implementation planning; regulatory review and approval; and continuous performance monitoring infrastructure with concept drift detection.

While this proof-of-concept establishes technical feasibility and identifies the most promising architecture among the tested alternatives, the substantial gap between current evidence and safe clinical deployment must be acknowledged transparently. AI-assisted triage holds potential to improve patient safety, enhance consistency, reduce cognitive burden on nurses during surge periods, and support quality assurance—but realizing these benefits responsibly requires rigorous validation, thoughtful implementation, continuous monitoring, and humility about current limitations. This study provides a foundation and road map for that essential future work.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TRIPOD-AI fullfilled checklist.
[PDF File (Adobe PDF File), 302 KB - medinform_v14i1e83318_app1.pdf]

Multimedia Appendix 2
Code for gold-standard.
[ZIP File (Zip Archive), 32 KB - medinform_v14i1e83318_app2.zip]

References

1. Taboulet P, Moreira V, Haas L, Porcher R, Braganca A, Fontaine J, et al. Triage with the French Emergency Nurses Classification in Hospital scale: reliability and validity. European Journal of Emergency Medicine 2009;16(2):61-67. [doi: 10.1097/mej.0b013e328304ae57]
2. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000 May;53(5):459-468. [doi: 10.1016/s0895-4356(99)00206-1] [Medline: 10812317]
3. Moll HA. Challenges in the validation of triage systems at emergency departments. J Clin Epidemiol 2010 Apr;63(4):384-388. [doi: 10.1016/j.jclinepi.2009.07.009] [Medline: 19875271]
4. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med 2018 May;71(5):565-574.e2. [doi: 10.1016/j.annemergmed.2017.08.005] [Medline: 28888332]
5. Ilicki J. Challenges in evaluating the accuracy of AI-containing digital triage systems: a systematic review. PLoS One 2022 Dec 27;17(12):e0279636 [FREE Full text] [doi: 10.1371/journal.pone.0279636] [Medline: 36574438]
6. Hang L, Loïc V, Jibril F, Vincent S, Maximin C, Benjamin L, et al. FlauBERT: unsupervised language model pre-training for French. Paris, France: European Language Resources Association; 2020 Presented at: The Twelfth Language Resources and Evaluation Conference; May 11-16, 2020; Marseille, France.
7. Brossard C, Goetz C, Catoire P, Cipolat L, Guyeux C, Gil Jardine C, et al. Predicting emergency department admissions using a machine-learning algorithm: a proof of concept with retrospective study. BMC Emerg Med 2025 Jan 06;25(1):3-11 [FREE Full text] [doi: 10.1186/s12873-024-01141-4] [Medline: 39762754]
8. Bardes A, Ponce J, LeCun Y. Vicreg: variance-invariance-covariance regularization for self-supervised learning. arXiv. Published online May 11, 2021. 2022. URL: https://arxiv.org/abs/2105.04906 [accessed 2026-02-25]
9. Lansiaux E, Baron M, Vromant A. Navigating the landscape of medical triage: unveiling the potential and challenges of large language models and beyond. Am J Emerg Med 2024 Apr;78:224. [doi: 10.1016/j.ajem.2024.02.013] [Medline: 38395695]
10. Lansiaux E, Ibrahim ES, Minaz A, Itani R, Kukkadi R, Obina OG, et al. Medical triage review: from battlegrounds darkness to the machine light. SSRN. Published online February 13, 2024. 2024. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4713040 [accessed 2026-02-25]
11. Guerra-Adames A, Avalos-Fernandez M, Doremus O, Gil-Jardiné C, Lagarde E. Uncovering judgment biases in emergency triage: a public health approach based on large language models. 2025 Dec Presented at: The 4th Machine Learning for Health Symposium; December 15-16, 2025; Vancouver, BC, Canada p. 402-439.

XSL•FO
RenderX

12. Avalos M, Cohen D, Russon D, Davids M, Doremus O, Chenais G, et al. Detecting human bias in emergency triage using LLMs. 2024 May 13 Presented at: The International FLAIRS Conference Proceedings; May 18-21, 2024; Sandestin Beach, FL. [doi: 10.32473/flairs.37.1.135586]
13. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun 19;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
14. Ahmadian R, Ghatee M, Wahlström J. Superior scoring rules for probabilistic evaluation of single-label multi-class classification tasks. International Journal of Approximate Reasoning 2025 Jul;182:109421. [doi: 10.1016/j.ijar.2025.109421]
15. Statistique annuelle des etablissements de Santé. DRESS. URL: https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/00-la-statistique-annuelle-des-etablissements-sae [accessed 2026-02-25]
16. Lansiaux E. TIAEU: Triage Intelligent A l'Entrée des Urgences. Health Data Hub. 2023. URL: https://health-data-hub.fr/projets/tiaeu-triage-intelligent-lentree-des-urgences-0 [accessed 2026-02-25]
17. Circulaire DHOS/E 3/PMSI n°-433 du 10 septembre relative au projet de mise en oeuvre progressive d?un financement propre aux urgences dans les établissements de santé. Ministère de la Santé. 2001 Jul. URL: https://sante.gouv.fr/fichiers/bo/2001/01-42/a0422740.htm [accessed 2026-02-25]
18. Milokov T, Chen K, Corrado GS, Dean J. Efficient estimation of word representations in vector space. 2013 Presented at: International Conference on Learning Representations; May 2-4, 2013; Scottsdale, AZ.
19. Ramachandran P, Zoph B, Le QV. Searching for activation functions. 2018 Presented at: International Conference on Learning Representations; April 30 to May 3, 2018; Vancouver, BC, Canada.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837. [doi: 10.2307/2531595]
21. Derczynski L. Complementarity, F-score, and NLP evaluation. 2005 Feb Presented at: The Tenth International Conference on Language Resources and Evaluation (LREC 16); May 23-28, 2016; Portorož, Slovenia p. 261-266. [doi: 10.1007/s10579-005-2692-5]
22. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. BMC Med Inform Decis Mak 2021 Dec 18;21(1):351-359 [FREE Full text] [doi: 10.1186/s12911-021-01722-4] [Medline: 34922532]
23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: NIPS'17: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 4768-4777.
24. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 2019 Jul;31(7):1235-1270. [doi: 10.1162/neco_a_01199] [Medline: 31113301]
25. Taboulet P, Vincent-Cassy C, Squara P, Resche-Rigon M. Validité de la FRENCH, l'échelle de tri des urgences hospitalières élaborée par la Société française de médecine d'urgence. Ann Fr Med Urgence 2019 Jan 15;9(1):10-16. [doi: 10.3166/afmu-2018-0099]
26. Jing M, Wang S, Ma L, Chu Z, Zhang JY, Shi XL, et al. Time-llm: time series forecasting by reprogramming large language models. 2023 Presented at: International Conference on Learning Representations; May 1-5, 2023; Kigali, Rwanda.
27. Ogut E. Graph-theoretical mapping of cortical and subcortical network alterations in preclinical neurodegeneration. Discov Neurosci 2025 Sep 15;20(1):e1. [doi: 10.1186/s13064-025-00214-9]
28. Ogut E. Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. Clin Pract 2025 Sep 16;15(9):169 [FREE Full text] [doi: 10.3390/clinpract15090169] [Medline: 41002784]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
CE: Conformité Européenne
CHU: Centre Hospitalier Universitaire
CLS: classification
CNIL: Commission Nationale de l'Informatique et des Libertés
ED: emergency department
FDA: Food and Drug Administration
FN: false negative
FP: false positive
FRENCH: French Emergency Nurses Classification in Hospital
GEMSA: Groupes d'Etude Multicentrique des Services d'Accueil
ICU: intensive care unit
JEPA: Joint Embedding Predictive Architecture
LLM: large language model

XSL•FO
RenderX

LoRA: low-rank adaptation
MAE: mean absolute error
NLP: natural language processing
RMSE: root mean square error
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanations
TP: true positive
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
VICReg: variance-invariance-covariance regularization
XGBoost: Extreme Gradient Boosting

Edited by A Benis; submitted 11.Sep.2025; peer-reviewed by AS Mutluer, E Ogut, J Kim; comments to author 11.Nov.2025; revised version received 14.Nov.2025; accepted 22.Dec.2025; published 10.Mar.2026.

Please cite as:
Lansiaux E, Azzouz R, Chazard E, Vromant A, Wiel E
Artificial Intelligence Models for Predicting Triage in Emergency Departments: Seven-Month Retrospective Comparative Study of Natural Language Processing, Large Language Model, and Joint Embedding Predictive Architectures
JMIR Med Inform 2026;14:e83318
URL: https://medinform.jmir.org/2026/1/e83318
doi:10.2196/83318
PMID:

©Edouard Lansiaux, Ramy Azzouz, Emmanuel Chazard, Amélie Vromant, Eric Wiel. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Application of a Large Visual Language Model on Tongue Image Description Generation and Physical Constitution Reasoning in Traditional Chinese Medicine (TongueVLM): Model Development and Validation Study

Chengdong Peng[1,2], MSc, PhD; Jun Gao[1], Prof Dr, PhD; Nuo Yang[2,3], BE; Yong Wang[2,3], BE; Renming Chen[2], MSc; Changwu Dong[4], Prof Dr Med

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui, China
[2]Artificial Intelligence Laboratory, Hefei Yunzhen Information Technology Co, Ltd, Hefei, Anhui, China
[3]School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, China
[4]The Second Clinical Medical School, Anhui University of Traditional Chinese Medicine, Hefei, Anhui, China

Corresponding Author:
Chengdong Peng, MSc, PhD
School of Computer Science and Information Engineering
Hefei University of Technology
Number 193 Tunxi Road
Hefei, Anhui, 230009
China
Phone: 86 055165337378 ext 86
Email: 2020020018@mail.hfut.edu.cn

Abstract

Background: In the field of traditional Chinese medicine (TCM), diagnostic work based on tongue images to recognize the physical constitution is a process of collecting clinical information, reasoning, and combining the patient's tongue image features with questioning. It is necessary to simulate the recognition of pathological information of tongue images by TCM practitioners and professional dialogue based on tongue image features, which helps to develop an intelligent interactive system for TCM diagnosis.

Objective: This study aimed to develop and validate a vertical model of the TCM domain with TCM's understanding and reasoning capability for tongue images.

Methods: A TongueVLM multimodal large model is designed, which includes a visual encoder module, a modal fusion module, and a language decoder module. First, the visual encoder based on the CLIP-ViT (Contrastive Language-Image Pre-Training With Vision Transformer) pretrained model is used for image patch, dimensionality reduction, and migration learning, which maps the high-dimensional tongue features into low-dimensional language encoding vectors. Further, a modal fusion module with a residual architecture is applied to map visual features to a natural language word embedding space, realizing the conceptual alignment between visual encoding and TCM terminology. Finally, fine-tuning of visual instructions is performed based on the LLaMA (large language model meta artificial intelligence), and a TCM-domain large language model with 7B parameters is trained.

Results: The constructed multimodal dataset has 3 test datasets, and experiments are conducted using 3000 samples from each test dataset, respectively. Experimental results indicate that the TongueVLM model outperforms general-purpose large models on all 3 tasks. On the multimodal test dataset, the TongueVLM model achieved accuracy rates of 79.8%, 78.6%, and 60.7% in evaluation tasks respectively, it achieves 9.1%, 8.4%, and 1.1% in greater accuracy than LLaVA-OneVision, and is 7.5%, 7%, and 5.9% more accurate than Qwen2.5-VL-7B, with the text generation time being around 24 tokens per second.

Conclusions: The TongueVLM model, which achieves tongue image description generation and physical constitution reasoning in TCM, is suitable for the application of a Chinese medicine intelligent diagnosis system.

(JMIR Med Inform 2026;14:e87237) doi:10.2196/87237

XSL•FO
RenderX

KEYWORDS

vision-language model; tongue image description generation; visual constitution reasoning; visual dialogue; intelligent diagnosis of traditional Chinese medicine (TCM)

Introduction

In recent years, the research focus on computer vision has gradually shifted toward innovative transformers and their variant architectures. These studies have made significant progress in the 3 basic tasks of image classification, object detection, and image segmentation, along with in-depth experiments on visual-textual multimodal data validity. The visual transformer model, represented by the vision transformer, approaches or even surpasses the convolutional neural network approach in several benchmark tests [1-14]. The key technique of the vision transformer is to construct image-to-vector transformations and effectively maintain the features of the image, bridging the gap between language and vision. Lava family [15-17], GPT-4v [18], and Gemini 1.5 [19] visual-language multimodal modeling have even ushered artificial intelligence-generated content into a new era of visual applications, and the field of medicine has received increasing attention from researchers [20-26].

In the field of traditional Chinese medicine (TCM), diagnostic work based on tongue images to recognize the physical constitution is a process of collecting clinical information, reasoning, and combining the patient's tongue image features with questioning. It is necessary to simulate the recognition of pathological information of tongue images by TCM practitioners and professional dialogue based on tongue image features, which helps to develop an intelligent interactive system for TCM diagnosis. However, the following limitations still exist in the current investigation: (1) high cost of acquisition equipment and collection of clinical data, mostly relying on publicly available datasets, and prominent data sample noise problems; (2) lack of qualitative or quantitative specialized labeling data; (3) lack of validation and analysis of the visual encoding results at the modal fusion stage; and (4) the evaluation metrics of the trained models focus on technical performance and ignore professional human judgments.

Previous studies have shown that, based on the image recognition model in the field of TCM applications, they fused the traditional image features of the tongue with deep tongue features and combined them with a machine learning model to construct a physical constitution identification, which compensated for the lack of information in a single modality and improved the accuracy of constitution classification [27-29]. In terms of training a vertical model of TCM based on a large language model (LLM), we constructed a pretraining dataset and an instruction fine-tuning dataset oriented to TCM and adopted a 2-phase training method of continuous pretraining and supervised fine-tuning to develop an LLM of TCM [30-35]. This model outperforms the existing models in several evaluation metrics and can play a role in TCM consultation.

Medical applications based on multimodal large models are also favored by researchers. A dataset containing 5.3 million+ image-text pairs was used to train the UniMed-CLIP model, which outperforms the existing health care visual language models (VLM) with 0-shot performance [36]. Multimodal health care models, which introduce evaluation metrics and frameworks and perform well in multimodal Q&A (question and answer) and text-based Q&A tasks [33,37]. BiomedCLIP with LLaMA-3 (large language model meta artificial intelligence) to realize a lightweight and efficient medical visual Q&A model [38]. Med-VQA interpretability by localizing image preference regions [39] enhanced the performance of biomedical VLMs by using instruction fine-tuning and multi-image training, respectively [40,41]. Aligned visual features with medical concepts and Med-VQA, enhanced the generation of large models, and achieved excellent performance on public datasets [42-44].

The primary aim of this study is to develop and validate TongueVLM, a specialized VLM for TCM tongue diagnosis. Specifically, our objectives were to (1) design a novel multimodal architecture that effectively aligns visual features of tongue images with domain-specific TCM textual knowledge; (2) construct a high-quality, multimodal dataset for supervised study, and end-to-end training on the TCM dataset; and (3) evaluate the performance of TongueVLM against established baseline models, and validation multimodal capabilities across 3 tasks of tongue feature description generation, physical constitution reasoning, and multiround dialogue.

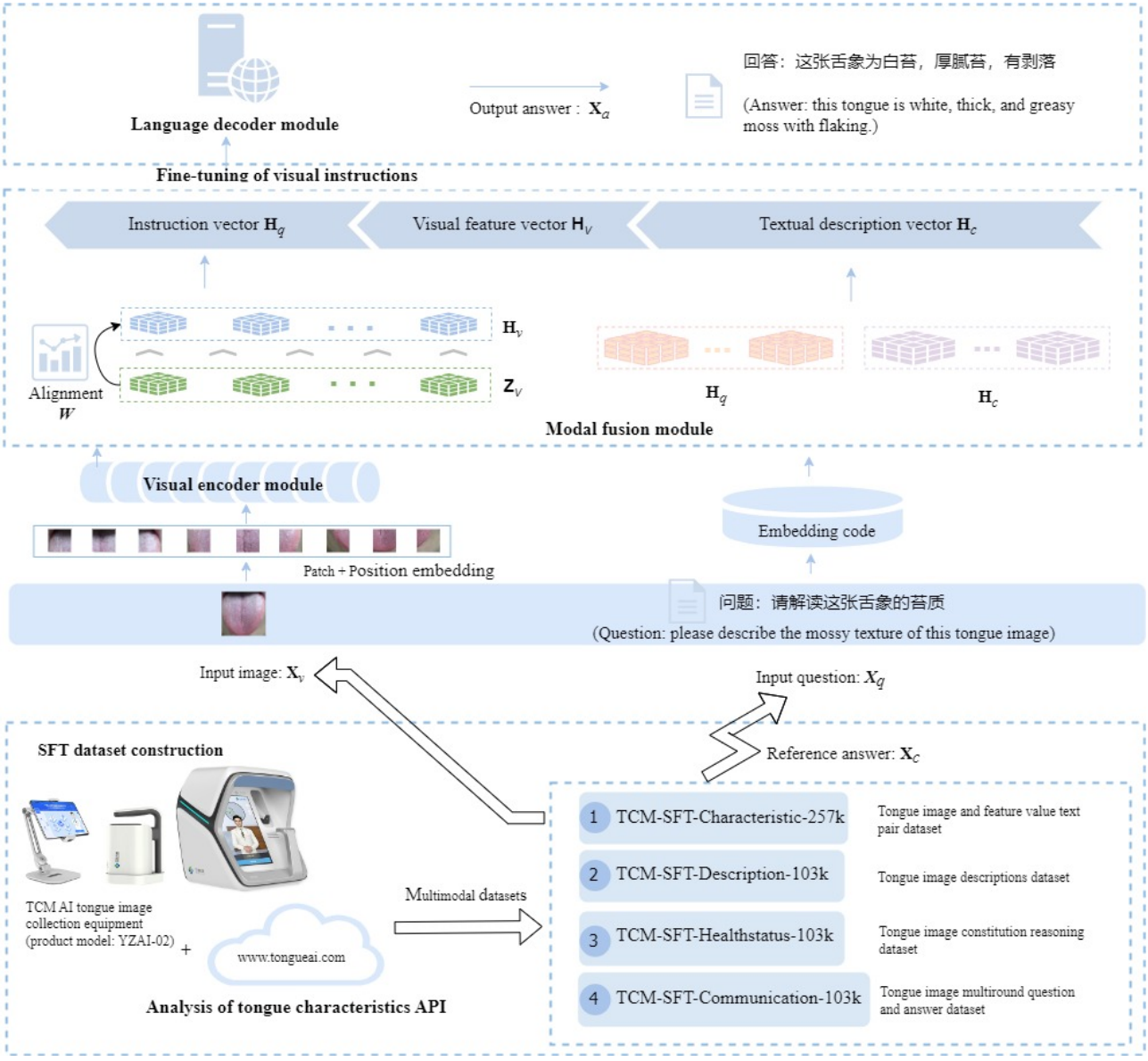
Methods

Overall Architecture of TongueVLM

Our TongueVLM model uses a hybrid architecture combining the visual encoder module, modal fusion module, and language decoder module. First, the visual encoder based on the CLIP-ViT (Contrastive Language-Image Pre-Training With Vision Transformer) pretrained model is used for image patch, dimensionality reduction, and migration learning. Further, a multimodal fusion module with a residual architecture is applied to map visual features to the natural language word embedding space, realizing the conceptual alignment between visual encoding and TCM terminology. Finally, a visual-textual instruction dataset is constructed to perform multimodal instruction fine-tuning based on the LLaMA general-purpose LLM. As a result, a vertical model of the TCM domain with TCM understanding and reasoning capability for tongue images is obtained. The realization procedure is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Network structure and computation of a large visual-language model. Includes a visual encoder module, a modal fusion module, and a language decoder module. First, the visual encoder based on the CLIP-ViT pretrained model is used for image patch, dimensionality reduction, and migration learning. Further, a modal fusion module with a residual architecture is applied to map visual features to a natural language word embedding space. Finally, fine-tuning of visual instructions is performed based on the LLaMA. AI: artificial intelligence; API: application programming interface; CLIP ViT: Contrastive Language Image Pre Training With Vision Transformer; LLaMa: large language model meta artificial intelligence; SFT: supervised fine-tuning; TCM: traditional Chinese medicine.



Model Design

Phase 1: Visual Encoder Module

In the first phase, the visual encoder module of the network was designed. Applying tongue images with TCM feature description text pair data in CLIP-ViT pretrained model migration learning. Semantically rich text is used as training labels to project the image feature representations into a kind of encoding that is like the text encoding space, as shown in Figure S1 in Multimedia Appendix 1.

First, we speak of image patch dimensionality reduction preprocessing. Scale the input image $\mathbf{x}_v \in (H,W,C)$ to $\mathbf{x} \in (h,w,C)$ (h is the height, w is the width, and C is the number of channels). Sequence of blocks that transform x into $\mathbf{x}_p \in (N^2,p^2C)$ for image patch operations. The number of blocks is $N^2=(h\times w)/p^2$, and the block dimension is (p^2C); where p is the image patch size. A sequence of blocks is converted into a patch embedding vector $\mathbf{E}_{patch}' \in (N^2,dim)$ by a linear transformation, where dim=1024, p=14, and N= 24. $\boldsymbol{x}_0$ is the zeroth encoding vector taken at the output layer of the transformer encoder, which serves as the visual feature encoding. By adding a learnable embedding vector before the image patch embedding vector E_{patch}, the image block embedding vector becomes $\mathbf{E}_{patch} \in (N^2+1,dim)$. The absolute position encoding algorithm is used to create a learnable position embedding vector $\mathbf{E}_{pos} \in (N^2+1,dim)$ for the block position. The elementwise summation operation on $\mathbf{E}_{patch}$ and $\mathbf{E}_{pos}$ yields an embedding vector $\mathbf{z} \in (N^2+1,\mathrm{dim})$ that incorporates spatial positional encoding and visual representations.

XSL•FO
RenderX

Second, transformer encoder processing. The encoder consists of a stack of **L**=24 encoding layers, and the encoder inputs are image feature embedding vectors $z \in (N^2+1, dim)$. Each encoding layer consists of 2 unitary structures connected: the first unitary structure consists of a multihead self-attention layer and a layer normalization, as well as a residual connectivity module, and the second unitary structure consists of a feedforward fully connected layer multilayer perceptron and a layer normalization, as well as a residual connectivity module.

Phase 2: Modal Fusion Module

The modal fusion module is designed with a multilayer linearly transformed residual architecture, which is a bridging module between the visual model and the language generation model. The tongue visual encoding features are projected into the natural language word embedding vector feature space to solve the conceptual alignment problem of visual encoding with TCM terminology.

The modal fusion module design uses a combination of multiple linear layers and GELU (Gaussian Error Linear Unit) activation functions, as shown in the middle region of Figure S2 in Multimedia Appendix 1. This design allows the modal fusion layers to transform visual features while maintaining essential nonlinear characteristics. In addition, the introduction of the residual structure enables the model to realize efficient feature transformation and projection while fully preserving the integrity of the original visual features.

Splice visual encoding with textual word-embedded encoding to form a sequence of visual-textual instructions. First, the visual encoding feature $\mathbf{Z}_v$ is converted into a visual feature vector $\mathbf{H}_v$, and the textual instruction and textual description are converted into an instruction vector $\mathbf{H}_q$ and a textual description vector $\mathbf{H}_c$ by word embedding encoding, respectively. Thereafter, the vectors are spliced in the order of instruction, visual, and textual description to form the visual-textual fusion instruction sequence data.

Phase 3: Language Decoder Module

The modal fusion module connects the visual encoder and language decoder for end-to-end fine-tuned training on the constructed visual-textual data for a variety of downstream generative output tasks, such as visual quizzing, image description generation, and visual dialogue. The language decoder module chooses LLaMA, a pretrained model with good support for Chinese, which consists of a 32-layer transformer decoder layer, and each decoder layer contains an RMSNorm normalization layer, a grouping attention mechanism, and a multilayer perceptron projection layer. The network architecture is shown on the right side of Figure S2 in Multimedia Appendix 1.

During training, each input image $\mathbf{X}_v$ is associated with a few rounds of dialogue data, where each round of question description can be defined as follows.

The visual encoder module extracts the visual feature V_x (the feature representation obtained by encoding the input image using a visual encoder) and caches it into the context, and the subsequent t-th round (t>1) of training data only needs to input the new instruction and answer . The model automatically fuses the historical dialogue information with the initial visual feature $\mathbf{V}_x$ and builds a spliced sequence of instructions to realize multiple rounds of dialogue training through cross-round dependency.

The instruction sequence of multiple rounds and the corresponding output reply text are temporally spliced into a uniform sequence , the autoregressive training objective is used for instruction tuning, and the maximum likelihood function optimization model is formulated as follows:

where θ is a trainable parameter (the training process is to adjust to maximize the likelihood function) and $\mathbf{X}_{instruct,<i}$,$\mathbf{X}_{a,<i}$ is the instruction sequence and answer sequence of all rounds before the current prediction x_i, respectively. $p(\mathbf{X}_a|\mathbf{X}_v,\mathbf{X}_{instruct})$ is the probability of predicting $\mathbf{X}_a$ under the current $\mathbf{X}_v$ and $\mathbf{X}_{instruct}$ is based on the premise of guaranteeing the accuracy of the result of $\mathbf{X}_a$under the premise that the prediction results of all rounds before the current round takes the maximum probability value.

In other words, the model uses the information accumulated in the history of the previous *t*-1 rounds of the conversation (especially the context of the textual instructions) to accurately estimate the probability distribution of the generated results in the t-th round. In this way, the model can maintain its understanding of the context over consecutive rounds of interactions and accordingly generate accurate responses associated with the content of the previous rounds.

Datasets and Preprocessing

The 90,000 tongue images collected by the TCM artificial intelligence tongue image collection equipment (product model: YZAI-02) jointly developed by Hefei Cloud Diagnostics Information Technology Co and Anhui University of TCM were used, which consists of a standard LED ring light source, a portable collection device, an ultrahigh-definition camera, and a host computer, as well as supporting software. The tongue feature recognition API (application programming interface) was applied to preprocess the collected tongue images with color correction, tongue region segmentation, and tongue moss and tongue texture separation, after which tongue feature analysis was performed on the tongue images to obtain more than 40 types of qualitative or quantitative text containing tongue color, tongue texture, moss color, moss texture, and sublingual collateral data and 10 types of TCM constitution [45].

The TongueVLM training datasets are constructed, including TCM feature description data (traditional Chinese medicine–supervised fine-tuning, TCM-SFT (TCM–supervised fine-tuning)-Characteristic-257k, and TCM-SFT-Description-103k), multiround Q&A data

XSL•FO
RenderX

(TCM-SFT-Healthstatus-103k), and physical constitution reasoning data (TCM-SFT-Communication-103k). The methodology for constructing the dataset is described in Multimedia Appendix 1. The distribution of the tongue image features in the dataset is shown in Figures S3 and S4 in Multimedia Appendix 1.

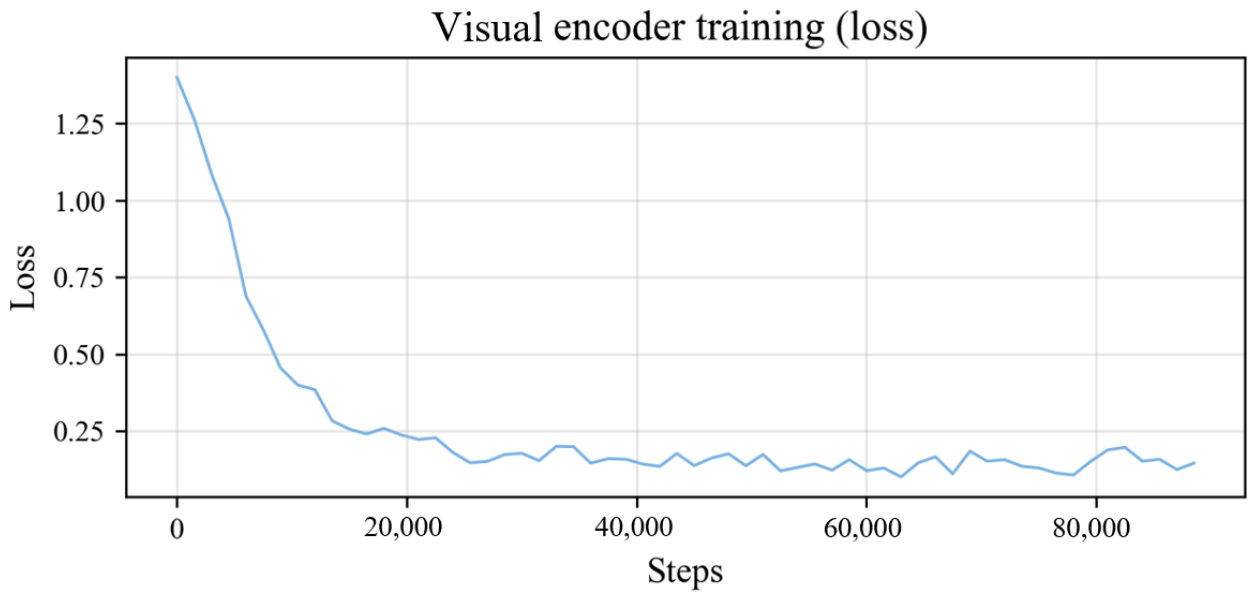
Model Training Strategy

Stage 1: Pretraining of the Visual Encoder

TCM tongue characterization datasets (TCM-SFT-Characteristic-257k) of 257,000 tongue images and characterization texts are constructed, and the zeroth encoding vector used for learning by comparing the similarity between the tongue images and their corresponding tongue manifestation characterization is used for training the fine-tuning of the tongue visual encoder.

The process to obtain the visual encoder model is as follows: first, the training parameters openai/clip-vit-large-patch14-336 (OpenAI Inc, pretrained weights for the clip model on 400 million text-image pairs) were used as the initial weights for the visual encoder and the text transformer encoder. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-3, the training batch size was set to 8, the learning rate was 5e-5, the number of training epochs was 3, and the maximum sequence length was 77. Second, the loss function uses the cosine similarity $Cos(I_m,T_n),m\in[1\ldots N^2],n\in[1\ldots N^2]$, where I_m is the embedding vector encoded by visual features and T_n is the embedding vector encoded by textual semantics. The optimization objective is to maximize the cosine similarity $Cos(I_j,T_j)$ between correct image-text pair embeddings while minimizing the cosine similarity $Cos(I^j,T^k),j\neq k$ of incorrect image-text pair embeddings. The results show that the loss value stabilizes after 3 epochs of training, and the loss curves are shown in Figure 2.

Figure 2. Visual encoder training loss and dynamic learning rate curves. The loss curve shows that after 3 rounds of training, the training loss of modal fusion has gradually stabilized, indicating that the model is beginning to converge.



Stage 2: Training of the Fusion Modal

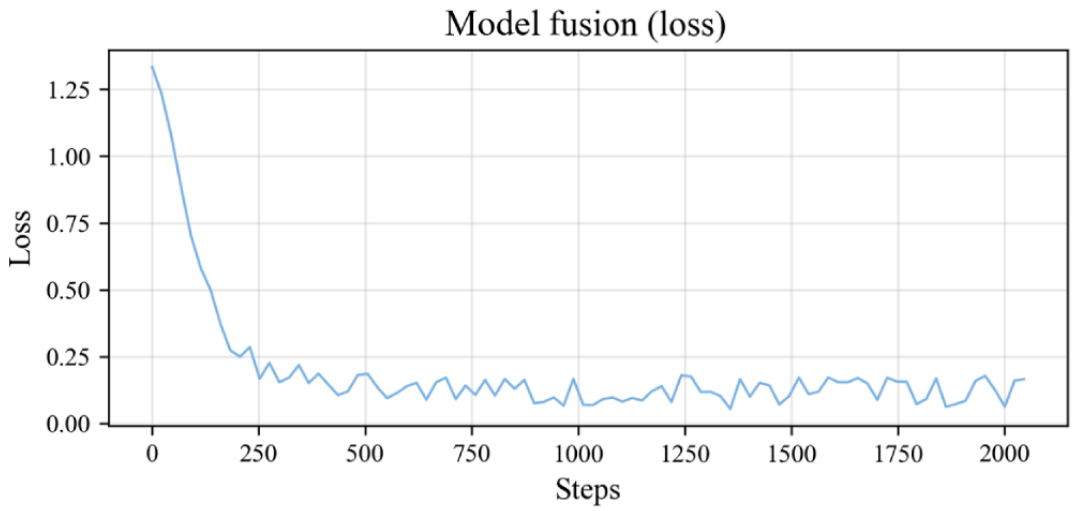
TCM tongue description datasets (TCM-SFT-Description-103k) of 103,000 tongue images and description text are constructed, the visual encoder of TongueVLM and the network weights of the LLM are frozen, the visual-textual multimodal instruction data are fed into the LLM, and the modal fusion module is fine-tuned for training.

The process to obtain the modal fusion model is, first, the training method and parameters are as follows: the pretrained visual encoder weight parameters are loaded, the modal fusion layer weight parameters are randomly initialized, and only the modal fusion layer weights are allowed to participate in the gradient update. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-4, the training batch size on each GPU is 8, the gradient accumulation step is 8, the learning rate is 2e-5, and the total number of iteration rounds is 3. Spend 30 hours to complete all the training on the Ubuntu (version 20.04.6, Canonical Ltd) platform using 2×RTX4090. Second, the loss function adopts the cross-entropy loss , where C is the total number of word lists, L is the length of the prediction sequence, $S_{n,i}$ denotes the probability that the word in the ith position is the nth word in the lexicon based on the previously known sequence, and $y_{n,i}$ denotes the ground-truth labels in the ith position in the prediction sequence (expressed as 1-hot encoding; ie, if the corresponding word in the position is the nth vocabulary word in the lexicon, only the first element of the 1-hot vector is 1 and the rest are 0), the goal is to minimize the loss; when the probability of each word in the prediction sequence is infinitesimally close to the actual vocabulary 1-hot encoding, that is, for any position i in the prediction sequence, $S_{n,i}\approx y_{n,i}$, the loss will be minimized, and

XSL•FO
RenderX

at this time, the model is considered to be in the optimal state. The results show that the loss value stabilizes after 3 epochs of training, and the loss curves are shown in Figure 3.

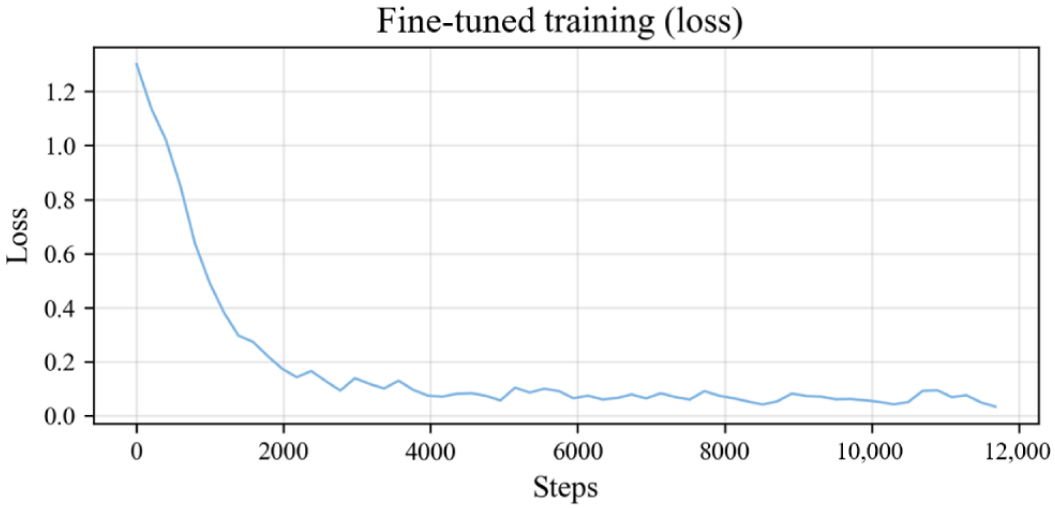
Figure 3. Loss curve of the modal fusion model. It shows that after 3 training epochs, the model enters a stable convergence state, and the loss curve for its modal fusion training stabilizes.



Stage 3: End-to-End Fine-Tuning of TongueVLM

The training data are the TCM multimodal dataset (training set 510,000; validation set 25,000). First, of the training method and parameters, load the pretrained visual encoder weight parameters and modal fusion layer weight parameters, freeze the weight parameters of the visual encoder, and allow only the modal fusion layer and the language decoder weights to participate in the gradient update. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-4, the training batch size is set to 8 on each GPU, the number of gradient accumulation steps is 8, the learning rate is 2e-5, and the total number of iteration rounds is 3. It takes 135 hours to complete all the training on the Ubuntu (version 20.04.6, Canonical Ltd.) platform using 2×RTX4090. Second, of loss function, the cross-entropy loss function is adopted, and the loss curve is shown in Figure 4. After 4000 iterations, the loss tends to stabilize, and the loss fluctuation is very small when the number of iterations is 8000; currently, the model has converged.

Figure 4. Fine-tuned training loss curve of TongueVLM. It shows that with the increase of training epochs, the fine-tuning training loss of TongueVLM gradually tends to stabilize, indicating the effectiveness of the optimization algorithm and the basic convergence of the model after about 8000 iterations.



Ethical Considerations

This study was conducted in accordance with the ethical standards of the Declaration of Helsinki and was approved by the ethics committee of the Second Clinical Medical School of Anhui University of Chinese Medicine (Anhui Acupuncture and Hospital). The data used in this research were sourced from 2 distinct origins. First, for data from the Yunzhen 365 app, the data consisted of deidentified user-uploaded tongue images and associated text. All users of the app provided informed consent upon registration, agreeing to the use of their anonymized data for research purposes in accordance with the app's terms of

XSL•FO
RenderX

service and privacy policy. Second, for data from hospital records, the clinical data were obtained with approval from the relevant hospital's institutional review board (2022-zjks-25). Patient confidentiality was strictly maintained, and all personal identifiers were removed before analysis.

All procedures involving human data were reviewed and approved by the appropriate ethics committees. No additional identifiable personal information was accessed or used in this study. Participants received no compensation for their involvement in this study.

Results

Overall Performance of TongueVLM on Benchmark Tasks

The performance evaluation of the TongueVLM on the TCM multimodal test dataset is as follows.

To determine whether the answer texts generated by professional TCM practitioners are close in meaning to the standard answer texts in the corresponding test set, the evaluation is divided into 3 results; according to the degree of similarity in meaning, we define a score value Score for it, in descending order: consistent (Score=100), similar (Score=50), and irrelevant (Score=0); the accuracy of TongueVLM on the test set can be expressed as follows: , where N is the number of test datasets and $Score_i$ is the score of the i-th test dataset.

To directly verify the effectiveness of the improved methods proposed in this study for the field of TCM, we selected visual multimodal models with similar architectures for comparison: LLaVA-1.5 (released in October 2023; Meta Platforms, Inc), LLaVA-1.6 (released in January 2024; Meta Platforms, Inc), and LLaVA-OneVision (released in August 2024; Meta Platforms, Inc). To demonstrate that good results can be achieved in the field of TCM through targeted model design without relying on massive general data, we also compared it with the current open-source multimodal large model benchmark Qwen2.5-VL-7B (released in August 2024; Alibaba Cloud). The accuracy and performance are shown in Table 1. The experimental results show that the TongueVLM model outperforms the general-purpose large model on all 3 tasks; it is 9.1%, 8.4%, and 1.1% more accurate than LLaVA-OneVision and 7.5%, 7%, and 5.9% more accurate than Qwen2.5-VL-7B. The generation time efficiency of each large model is similar, with the generation time of text being 24 tokens per second (approximately 40 words per second, given the word-to-token ratio in our dataset). Due to the optimization of the lexicon in the field of TCM, TongueVLM inference speed is slightly faster than the baseline model.

Table 1. Evaluation indicators for the output of the model. Shows the evaluation metrics of the model output. Experiments are conducted using the test dataset of 3 multimodal datasets, with each test dataset containing 3000 samples with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Model	TCM-SFT[a]-Description-103k			TCM-SFT-Healthstatus-103k			TCM-SFT-Communication-103k		
	Test result details	Accuracy (%)	Mean time (SD, s)	Test result details	Accuracy (%)	Mean time (SD,s)	Test result details	Accuracy (%)	Mean time (SD,s)
CNN[b] parallel computing	167,58,2775/3000	79	1.48 (>6)	—[c]	—	—	—	—	—
LLaVA-1.5	1730,973,297/3000	26.1	0.27 (1.25)	1828,920,252/3000	23.7	0.24 (1.5)	1174,1653,173/3000	33.3	0.64 (3)
LLaVA-1.6	568,1285,1147/3000	59.7	0.21 (1.55)	545,1241,1214/3000	61.2	0.21 (1.5)	690,1094,1216/3000	58.8	0.62 (3)
LLaVA-OneVision	393,974,1633/3000	70.7	0.17 (1.54)	423,941,1636/3000	70.2	0.25 (1.5)	635,1156,1209/3000	59.6	0.86 (3)
Qwen2.5-VL-7B	412,836,1752/3000	72.3	0.18 (1.48)	420,866,1714/3000	71.6	0.15 (1.4)	519,1672,809/3000	54.8	0.69 (3)
TongueVLM	305,599,2096/3000	79.8	0.11 (1.25)	353,577,2070/3000	78.6	0.20 (1.5)	428,1502,1070/3000	60.7	0.54 (3)

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

[b]CNN: convolutional neural network.

[c]Not available.

Ablation Studies on the 3-Stage Training Strategy

Visual Encoding Feature Layer Comparison Experiment

The visual encoder extracts image features and outputs a total of 24 hidden layers (disregarding the hidden layer feature values before being input into the visual encoder), and it is necessary to select the appropriate layer as the visual feature information input. Therefore, comparison experiments were conducted on 3 datasets, and the last 5 hidden layers (20th to 24th) features were selected as the visual feature information to train the TongueVLM model; the accuracies on the test datasets are shown in Table 2 below. The experimental results show that the model performs best in all tasks when the penultimate layer of features output from the visual encoder, that is, the 23rd layer of features, is selected as the final visual feature information. These findings reveal that the visual features extracted at that depth level have significant advantages for multimodal intelligent analysis related to TCM tongue diagnosis.

XSL•FO
RenderX

Table 2. Accuracy of different hidden layers as a visual encoder in the test dataset. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation. It shows the accuracy of different hidden layers as visual encoders in the test dataset, with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets).

Hidden layers	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
Layer-24	54,102,344/500	79	44,29,427/500	88.3	60,29,411/500	85.1
Layer-23	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7
Layer-22	52,105,343/500	79.1	45,37,418/500	87.3	63,25,412/500	84.9
Layer-21	52,114,334/500	78.2	45,49,406/500	86.1	63,32,405/500	84.2
Layer-20	55,117,328/500	77.3	48,48,404/500	85.6	65,47,388/500	82.3

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Modal Fusion Module Structure Experiment

To verify the reasonableness of the structure of the modal fusion module, 3 different neural network structures, namely, numbers 1 (Linear + GELU), 2 (Linear + GELU + Linear + GELU + Linear), and 3 (Linear + GELU + Linear + GELU + Linear + Linear + Res), are designed, and initial weights are randomized with the same datasets; the accuracy on the test datasets is shown in Table 3. The results show that the difference in the effect of the 3 structures on the final performance is not significant. However, the use of the fusion structure of group number 3, containing multilayer linear transformation, together with a multilayer nonlinear activation function and the introduction of residual connections, slightly outperforms the performance of the other schemes. Therefore, group number 3 is chosen as the neural network structure for modal fusion learning.

Table 3. Comparative analysis of the modal fusion neural network structure in terms of accuracy on the test datasets. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Number	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
1	55,99,346/500	79.1	46,26,428/500	88.2	62,27,411/500	84.9
2	54,98,348/500	79.4	45,29,426/500	88.1	60,27,413/500	85.3
3	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Modal Fusion Module Pretraining Experiment

The weight parameters of the modal fusion are also fine-tuned for training during the overall training of the TongueVLM model; thus, the pretraining of the modal fusion is compared to that of number 1, the pretraining is fine-tuned individually, and number 2 is randomly initialized in TongueVLM without pretraining. The accuracy of the 3 test datasets is shown in Table 4 below. The results show that the accuracy improvement using the individually fine-tuned pretraining approach is significant, with 2.3%, 5%, and 3.5% improvement for the 3 datasets, respectively. Fine-tuning the modal fusion layer using the Chinese interpretation feature description data allows the model to focus more on constructing and optimizing higher-order correspondences between visual-language features.

Table 4. Evaluation of pretraining methods for modal fusion modules. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Numbers	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
1	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7
2	53,119,328/500	77.5	46,73,381/500	83.5	63,52,385/500	82.2

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Analysis of Module Effectiveness and Interpretability

In this experiment, the visualization of the projection of the attention weight values of the multihead self-attention layer during the forward computation of the visual encoder to the image is investigated, and pseudo-color maps of the regional thermal response during the visual encoding process for typical texture feature locations on the surface of the tongue image are generated.

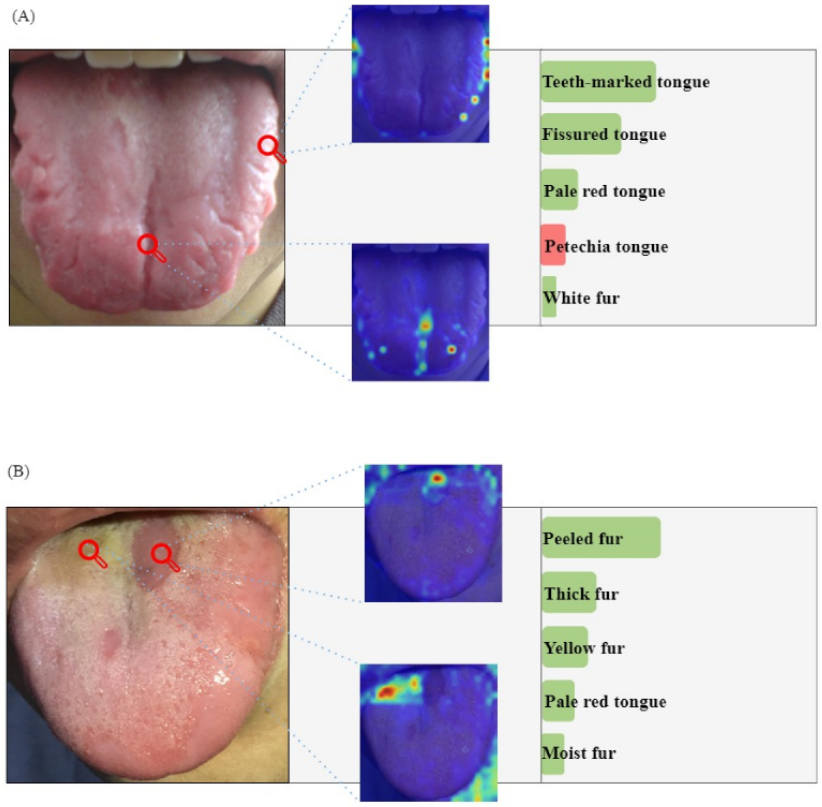
XSL•FO
RenderX

The multihead attention weights during the forward propagation of the cached visual encoder are Atten_weights$\in R^{(num_heads,dim,dim)}$, where num_heads$\in$(1,2,…,16) is the number of heads of multihead attention, and *dim*=577 is the length of the encoding sequence. The 0th position of the encoding sequence is the classification head, and the $E_i \in R^{(dim-1,dim-1)}$ vector represents the correlation between the 24 × 24 patches of the image in the ith attention header.

As shown in Figure 5, feature locations were selected in the tongue image; Figure 5A for fissured and teeth-marked tongues, and Figure 5B for peeled fur and yellow moss. The attention weights for the specified location are projected into the pseudo-color map of the original image. The text labels of the top 5 tongue features are obtained as output, with red indicating incorrect prediction and green indicating correct prediction, and the length of the color indicates the magnitude of its probability.

Through the feature attention visualization map, it can be seen that the visual encoder for tongue texture features focuses on localized regional positional responses, while the tongue color feature focuses on the overall area response of the image, so that the visual encoder can learn the effective feature information on the tongue, which provides good support for the subsequent feature fusion.

Figure 5. Visualization of the thermal response of the visual encoder's multihead attention. Presents the top 5 categories and their confidence levels of tongue image features predicted by the model in text form. Green represents correctly predicted labels, red represents incorrectly predicted labels, and the length of the color bars intuitively indicates the probability of the prediction.



Validation of Multimodal Capabilities Across Tasks

Tongue Image Description Generation Capability

In the task of describing the tongue images, we observe that LLaVA-1.6, LLaVA-OneVision, and Qwen2.5-VL-7B all generate output answers according to structured texts. Moreover, the text of LLaVA-1.6's response to the tongue surface image was intermingled with the description of the tongue sublingual veins. This type of response may be a language hallucination of the LLM. The output text of the TongueVLM is relatively short, but in terms of content, it focuses mainly on image color, texture, and morphological information. Representative examples of model outputs are shown in Table S1 in Multimedia Appendix 1. The tongue feature interpretation of the TongueVLM model is closer to the results of visual model recognition, which fully demonstrates that the TongueVLM model, which has been fine-tuned with the tongue multimodal datasets, solves the conceptual alignment between the visual encoding of the images and the TCM terminology.

Physical Constitution Reasoning and Analysis Capability

In the task of physical constitution reasoning about the tongue image, we observe that the output of the LLaVA-1.5 model completely deviates from the TCM context and seems to lack an understanding of the concept of TCM physique, and the LLaVA-1.6 model outputs the reasoning thinking about TCM physique through the tongue image. The LLaVA-OneVision model misjudges the tongue feature as a red tongue, but can reason about the physical constitution in combination with the tongue feature. Of course, the output is also wrong.

XSL•FO
RenderX

QWen2.5-VL-7B still follows some kind of templated output content, which is accurate in understanding the tongue color and color of the tongue fur features, but misclassifies the tongue shape as fat and large and dentate to the extent that the analysis of the physical constitution is not accurate enough.

In contrast, the TongueVLM model shows professional advantages in the TCM field in terms of effectively recognizing the tongue features and then briefly outputs the results through the tongue features, but its language expression level is slightly rough and needs to be further improved. Representative examples of model outputs are shown in Table S2 in Multimedia Appendix 1.

Tongue Image Dialogue Capability

In the multiround dialogue task of TCM with tongue images, we observe that all the models can accurately recognize the teeth-marked tongue feature in the tongue image in the first round of dialogue, and LLaVA-1.6 not only interprets the tongue feature dimension but also associates more scientific knowledge with the physical condition. However, in the second round of dialogue, LLaVA-1.5 and LLaVA-1.6 gave vague answers of "possibly related." Moreover, LLaVA-OneVision and Qwen2.5-VL-7B were able to provide definite answers, but they were not related to the teeth-marked tongue feature in the previous round of dialogue. The TongueVLM model combines the content of the teeth-marked tongue feature and provides the exact answer; thus, the TongueVLM model still outperforms the other models. Representative examples of model outputs are shown in Table S3 in Multimedia Appendix 1.

The experimental results show that the TongueVLM model outperforms the general-purpose large model on all 3 tasks. First, in terms of tongue feature descriptions, the TongueVLM model is closest to the recognition results of the visual model and includes color and morphological features and local positional features, and the parallel computation time of multiple visual models is greater than that of the large model, which fully reflects the image feature retention ability and computational performance of the TongueVLM visual encoder. Second, in terms of physical constitution reasoning, it is obvious that LLaVA-1.5 does not have expertise in TCM. LLaVA-1.6, LLaVA-OneVision, and Qwen2.5-VL-7B are generated by interpreting the text according to the unified paradigm; it cannot distinguish between tongue photos and sublingual photos, and identifying the positive and negative features is difficult. The TongueVLM model, on the other hand, quantitatively analyzes and localizes the positive features of the tongue image and can reason about the physical constitution based on the tongue image features, with an accuracy of 78.6%. The traditional method requires multiple visual models to extract the tongue features and then input the feature data into decision trees, support vector machines, and other classifiers for supervised learning of somatic qualities, which is less efficient to execute. Third, in terms of multiround Q&As, the TongueVLM model's interactive answers are brief, the contextual correlation is stronger, and the language discourse is not as strong as that of LLaVA-1.6, but its overall answers are more precise and concise, and language generation is not an advantage of the visual model.

Discussion

Principal Findings

This study proposes a TongueVLM network architecture design and a 3-stage training framework, achieving state-of-the-art performance on multimodal tongue image diagnosis tasks. Crucially, ablation experiments confirm that the staged training strategy is essential for stable optimization, while visual analysis demonstrates that the model spontaneously focuses on tongue regions relevant to clinical diagnosis.

The TongueVLM multimodal model proposed in this study is characterized by the following key features. First, the visual encoder characterizes the image features with the language encoding of low-dimensional vectors and approximates the feature extraction capability of the visual model by migration learning of TCM images from the CLIP-ViT pretrained model, which is a key step in the visual-to-large language model. Second, the modal fusion module splices visual and textual instruction data, realizes a seamless connection between natural language and image feature information, and unifies the multimodal data format input to the large model. Third, the TongueVLM model is fine-tuned and trained end-to-end on the TCM dataset. The model exhibits high accuracy in all 3 tasks of tongue feature description generation, physical constitution reasoning, and multiround dialogue.

Limitations and Future Work

This study has several limitations. First, the current data modality only includes tongue image and text-based question-answering data, failing to encompass the complete information system of TCM's "observation, auscultation, inquiry, and palpation." Second, the dataset's diversity requires enhancement, with shortcomings in ethnic distribution and standardization of image acquisition environments, which may compromise the model's generalizability. Third, evaluation of generated texts (eg, tongue pattern descriptions) still relies on expert judgments with inherent subjectivity, lacking objective, fine-grained automated assessment criteria. Finally, the model's generalization capability for clinically rare tongue patterns remains to be validated.

Future research may explore the following directions: first, in terms of model architecture, a gating mechanism is introduced to achieve adaptive attention regulation. Meanwhile, it is considered to incorporate modules such as depth-wise separable convolution to compensate for the shortcomings of self-attention mechanisms in capturing local detail features. Second, expand multimodal data sources by systematically integrating comprehensive information from facial diagnosis, pulse diagnosis, olfactory diagnosis, and physical constitution identification to construct a full-disease-course diagnostic model more aligned with TCM's holistic perspective. Third, fuse multidimensional data acquisition technologies such as 400-1000 nm spectral data and thermal infrared imaging data, and 3D tongue morphology to establish a more refined quantitative tongue pattern characterization system. Fourth, explore automated evaluation frameworks based on LLMs to achieve fine-grained quantitative assessments of the professionalism, consistency, and logical coherence of generated content. Fifth,

XSL•FO
RenderX

advance prospective trials in collaboration with clinical institutions to validate the model's auxiliary diagnostic efficacy and clinical applicability in real-world diagnostic settings. Sixth, currently, TongueVLM is built upon a 7B-parameter model, and its deployment still imposes certain computational requirements. We will explore parameter-efficient fine-tuning techniques (eg, low-rank adaptation) and model quantization. Finally, the model will incorporate continuous learning mechanisms to iteratively improve its tongue image recognition accuracy and diagnostic reasoning ability, thereby advancing research on artificial intelligence–assisted TCM tongue diagnosis.

Conclusions

Intelligent diagnosis and treatment technology in the field of TCM is often based on knowledge graphs, deep learning visual models, and machine learning approaches. Traditional supervised learning is limited by the quality and size of the training data, as well as the difficulty of fusion with natural language and its limited generalizability. The findings demonstrate that the TongueVLM model achieves ideal results in terms of understanding tongue images, constitutive reasoning, and dialogue abilities, and provides a foundational framework suitable for future integration into a Chinese medicine intelligent diagnosis system.

Funding

We are thankful for funding support from the Project of Plans for Major Provincial Science & Technology (202303a07020008) and Xin'an Institute of Medicine and Modernization of TCM Research and Development Project: "Leading the Charge with Open Competition" (2024CXMMTCM002).

Data Availability

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to this study's conception and design. Material preparation, data collection, and analysis were performed by CP, JG, NY, YW, RC, and CD. The first draft of this paper was written by CP, and all authors commented on previous versions of this paper. All authors read and approved this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables (S1-S3) and figures (S1-S4), formulas, and example data.
[DOCX File , 33112 KB - medinform_v14i1e87237_app1.docx]

References

1. Pak S, Son HJ, Kim D, Woo JY, Yang IK, Hwang HS, et al. Comparison of CNNs and transformer models in diagnosing bone metastases in bone scans using Grad-CAM. Clin Nucl Med 2025;50(7):596-604. [doi: 10.1097/RLU.0000000000005898] [Medline: 40237349]
2. Zhang Y, Wang J, Górriz JM, Wang S. Deep learning and vision transformer for medical image analysis. J Imaging 2023;9(7):147 [FREE Full text] [doi: 10.3390/jimaging9070147] [Medline: 37504824]
3. Nurgazin M, Tu NA. A comparative study of vision transformer encoders and few-shot learning for medical image classification. 2023 Presented at: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2025 Oct 19; Honolulu, Hawaii p. 2505-2513. [doi: 10.1109/iccvw60793.2023.00265]
4. Dayan B. Lung disease detection with vision transformers: a comparative study of machine learning methods. arXiv Preprint posted online on Nov 18, 2024. [doi: 10.48550/arXiv.2411.11376]
5. Falqueto P, Sanfeliu A, Palopoli L, Fontanelli D. Learning priors of human motion with vision transformers. In: IEEE. 2025 Presented at: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC); July 02-04, 2024; Osaka, Japan p. 1-9. [doi: 10.1109/COMPSAC61105.2024.00060]
6. Zhang H, Ramachandra R, Raja K, Busch C. Generalized single-image-based morphing attack detection using deep representations from vision transformer. 2024 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 17, 2024; Seattle, WA p. 1510-1518. [doi: 10.1109/cvprw63382.2024.00158]
7. Sheffield B, Ellen J, Whitmore B. On vision transformers for classification tasks in side-scan sonar imagery. arXiv Preprint posted online on Sep 18, 2024. [doi: 10.48550/arXiv.2409.12026]
8. Gao W, Han B, Sun Z, Yan Y, Ye Y, Feng J, et al. A novel system for precise identification and explainability analysis based on multimodal learning combining laser-induced breakdown spectroscopy and laser-induced plasma acoustic signals. Talanta 2025;293:128182. [doi: 10.1016/j.talanta.2025.128182] [Medline: 40252502]

XSL•FO
RenderX

9. Singh G. Comparative analysis of vision transformers and traditional deep learning approaches for automated pneumonia detection in chest X-Rays. arXiv Preprint posted online on Jul 11, 2025. [doi: 10.48550/arXiv.2507.10589]
10. Alqutayfi A, Almattar W, Al-Azani S, Khan FA, Qahtani AA, Alageel S, et al. Explainable disease classification: exploring Grad-CAM analysis of CNNs and ViTs. JAIT 2025;16(2):264-273. [doi: 10.12720/jait.16.2.264-273]
11. Zhang W, Belcheva V, Ermakova T. Interpretable deep learning for diabetic retinopathy: a comparative study of CNN, ViT, and hybrid architectures. Computers 2025;14(5):187. [doi: 10.3390/computers14050187]
12. Amangeldi A, Taigonyrov A, Jawad MH, Mbonu CE. CNN and ViT efficiency study on tiny ImageNet and DermaMNIST datasets. arXiv Preprint posted online on Feb 13, 2026. [doi: 10.48550/arXiv.2505.08259]
13. Mehdipour S, Mirroshandel SA, Tabatabaei SA. Vision transformers in precision agriculture: a comprehensive survey. Intell Syst Appl 2026;29:200617. [doi: 10.1016/j.iswa.2025.200617]
14. Agarwal A, Gearon J, Rank R, Chenevert E. Fighting fires from space: leveraging vision transformers for enhanced wildfire detection and characterization. arXiv Preprint posted online on Apr 18, 2025. [doi: 10.48550/arXiv.2504.13776]
15. Haotian L, Chunyuan L, Qingyang W, Lee YL. Visual instruction tuning. 2023 Presented at: Proceedings of the 37th International Conference on Neural Information Processing Systems; December 10-16, 2023; New Orleans, LA p. 1516. [doi: 10.5860/choice.189890]
16. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. 2024 Presented at: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 16-22, 2024; Seattle, WA p. 26286-26296. [doi: 10.1109/cvpr52733.2024.02484]
17. Li B, Zhang Y, Guo D. LLaVA-OneVision: easy visual task transfer. arXiv Preprint posted online on Oct 26, 2024. [doi: 10.48550/arXiv.2408.03326]
18. Yang Z, Li L, Lin K. The dawn of LMMs: preliminary explorations with GPT-4V(ision). arXiv Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2309.17421]
19. Gemini T, Georgiev P, Lei V. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2403.05530]
20. Dehghani M, Djolonga J, Mustafa B. Scaling vision transformers to 22 billion parameters. arXiv Preprint posted online on Feb 10, 2023. [doi: 10.48550/arXiv.2302.05442]
21. Xuyan H, Meng S, Chengxing S, Haoxuan L, Jianlin Z. Visual-language reasoning large language models for primary care: advancing clinical decision support through multimodal AI. Vis Comput 2025;41(13):11327-11348. [doi: 10.1007/s00371-025-04109-y]
22. Thapa S, Naseem U, Zhou L. Vision-language models for biomedical applications. 2024 Presented at: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications; Oct 28-Nov 1, 2024; Melbourne, Victoria, Australia p. 1-2. [doi: 10.1145/3689096.3690770]
23. Raminedi S, Shridevi S, Won D. Multi-modal transformer architecture for medical image analysis and automated report generation. Sci Rep 2024;14(1):19281 [FREE Full text] [doi: 10.1038/s41598-024-69981-5] [Medline: 39164302]
24. AlShibli A, Bazi Y, Rahhal MMA, Zuair M. Vision-BioLLM: large vision language model for visual dialogue in biomedical imagery. Biomed Signal Proc Control 2025;103:107437. [doi: 10.1016/j.bspc.2024.107437]
25. Kalpelbe BC, Adaambiik AG, Peng W. Vision language models in medicine. arXiv Preprint posted online on Feb 24, 2025. [doi: 10.48550/arXiv.2503.01863]
26. Peng C, Zhang K, Lyu M, Liu H, Sun L, Wu Y. Scaling up biomedical vision-language models: fine-tuning, instruction tuning, and multi-modal learning. J Biomed Inform 2025;171:104946. [doi: 10.1016/j.jbi.2025.104946] [Medline: 41138953]
27. Liu Y, Fan L, Zhao M, Wei D, Zhao M, Dong Y, et al. Study on a traditional chinese medicine constitution recognition model using tongue image characteristics and deep learning: a prospective dual-center investigation. Chin Med 2025;20(1):84 [FREE Full text] [doi: 10.1186/s13020-025-01126-w] [Medline: 40506765]
28. Jiatuo XU, Tao J, Shi L. Research status and prospect of tongue image diagnosis analysis based on machine learning. Digital Chin Med 2024;7(1):3-12. [doi: 10.1016/j.dcmed.2024.04.002]
29. Tian-Yu GU, Zhuang-Zhi Y, Jie-Hui J. Classifying Chinese medicine constitution using multimodal deep-learning model. Chin J Integr Med 2024;30(2):163-170. [doi: 10.1007/s11655-022-3541-8]
30. Ye Q, Liu J, Chong D. Qilin-Med: Multi-stage knowledge injection advanced medical large language model. arXiv Preprint posted online on Apr 17, 2024. [doi: 10.48550/arXiv.2310.09089]
31. Jia Y, Ji X, Wang X, Zhang H, Meng Z, Zhang J, et al. Qibo: a large language model for traditional chinese medicine. Expert Syst Appl 2025;284:127672. [doi: 10.1016/j.eswa.2025.127672]
32. Yu S, Xu X, Xu F. Enhancing the traditional chinese medicine capabilities of large language model through reinforcement learning from AI Feedback. arXiv Preprint posted online on Nov 1, 2024. [doi: 10.48550/arXiv.2411.00897]
33. Xie J, Yu Y, Zhang Z. TCM-Ladder: a benchmark for multimodal question answering on traditional chinese medicine. arXiv Preprint posted online on Oct 24, 2025. [doi: 10.48550/arXiv.2505.24063]
34. Sun Z, Huang R, Feng J. DoPI: doctor-like proactive interrogation LLM for traditional chinese medicine. arXiv Preprint posted online on Jul 7, 2025. [doi: 10.48550/arXiv.2507.04877]
35. Wei S, Peng X, Wang Y, Shen T, Si J, Zhang W, et al. BianCang: a traditional chinese medicine large language model. IEEE J Biomed Health Inf 2025:1-12. [doi: 10.1109/jbhi.2025.3612415]

XSL•FO
RenderX

36. Khattak MU, Kunhimon S, Naseer M. UniMed-CLIP: towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv Preprint posted online on Dec 13, 2024. [doi: 10.48550/arXiv.2412.10372]
37. LASA Team, Xu W, Chan HP, Li L. Lingshu: a generalist foundation model for unified multimodal medical understanding and reasoning. arXiv Preprint posted online on Jun 13, 2025. [doi: 10.48550/arXiv.2506.07044]
38. Alsinglawi B, Mccarthy C, Webb S, Fluke C, Saidy NT. A lightweight large vision-language model for multimodal medical images. arXiv Preprint posted online on Apr 8, 2025. [doi: 10.48550/arXiv.2504.05575]
39. Wang Y, Liu J, Gao S, Feng B, Tang Z. V2T-CoT: from vision to text chain-of-thought for medical reasoning and diagnosis. Med Image Comput Comput Assisted Intervention – MICCAI 2025 2026;15964:658-668. [doi: 10.1007/978-3-032-04971-1_62]
40. Li C, Wong C, Zhang S. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv Preprint posted online on Jun 1, 2023. [doi: 10.48550/arXiv.2306.00890]
41. Yang X, Miao J, Yuan Y, Jiaze W, Qi D. Medical large vision language models with multi-image visual ability. Med Image Comput Comput Assisted Intervention – MICCAI 2025 2026;15964:402-412. [doi: 10.1007/978-3-032-04971-1_38]
42. Xing Q, Song Z, Zhang Y. MCA-RG: enhancing LLMs with medical concept alignment for radiology report generation. Med Image Comput Comput Assisted Intervention – MICCAI 2025 2026;15964:380-390. [doi: 10.1007/978-3-032-04971-1_36]
43. Yang Y, Ma T, Li R, Zheng X, Shan G. JingFang: an expert-level large language model for traditional chinese medicine clinical consultation and syndrome differentiation-based treatment. arXiv Preprint posted online on Jan 20, 2026. [doi: 10.48550/arXiv.2502.04345]
44. Xu Z, Li Q, Nie W, Wang W, Liu A. Structure causal models and LLMs integration in medical visual question answering. IEEE Trans Med Imaging 2025;44(8):3476-3489. [doi: 10.1109/tmi.2025.3564320] [Medline: 40299735]
45. Traditional Chinese Medicine Identification v2 API. AI Open Platform for TCM Tongue Diagnosis. Beijing: Hefei Cloud Diagnostics Information Technology Co; 2025. URL: https://www.ai-tongue.com/doc_n4t9r9w8/apiapp/composite/home.html [accessed 2026-02-24]

Abbreviations

API: application programming interface
CLIP-ViT: Contrastive Language-Image Pre-Training With Vision Transformer
GELU: Gaussian Error Linear Unit
LLaMA: large language model meta artificial intelligence
LLM: large language model
Q&A: question and answer
TCM: traditional Chinese medicine
TCM-SFT: traditional Chinese medicine–supervised fine-tuning
VLM: visual language model

Edited by A Benis; submitted 06.Nov.2025; peer-reviewed by Y Liu, M Fang; comments to author 05.Dec.2025; revised version received 17.Dec.2025; accepted 17.Feb.2026; published 12.Mar.2026.

Please cite as:
Peng C, Gao J, Yang N, Wang Y, Chen R, Dong C
Application of a Large Visual Language Model on Tongue Image Description Generation and Physical Constitution Reasoning in Traditional Chinese Medicine (TongueVLM): Model Development and Validation Study
JMIR Med Inform 2026;14:e87237
URL: https://medinform.jmir.org/2026/1/e87237
doi:10.2196/87237
PMID:

©Chengdong Peng, Jun Gao, Nuo Yang, Yong Wang, Renming Chen, Changwu Dong. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Emotion Expression in Breast Cancer Support Seeking: Empirical Study of an Online Community

Anqi Xu[1], PhD; Ahmad Aljanaideh[1], PhD; Jennifer Xu[1], PhD; Haijing Hao[1], PhD

Department of Computer Information Systems, Bentley University, Waltham, MA, United States

Corresponding Author:
Anqi Xu, PhD
Department of Computer Information Systems
Bentley University
175 Forest Street
Waltham, MA, 02452-4713
United States
Phone: 1 7818912584
Email: axu@bentley.edu

Abstract

Background: Breast cancer affects millions of women and presents not only medical challenges but also emotional, financial, and social burdens. Beyond clinical treatment, patients increasingly turn to online cancer communities (OCCs) for informational support, emotional support, and shared coping strategies. OCCs help patients manage daily life and reduce psychological distress through shared experiences and empathetic engagement. Within these communities, emotional expressions serve as critical cues through which patients communicate their situations and needs with other OCC members.

Objective: This study explores the relationship between emotions expressed in patients' initial posts and community reactions and engagement in OCCs. Based on the emotions as social information (EASI) model and Plutchik's Wheel of Emotions, this study investigates whether different emotions function as distinct social signals that lead to different patterns of member response, and whether emotions defined as opposites elicit opposing or similar engagement behaviors.

Methods: This study examines how the 8 primary emotions expressed in an initial post—surprise, anticipation, joy, sadness, trust, disgust, fear, and anger—distinctively influence responses of members in OCCs. We collected data from a breast cancer community from 2002 to 2017 and analyzed 23,633 threads from 9137 patients with breast cancer. Using BERT (Bidirectional Encoder Representations From Transformers), we extract emotion scores from patients' initial posts, measure community engagement across 5 response dimensions, and empirically analyze how different emotions are associated with user response behaviors.

Results: The analysis shows that certain emotions, such as joy and anticipation, consistently elicit significant effects across all measured response categories (P<.001). Most other emotions, except disgust, also demonstrate significant effects in most categories (3 or 4 of 5 categories; P<.05). Consistent with the EASI perspective, different emotional expressions influence how community members allocate attention, effort, and timing in their responses in distinct ways. Moreover, for most pairs of opposite emotions (eg, surprise vs anticipation, joy vs sadness, and anger vs fear), the impacts on the 5 types of community responses move in parallel rather than opposite directions.

Conclusions: By integrating Plutchik's emotion framework with the EASI model, this study advances understanding of how emotional expressions function as social information and signals in OCCs. The findings show that emotions expressed in patients' initial posts influence community engagement not only through emotional valence, but also through the situational context that community members can infer from the initial posts. In addition, modeling emotions as continuous intensity scales helps reveal how emotional strength amplifies engagement across multiple dimensions, including participation, effort, timing, and topical relevance. These insights extend prior OCC research beyond polarity-based sentiment analysis and offer practical implications for patients, community platforms, and health care practitioners seeking to better support peer interaction and psychosocial care in digital health environments.

(JMIR Med Inform 2026;14:e83674) doi:10.2196/83674

KEYWORDS

breast cancer; online cancer community; emotion; patient engagement; machine learning

XSL•FO
RenderX

Introduction

Cancer is a life-threatening medical condition that can overwhelm patients with a wide range of challenges, including physical symptoms, emotional distress, financial difficulties, social isolation, etc [1-3]. For example, breast cancer, one of the most diagnosed cancers in women [4,5], traumatizes, threatens, debilitates, and devastates many female patients all over the world. Men could also endure breast cancer, with a likelihood or frequency much lower than that for women [6]. According to Breastcancer.org [7], about 32% of all new cancer cases each year are expected to be breast cancer, and it is estimated that 13% of American women will develop invasive breast cancer at some point in their lives. Although breast cancer typically has higher survival and recovery rates than some other types of cancer (eg, lung cancer and liver cancer), it usually requires ongoing care over an extended period of time [8].

The complexity of cancer care and the need for comprehensive support motivate society to provide patients with various forms of assistance in addition to medical treatments and care from physicians. Online cancer communities (OCCs) play a vital role in providing patients with cancer with technology-powered solutions to information sharing and social support [9-11]. In these communities, patients engage in interactions with other members through posting messages to share and obtain timely information (eg, treatments, symptoms, and coping strategies), medical or nonmedical experiences, care and empathy, encouragement, and companionship [10,12-18]. The tips they share about managing daily life with cancer, ranging from strategies for dealing with side effects to general nutrition advice, can help improve patients' quality of life. In this sense, engaging in OCCs can contribute to better physical health outcomes and treatment adherence.

Previous research has explored online community engagement from multiple perspectives, such as active vs passive participation, churn behavior, life span in the community, the role of moderators in improving engagement, and community members' responses to different types of initial posts [19-25]. In particular, it has been shown that other community members' responses to initial posts are important to both the initial posters and the community as a whole, because these responses not only improve the initial posters' satisfaction [26] and empowerment [27] but also contribute new information resources that help others [28,29]. Prior research on community members' responses has examined the initial poster's characteristics [21] and content [23,30], various types of social support expressed in the post content [19-22], and how participation can improve community members' emotional states [10,16,31,32]. The emotional support and sense of belonging that patients receive and develop through interacting with other members in OCCs can reduce patients' feelings of isolation and anxiety. These supports can also potentially lower the risk of depression among patients with cancer, which is crucial to their health and recovery [33-36]. Research has shown that emotions can affect a patient's acceptance of the prescribed treatments or therapeutic plan, and influence their responses to the associated stresses and anxieties [37].

Importantly, emotional expressions in OCCs are not merely reflections of patients' internal states but play a central role in affecting social interaction. Prior research conceptualizes emotional expressions as social information that helps others interpret ambiguous situations and decide whether and how to respond. In particular, the emotions as social information (EASI) model states that emotional expressions convey situational appraisals by signaling how individuals evaluate their situations, which allows observers to interpret ambiguous contexts and adjust their responses through inferential and affective processes [38]. Building on this view, Van Kleef et al [39] further argue that emotional expressions function as social signals that guide and coordinate others' behavioral responses in social interaction, with different emotions signaling distinct behavioral responses. These perspectives are especially relevant in digitally mediated health communities, where contextual cues are limited and emotional expressions often serve as primary signals of patients' needs, concerns, and circumstances.

However, most research on emotions in online communities polarized sentiments as being either positive or negative [31,40-45]. Such a dichotomy oversimplifies the complexity of human emotions, especially in situations where one must deal with highly stressful conditions such as cancer. Indeed, emotions can have multiple dimensions and are not always definitively positive or negative. For example, anger is commonly considered a negative emotion; yet, some studies have discovered its positive aspects [46]. Cancer can also evoke fear and sadness in patients when first diagnosed with cancer, and patients may later feel joy and relief when their conditions improve due to treatments. Patient emotions can also change, and their emotional intensity can vary over time [47,48], ranging from mild to strong. Patients experience these ups and downs and various emotions with changing intensities throughout their cancer management journey, and many choose to share their feelings on OCCs. It is important to recognize such complexity in both emotion types and emotion intensities expressed in patient posts and to examine how community members respond to posts that carry emotions.

To capture emotional complexity, we use Plutchik's Wheel of Emotions [49,50] to examine a more comprehensive representation of patient emotions in the context of OCCs. This framework has been widely applied to analyze emotional expressions in online text [51-53]. Figure 1 presents Plutchik's Wheel of Emotions, which consists of 8 distinct emotions: surprise, anticipation, joy, sadness, trust, disgust, fear, and anger [49]. In terms of community members' responses, according to the EASI model and the derived social signal perspective, emotional expressions function as social information and social signals in online interaction, and different types of emotions may have varying impacts on how community members interpret and respond to patient posts. Drawing on Plutchik's identification of distinct primary emotions and the EASI perspective, and taking emotion intensity into consideration, this study asks the following research question: RQ1: in OCC settings, are the type and intensity of emotions expressed in patients' initial posts associated with different patterns of community engagement and response?

XSL•FO
RenderX

Figure 1. Plutchik's 8 primary emotions.

Plutchik arranges the emotions in a circular format, with each primary emotion having an opposite emotion directly across from it on the wheel, forming 4 pairs of opposing emotions [49,50]. We modify the original Plutchik's Wheel of Emotions [49] and Plutchik's Primary Emotion Dyads [54] to show the 8 primary emotions and their opposing counterparts. This study only focuses on the 8 primary emotions in Plutchik's theory and does not consider other emotions, such as the secondary emotions that are the combinations of the primary emotions, as they are not the focus of this study. This structure reflects Plutchik's conceptualization of emotions as adaptive reactions to 4 fundamental life problems: territoriality, temporality, identity, and hierarchy [48,55]. Each life problem can be evaluated as either an opportunity or a threat, and these alternative evaluations activate different adaptive functions and corresponding emotional reactions. Emotions thus reflect not only affective states but also action readiness, such as acceptance vs rejection or protection vs confrontation. Accordingly, territoriality may evoke surprise or anticipation; temporality may evoke sadness or joy; hierarchy may evoke fear or anger, and identity may evoke disgust or trust [49].

Within the cancer context, the territoriality problem refers to patients' attempts to explore their uncertain health states. As cancer outcomes are often unpredictable, patients may experience anticipation while awaiting scan results or surprise when confronted with unexpected findings. Temporality-related situations reflect concerns about the limited duration of life, which activate patients' emotions related to life longevity (joy) or loss (sadness) as the illness and treatment progress over time. The identity problem concerns inclusion and exclusion (eg, who or what is accepted or rejected), which may reflect trust or disgust toward one's diagnosis, treatment options, or the opinions and support of others. Finally, hierarchy-related threats, such as the perceived power imbalance between oneself and an uncontrollable disease or condition, may elicit patients' fear-oriented avoidance or anger-driven confrontation.

Plutchik's Wheel of Emotions suggests that the same underlying circumstances can lead to opposing emotions through alternative adaptive appraisals. Hence, the same cancer-related situation can elicit opposite emotions, depending on how patients cognitively evaluate the situation and the adaptive reactions that are activated. In contrast, existing EASI-based research has mostly examined individual emotions separately (eg, different emotions signal distinct behavioral responses from observers) [39] and has not explicitly considered how pairs of opposite emotions operate in social interaction, especially in contexts where opposite emotions may be triggered by the same or highly similar situations. If the same situation can produce different emotional expressions, it is unclear whether observers interpret these emotions as signaling different situational contexts, which leads to different responses, or as alternative emotional reactions to the same underlying situation, which leads to similar responses. Although prior studies have linked individual emotions or polarized sentiments to engagement [31,40-45], they have not examined opposing emotions as social signals in breast cancer online communities. As a result, it remains theoretically unresolved whether opposite emotions function as equivalent or opposing social signals in affecting community members' engagement behaviors, such as whether to respond, how quickly to respond, how much effort to invest, and how closely responses align with patients' expressed concerns. To directly address this unresolved issue, we ask the following research question: RQ2: Do opposite emotions elicit opposing or similar engagement behaviors and responses from community members?

By addressing these 2 research questions, we seek to ascertain whether community members' responses are driven primarily by the specific emotion expressed (ie, treating opposite emotions

XSL•FO
RenderX

as distinct social signals) or by the underlying situational context inferred from the post (ie, treating opposite emotions as alternative expressions of the same situation).

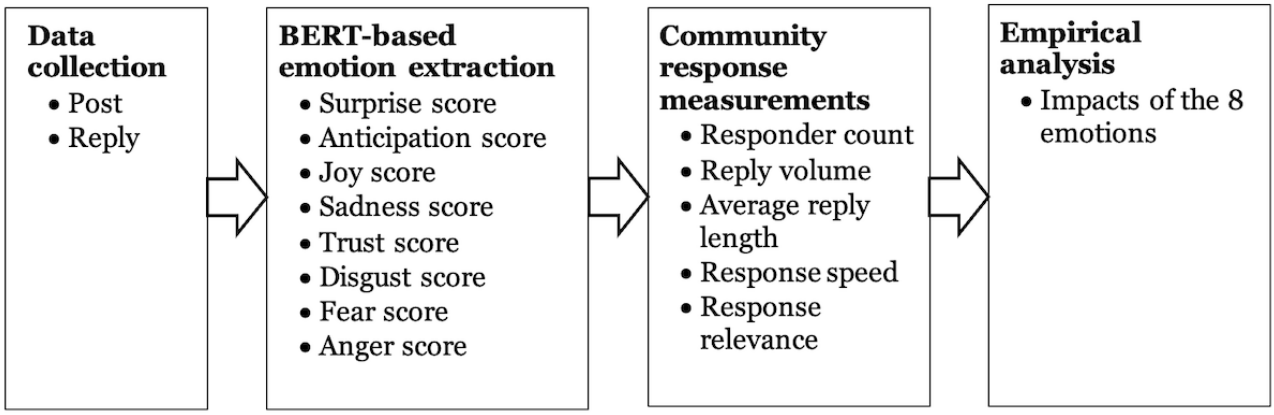
In summary, this study is aimed at providing a comprehensive and fine-grained understanding of how different types of emotions and their intensities expressed in patients with cancer's initial posts stimulate community responses in OCCs. By examining 8 primary emotions, this study moves beyond polarity-based sentiment analysis and uncovers the emotional mechanisms that affect patient engagement, peer support, and interaction dynamics in OCCs.

Methods

Study Design

Figure 2 presents the overall process of our study, which includes 4 stages: collecting posts and replies from an OCC (a breast cancer community), extracting emotion scores from initial posts, gauging engagement behaviors by analyzing community responses, and empirically analyzing the effects of different emotions on user response behavior.

Figure 2. The overall approach. BERT: Bidirectional Encoder Representations From Transformers.



Ethical Considerations

Our study uses publicly available data from an online breast cancer community, community.breastcancer.org. The research protocol was reviewed by the Institutional Review Board at Bentley University and was determined to be exempt from further review under the US Department of Health and Human Services regulations for the protection of human research participants (45 CFR 46.102[e][2][ii]), which applies to research involving observation of public behavior where the information recorded does not allow identification of the individuals.

As this study analyzed publicly accessible user-generated content and did not involve direct interaction with participants, informed consent was not required. To protect participant privacy, all data were deidentified before analysis. Usernames and other potentially identifiable information were removed, and findings are reported only in aggregated form to prevent identification of individual users.

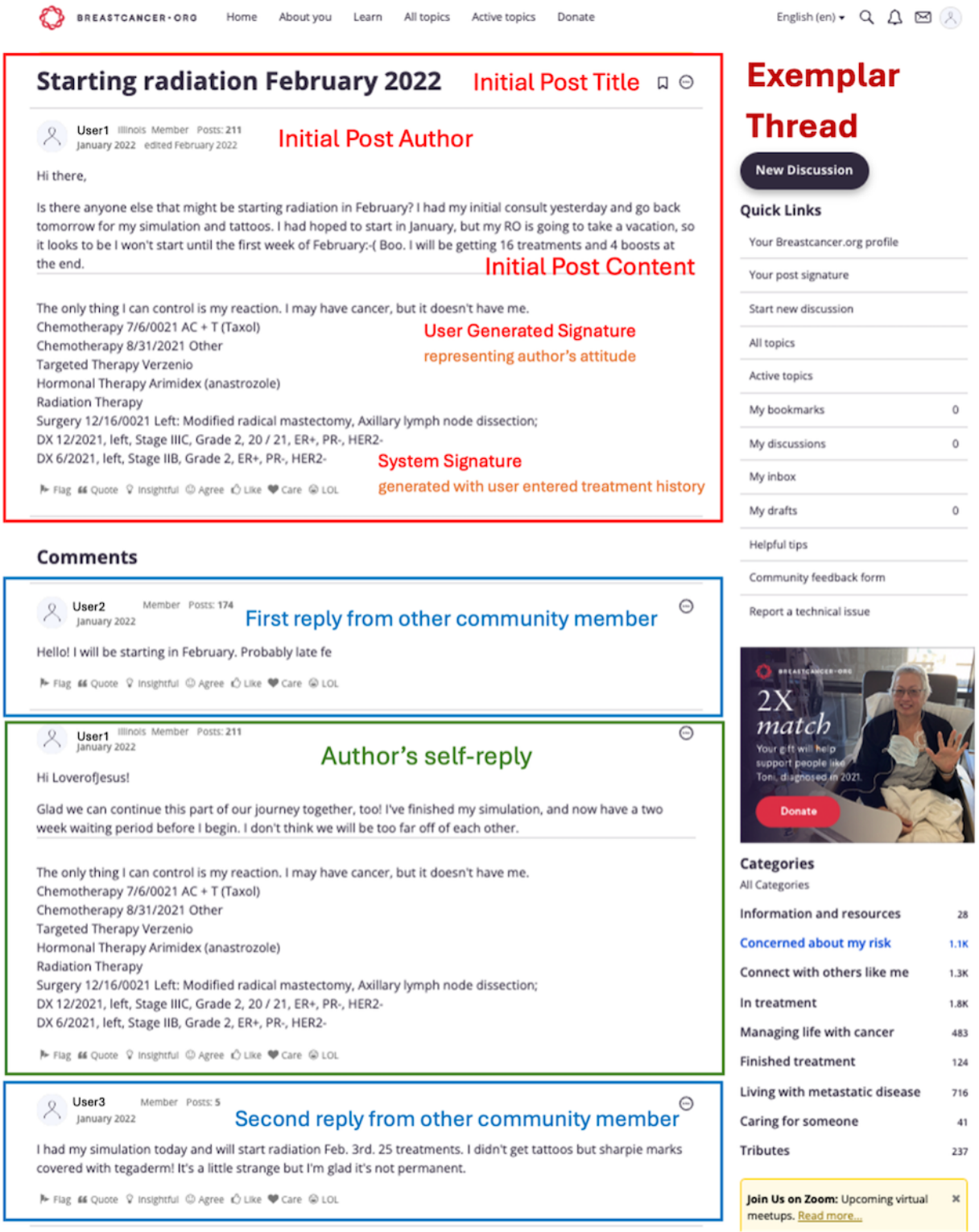
Participants did not receive compensation because this study relied solely on publicly available online posts and did not involve recruitment or intervention.

Data Collection

community.breastcancer.org is a prominent OCC for patients with breast cancer and covers a range of topics relevant to breast cancer, and provides a structured platform for patients to interact and exchange information. In this OCC, a patient in need of help typically starts a conversation (a.k.a., thread) by posting a message (a.k.a., initial post). Other members in the community can respond by adding their replies to the same thread. The initial poster may also post self-replies. Following each message (eg, the initial post and the replies) are the authors' signatures. There are 2 types of signatures—user-generated signatures and system signatures. User-generated signatures are customized texts that members choose to display, while system signatures are automatically generated using the member's self-reported diagnosis and treatment history. Figure 3 depicts a thread in community.breastcancer.org.

XSL•FO
RenderX

Figure 3. Exemplar thread from an online breast cancer community.



The forum consists of numerous threads, with each thread comprising an initial post followed by a chronological sequence of replies to that post. One patient may initiate multiple initial posts, and each initial post leads to 1 thread. Our dataset spans from October 21, 2002, to November 19, 2017, containing 101,171 threads initiated by 25,481 patients, with a total of 1,949,894 replies by the time of data collection. For the purposes of this study, we excluded threads with no replies and those whose initial posts were predominantly informational with minimal emotional expression. To identify informational posts and to further ensure that only emotionally expressive content remained, we applied a pretrained BERT (Bidirectional Encoder Representations From Transformers)–based sentiment analysis model, which classified each initial post as positive, negative, or neutral. Only threads whose initial posts were labeled as positive or negative were kept, while threads with neutral sentiment were removed because they contained little or no emotional content. After this screening and sentiment-based

XSL•FO
RenderX

filtering, the final dataset for analysis comprises 23,633 threads initiated by 9137 patients.

Emotion Extraction

We capture both the presence and strength of emotions expressed in the initial posts. We apply the BERT [56] model to extract the 8 emotions and explore their effects on 5 community response categories (detailed description in section Community Response Measurements). As it was impractical to annotate emotions in all 23,633 initial posts manually, we used a machine learning–based approach to extract emotions from the posts. For the 8 emotions, the score for each emotion ranges from 0 to 10, where 0 indicates that the emotion is not expressed in the post, and 10 indicates the highest level of emotional strength. The details of the extraction process are described in Multimedia Appendices 1 and 2, and the distribution of the estimated scores for each emotion is presented in Table 1.

Table 1. Distribution of the predicted emotion scores.

Emotion	Observations	Mean (SD)	Minimum	Maximum
Surprise	23,633	1.66 (0.46)	0.08	2.73
Anticipation	23,633	4.46 (0.86)	0.81	5.69
Joy	23,633	1.78 (1.31)	0.02	5.22
Sadness	23,633	3.28 (1.09)	0.02	4.86
Trust	23,633	3.37 (0.69)	1.83	5.68
Disgust	23,633	2.52 (1.01)	0	3.84
Fear	23,633	3.76 (1.42)	0	5.23
Anger	23,633	1.6 (0.94)	0	3.19

Community Response Measurements

In this study, the unit of analysis is per thread, and we examine community members' responses to the initial post within each thread in 5 categories and summarize them in Table 2.

Table 2. Definition and explanation of community response categories.

Community response category	Definition	Explanation
Responder count	The number of distinct participants, aside from the initial poster, contributing to the thread within 7 days.	A higher number of responders can make the original poster feel that more community members care about them and are interested in the topic. Additionally, increased participants generally demonstrate that the community is active and welcoming.
Reply volume	The number of replies to a thread within 7 days, excluding the original poster's self-replies.	This metric provides the original posters with a sense of the level of interest and attention generated by their initial post. A higher reply volume indicates greater engagement and interest in the topic.
Average reply length	The mean word count in replies, excluding self-replies by the initial poster, over the established time intervals (eg, 7 days).	This metric offers an estimate of the depth and quality of engagement from participants. Longer replies may indicate that participants are deeply engaging with the topic.
Response promptness	The time at which a thread receives a designated number of replies, excluding the initial poster's self-replies. Specifically, it measures the time in seconds taken (in logarithm) for a thread to reach the first 10 replies.	Response promptness reflects how quickly the thread obtains initial responses. For the original posters, a quicker response can indicate immediate interest or urgency regarding their posts. This can be particularly important for time-sensitive topics or when the poster is seeking rapid feedback from the community.
Response relevance	The average topical relevance between the original post and all replies received within 7 days.	This measure evaluates whether community replies meaningfully address the concerns or information expressed in the original post. To compute this score, Sentence-BERT[a] is used to generate embeddings for the original post and for all replies within a fixed time window (eg, 7 days). Semantic similarity (eg, cosine similarity, range from –1 to 1) is calculated for each reply-post pair. The response relevance is then computed as the average similarity of all replies falling within that time period. Higher scores indicate that replies are strongly aligned with the poster's topic, while lower scores suggest off-topic, generic, or mismatched responses.

[a]BERT: Bidirectional Encoder Representations From Transformers.

XSL•FO
RenderX

Empirical Analysis

We use the following ordinary least squares regression model (equation 1) to evaluate the impact of expressed emotion intensity in an initial post on subsequent community responses. The independent variables are the emotion-intensity scores for the 8 primary emotions. The model additionally controls for the medical context described in the post, user attributes, prior user participation behavior, and a range of linguistic and content-based features of the post. The dependent variables, independent variables, and all control variables are summarized in Table 3.

$$\text{CommunityResponses} = \alpha_0 + \alpha_1\text{PostEmotions} + \alpha_2\text{PostHealth} + \alpha_3\text{AuthorHealthHistory} + \alpha_4\text{AuthorCommunityActivity} + \alpha_5\text{PostChar} + \varepsilon \quad (1)$$

Table 3. Summarization of the dependent variable, independent variable, and control variables.

Variables	Description
Dependent variable	
CommunityResponses	In our analysis, we treat each of the 5 categories of community response measures as a dependent variable individually. Thus, we conduct 5 separate analyses on our data: 1 for responder count, 1 for reply volume, 1 for average reply length, 1 for response promptness, and 1 for response relevance.
Independent variable	
PostEmotions	We take the independent variable in the form of scores of the 8 emotion dimensions expressed in the initial post, for example, surprise, anticipation, joy, sadness, trust, disgust, fear, and anger.
Control variables	
PostHealth	PostHealth is a set of features regarding the number of diseases, diagnoses, drugs, examinations, procedures, and symptoms mentioned in the initial post. We used an LLM[a], gpt-3.5-turbo-0125 [57], to identify and count those terms. Multimedia Appendix 3 provides more details about the identification process. These features can indicate the complexity of the patients' conditions expressed in the initial post, which may be related to how other community members respond to the initial post and lead to more targeted, in-depth discussions.
AuthorHealthHistory	AuthorHealthHistory is a group of features containing the initial poster's medical histories, including the number of diagnoses, surgeries, chemotherapy, radiation, target-therapy, and hormonal-therapy the initial poster had before starting this thread. This information is retrieved from the system-generated signature entered by the member. These features indicate the complexity of the patients' overall health conditions, which may affect how community members respond to the author, fostering more specific and detailed discussions.
AuthorCommunityActivity	AuthorCommunityActivity represents the community participation behaviors of the initial post's author, including the number of threads the author has initialized before starting the focal one, and the number of replies the author has written to others (excluding self-replies) before initializing this focal post. The reason for including this control variable is that prior research suggests that user behavior in social communities can be influenced by others' actions and history [43,58,59]. Community members who have seen the author's previous contributions may respond differently based on prior interaction and established reputation within the community.
PostChar	PostChar represents additional characteristics of the initial post, including the number of images in the post, whether the initial poster has signatures, the day of the week the post is created, and other linguistic features, which are extracted using LIWC[b] [60]. Examples of LIWC features include word count of the post, number of "I," "we," and "you" in the post, and number of family, friend, body, money, and death-related words. These features represent the initial post's content and visibility, which further impact how community members interpret and interact with the initial post.

[a]LLC: large language model.

[b]LIWC: linguistic inquiry and word count.

Results

The results of the analysis are presented in Table 4 regarding the effects of the independent variable (ie, the 8 primary emotions) on subsequent community responses across the 5 categories. Multimedia Appendix 4 reports the comprehensive results for other variables included in equation 1. The numbers in the parentheses indicate the SEs. In Table 4, columns 1 and 2 report the impacts of emotions on the responder count and the number of replies, respectively, which show significant similarities across the 2 categories. Emotions such as surprise (P<.001 for responder count and P=.03 for number of replies) and anticipation (P<.001 for both responder count and number of replies) negatively influence the responder count and reply volume, indicating that higher levels (greater intensity) of these emotions in the initial post significantly decrease the number of participants and the number of replies in the thread. Conversely, emotions such as joy (P<.001 for both responder count and number of replies), sadness (P=.003 for responder count and P=.02 for reply volume), trust (P<.001 for responder count and P=.047 for reply volume), fear (P<.001 for responder count and P=.02 for reply volume), and anger (P<.001 for both responder count and reply volume) positively affect the number of responders and replies, with increases in emotional intensity associated with more active participation. Disgust has no significant impact on the responder count but significantly increases the reply volume. Among all emotions, sadness has the most substantial effect, with each unit increase in sadness intensity attracting an average of 1.68 (SD 0.57) additional users to respond with 2.19 (SD 0.87) more replies in the thread.

XSL•FO
RenderX

Column 3 demonstrates that the stronger expressions of anticipation (P<.001) and anger (P<.001) lead to longer replies, whereas higher-intensity presence of joy (P<.001) and sadness (P=.02) in the initial post tends to result in shorter replies. Anger and sadness exert the strongest but opposite effects on the average reply length. Column 4 shows that surprise (P<.001), anticipation (P<.001), and trust (P<.001) positively affect response promptness, meaning that greater emotional intensity in these categories delays the time it takes for an initial post to receive a certain number of replies (in this case, 10 replies). In contrast, posts expressing a higher level of intensity in joy (P<.001), fear (P<.001), and anger (P<.001) prompt quicker responses from other community members. Column 5 indicates that, when the original post expresses higher-intensity anticipation (P<.001) and fear (P=.007), community members are more likely to respond in a way that remains topically consistent with the post. However, greater intensity of joy (P<.001) is related to lower topical relevance, which suggests that replies tend to diverge from the original post's content.

Table 4. Empirical results of emotions (independent variables only).[a]

Independent variables	Responder count (1)	Reply volume (2)	Average reply length (3)	Response promptness (4)	Response relevance (5)
Surprise					
Coefficient value	–0.70[a]	–0.61[b]	1.60	0.15[a]	0.01
SE	0.18	0.27	1.25	0.04	0.003
Anticipation					
Coefficient value	–1.06[a]	–1.04[a]	4.67[a]	0.21[a]	0.03[a]
SE	0.12	0.18	0.84	0.03	0.002
Joy					
Coefficient value	0.94[a]	1.16[a]	–3.73[a]	–0.17[a]	–0.01[a]
SE	0.10	0.16	0.73	0.02	0.002
Sadness					
Coefficient value	1.68[c]	2.19[b]	–9.30[b]	–0.00	–0.02
SE	0.57	0.87	4.04	0.13	0.01
Trust					
Coefficient value	0.75[a]	0.62[b]	–2.47	0.17[a]	–0.003
SE	0.19	0.29	1.32	0.04	0.003
Disgust					
Coefficient value	0.17	0.54[b]	0.45	–0.06	0.002
SE	0.17	0.26	1.18	0.04	0.003
Fear					
Coefficient value	0.66[a]	0.40[c]	1.25	–0.10[a]	0.01[c]
SE	0.10	0.15	0.69	0.02	0.002
Anger					
Coefficient value	0.59[a]	0.81[c]	9.17[a]	–0.15[a]	–0.003
SE	0.14	0.22	0.99	0.03	0.003

[a]P<.001.

[b]P<.05.

[c]P<.01.

Table 5 summarizes the control variables.

XSL•FO
RenderX

Table 5. Control variables.

Variables	Source
PostHealth	See Multimedia Appendix 4
AuthorHealthHistory	See Multimedia Appendix 4
AuthorCommmunityActivity	See Multimedia Appendix 4
PostChar	See Multimedia Appendix 4

To ensure the validity of our findings, we expanded our analysis to include various time windows (1, 2, 3, 10, and 30 days) and to compare the effects of different emotions. The results illustrate that, across all time intervals, sadness predominantly affects responder count, reply volume, and average reply length. Higher-intensity expressions of anger also have a strong impact on average reply length, while increases in anticipation intensity exert the most substantial influence on response promptness and response relevance among all emotions. Joy remains a relatively influential emotion across all 5 response categories. The detailed analysis process is demonstrated in Multimedia Appendix 5.

Discussion

Our findings show that both the type and intensity of emotions expressed in patients' initial posts are associated with different levels of community engagement and different response behaviors in OCCs, and that opposite emotions elicit similar rather than opposing responses, which confirms our hypothesis that the emotions shaping community engagement in OCCs extend beyond a simple positive-negative polarization. Specifically regarding RQ1, higher intensity of joy, sadness, trust, fear, and anger increase responder participation and the volume of reply, while surprise and anticipation decrease them. Sadness has the largest overall effect. Stronger expressions of anticipation and anger lead to longer replies, whereas higher-intensity joy and sadness are associated with shorter replies. Surprise, anticipation, and trust slow response speed, whereas joy, fear, and anger accelerate it. In addition, higher-intensity anticipation and fear increase topical relevance in replies, while joy reduces it. Furthermore, with respect to RQ2, this study also shows that for most opposing emotion pairs (eg, surprise vs anticipation, joy vs sadness, and fear vs anger), their effects on engagement move in parallel rather than in opposing directions. These results indicate that emotional expressions convey situational appraisals (social information) and simultaneously serve as social signals that affect how community members allocate attention, effort, and timing when providing support.

Across the 5 response dimensions, our results show that different emotional expressions influence multiple dimensions of community engagement in distinct ways, for example, whether members respond, when they respond, how much effort they invest, and how relevant their responses are. Certain emotions, including joy, sadness, anticipation, and anger, exert particularly strong effects across responder count, reply volume, reply length, response speed, and topical relevance, although the direction and magnitude of these effects vary by dimension. These findings support RQ1 by demonstrating that emotional expressions do not affect engagement uniformly; instead, different types of emotions guide attention, effort, timing, and content alignment in qualitatively different ways. This pattern is consistent with the EASI framework, which emphasizes that emotions function as social information and social signals that guide observers' interpretations and behavioral decisions.

At the same time, our results indicate that emotional intensity amplifies signaling strength but does not override situational meaning. As emotional expressions are measured as continuous intensity scores, our analysis captures how increasing emotional strength alters engagement behaviors. Higher intensity generally increases the likelihood and scale of community response, but these effects are conditional on emotion type and response dimension. Emotional intensity, therefore, enhances the importance and clarity of emotional signals, which makes situational appraisals more visible to observers, without changing the situational context conveyed by a given emotion. This finding extends prior work that treats emotions categorically or dichotomously by showing that intensity complements, rather than replaces, the informational role of emotion type.

Regarding RQ2, our findings reveal that emotions defined as opposites in Plutchik's framework do not consistently function as opposing social signals and elicit opposite engagement behaviors in OCCs. Instead, for most opposing emotion pairs (eg, surprise vs anticipation, joy vs sadness, and anger vs fear), the impacts on the 5 types of community responses move in parallel rather than opposing directions. A straightforward interpretation might be that opposite emotions receive similar reactions or community engagement because they are expressed with comparable intensity. However, our descriptive statistics do not support this explanation. As shown in our descriptive statistics, opposite emotions differ substantially in both their prevalence and intensity distributions. For example, sadness is expressed more frequently and with higher average intensity than its opposite emotion, joy; likewise, fear and anger also show distinct levels of occurrence and magnitude. As these opposite emotions do not share similar intensity patterns, the parallel effects observed in our models cannot be attributed to intensity alone. Instead, our results suggest that community members respond primarily to the situational context, such as facing uncertainty, confronting threats, or processing important outcomes, conveyed by emotional expressions rather than to emotional valence. For instance, both surprise and anticipation are about patients' reactions to unpredictable situations; both joy and sadness are related to patients' important milestones or accomplished events; both fear and anger are linked to potential threats to patients [50]. When emotions share similar appraisal meanings, they can elicit similar patterns of engagement even

XSL•FO
RenderX

when they are theoretically defined as opposites. In addition, not all opposing emotion pairs behave identically; for example, 2 opposing pairs of emotions, namely surprise vs anticipation, and joy vs sadness, demonstrate completely opposite impacts across all community response categories. For example, initial posts expressing surprise or anticipation tend to receive a delayed and reduced number of replies from fewer responders. In contrast, posts expressing joy or sadness prompt quicker and more numerous replies. These contrasting response patterns may be caused by the situations described in the initial posts, for example, an unpredictable scenario vs a fact. Specifically, when other community members encounter descriptions of unpredictable situations, which are associated with surprise and anticipation, they may need more time to come up with thoughtful responses. In contrast, accomplished facts, which cause gains or losses to patients and further evoke joy or sadness, tend to elicit more immediate and numerous replies as the community members may be clearer about how to react, for example, celebrating successes or providing comfort for difficult issues. These patterns further support the EASI framework by showing that emotional expressions guide observers' interpretations of situational context and appropriate response strategies.

Overall, this study contributes to health care, psychological, and sociological research by clarifying how emotions serve as social information and social signals in high-stress health contexts. First, this study extends the EASI model by showing that emotions defined as opposites do not necessarily function as opposing social signals in OCCs. Opposite emotions can elicit similar engagement behaviors not because they share comparable intensity, but because they convey similar situational context, such as uncertainty, threat, or outcome significance, from the perspective of community members. This finding refines EASI by demonstrating that observers' responses are driven primarily by inferred situational context rather than emotional valence alone. In this sense, emotional expressions help community members infer underlying circumstances and determine appropriate forms of engagement, consistent with prior work showing that post content influences interaction patterns in online health communities [20,61,62]. Second, by modeling emotions as continuous intensity scales, this study provides a more precise account of how emotional communication affects support dynamics in OCCs. By integrating Plutchik's emotion framework with the EASI perspective, we show that these effects are caused by the combination of emotion type and emotional intensity, offering insights beyond what polarity-based sentiment analysis has discovered. Emotional intensity clearly matters, but it is not independent of emotion type or appraisal meaning; instead, it amplifies the signaling strength of emotions and influences participation, effort, timing, and topical relevance in community responses. Third, our findings confirm the key tenets of the EASI model by demonstrating that different emotions expressed in patients' initial posts are associated with distinct patterns of community engagement and response. When different emotions are expressed, community members react differently in terms of whether they respond, how quickly they respond, how much effort they invest, and how closely their replies align with patients' concerns.

Our findings also have practical implications for patients, OCC platforms, and health care practitioners by clarifying how emotional expressions influence the distribution, timing, and nature of peer support in digitally mediated health communities. From the patient's perspective, these findings help set realistic expectations about how OCCs may react and what responses they may receive. Patients may benefit from understanding that variations in response volume, speed, or depth are often a natural outcome of community dynamics rather than a reflection of the importance, legitimacy, or severity of their situation. For example, posts expressing certain emotions (eg, joy, sadness, anger, or anticipation) may be more likely to receive rapid or numerous responses, whereas others may elicit fewer or delayed replies. Increased awareness of these patterns can reduce feelings of neglect, self-blame, or disappointment and help patients interpret their online experiences more constructively. Such awareness may also encourage patients to engage more confidently with peer support resources throughout their cancer journey.

From the OCC platform and community management perspective, the findings provide actionable guidance for improving the effectiveness and equity of social support. OCCs can train moderators to recognize posts that are less likely to attract spontaneous engagement and proactively highlight or promote these posts to the community. More importantly, early prediction of low-engagement posts based on the emotions expressed enables scalable neglect-mitigation strategies. Given the large volume and length of user-generated posts in OCCs, manual monitoring is often infeasible. Emotion-aware models that predict engagement early can help identify posts at risk of being overlooked and support timely intervention. Such systems allow moderators to deploy targeted strategies aligned with the emotions expressed in the posts, such as prioritizing reassurance-oriented responses for posts expressing anticipation, or providing follow-up explanations for posts expressing surprise. By intervening strategically and broadly, OCCs can improve participation, encourage more balanced information exchange, and strengthen inclusive and supportive community environments [5,63,64].

From the health care practitioner's perspective, our findings can inform how clinicians guide patients in using online support resources as part of psychosocial support. By understanding the relationship between emotional expression and community response patterns, health care providers can better prepare patients for their online interactions, help them interpret engagement outcomes, and reduce anxiety associated with delayed or limited responses. Clinicians may incorporate such guidance into consultations to encourage appropriate and beneficial engagement with OCCs.

In addition, health care providers, OCC facilitators, and technical teams can collaborate to develop emotion-aware systems that automatically identify posts expressing heightened distress, uncertainty, or fear. These systems can support early detection and prioritization of patients who may benefit from timely psychosocial support, counseling, or patient navigation services. By complementing existing screening and triage processes, such approaches can enhance care coordination and improve the overall quality of patient support.

XSL•FO
RenderX

This study has limitations. One limitation of this study is that our data come from a single long-standing cancer community with its own communication norms, which limits the generalizability of our findings to other health conditions or support platforms. In addition, emotional expressions in online posts may not fully represent patients' true emotional states. Individuals often moderate, soften, or mask their emotions when communicating in a public forum, which means that the emotional intensity captured in their written posts may underestimate or misrepresent their actual feelings. As our analysis relies on expressed emotion as the independent variable, this moderation can weaken the correspondence between true emotional experience and textual emotional signals, which may potentially weaken the observed relationships between emotional expression and community responses. Furthermore, different community members may interpret the same emotional expression in varied ways based on their personal experiences, coping styles, or communication preferences, which introduces natural variability in how emotions are understood and reacted to within the community.

Future research can improve these limitations in several ways. First, expanding the analysis to other health conditions, support platforms, or culturally diverse communities would help assess the generalizability of these emotional dynamics and reveal whether different patient populations exhibit distinct emotional communication patterns. Second, studies may incorporate complementary qualitative or mixed method approaches to better understand how patients choose to express or withhold emotions in OCCs, and how these choices affect the interpretation and impact of their posts. Third, future work could examine how emotions evolve throughout the entire conversation rather than focusing solely on the initial post, offering deeper insight into the dynamics of emotional exchange and support over time. Fourth, researchers might explore how individual differences influence both emotional expression and community responses, such as for patients with different illness stages, coping strategies, or social support needs. Finally, investigating how emotions and other forms of content (eg, informational requests, narratives, or disclosures) interact to affect peer support could broaden our understanding of the mechanisms underlying engagement in online health communities and inform the design of more responsive and patient-centered support systems.

Acknowledgments

Grammarly (Superhuman Platform) and ChatGPT (OpenAI) were used to correct grammatical issues and typographical errors. The authors wrote all creative content and original text.

Data Availability

The data used in this study are publicly accessible through Breastcancer.org.

Funding

This work was partially supported by the Faculty Affairs Committee (FAC) research funding at Bentley University for data annotation. The article processing fee for this publication was supported by the Provost's Office at Bentley University. The funders were not involved in this study's design, data collection, analysis, interpretation of the results, or the writing of this paper.

Authors' Contributions

Conceptualization: AX, JX, HH
Data curation: AX
Formal analysis: AX, AA, JX
Investigation: AX
Methodology: AX, AA, JX, HH
Project administration: AX
Resources: AX
Validation: AX
Visualization: AX
Writing – original draft: AX, AA
Writing – review & editing: AX, AA, JX, HH
AX, AA, and JX worked on the software.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Emotion extraction process.
[DOCX File , 23 KB - medinform_v14i1e83674_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Instructions to annotators.
[DOCX File , 17 KB - medinform_v14i1e83674_app2.docx]

Multimedia Appendix 3
Medical-term extraction.
[DOCX File , 21 KB - medinform_v14i1e83674_app3.docx]

Multimedia Appendix 4
Empirical results.
[DOCX File , 37 KB - medinform_v14i1e83674_app4.docx]

Multimedia Appendix 5
Process for robustness check.
[DOCX File , 89 KB - medinform_v14i1e83674_app5.docx]

References

1. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel approach to cluster patient-generated data into actionable topics: case study of a web-based breast cancer forum. JMIR Med Inform 2018 Nov 29;6(4):e45 [FREE Full text] [doi: 10.2196/medinform.9162] [Medline: 30497991]
2. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. J Biomed Inform 2017;69:1-9 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.012] [Medline: 28323113]
3. He Y, Pang Y, Liu Y, Li Z, Wang Y, Zhang Y, et al. Companionship and sharing create social connections of an online community-based intervention for patients with cancer receiving outpatient care: pilot study. JMIR Cancer 2025;11:e64977-e64977 [FREE Full text] [doi: 10.2196/64977] [Medline: 40774327]
4. Chen W, Lu Y, Qiu L, Kumar S. Designing personalized treatment plans for breast cancer. Inf Syst Res 2021;32(3):932-949. [doi: 10.1287/isre.2021.1002]
5. Yoo W, Namkoong K, Choi M, Shah DV, Tsang S, Hong Y, et al. Giving and receiving emotional support online: communication competence as a moderator of psychosocial benefits for women with breast cancer. Comput Hum Behav 2014;30:13-22 [FREE Full text] [doi: 10.1016/j.chb.2013.07.024] [Medline: 24058261]
6. Giordano SH, Cohen DS, Buzdar AU, Perkins G, Hortobagyi GN. Breast carcinoma in men. Cancer 2004;101(1):51-57 [FREE Full text] [doi: 10.1002/cncr.20312] [Medline: 15221988]
7. Breast cancer facts and statistics. Breastcancer.org. 2026. URL: https://www.breastcancer.org/facts-statistics [accessed 2026-03-07]
8. Ferlay J, Shin H, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127(12):2893-2917. [doi: 10.1002/ijc.25516] [Medline: 21351269]
9. Ziebland S, Chapple A, Dumelow C, Evans J, Prinjha S, Rozmovits L. How the internet affects patients' experience of cancer: a qualitative study. BMJ 2004;328(7439):564 [FREE Full text] [doi: 10.1136/bmj.328.7439.564] [Medline: 15001506]
10. Yan L, Tan Y. Feeling blue? Go online: an empirical study of social support among patients. Inf Syst Res 2014;25(4):690-709. [doi: 10.1287/isre.2014.0538]
11. Haldar S, Mishra SR, Kim Y, Hartzler A, Pollack AH, Pratt W. Use and impact of an online community for hospital patients. J Am Med Inform Assoc 2020;27(4):549-557 [FREE Full text] [doi: 10.1093/jamia/ocz212] [Medline: 31986197]
12. Neal L, Lindgaard G, Oakley K. Online health communities. : ACM Press; 2006 Presented at: CHI '06 Extended Abstracts on Human Factors in Computing Systems - CHI EA '06; April 21, 2006; New York p. 444-447. [doi: 10.1145/1125451.1125549]
13. Leung L. Loneliness, social support, and preference for online social interaction: the mediating effects of identity experimentation online among children and adolescents. Chin J Commun 2011;4(4):381-399. [doi: 10.1080/17544750.2011.616285]
14. Huang KY, Chengalur-Smith I, Pinsonneault A. Sharing is caring: social support provision and companionship activities in healthcare virtual support communities. MISQ 2019;43(2):395-423. [doi: 10.25300/misq/2019/13225]
15. Zhang S, Bantum EOC, Owen J, Bakken S, Elhadad N. Online cancer communities as informatics intervention for social support: conceptualization, characterization, and impact. J Am Med Inform Assoc 2017;24(2):451-459 [FREE Full text] [doi: 10.1093/jamia/ocw093] [Medline: 27402140]
16. Attai DJ, Cowher MS, Al-Hamadani M, Schoger JM, Staley AC, Landercasper J. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. J Med Internet Res 2015;17(7):e188 [FREE Full text] [doi: 10.2196/jmir.4721] [Medline: 26228234]

XSL•FO
RenderX

17. Suarez NRE, Morrow AS, LaVecchia CM, Dugas M, Carnovale V, Maraboto A, et al. Connected and supported: a scoping review of how online communities provide social support for breast cancer survivors. J Cancer Surviv 2024:1-12. [doi: 10.1007/s11764-024-01660-w] [Medline: 39196462]
18. Feng X, Hu Y, Pfaff H, Liu S, Wang H, Wang Y. Emotional or informational? Understanding cancer patients' multifaceted help-seeking behaviors in online health communities. Soc Sci Med 2025;371:117902. [doi: 10.1016/j.socscimed.2025.117902] [Medline: 40056535]
19. Wang X, Zhao K, Street N. Social support and user engagement in online health communities. In: International Conference on Smart Health. Cham: Springer; 2014:97-110.
20. Wang Y, Kraut RE, Levine JM. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. J Med Internet Res 2015;17(4):e99 [FREE Full text] [doi: 10.2196/jmir.3558] [Medline: 25896033]
21. Lee YJ, Kamen C, Margolies L, Boehmer U. Online health community experiences of sexual minority women with cancer. J Am Med Inform Assoc 2019;26(8-9):759-766 [FREE Full text] [doi: 10.1093/jamia/ocz103] [Medline: 31361002]
22. Wang X, Zhao K, Street N. Analyzing and predicting user participations in online health communities: a social support perspective. J Med Internet Res 2017;19(4):e130 [FREE Full text] [doi: 10.2196/jmir.6834] [Medline: 28438725]
23. Xu A, Gao Y. Supporting the care to breast cancer patients with unique needs: evidence from online community members' responses. Int J Med Inform 2025;193:105695. [doi: 10.1016/j.ijmedinf.2024.105695] [Medline: 39546951]
24. Deng D, Rogers T, Naslund JA. The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: a qualitative exploratory study. J Technol Behav Sci 2023;8(2):128-139 [FREE Full text] [doi: 10.1007/s41347-023-00302-9] [Medline: 36810998]
25. Huh J, Marmor R, Jiang X. Lessons learned for online health community moderator roles: a mixed-methods study of moderators resigning from webMD communities. J Med Internet Res 2016;18(9):e247 [FREE Full text] [doi: 10.2196/jmir.6331] [Medline: 27608721]
26. Vlahovic T, Wang YC, Kraut R. Support matching and satisfaction in an online breast cancer support community. 2014 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2014; Ontario, Toronto, Canada p. 1625-6134. [doi: 10.1145/2556288.2557108]
27. Ruiz-Romeo M, Ciria-Suarez L, Medina JC, Serra-Blasco M, Souto-Sampera A, Flix-Valle A, et al. Empowerment among breast cancer survivors using an online peer support community. Support Care Cancer 2024;33(1):56 [FREE Full text] [doi: 10.1007/s00520-024-09119-5] [Medline: 39730856]
28. Bateman PJ, Gray PH, Butler BS. The impact of community commitment on participation in online communities. Inf Syst Res 2011;22(4):841-854. [doi: 10.1287/isre.1090.0265]
29. Lakhani KR, Von Hippel E. How open source software works: “free” user-to-user assistance. Res Policy 2003;32(6):923-943. [doi: 10.1016/s0048-7333(02)00095-1]
30. Xu A, Liu X, Hu P. Examining effects of fit between patient need and social support: a deep learning based multi-label classification approach. 2019 Presented at: ICIS 2019 Proceedings; Nov 12, 2019; Munich, Germany p. 1-9 URL: https://aisel.aisnet.org/icis2019/is_health/is_health/13
31. Zhang S, Bantum E, Owen J, Elhadad N. Does sustained participation in an online health community affect sentiment? AMIA Annu Symp Proc 2014;2014:1970-1989 [FREE Full text] [Medline: 25954470]
32. Zhao J, Gou L, Wang F. PEARL: an interactive visual analytic tool for understanding personal emotion style derived from social media. 2014 Presented at: IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings; 2014; New York City p. 203-212. [doi: 10.1109/vast.2014.7042496]
33. Butow P, Sharpe L, Thewes B, Turner J, Gilchrist J, Beith J. Fear of cancer recurrence: a practical guide for clinicians. Oncol J 2018;32(1):32-38 [FREE Full text] [Medline: 29447419]
34. Derogatis LR, Abeloff MD, Melisaratos N. Psychological coping mechanisms and survival time in metastatic breast cancer. JAMA 1979;242(14):1504-1508. [doi: 10.1001/jama.1979.03300140020016]
35. Hendricks L, Bore S, Aslinia D. The effects of anger on the brain and body. Nat Forum J Counsel Addict 2013;2:1-12. [doi: 10.1007/s11377-016-0110-8]
36. Pressman SD, Cohen S. Does positive affect influence health? Psychol Bull 2005;131(6):925-971. [doi: 10.1037/0033-2909.131.6.925] [Medline: 16351329]
37. Taylor SE, Stanton AL. Health Psychology. New York: McGraw-Hill Education; 2020.
38. Van Kleef GA. How emotions regulate social life: the emotions as social information (EASI) model. Curr Dir Psychol Sci 2009;18(3):184-188. [doi: 10.1111/j.1467-8721.2009.01633.x]
39. Van Kleef GA, De Dreu CKW, Manstead ASR. An interpersonal approach to emotion in social decision making: the emotions as social information model. Adv Exp Soc Psychol 2010;42:45-96. [doi: 10.1016/s0065-2601(10)42002-x]
40. Ofek N, Caragea C, Biyani P. Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. 2013 Presented at: Proceedings-International Conference on Social Intelligence and Technology, SOCIETY; 2013; CA. [doi: 10.1109/society.2013.20]

XSL•FO
RenderX

41. Portier K, Greer GE, Rokach L, Ofek N, Wang Y, Biyani P, et al. Understanding topics and sentiment in an online cancer survivor community. J Natl Cancer Inst Monogr 2013;2013(47):195-198. [doi: 10.1093/jncimonographs/lgt025] [Medline: 24395991]
42. Bui N, Yen J, Honavar V. Temporal causality analysis of sentiment change in a cancer survivor network. IEEE Trans Comput Soc Syst 2016;3(2):75-87 [FREE Full text] [doi: 10.1109/TCSS.2016.2591880] [Medline: 29399599]
43. Zhao K, Yen J, Greer G, Qiu B, Mitra P, Portier K. Finding influential users of online health communities: a new metric based on sentiment influence. J Am Med Inform Assoc 2014;21(e2):e212-e218 [FREE Full text] [doi: 10.1136/amiajnl-2013-002282] [Medline: 24449805]
44. Rodrigues RG, das Dores RM, Camilo-Junior CG, Rosa TC. SentiHealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. Int J Med Inform 2016;85(1):80-95. [doi: 10.1016/j.ijmedinf.2015.09.007] [Medline: 26514078]
45. Qiu B, Zhao K, Mitra P. Get online support, feel better? Sentiment analysis and dynamics in an online cancer survivor community. 2011 Presented at: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing; Oct 9-11, 2011; Boston, MA p. 274-281. [doi: 10.1109/passat/socialcom.2011.127]
46. Hess U. Anger is a positive emotion. In: Parrott WG, editor. The Positive Side of Negative Emotions. New York, NY: The Guilford Press; 2014:55-75.
47. Kennifer SL, Alexander SC, Pollak KI, Jeffreys AS, Olsen MK, Rodriguez KL, et al. Negative emotions in cancer care: do oncologists' responses depend on severity and type of emotion? Patient Educ Couns 2009;76(1):51-56 [FREE Full text] [doi: 10.1016/j.pec.2008.10.003] [Medline: 19041211]
48. TenHouten WD. From primary emotions to the spectrum of affect: an evolutionary neurosociology of the emotions. Neurosci Soc Sci: Missing Link 2017:141-167. [doi: 10.1007/978-3-319-68421-5_7]
49. Plutchik R. The nature of emotions. Am Sci 2001;89(4):344-350. [doi: 10.1511/2001.4.344]
50. Plutchik R. A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H, editors. Theories of Emotion. New York: Academic Press; 1980:3-33.
51. Weismayer C, Gunter U, Önder I. Temporal variability of emotions in social media posts. Technol Forecast Soc Change 2021;167:120699. [doi: 10.1016/j.techfore.2021.120699]
52. Mohsin MA, Beltiukov A. Summarizing emotions from text using Plutchik's Wheel of Emotions. 2019 Presented at: Proceedings of the 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019); May 28-29, 2019; Ufa, Russia p. 291-294. [doi: 10.2991/itids-19.2019.52]
53. Pröllochs N, Bär D, Feuerriegel S. Emotions in online rumor diffusion. EPJ Data Sci 2021;10(1):51. [doi: 10.1140/epjds/s13688-021-00307-5]
54. Plutchik R. Integration, differentiation, and derivatives of emotion. Evol Cognit 2001;7(2):113-205 [FREE Full text]
55. Plutchik R. In: Calhoun J, editor. Universal Problems of Adaptation: Hierarchy, Territoriality, Identity, and Temporality. New York: Environment and Population; 1983:223-226.
56. Devlin J, Chang MW, Lee K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference; 2018; North American p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
57. Models. OpenAI Developers. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo [accessed 2026-03-07]
58. Mehra A, Dixon AL, Brass DJ, Robertson B. The social network ties of group leaders: implications for group performance and leader reputation. Organ Sci 2006;17(1):64-79 [FREE Full text] [doi: 10.1287/orsc.1050.0158]
59. Centola D. The spread of behavior in an online social network experiment. Science (1979) 2010;329(5996):1194-1197. [doi: 10.1126/science.1185231] [Medline: 20813952]
60. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The Development and Psychometric Properties of LIWC2015. United States: University of Texas at Austin; 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2026-03-10]
61. Wang X, Zhao K, Street N. Analyzing and predicting user participations in online health communities: a social support perspective. J Med Internet Res 2017;19(4):e130 [FREE Full text] [doi: 10.2196/jmir.6834] [Medline: 28438725]
62. Biyani P, Caragea C, Mitra P. Identifying emotional and informational support in online health communities. 2018 Presented at: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers; 2014; Dublin, Ireland p. 827-836. [doi: 10.1609/aaai.v32i1.12170]
63. Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. J Biomed Inform 2013;46(6):998-1005 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.011] [Medline: 24025513]
64. Camacho N, Nam H, Kannan PK, Stremersch S. Tournaments to crowdsource innovation: the role of moderator feedback and participation intensity. J Mark 2019;83(2):138-157. [doi: 10.1177/0022242918809673]

XSL•FO
RenderX

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
EASI: emotions as social information
OCC: online cancer community

Edited by A Benis; submitted 07.Sep.2025; peer-reviewed by M Lee, D Hu; comments to author 16.Nov.2025; revised version received 27.Jan.2026; accepted 17.Feb.2026; published 13.Apr.2026.

Please cite as:
Xu A, Aljanaideh A, Xu J, Hao H
Emotion Expression in Breast Cancer Support Seeking: Empirical Study of an Online Community
JMIR Med Inform 2026;14:e83674
URL: https://medinform.jmir.org/2026/1/e83674
doi:10.2196/83674
PMID:41973505

©Anqi Xu, Ahmad Aljanaideh, Jennifer Xu, Haijing Hao. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study

Sandra L Taylor[1,2], PhD, MS; Melissa Jost[3], MS; Scott MacDonald[3], MD; Yunyi Ren[1], MS; Shelley Hilton[4], BS; Sadie Davenport[5], MLIS; Debbie Aizenberg[6], MD, MBA; Bruce Hall[3], MD, MBA, PhD; Courtney R Lyles[1,2], PhD; Jason Y Adams[7,8], MS, MD

[1]Department of Public Health Sciences, School of Medicine, University of California, Davis, 4480 2nd Avenue, Suite 4152, Sacramento, CA, United States

[2]Clinical and Translational Science Center, University of California, Davis, Sacramento, CA, United States

[3]Department of Clinical Informatics, University of California, Davis, Sacramento, CA, United States

[4]IT Clinical Applications, University of California, Davis, Sacramento, CA, United States

[5]Blaisdell Medical Library, University of California, Davis, Sacramento, CA, United States

[6]Department of Otolaryngology, Head and Neck Surgery, School of Medicine, University of California, Davis, Sacramento, CA, United States

[7]Department of Internal Medicine, School of Medicine, University of California, Davis, Sacramento, CA, United States

[8]IT Enterprise Analytics and Data Services, University of California, Davis, Sacramento, CA, United States

Corresponding Author:
Sandra L Taylor, PhD, MS
Department of Public Health Sciences, School of Medicine, University of California, Davis, 4480 2nd Avenue, Suite 4152, Sacramento, CA, United States

Abstract

Background: Physicians routinely document specifics of patient encounters in clinic visit notes, a critical but potentially time-consuming task. Ambient listening artificial intelligence (AI) technology is being integrated into clinical workflows to reduce documentation burden by creating draft visit notes. While this technology is promising, it is not perfect, and the potential for patient harm needs to be understood and mitigated. We developed and piloted an efficient, standardized approach to evaluating AI-generated notes for safety concerns in ambulatory care visits.

Objective: The objective of this quality improvement project was to develop and pilot an efficient, standardized, and scalable approach to evaluating AI-generated notes for safety concerns in ambulatory care visits.

Methods: During a 2-month pilot (July to August 2024), 31 physicians across multiple specialties used an ambient listening AI scribe to assist with the creation of 7545 clinic notes. A novel survey instrument was developed to assess note quality, focusing on 4 error types: accidental inclusions, accidental omissions, hallucinations, and bias. Physicians evaluated 356 (4.7%) AI-generated notes. Where an error was present, physicians rated its severity based on its potential to cause patient harm if it was not corrected, on a 0 to 5 scale. Additionally, a vendor-reported metric on the percentage of note content edited by physicians was analyzed.

Results: Of the 356 evaluated notes, accidental omissions were the most frequent error (n=64, 18%), followed by hallucinations (n=41, 11.5%), and accidental inclusions (n=33, 9.3%). Bias was rare (n=4, 1.1%). Most (119/142, 83.8%) errors were rated as mild to moderate (severity 1 - 3), with only 19 (5.3%) notes containing errors rated as posing serious or imminent risk (severity 4 - 5). Editing metrics across all AI-created notes showed a median of 9.0% (IQR 2.5%-21.9%) of AI-generated words were changed, with 14.9% (143/960) of notes left entirely unedited. Physician editing practices varied widely, with average percentages of AI-generated words changed ranging from 1.9% to 69.3% (median 9.0%, IQR 2.5%-21.9%).

Conclusions: AI-generated clinical notes were generally of high quality, with 94.7% (337/356) free from significant errors. However, because a small number contained errors that carried the risk of serious harm if not corrected, careful clinician review of notes remains imperative. Prior to deploying an AI scribe, organizations should pilot the technology and include an efficient review process to understand the nature and type of errors common at their organization. This pilot provides a scalable model for other health systems seeking to implement AI scribe technology responsibly.

(JMIR Med Inform 2026;14:e86474) doi:10.2196/86474

KEYWORDS

generative artificial intelligence; generative AI; artificial intelligence scribe; AI scribe; clinical documentation; ambient listening; quality of care

XSL•FO
RenderX

Introduction

Generative artificial intelligence (AI) is infiltrating and transforming many aspects of medicine, with specific emphasis on clinician-based tasks such as note creation and after-visit summaries, drafting responses to patient messages, and summarizing patient charts—all of which could reduce the clinical workload [1]. The specific uptake of AI scribes for note creation has some of the highest rates of implementation in the United States. These AI scribes use ambient listening generative AI technology to generate draft notes [2,3], capturing patient-physician conversations during encounters and converting them to documentation that is reviewed, modified, and approved by the physician. To date, the literature related to AI scribes has signaled improvements in time per note for physicians, with a potential for reducing workload and burnout [4,5].

While capitalizing on the potential benefits of AI scribes, the potential for patient harm also needs to be understood and mitigated. Generative AI is not perfect, and errors embedded in a patient's medical record could impact care [6]. Indeed, concerns have been raised about the use of generative AI to create physician responses to patient inquiries [7], and more broadly, there is a need for rigorous validation of large language models in clinical practice [8]. Real-world validation of generative AI is essential to avoiding patient harm [4], but efficient assessment tools that capture the most critical errors are not yet widely available.

Therefore, as part of a quality improvement pilot program using an ambient listening AI scribe to assist in preparing clinic visit notes, the University of California, Davis Health (UCDH) developed a novel survey instrument to assess the quality of AI-generated notes and identify errors with potential to impact patient care and outcomes. The overall goal of the quality assessment was to ensure patient safety prior to widespread deployment of the AI scribe, with the following specific objectives: develop an efficient, scalable, and standardized approach to assess the quality of AI scribe–generated clinical visit notes; identify the type and frequency of errors introduced by the AI scribe; quantify the severity of different errors and potential for patient harm, and develop a long-term monitoring approach.

Methods

Study Design

The UCDH system is an academic medical center with a 653-bed teaching hospital and the only level I adult and pediatric trauma center in inland Northern California, serving a 33-county region of about 6 million residents with more than 1.9 million patient encounters annually. UCDH conducted an ambulatory care pilot program evaluating use of an AI scribe tool to assist with preparation of clinic visit notes. The pilot program ran for 2 months (July to August 2024), during which the physicians in the pilot had the option of using the ambient listening tool. We first summarized the total number of notes created by the AI scribe technology during the pilot period, overall and by physician, followed by a secondary subanalysis of AI scribe note quality.

Ethical Considerations

This study was determined by the University of California, Davis institutional review board to not be human subjects research and therefore was exempt from review (2367684 - 1). Because this study was exempt from review and patient information was aggregated, study-specific informed consent and privacy protections were not necessary. Physicians were not compensated for participating in the pilot program.

Note Quality Assessment

All physicians in the pilot agreed to assess the quality of draft notes produced by the AI scribe technology using a novel standardized assessment instrument for a subset of their clinic encounters. To minimize assessment burden, physicians were asked to evaluate 10 draft notes on 2 different days during a 3-week period. This number was selected to represent most notes in their outpatient practice on these selected days, and each clinician received regular reminders during the pilot to complete their assessments. Note types included history and physical notes and progress notes; note type was not captured as part of the assessment.

Detailed quality ratings by physicians in the pilot were captured using a novel standardized assessment instrument accessible via an online Qualtrics (Qualtrics, LLC) survey (Figure S1 in Multimedia Appendix 1). The survey was developed in collaboration with our institution's AI Oversight Committee and included four components of quality deemed most relevant to patient safety, specifically whether the AI scribe note had the following: (1) accidental inclusions, (2) accidental omissions, (3) hallucinations, or (4) bias in the draft language (Textbox 1). These components were informed by the Physician Documentation Quality Instrument (PDQI-9) and known deficiencies in AI and large language model summarization tools [9,10]. On the survey, physicians reported if any of these types of errors were present. If so, the physician rated the severity of the error on a scale of 0 to 5, with 0 representing negligible risk to the patient, physician, or health system and 5 representing potential for serious and imminent risk of harm. Within this guidance, severity ratings were based on individual physicians' clinical judgment in the context of the visit. Table 1 provides examples of errors reported for each severity rating level.

XSL•FO
RenderX

Textbox 1. Components of note quality assessment.

Accidental inclusion

- Does the draft note contain inaccurate, real information discussed in the visit that was accidentally included or misattributed by the AI?

Accidental omission

- Does the draft omit information that you would have included in the note if you had written/dictated it without the AI?

Hallucination

- Does the draft note contain inaccurate, undiscussed, hallucinated information that was made up by the AI?

Bias

- Does the note appear to either include or omit information that might increase the risk that vulnerable populations/protected classes of patients are treated unfairly?

Table . Examples of errors in artificial intelligence (AI)–generated notes at each severity rating level.

Severity rating	Examples of AI documentation mistakes
0	• Misattributed a medication dosage statement to the parent, although it was part of the clinician's discussion of risks and benefits • Inserted advice on substance abuse that the clinician did not discuss
1	• Included irrelevant details (eg, ability to use computer) • Attributed a medication side effect comment to the patient when it was stated by the clinician • Missed noting supplement use and risk-benefit discussion • Missed noting spine clinic appointment (instead wrote referral would be placed)
2	• Electrocardiogram results reported as echocardiogram results • Incorrect timeline for COVID-19 exposure • Ovarian adenoma incorrectly documented rather than ovarian thecoma or fibroma • Medication instructions simplified incorrectly ("take for one month" instead of conditional continuation) • Anxiety medication documented for the patient instead of the spouse • Misinterpretation of recommended blood test timing ("have blood drawn in 1 week" instead of 1 week before follow-up)
3	• Included history of osteoporosis without supporting evidence • Attributed discontinuation of a medication stopped years ago to one stopped yesterday • Incorrectly stated that the patient was on methadone • Missed discussion on home safety and assisted living recommendation • Added the phrase "increase dose if symptoms improve," which was contrary to the documented treatment plan
4	• Incorrectly added directive to start aspirin for ankle pain, which was not the proposed plan • Incorrectly stated that the patient has diabetes • Did not include the patient's discussion of weight loss, fatigue, and medication change • Improper medication dose documented in plan
5	• Attributed a history provided by the patient to an incorrect diagnosis • Incorrectly stated that the patient should continue current metformin dosing • Did not document computed tomography scan discussed by the clinician

Using these surveys, we summarized the proportion of each error type during the pilot. Furthermore, we calculated a quality score as follows. For each component of quality, a score ranging from 0 to 1 was calculated as 1 – (0.2 × severity) such that higher severity errors resulted in higher point deductions. The overall quality score was then calculated as the sum of the 4

XSL•FO
RenderX

component scores, yielding a score ranging from 0 (lowest quality) to 4 (highest quality). We also linked these quality ratings to patient characteristics for each note to report on the patient distribution in the sample.

The time between the clinic visit and the quality assessment review was calculated as the difference between the time that the AI scribe recording ended and when the physician started to review the note. Data for the quality assessment were collected separately from the AI scribe recording time data, which came from the vendor. We were not able to link the quality assessments and AI scribe data for 9 visits, and recording times were not available for 4 of the assessed notes.

For each note, the vendor calculated a measure of the percentage of the note that was retained, calculated as 1 – (number of word-level edits / number of words in the final completed note). Word-level edits are defined as additions, substitutions, and deletions. The vendor provided the average percentage retained for each physician during the pilot period. Due to technical issues, this measure could only be obtained at the note level from the vendor for 960 notes. We converted the percentage retained metric to the percentage of words changed.

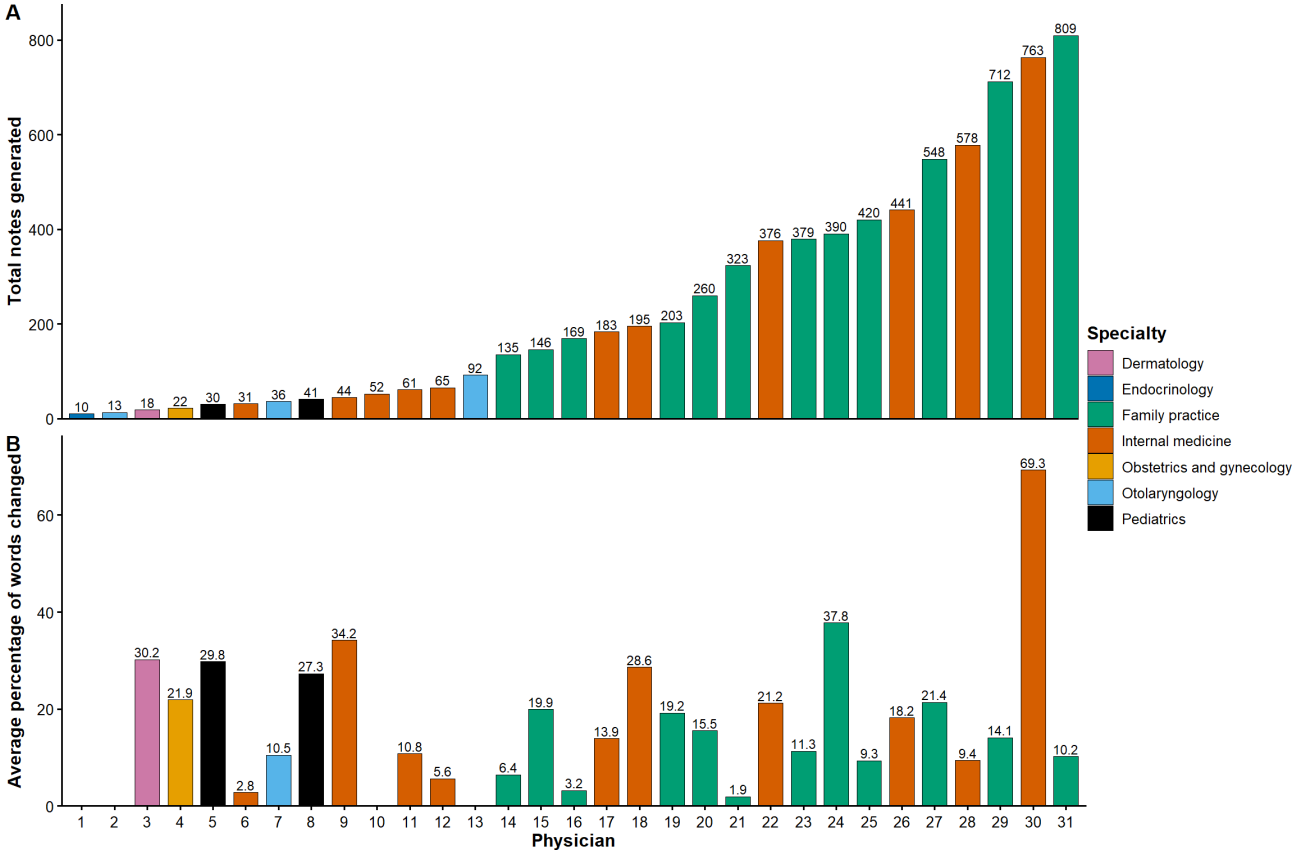
Results

A total of 31 physicians volunteered for the pilot program. Specialties of physicians in the pilot program included family practice (n=12, 38.7%); internal medicine (n=11, 35.5%); otolaryngology (n=3, 9.7%); pediatrics (n=2, 6.5%); and dermatology, obstetrics and gynecology, and endocrinology (n=1 each, 3.2%).

Quality Assessment

Over the pilot period, a total of 7545 notes were generated using the AI scribe by the 31 physicians in the pilot program. There was substantial variability in the number of notes generated across physicians, with a median (IQR) of 169.0 (42.5-384.5) notes, ranging from 10 to 809 (Figure 1).

Figure 1. (A) Total number of notes generated during pilot period by each physician and (B) average percentage of words changed in artificial intelligence (AI) scribe–generated notes by each physician. Due to technical issues, the percentage of words changed was not recorded for some physicians.



Physicians evaluated the quality of 356 (4.7%) of 7545 notes, with a median number of notes evaluated by a physician of 11.0 (IQR 8.5-13.0). Family medicine and internal medicine physicians evaluated most of the notes, 182 (51.1%) and 115 (32.3%), respectively, as these specialties comprised three-quarters (23/31, 74.2%) of the participating physicians. Entering these data for the quality assessments into Qualtrics required a median of only 38.0 (IQR 18.0-96.2) seconds. Physicians usually conducted the quality assessments on the same day as the clinic visit, yielding a median time between the clinic visit and the quality assessment of 3.9 (IQR 1.6-21.1) hours. Among these 356 notes that were scored by physicians, aggregated patient characteristics are summarized in Table S1 in Multimedia Appendix 2, demonstrating a wide representation among patients seen in the health care system.

Error Frequency and Severity

Among the 356 notes, accidental omissions were the most prevalent quality concern, occurring in 18% (n=64) of the notes, followed by hallucinations (n=41, 11.5%) and accidental inclusions (n=33, 9.3%; Table 2). Bias was only reported in 4

XSL•FO
RenderX

notes (1.1%). In general, quality concerns tied to omissions, hallucinations, or accidental inclusions were rated as mild to moderate (severity ratings of 1 - 3). Severity scores of 4 to 5, representing potentially serious or imminent risk of harm if draft errors were not corrected by clinicians, were infrequent—with only 7 (2%) draft notes in this range for accidental omissions, 9 (2.5%) for hallucinations, 5 (1.4%) for accidental inclusions, and 2 (0.6%) for bias.

Finally, the combined note quality score (accounting for severity) was high, with a median quality score of 4.0 (IQR 3.8-4.0). Two-thirds (n=242, 68%) of the notes received a score of 4.0, indicating no quality concerns. A small number of notes (n=11, 3.1%) had an overall score less than 3. Note quality was similar across the specialties (Table 3; Table S2 in Multimedia Appendix 1).

Table . Summary of concern severity ratings for quality assessments of visit notes (N=356).

Severity score	Inclusion, n (%)	Omission, n (%)	Hallucination, n (%)	Bias, n (%)
5 (severe)	1 (0.3)	3 (0.8)	3 (0.8)	—[a]
4	4 (1.1)	4 (1.1)	6 (1.7)	2 (0.6)
3	9 (2.5)	12 (3.4)	14 (3.9)	1 (0.3)
2	7 (2.0)	17 (4.8)	10 (2.8)	—
1	12 (3.4)	28 (7.9)	8 (2.2)	1 (0.3)
0 (no concerns)	323 (90.7)	292 (82.0)	315 (88.5)	352 (98.9)

[a]No notes were identified with bias with this severity score.

Table . Summaries of total scores from the note quality assessments and the percentage of each note edited, as reported by the vendor, by specialty.

Specialty[a]	Total score (n=356)		Percentage edited (n=960)	
	Notes, n (%)	Median (IQR; range)	Notes, n (%)[b]	Median (IQR; range)
Dermatology	11 (3.1)	4.00 (3.50-4.00; 2.60 - 4.00)	0 (0)	—[c]
Family practice	182 (51.1)	4.00 (3.80-4.00; 1.80 - 4.00)	571 (59.5)	8.3 (2.4-17.3; 0.0 - 63.5)
Internal medicine	115 (32.3)	4.00 (3.60-4.00; 1.40 - 4.00)	364 (37.9)	9.5 (2.1-25.5; 0.0 - 100.0)
Obstetrics and gynecology	4 (1.1)	4.00 (3.80-4.00; 3.20 - 4.00)	0 (0)	—
Otolaryngology	26 (7.3)	4.00 (4.00-4.00; 4.00 - 4.00)	21 (2.2)	34.4 (21.2-46.8; 8.9 - 91.2)
Pediatrics	18 (5.1)	3.80 (3.45-4.00; 3.00 - 4.00)	4 (0.4)	22.4 (16.2-33.0; 10.7 - 51.7)

[a]No quality assessments were conducted by the endocrinologist.

[b]Due to technical difficulties with the data from the vendor, the percentage edited was only obtained at the note level for a portion of all notes and was not available for any notes for dermatology, obstetrics and gynecology, or endocrinology.

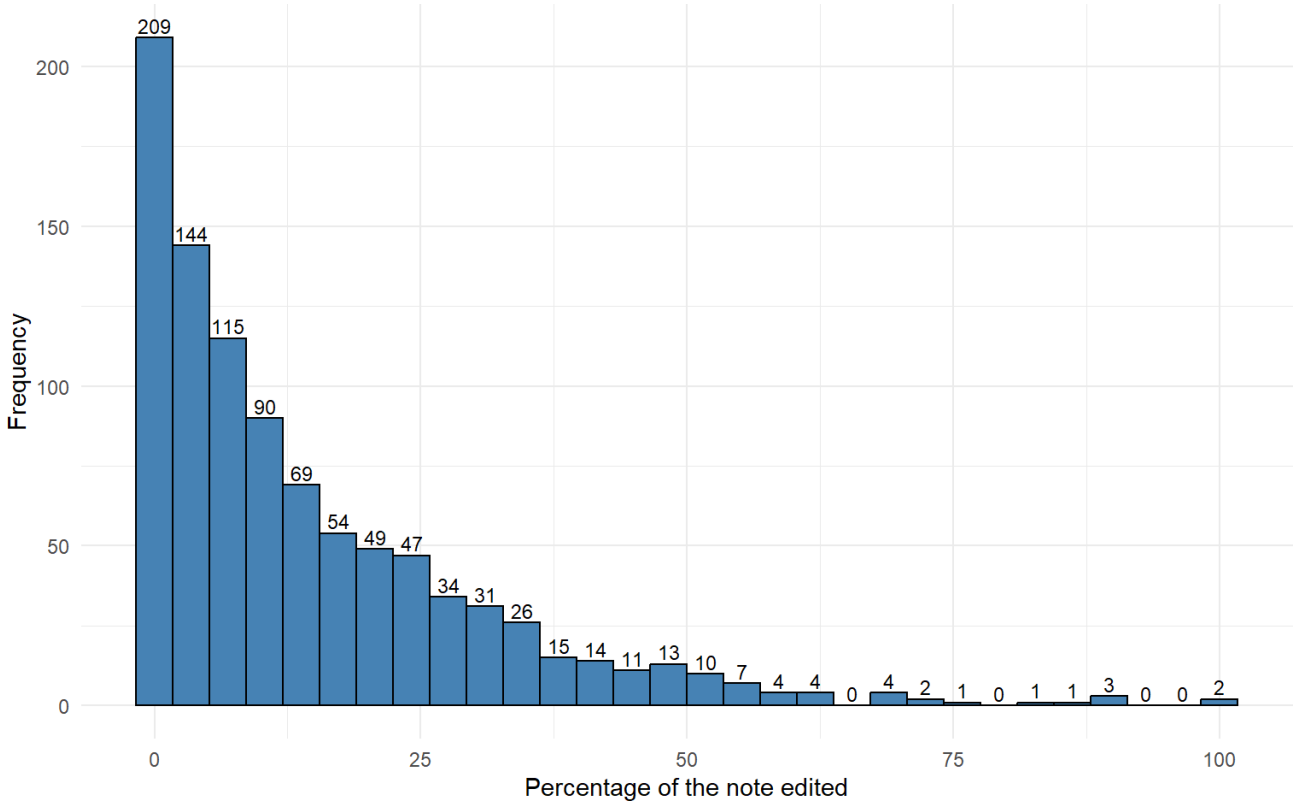
[c]Percentage edited was not available for any artificial intelligence–generated notes for this specialty.

Percentage Edited

Of the 7545 AI-created notes, the vendor metric of percentage of words retained was reported for 960 (12.7%) notes. Most of the AI-drafted note content was left unedited by physicians, with a median 9.0% (IQR 2.5%-21.9%; Figure 2) of AI-generated words changed. Notably, 14.9% (143/960) of the AI-generated notes were not edited at all. Individual editing practices varied substantially by physician, with the average percentage of words changed by physician ranging from 1.9% to 69.3% (Figure 1). At the note level, the percentage edited was similar for family practice (median 8.3%, IQR 2.4%-17.3%) and internal medicine (median 9.5%, IQR 2.1%-25.5%) but higher for otolaryngology (median 34.4%, IQR 21.2%-46.8%) and pediatrics (median 22.4%, IQR 16.2%-33.0%; Table 3).

XSL•FO
RenderX

Figure 2. Distribution of percentage of words changed in artificial intelligence–generated clinic visit notes. The x-axis indicates the percentage of the note that was changed, and the y-axis indicates the number of notes with that level of editing.



Discussion

Consistent with other studies, we found that most (337/356, 94.7%) AI scribe platform–generated notes were of high quality and free from significant errors [11,12]. However, accidental omissions were relatively common, identified in 18% (64/356) of the notes evaluated. Most errors were rated as low risk for patient harm; however, 5.3% (19/356) of AI-drafted notes contained an error deemed to pose a risk of serious harm if left uncorrected. Thus, careful clinician review of notes remains imperative.

Because we prioritized patient safety and specifically focused on identifying errors or bias in AI-generated notes, our evaluation tool to capture physician ratings was very brief, allowing rapid assessment of a large number of notes. This contrasts with the longer PDQI-9 instrument that evaluates overall document quality. Natural language processing metrics have also been used to evaluate note quality in previous studies [13], but these metrics do not explicitly identify deviations that could impact patient safety. Our note quality assessment approach is therefore unique in rating the severity of risk posed to patients if the error is not corrected.

Several limitations were encountered in implementing and evaluating this program. First, this was a pilot program, and clinicians volunteered as early adopters of AI scribe technology within the health care system. Second, although the assessment rubric was short, most physicians in the pilot assessed fewer than the requested 20 notes, despite repeated outreach from the project team. Thus, our method of including participants in the pilot and their differential response rates could have biased our results. Third, because physicians reviewed notes from their own clinic visits, identification and rating of the severity of the errors could be biased due to the physician's individual perspective as well as the passage of time between the visit and note review. However, physicians tended to review notes the same day as the clinic visit, which would reduce recall bias. Finally, to achieve the objectives of our quality improvement project, we did not seek to develop a rigorously validated instrument. Severity ratings were subjective, based on individual physician's clinical judgment within the context of a particular patient, such that scores could vary among institutions and clinical departments, thus requiring local calibration. Future studies and improvements could include rigorous development of the instrument through multiple reviewer assessments of AI-generated notes relative to recorded transcripts, standardization of severity ratings, and external validation.

In the future, ongoing monitoring of the quality of AI-generated notes will be important to identify changes in the underlying algorithm or practices that could degrade note quality. While our assessment tool is time efficient, it still requires human review, which might impede its use for long-term monitoring. Importantly, we combined direct physicians' ratings with vendor-reported metrics that are more easily obtainable, specifically the percentage of each note that was edited, as a secondary indicator of note quality. Using direct feedback and reporting from end users as well as automatically generated information from platforms will be important for a long-term monitoring approach. Given the occurrence of errors in AI-generated notes, monitoring trends, both in the percentage of notes that are heavily edited or notes that are not edited, will facilitate early identification of changes in physician vigilance

XSL•FO
RenderX

in note review. Finally, as our health care system moves toward broader implementation of an AI scribe platform, we are investigating the standardization of quality monitoring by correlating vendor metrics with physician ratings. Given that physician behavior can change over time when adopting AI tools into routine practice, and the performance of the tools themselves may drift, our health system has prioritized long-term monitoring of quality and safety.

In summary, we have outlined a pilot program that might be a generalizable model for other health care systems. Our focused priorities on quality and safety of AI-generated notes, while also maintaining feasibility in the evaluation, were essential for piloting and will likely remain so for broad-scale implementation. Furthermore, we continually engage with the AI vendor to inform their automated metrics that align with our ongoing evaluation priorities. Moving forward, we also plan to implement these approaches and our findings into clinician training in the use of AI technology. With the promise of AI, particularly tools aimed at reducing physician workload and burnout, there is a pathway for rapid deployment while maintaining high standards of patient care.

Funding

Funding was provided by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant UL1 TR001860.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MJ, SM, CRL, JYA
Data curation: SH
Formal analysis: SLT, YR
Methodology: SLT, MJ, SM, JYA
Project administration: MJ, DA, BH
Resources: MJ, DA, BH
Supervision: SLT
Visualization: SLT, YR
Writing—original draft: SLT, CRL, JYA
Writing—review and editing: SLT, MJ, SM, SD, DA, BH, CRL, JYA

Conflicts of Interest

None declared.

Multimedia Appendix 1
Note quality survey user interface and error types by specialty.
[DOCX File, 191 KB - medinform_v14i1e86474_app1.docx]

Multimedia Appendix 2
Patient demographics.
[DOCX File, 16 KB - medinform_v14i1e86474_app2.docx]

References

1. Lee C, Vogt KA, Kumar S. Prospects for AI clinical summarization to reduce the burden of patient chart review. Front Digit Health 2024;6:1475092. [doi: 10.3389/fdgth.2024.1475092] [Medline: 39575412]
2. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
3. Shah SJ, Devon-Sand A, Ma SP, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc 2025 Feb 1;32(2):375-380. [doi: 10.1093/jamia/ocae295] [Medline: 39657021]
4. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J 2024 Sep;11(3):100157. [doi: 10.1016/j.fhj.2024.100157] [Medline: 39371531]
5. Albrecht M, Shanks D, Shah T, et al. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open 2025 Feb 21;8(1):ooaf013. [doi: 10.1093/jamiaopen/ooaf013] [Medline: 39991073]

XSL•FO
RenderX

6. Mess SA, Mackey AJ, Yarowsky DE. Artificial intelligence scribe and large language model technology in healthcare documentation: advantages, limitations, and recommendations. Plast Reconstr Surg Glob Open 2025 Jan 16;13(1):e6450. [doi: 10.1097/GOX.0000000000006450] [Medline: 39823022]
7. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381. [doi: 10.1016/S2589-7500(24)00060-8] [Medline: 38664108]
8. de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health 2024 Jul;6(7):e441-e443. [doi: 10.1016/S2589-7500(24)00111-0] [Medline: 38906607]
9. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174. [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
10. Gerke S, Simon DA, Roman BR. Liability risks of ambient clinical workflows with artificial intelligence for clinicians, hospitals, and manufacturers. JCO Oncol Pract 2026 Mar;22(3):357-361. [doi: 10.1200/OP-24-01060] [Medline: 40749149]
11. Palm E, Manikantan A, Mahal H, Belwadi SS, Pepin ME. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. Front Artif Intell 2025;8:1691499. [doi: 10.3389/frai.2025.1691499] [Medline: 41199808]
12. Cain CH, Davis AC, Broder B, et al. Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. NEJM AI 2025 Mar 27;2(4). [doi: 10.1056/AIcs2400977]
13. Gebauer S. Benchmarking and datasets for ambient clinical documentation: a scoping review of existing frameworks and metrics for AI-assisted medical note generation. medRxiv. Preprint posted online on Jan 29, 2025. [doi: 10.1101/2025.01.29.25320859]

Abbreviations

AI: artificial intelligence
PDQI-9: Physician Documentation Quality Instrument
UCDH: University of California, Davis Health

Edited by A Benis; submitted 24.Oct.2025; peer-reviewed by AJ Lisi, D Chartash; revised version received 29.Jan.2026; accepted 06.Mar.2026; published 17.Apr.2026.

Please cite as:
Taylor SL, Jost M, MacDonald S, Ren Y, Hilton S, Davenport S, Aizenberg D, Hall B, Lyles CR, Adams JY
Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study
JMIR Med Inform 2026;14:e86474
URL: https://medinform.jmir.org/2026/1/e86474
doi:10.2196/86474

© Sandra L Taylor, Melissa Jost, Scott MacDonald, Yunyi Ren, Shelley Hilton, Sadie Davenport, Debbie Aizenberg, Bruce Hall, Courtney R Lyles, Jason Y Adams. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of an Ambient AI Scribe Among Clinicians and Patients: Real-World Prospective Observational Time-Motion Study

Jonathan Yue En Tan[1*], BEng; Iffat Bin Mohammad Rafi[1*], BEng; Gerald Gui Ren Sng[2,3], MBBS, MRCP, MMed, MPH; Joshua Yi Min Tung[2,4], MBBS, MRCS, MPH; Daniel Yan Zheng Lim[2,5], MBBS, MMed, MRCP, Mtech; Jasmine Chiat Ling Ong[6,7], PharmD; Gek Hsiang Lim[8], MSc; Eileen Yi Lee Lew[9], PhD; Kuok Wei Chia[1], BSc; Mark Kok Hong Heng[9], MTech

[1]Department of Future Health System, Singapore General Hospital, 1 Hospital Blvd, SingHealth Tower, Singapore, Singapore
[2]Data Science and Artificial Intelligence Laboratory, Singapore General Hospital, Singapore, Singapore
[3]Department of Endocrinology, Singapore General Hospital, Singapore, Singapore
[4]Department of Urology, Singapore General Hospital, Singapore, Singapore
[5]Department of Gastroenterology and Hepatology, Singapore General Hospital, Singapore, Singapore
[6]Division of Pharmacy, Singapore General Hospital, Singapore, Singapore
[7]Duke-NUS AI + Medical Sciences Initiative, Duke-NUS Medical School, Singapore, Singapore
[8]Health Services Research Unit, Singapore General Hospital, Singapore, Singapore
[9]Office of Insights and Analytics, SingHealth, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Jonathan Yue En Tan, BEng
Department of Future Health System, Singapore General Hospital, 1 Hospital Blvd, SingHealth Tower, Singapore, Singapore

Abstract

Background: Ambient artificial intelligence scribes are increasingly used to reduce clinician burnout and cognitive load, although their impact on documentation time remains inconsistent across studies. Most existing real-world impact studies have been conducted in the United States and rely on electronic health record time stamps, which may not accurately reflect actual documentation time.

Objective: This study aimed to quantify the impact of ambient scribe technology on clinician efficiency and patient engagement in an outpatient setting using direct observational methods and patient surveys.

Methods: We conducted a prospective, within-clinician quality improvement study at a large academic medical center in Singapore from December 2024 to May 2025. Nine clinicians participated in matched observation sessions with and without an in-house ambient scribe tool. Five trained observers used standardized time-motion methodology to capture documentation time, proportion of eye contact, and consultation duration across 169 consultations. Patient surveys were administered following consultations using the ambient scribe. Linear mixed-effects models were used to account for clustering within clinicians.

Results: Ambient scribe use was associated with a 15.0% reduction in documentation time per consultation (mean 5.3, SD 2.7 minutes to 4.5, SD 2.4 minutes; *P*=.04); a 10.6% increase in proportion of eye contact time (mean 69.6%, SD 18.6% to 77.1%, SD 17.7%; *P*=.009); and no significant change in consultation duration (mean 11.5, SD 6.9 minutes vs 10.9, SD 5.6 minutes; *P*=.42) or total cycle time per patient (mean 13.7, SD 7.7 minutes vs 13.2, SD 6.3 minutes; *P*=.57), where total cycle time includes preclerking, consultation, and postconsultation documentation. Effects were consistent across new and follow-up patients. Patient acceptance was favorable: of 39 surveyed patients, 27 (69.2%) agreed that their physician focused on them more during the consultation, and none expressed discomfort with the technology.

Conclusions: This evaluation of ambient scribe use in an Asian health care setting demonstrates that experienced users achieved reduced documentation time and improved patient engagement without affecting consultation duration or total time per patient. The technology was also well accepted by patients. Taken together, these findings suggest that ambient scribes reallocate clinician effort toward patient interaction rather than enabling faster patient turnover, supporting their implementation across diverse health care settings.

(JMIR Med Inform 2026;14:e85580) doi:10.2196/85580

XSL•FO
RenderX

KEYWORDS

ambient intelligence; generative artificial intelligence; clinical documentation; physician-patient interaction; time-motion study; eye contact; ambient scribe; artificial intelligence; AI

Introduction

Ambient scribes use artificial intelligence (AI) to convert verbal clinician-patient interactions into structured clinical documentation. This is one of the fastest-adopted uses of generative AI in health systems worldwide [1].

Recent real-world studies across health systems in the United States demonstrate consistent improvements in clinician well-being and patient interactions with the use of ambient scribes, with studies reporting significant reductions in cognitive load, task burden, and burnout scores alongside enhanced patient experience during consultations [2-7]. However, reported time savings vary considerably between organizations. While some implementations report substantial reductions in documentation time and after-hours charting, others show minimal productivity gains [8,9].

To our knowledge, no comprehensive real-world impact studies have been conducted outside of the United States, leaving questions about the technology's effectiveness in other health care contexts unanswered. Documentation burdens and clinical workflows may differ between countries, potentially affecting the magnitude of time savings and productivity gains achievable through the use of clinical ambient scribes. Additionally, no published studies have evaluated clinical ambient scribe outcomes in multilingual clinical environments despite the prevalence of such settings in many health care systems worldwide. Shah et al [7] have previously highlighted limited functionality with non–English-speaking patients as a specific barrier to ambient scribe adoption. As digital health tools increasingly permeate global health care systems, understanding their implementation in linguistically diverse environments is important for ensuring equitable and effective deployment.

From a methodological perspective, existing studies have relied primarily on electronic health record (EHR) time stamp analysis to measure documentation time and productivity outcomes [4-6,10]. While convenient for large-scale analysis, EHR-based metrics provide only proxy measures that may not accurately reflect the actual time spent on documentation tasks. These metrics often rely on time stamps that capture when a note window is open, which may overestimate documentation time if clinicians multitask or leave notes open [11]. Direct time-motion observation, though more resource intensive, offers superior accuracy in distinguishing active documentation from passive or multitasking periods [12].

To address these evidence gaps, we aimed to determine the effectiveness and acceptability of ambient scribe technology when compared to the standard of care in a multilingual health care setting using direct observational methods.

Methods

Study Design

We conducted a prospective within-clinician observational study to evaluate the real-world clinical impact of an in-house ambient AI scribe in outpatient clinics of Singapore General Hospital, a quaternary referral center with more than 1700 beds that is the largest academic medical center in Singapore [13]. This study was conducted over 5 months between December 27, 2024, and May 30, 2025. It builds on a previous pilot study at the same institution that demonstrated the ambient scribe tool's ability to generate high-quality clinical notes comparable to those written by clinicians [14].

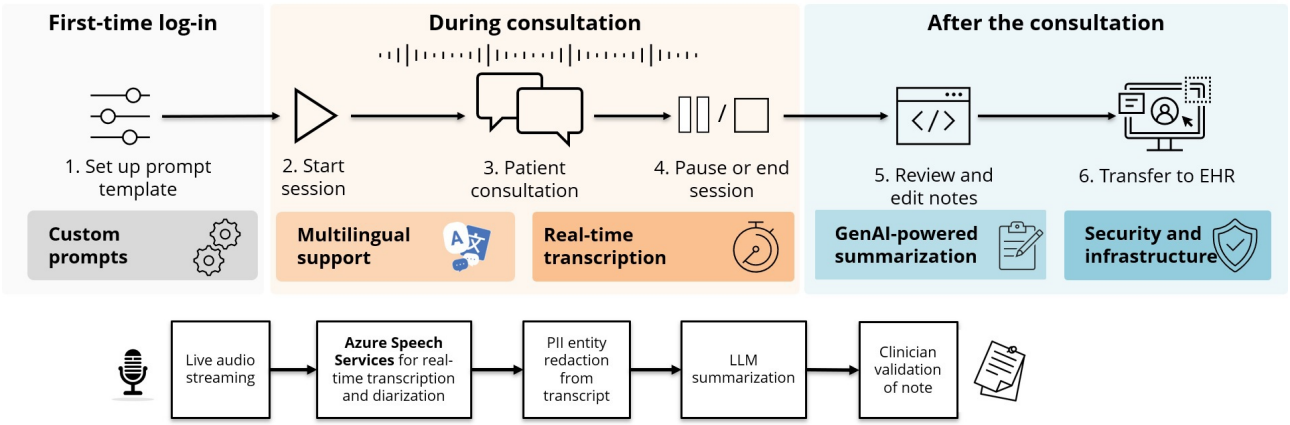
Ambient Scribe Technology

Note Buddy, the ambient scribe tool used in this study, was developed in-house using automatic speech recognition and large language model technology. This tool supports transcription in 5 languages commonly used in Singapore: English, Mandarin, Malay, Tamil, and Cantonese. However, during the study period, only English, Mandarin, and Malay consultations were observed.

Note Buddy can process conversations of up to approximately 96,000 words (based on token limits), identifying speakers and languages at the word level. It leverages GPT-4o's (OpenAI) native multilingual capabilities to process mixed-language transcripts without pretranslation or model fine-tuning. Note Buddy then reformats and summarizes transcripts into clinical consultation notes. On the basis of routine system use, note generation (from end of consultation to completed draft) typically requires 10 to 30 seconds, with duration varying by consultation length and prompt complexity. The system processes audio in real time without storing recordings (Figure 1).

XSL•FO
RenderX

Figure 1. Workflow of Note Buddy use in clinical practice. Users configure prompt templates that specify the structure and formatting of the generated notes. During consultations, the system transcribes clinician-patient interactions in the supported languages. After each consultation, the ambient scribe processes the transcript using the predefined prompt to generate a draft note, which clinicians review, edit if needed, and manually transfer into the electronic health record (EHR). GenAI: generative artificial intelligence; LLM: large language model; PII: personally identifiable information.



Users can customize the generated notes by creating prompt templates specifying note structure and clinical terminology, providing example notes for guidance, or selecting from preconfigured prompts in the prompt library. As the tool was not integrated into the EHR during the study period, clinicians had to manually transfer the AI-generated notes from the ambient scribe into the EHR.

Participants and Setting

This study was conducted in an outpatient clinic setting where consultations naturally occur in multiple languages according to patient preferences. We recruited clinicians who had generated at least 30 notes using the ambient scribe tool in the 3 months prior to study commencement.

Clinicians unable to participate in 2 consecutive clinic sessions were excluded. Due to the time-intensive nature of direct observation, we intentionally focused on experienced clinician users who had achieved proficiency with the system and were most likely to demonstrate its optimal impact. From 27 eligible clinicians, 9 (33.3%) were selected through convenience sampling based on scheduling feasibility and observer availability.

Each clinician selected 2 clinic sessions for observation: one session using the ambient scribe tool and one using standard documentation methods. Clinicians were asked to select sessions with a similar patient case mix and clinic characteristics. Standard documentation methods consisted of manual typing, with some clinicians potentially using built-in EHR features such as smart phrases. The use of such features was not systematically recorded. Each clinician decided whether to use the ambient scribe tool in the first or second observed session; this was not randomized.

The interval between the 2 clinic sessions was kept within 2 weeks to minimize changes in practice patterns over time. Scheduled appointments remained unchanged to reflect the natural distribution of new and follow-up patients in the real-world setting. The total number of patients per clinic session ranged from 8 to 15.

Data Collection: Observation Methodology

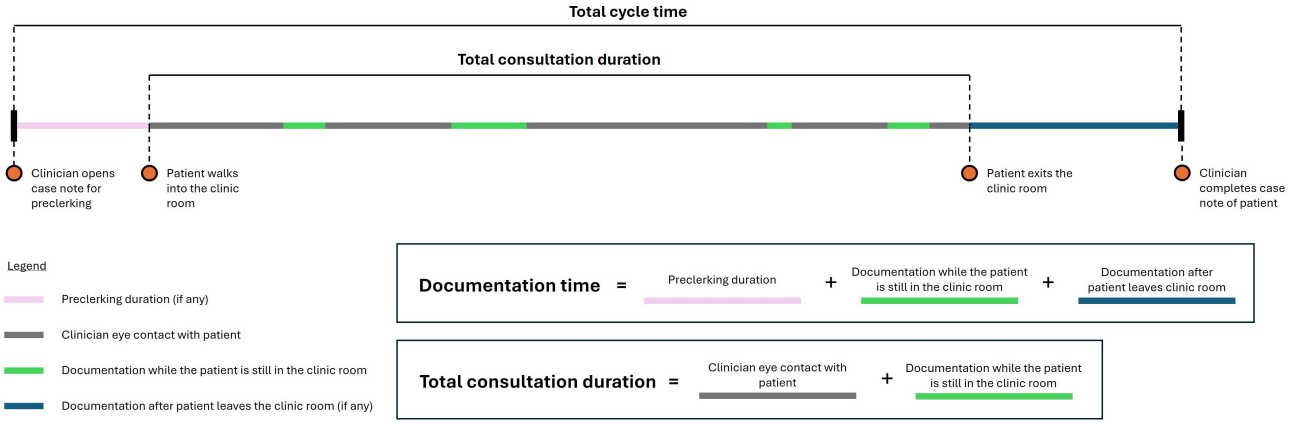
Five observers carried out the observations, with each observer consistently assigned to the same clinician across both clinic sessions. All observations were conducted in person, with the observer present in the consultation room during scribe and nonscribe sessions.

All 5 observers participated in a joint training session to minimize interobserver variability. This included instruction on using a macro-enabled Microsoft Excel worksheet for time-motion data collection. The session concluded with group practice on mock consultations until a consensus was reached on measurement definitions and recording techniques.

Using the standardized macro-enabled Microsoft Excel worksheet, observers recorded real-time data on 3 key metrics and 1 derived metric. Consultation duration was calculated as the time between patient entry and exit from the consultation room. Proportion of eye contact was defined as the percentage of the consultation duration for which the clinician was looking directly at the patient. Documentation time was measured as the total time that the clinician interacted with the computer across all phases of the patient encounter: preclerking (if any), during the consultation, and immediately after the consultation, as illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Schematic representation of time components measured in the observational time-motion study. Documentation time is the time taken for documentation performed before, during, or immediately after the consultation, defined as the total time the clinician interacted with the computer. Consultation duration is the time between patient entry and exit from the consultation room. Total cycle time is derived from the start of preclerking (if any) to completion of documentation. All time points were recorded in real time by trained observers using a standardized macro-enabled Microsoft Excel worksheet.



Total cycle time was calculated as a derived metric from preclerking start (if any) to documentation completion, as illustrated in Figure 2. This represents the complete time investment per patient encounter and directly reflects potential impact on clinic throughput.

As this ambient scribe was designed for real-time documentation, observers recorded documentation completed within the observation window, which required clinicians to finish notes during or immediately after consultations rather than deferring to later periods. In the sessions with ambient scribe use, documentation time included the time spent reviewing, editing, and manually transferring AI-generated notes into the EHR, as well as the time taken for the ambient scribe to generate the note. Observers did not separately track the time spent on each component.

Data Collection: Sample Size Explanation

Due to the resource-intensive nature of direct observation and the pragmatic evaluation design, we determined the sample size based on clinician availability and feasibility constraints rather than formal power calculations. A post hoc approximation using 2-tailed paired *t* test assumptions (effect size=0.5; α=.05; power=0.95) indicated a minimum of 54 consultations per group (total 108 consultations). This simplified calculation does not account for the clustering of consultations within clinicians, which is addressed by our mixed-effects models. It serves as a conservative proxy given the complexity of power calculations for such models. Our final sample of 169 consultations exceeded this threshold.

Patient Surveys

Patient surveys were conducted to evaluate patient perspectives on ambient scribe use. Four of the 9 participating clinicians were selected through convenience sampling for patient surveys based on surveyor availability and clinician consent. Surveys were conducted across selected clinic sessions, resulting in partial overlap with time-motion observations, although the 2 data collection activities were conducted independently. Surveys were administered verbally by independent surveyors from the Office of Patient Experience immediately following consultations in which the ambient scribe tool was used. The survey assessed 4 domains: perceived quality of the clinical encounter, clinician's attentiveness and engagement, comfort level with AI-assisted documentation, and likelihood of recommending the technology for other clinical encounters. Patients were excluded from surveys if they declined participation or were deemed clinically unsuitable by their physician, for example, due to having just received a poor prognosis.

Statistical Analysis

We used descriptive statistics to summarize the characteristics of participating clinicians and consultation sessions. We applied the Pearson chi-square test to assess whether the distribution of new vs follow-up patients differed between consultations conducted with and without the ambient scribe tool. Linear mixed-effects models were used to account for clustering of consultations within individual clinicians and unequal numbers of observations per clinician. These models included random intercepts for individual clinicians to control for baseline differences in practice patterns and fixed effects for ambient scribe use and patient type (new vs follow-up consultation). Statistical significance was set at P<.05, and all analyses were conducted using SPSS Statistics (version 26.0; IBM Corp).

This manuscript was prepared using the SQUIRE (Standards for Quality Improvement Reporting Excellence) reporting guidelines.

Ethical Considerations

This study evaluated process metrics of an existing ambient scribe tool without collecting identifiable patient data or implementing clinical interventions. In accordance with institutional governance policies, the study was determined to constitute a quality improvement initiative and, therefore, did not require formal institutional review board approval (SingHealth Policy for Governance on Data Use [Non-Research]; document no. SHS-DDS-DAG-1002 [internal policy document]). For routine clinical use of the ambient scribe, explicit patient consent was not required based on institutional guidance as the system functions solely as a documentation support tool and does not store audio recordings. No audio recordings, transcripts, or personally identifiable information

XSL•FO
RenderX

were retained for evaluation purposes, and analyses were limited to deidentified workflow and time-motion metrics. Verbal consent was obtained from patients participating in anonymous satisfaction surveys. Participation was voluntary, and no financial or material compensation was provided to clinicians or patients.

Results

Sample Characteristics

Nine clinicians participated in this study, representing 7 different specialties (Table 1). They had a mean of 20.8 (SD 7.8) years in practice. Each clinician completed matched observation sessions with and without ambient scribe use. Of the 9 clinicians, 4 completed their first observed session using the ambient scribe, and 5 completed their first session without it, resulting in an approximately even distribution of session order. In this study, all clinicians used their own preconfigured prompts and did not make any further modifications to the prompts.

Table . Characteristics of participating clinicians in the observational study of ambient scribe use (n=9).

Characteristics	Values
Duration in practice (y), mean (SD)	20.8 (7.8)
Observed consultations per clinician, mean (SD)	18.8 (6.5)
Clinicians per specialty, n	
Respiratory medicine	2
Endocrinology	2
Gastroenterology	1
Neurology	1
Rehabilitation medicine	1
Neonatology	1
General surgery	1

As shown in Table 2, a total of 169 consultations were observed across the study period: 78 (46.2%) conducted with the ambient scribe and 91 (53.8%) conducted without it. The unequal distribution reflected natural variation in scheduled appointments and patient attendance across the matched clinic sessions. Of these 169 consultations, 54 (32.0%) involved new patients, and 115 (68.0%) were follow-up appointments. Of the 78 consultations with ambient scribe use, 63 (80.8%) were in English, 14 (17.9%) were in Mandarin, and 1 (1.3%) was in Malay.

Table . Distribution of patient types (new vs follow-up) among consultations conducted with and without the artificial intelligence (AI) scribe. The proportions of patient types were similar across groups, with no statistically significant difference in distribution.

Patient type	Overall (n=169), n (%)	Without AI scribe use (n=91), n (%)	With AI scribe use (n=78), n (%)	*P* value[a]
New patients	54 (32.0)	28 (30.8)	26 (33.3)	.72
Follow-up patients	115 (68.0)	63 (69.2)	52 (66.7)	.72

[a]Chi-square test comparing distribution between groups.

Given the potential for variation in individual clinician response, we examined unadjusted clinician-level changes comparing sessions with and without ambient scribe use (Figure 3). Substantial heterogeneity was observed: 7 of 9 clinicians demonstrated reductions in documentation time (7.9% - 34.6%), 6 of 9 had shorter consultation durations (4.3% - 42.2%), and 8 of 9 showed increased eye contact (4.2% - 35.8%).

XSL•FO
RenderX

Figure 3. Unadjusted clinician-level outcome changes with artificial intelligence (AI) scribe use. Each dot represents the percentage change for an individual clinician comparing sessions with and without AI scribe use. The dashed line indicates no change (0%).

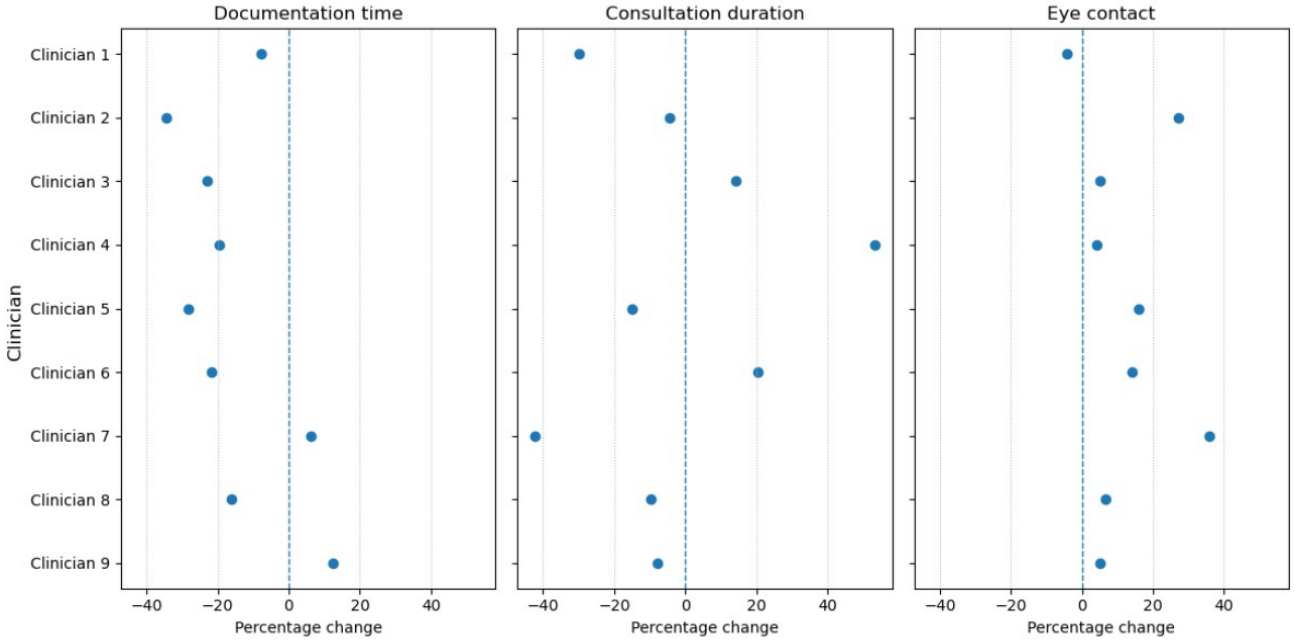


Table 3 summarizes the outcomes associated with the use of ambient scribes. Use of the tool was associated with 15.0% less time spent on documentation (from mean 5.3, SD 2.7 minutes to mean 4.5, SD 2.4 minutes per consultation; *P*=.04). The proportion of eye contact was 10.6% higher with ambient scribe use compared with standard documentation (from mean 69.6%, SD 18.6% to mean 77.1%, SD 17.7% of consultation duration; *P*=.009). Consultation duration was similar between conditions (mean 11.5, SD 6.9 minutes vs mean 10.9, SD 5.6 minutes; *P*=.42), and total cycle time per patient also did not differ significantly (mean 13.7, SD 7.7 minutes vs mean 13.2, SD 6.3 minutes; *P*=.57). The ambient scribe effects were consistent across new and follow-up patients for both documentation time and eye contact (interaction *P* values of .79 and .35, respectively), indicating uniform benefits across different consultation types.

Table . Comparison of outcomes between consultations conducted with and without artificial intelligence (AI) scribe use. Adjusted differences were calculated using linear mixed-effects models accounting for clinician-level clustering.

Outcome	Without AI scribe use (unadjusted), mean (SD)	With AI scribe use (unadjusted), mean (SD)	Adjusted difference (95% CI)	Change (%)	*P* value
Documentation time per consultation (min)	5.3 (2.7)	4.5 (2.4)	−0.8 (−1.5 to –0.05)	−15.0	.04
Proportion of eye contact time (%)	69.6 (18.6)	77.1 (17.7)	+7.4 (1.9 to 13.0)	10.6	.009
Consultation duration (min)	11.5 (6.9)	10.9 (5.6)	−0.7 (−2.6 to 1.1)	−6.1	.42
Total cycle time (min)	13.7 (7.7)	13.2 (6.3)	−0.6 (−2.6 to 1.4)	−4.4	.57

Crude independent-sample *t* tests treating all consultations as independent observations yielded similar effect sizes for eye contact (+7.5%; *P*=.008) and consultation duration (–0.67 minutes; *P*=.50). However, the reduction in documentation time reached statistical significance only in the adjusted mixed-effects analysis (crude: −0.74 minutes and *P*=.07; adjusted: –0.8 minutes and *P*=.04). This highlights the importance of accounting for clinician-level clustering.

Secondary Outcomes: Patient Experience

Patient surveys were conducted across 4 clinic sessions involving 4 of the 9 participating clinicians. Of 45 eligible patients in these sessions, 39 (86.7%) completed the surveys; 6 (13.3%) declined participation or were deemed unsuitable by their physician. Of the 39 surveyed patients, 18 (46.2%) were also included in the time-motion observations.

Age demographics are shown in Table 4, with 41.0% (16/39) of the patients aged 65 years and above. Survey language preferences were English (31/39, 79.5%), Mandarin and Chinese dialects (7/39, 17.9%), and Malay (1/39, 2.6%). Survey responses were collapsed into 3 analytic categories (“yes,” “neutral,” and “no”) based on evaluative meaning. Positive responses were coded as “yes,” neutral or indeterminate responses were coded as “neutral,” and negative responses were coded as “no.” The complete survey instrument is detailed in Multimedia Appendix 1.

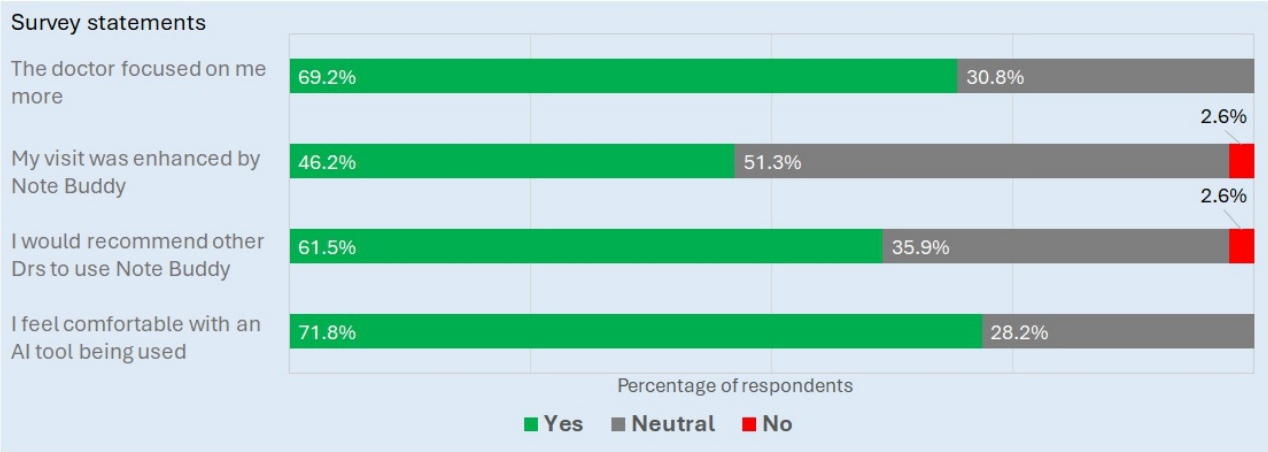
XSL•FO
RenderX

Table . Age distribution of patients who completed surveys about ambient scribe use during consultations.

Age group (y)	Patients (n=39), n (%)
0 - 19	1 (2.6)
20 - 29	3 (7.7)
30 - 39	7 (17.9)
40 - 49	5 (12.8)
50 - 64	7 (17.9)
≥65	16 (41.0)

Overall, patient acceptance of the AI tool was favorable (Figure 4), with 69.2% (27/39) agreeing that their physician focused on them more during the consultation and 61.5% (24/39) expressing willingness to recommend the technology for use by other physicians. While some patients (20/39, 51.3%) gave neutral responses and 2.6% (1/39) reported that the ambient scribe did not enhance their visit, none expressed discomfort with its use.

Figure 4. Patient perceptions of ambient artificial intelligence (AI) scribe use (Note Buddy). Postconsultation patient surveys assessed perceived clinician attentiveness, consultation experience, comfort with AI-assisted documentation, and willingness to recommend the technology. Responses were collapsed into "yes," "neutral," and "no" categories for presentation.



Analysis of the overlapping subset of patients who were both surveyed and included in time-motion observations (18/39, 46.2%) showed internal consistency between objective measurements and patient perceptions: 61.1% (11/18) of these patients reported that their physician focused on them more during the consultation, whereas measured eye contact was 15.2% higher compared to the same clinicians' sessions without the ambient scribe (mean 72.7%, SD 15.9% vs mean 63.1%, SD 16.9%; *P*=.07). This directional alignment between objective and subjective measures supports the validity of both metrics.

To assess potential selection bias in the patient survey data, we examined whether the effect of the ambient scribe differed between clinicians whose patients were surveyed and those whose patients were not. After adjusting for patient type (new vs follow-up), there was no evidence of a differential effect for either documentation time (interaction *P*=.71) or proportion of eye contact (interaction *P*=.62). These findings suggest that the patient survey results are representative of the broader clinician cohort.

Discussion

Principal Findings

To our knowledge, this is the first study of the real-world impact of an ambient scribe conducted in Asia. Use of an in-house ambient scribe tool was associated with reduced documentation time and increased eye contact, whereas consultation duration remained unchanged. This suggests that clinicians redirected time from clerical tasks toward patient engagement rather than shortening consultations. The overlap between surveyed patients and time-motion observations (18/39, 46.2% of patients) provides supportive internal consistency, with patients who experienced objectively measured increases in eye contact also reporting improved clinician attentiveness.

Allowing clinicians to choose in which clinic session to use the ambient scribe instead of randomizing may have introduced selection bias. However, patient mix and volume were comparable between the scribe and nonscribe sessions (ratio of new to follow-up patients: *P*=.72; patient volume per session: mean 9.0, SD 3.4 with scribe use vs mean 9.8, SD 4.1 without scribe use; *P*=.44), suggesting that any such bias was minimal.

Our findings align with those of recent US studies showing that time savings tend to be modest. Documentation time was estimated to have been reduced by 0.8 minutes (95% CI −1.5 to −0.05; *P*=.04) in the intervention group. Although the upper confidence limit approached zero, indicating potential sensitivity to individual outliers, our observed reduction of 0.8 minutes per consultation is consistent with that observed in prior observational studies. These include findings from Sutter Health

XSL•FO
RenderX

(drop from 6.2 to 5.3 minutes) [6], The Permanente Medical Group (0.4-minute reduction) [15], and Stanford Health Care (0.57-minute reduction) [10], as well as recent randomized controlled trials reporting similar results [4,16]. Additionally, our study uniquely demonstrates improved patient engagement through increased eye contact, which we were able to quantify because we used direct observation rather than EHR-based metrics.

Unsurprisingly, this real-world evaluation demonstrated heterogeneity in individual-level effects between clinicians. However, our mixed-effects model was specifically designed to account for this individual variation while estimating the overall treatment effect, which remained statistically significant despite the heterogeneity. This within-clinician comparison design controlled for individual clinician variability and practice patterns that could confound results. Although our linear mixed-effects model relied on only 9 clinician clusters, which can affect the reliability of variance estimates and increase the risk of type 1 errors in mixed-model analyses, this design allowed us to isolate the specific effects of ambient scribe implementation while accounting for the natural variation in consultation styles and documentation practices across specialists.

Our findings also demonstrate that benefits observed in Western health care systems can be realized in a different cultural and linguistic context. Of the 78 consultations that used the ambient scribe, 63 (80.8%) were conducted in English, 14 (17.9%) were conducted in Mandarin, and 1 (1.3%) was conducted in Malay, highlighting the tool's utility in multilingual environments. The favorable patient responses observed, particularly among older adults, who comprised 41% (16/39) of survey respondents, provide reassuring evidence of acceptance across age groups.

The reproducibility of modest time savings across studies—including our observed 0.8-minute reduction—raises important questions about clinical meaningfulness at the individual clinician and health system levels. Notably, our finding that consultation duration remained unchanged despite reduced documentation time indicates that the time savings did not result in shorter patient encounters. Furthermore, our finding that total cycle time also remained unchanged demonstrates that ambient scribes should not be expected to increase clinic throughput or patient volume at the system level. Recent randomized controlled studies suggest that the primary value of this technology lies in reducing clinician burnout [4,16]. Our results complement this by demonstrating measurable improvements in patient engagement through increased eye contact and positive patient survey responses. Taken together, these findings suggest that ambient scribes should be viewed primarily as quality-of-care interventions that shift clinician attention from clerical tasks to patient interaction rather than as productivity tools [15,17].

Limitations

This study has several limitations. First, the sample was limited to 9 clinicians in specialist outpatient clinics due to the resource-intensive nature of direct observation, which restricts generalizability to other specialties and primary care settings. We focused on experienced ambient scribe users to reflect real-world use patterns as prior research shows that regular users account for 89% of system activations [18]. This means that our findings reflect optimal rather than average implementation outcomes and cannot predict outcomes during initial adoption or among infrequent users. Because our clinicians were both experienced scribe users and relatively senior (mean 20.8, SD 7.8 years of practice), we cannot determine whether the observed benefits reflect scribe proficiency or preexisting EHR documentation efficiency. If our senior clinicians had already optimized their workflows through customized EHR templates and shortcuts, the observed time savings may underestimate the potential benefits for less efficient documenters. Conversely, new users may experience a learning curve that could temporarily reduce initial gains.

Second, the ambient scribe tool required manual transfer of notes into the EHR, which likely underestimates potential time savings compared to fully integrated systems and limits direct comparability with studies of integrated solutions.

Third, several potential biases may have influenced our findings. Clinicians self-selected the sessions in which they used the ambient scribe, introducing the possibility of selection bias. Observer expectancy and Hawthorne effects [19] may have influenced clinician behavior or the measurement of eye contact as observers and clinicians could not be blinded. Interrater reliability could not be assessed because multiple observers in the consultation room would have been intrusive; however, each clinician was consistently observed by the same individual across both conditions to minimize observer-related variability. Patient survey responses may have also been subject to social desirability bias as patients may have been reluctant to express negative views about the technology when surveyed immediately following their consultation within the clinical facility despite the use of independent surveyors. As the ambient scribe tool was developed in-house, some degree of institutional bias is possible, although objective measurement methods were used to minimize this risk.

Finally, this study did not specifically assess whether note accuracy was preserved in the context of the observed reductions in documentation time; nonetheless, the tool's clinical fidelity had been reported in a previous study [14]. The predominance of English-language consultations (63/78, 80.8%) limits conclusions about multilingual performance. Clinicians used different prompts, including customized ones, and we did not systematically analyze variation in prompt length or complexity. Such prompt variability may have influenced the time taken for the ambient scribe to generate a note and represents a potential confounder.

Future Directions

The observed heterogeneity in individual-level effect sizes in this real-world study highlights that the potential benefits of ambient scribe adoption may not be uniform even in experienced users. Future studies should evaluate the impact of ambient scribe technology across a broader range of clinicians and practice environments to identify which user profiles benefit the most and which may require additional support. Qualitative research into factors affecting adoption and sustained use will be important for ensuring long-term effectiveness and

XSL•FO
RenderX

scalability. Future work should systematically evaluate how prompt design—including specialty-specific or individualized prompts—affects the quality of AI-generated documentation and the time required for clinician review and editing. While our study suggested effectiveness in a multilingual patient population, most consultations (63/78, 80.8%) were conducted in English. Further work is needed to assess whether the same benefits are observed in non–English-language consultations as language may influence both scribe performance and outcomes.

Conclusions

This quality improvement study found that ambient scribe use among experienced users was associated with modest reductions in documentation time and improved patient engagement, with high patient acceptance. These findings quantify the impact of ambient scribes in a real-world outpatient setting using direct observational methods. Our findings demonstrate that these benefits can be achieved through in-house–developed solutions, offering health care systems an alternative to commercial products in cases in which data privacy, cost, or localization are key concerns.

Acknowledgments

The authors gratefully acknowledge the contributions of Singapore General Hospital's Office of Patient Experience for survey administration, the SingHealth Office of Insights and Analytics for developing the Note Buddy software, the Singapore General Hospital Specialist Outpatient Clinics for equipment setup, SingHealth's National Neuroscience Institute for time-motion study support, and Joshua Ohde from the Mayo Clinic for manuscript guidance and proofreading. The authors used Anthropic's Claude Sonnet 4 for proofreading and language editing during manuscript preparation. All content was subsequently reviewed and edited by the authors, who take full responsibility for the publication.

Funding

No funding was received for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JYET, IBMR, KWC
Data curation: JYET, IBMR
Formal analysis: GHL
Investigation: JYET, IBMR, EYLL
Methodology: JYET, IBMR, GHL
Supervision: KWC, MKHH
Writing—original draft: JYET, IBMR
Writing—review and editing: JYET, IBMR, GGRS, JYMT, DYZL, LGH, JCLO, EYLL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials, including technical details on the ambient artificial intelligence scribe (Note Buddy), the time-motion study instrument, and the patient survey instrument.
[DOCX File, 914 KB - medinform_v14i1e85580_app1.docx]

References

1. Adoption of artificial intelligence in healthcare delivery systems: early applications and impacts. Peterson Health Technology Institute. 2025 Mar 25. URL: https://phti.org/ai-adoption-early-applications-impacts/ [accessed 2025-09-22]
2. Prasad K, Frits M, Iannaccone C, et al. Clinician perceptions of virtual scribe use: a survey study. J Gen Intern Med 2025 Aug 5. [doi: 10.1007/s11606-025-09771-5] [Medline: 40762875]
3. Evans K, Papinniemi A, Ploderer B, et al. Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: perspectives of clinicians and patients. Musculoskelet Sci Pract 2025 Aug;78:103333. [doi: 10.1016/j.msksp.2025.103333] [Medline: 40328004]
4. Lukac PJ, Turner W, Vangala S, et al. A randomized-clinical trial of two ambient artificial intelligence scribes: measuring documentation efficiency and physician burnout. medRxiv. Preprint posted online on Jul 11, 2025. [doi: 10.1101/2025.07.10.25331333] [Medline: 40672471]

XSL•FO
RenderX

5. Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 3;8(2):e2460637. [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]
6. Stults CD, Deng S, Martinez MC, et al. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Netw Open 2025 May 1;8(5):e258614. [doi: 10.1001/jamanetworkopen.2025.8614] [Medline: 40314951]
7. Shah SJ, Crowell T, Jeong Y, et al. Physician perspectives on ambient AI scribes. JAMA Netw Open 2025 Mar 3;8(3):e251904. [doi: 10.1001/jamanetworkopen.2025.1904] [Medline: 40126477]
8. Shin HS, Williams H, Braykov N, Jahan A, Meller J, Orenstein EW. The influence of artificial intelligence scribes on clinician experience and efficiency among pediatric subspecialists: a rapid, randomized quality improvement trial. Appl Clin Inform 2025 Aug;16(4):1041-1052. [doi: 10.1055/a-2657-8087] [Medline: 40675605]
9. Liu TL, Hetherington TC, Stephens C, et al. AI-powered clinical documentation and clinicians' electronic health record experience: a nonrandomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2432460. [doi: 10.1001/jamanetworkopen.2024.32460] [Medline: 39240568]
10. Ma SP, Liang AS, Shah SJ, et al. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc 2025 Feb 1;32(2):381-385. [doi: 10.1093/jamia/ocae304] [Medline: 39688515]
11. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
12. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Dec 6;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
13. Who we are. Singapore General Hospital. URL: https://www.sgh.com.sg/about-sgh/who-we-are [accessed 2025-09-22]
14. Ong JHW, Tung JYM, Sng GGR, et al. A pilot study using ambient artificial intelligence scribes in clinical documentation in a urology outpatient clinic. BJU Int 2025 May 20;20. [doi: 10.1111/bju.16784] [Medline: 40390678]
15. Haberle T, Cleveland C, Snow GL, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc 2024 Apr 3;31(4):975-979. [doi: 10.1093/jamia/ocae022] [Medline: 38345343]
16. Afshar M, Baumann MR, Resnik F, et al. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI 2025 Dec;2(12):10. [doi: 10.1056/aioa2500945] [Medline: 41625485]
17. Kim E, Liu VX, Singh K. AI scribes are not productivity tools (yet). NEJM AI 2025 Nov 26;2(12). [doi: 10.1056/AIe2501051]
18. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catal Innov Care Deliv 2025 Mar 31;6(5). [doi: 10.1056/CAT.25.0040]
19. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ 2015 Sep 4;351:h4672. [doi: 10.1136/bmj.h4672] [Medline: 26341898]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
SQUIRE: Standards for Quality Improvement Reporting Excellence

Edited by A Coristine; submitted 09.Oct.2025; peer-reviewed by A Famotire, R Mamidipally; revised version received 24.Feb.2026; accepted 24.Feb.2026; published 31.Mar.2026.

Please cite as:
Tan JYE, Rafi IBM, Sng GGR, Tung JYM, Lim DYZ, Ong JCL, Lim GH, Lew EYL, Chia KW, Heng MKH
Impact of an Ambient AI Scribe Among Clinicians and Patients: Real-World Prospective Observational Time-Motion Study
JMIR Med Inform 2026;14:e85580
URL: https://medinform.jmir.org/2026/1/e85580
doi:10.2196/85580

© Jonathan Yue En Tan, Iffat Bin Mohammad Rafi, Gerald Gui Ren Sng, Joshua Yi Min Tung, Daniel Yan Zheng Lim, Jasmine Chiat Ling Ong, Gek Hsiang Lim, Eileen Yi Lee Lew, Kuok Wei Chia, Mark Kok Hong Heng. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Bilingual Arabic-English Ambient AI Scribe for Clinical Documentation: Prospective Evaluation Study

Umair Tahir Khan[1], MBChB, PhD; Ammar Tahir Khan[1], BSc; Waleed Aljaadi[1], MBBS; Razan Alhadlaq[2], SBFM, ABFM, MBBS; Zahran Baqashmer[1], MBBS; Yasin Alsafi[2], BA, MSc; Yousef Alomran[2], SSC-FM, ABFM, MBBS; Maha Al Rusaiyes[1,2], MSc; Muaddiyah Radif[3], BSN, MBA; Tahir Naeem Khan[1,4], MBBS, MPhil, PhD; Saleh Abdullah Saleh Altamimi[3], MHA, MD

[1]Sahl AI, Anas Ibn Malik, Riyadh, Saudi Arabia

[2]King Saud Medical City, Riyadh, Saudi Arabia

[3]Riyadh First Health Cluster, Riyadh, Saudi Arabia

[4]IQVIA, Livingston, United Kingdom

Corresponding Author:
Umair Tahir Khan, MBChB, PhD
Sahl AI, Anas Ibn Malik, Riyadh, Saudi Arabia

Abstract

Background: Medical ambient artificial intelligence (AI) scribes reduce documentation burden, but the current evidence is almost entirely from English systems. In the Arabic-speaking world, physicians converse mainly in Arabic and write clinical notes in English, adding cognitive burden. Due to scarce corpora in the Arabic language, the development of Arabic-enabled AI speech technologies has been challenging. Here, we address this gap by developing and evaluating a bilingual Arabic-English medical AI scribe.

Objective: This study aims to evaluate the feasibility and performance of a bilingual medical Arabic-English ambient AI scribe, Sahl AI, using a full end-to-end methodology from raw audio to clinical note.

Methods: A prospective, single-arm feasibility pilot study was conducted in 2 stages across outpatient clinics, inpatient services, and primary care clinics within Riyadh First Health Cluster: development and implementation. In stage 1 (development), consultation audios were collected and manually annotated to fine-tune the AI pipeline. Version 1 of Sahl AI underwent feasibility evaluation using a modified 9-item Physician Documentation Quality Instrument (PDQI-9) framework based on Likert-scale ratings. Subsequently, in stage 2 (implementation), the AI pipeline was fine-tuned, and version 2 was evaluated as a real-world deployment in family medicine clinics. Independent reviewers evaluated clinical notes. Physician experience was captured using targeted surveys.

Results: During stage 1, the AI pipeline was fine-tuned, producing version 1 of Sahl AI, which was tested in 64 clinical encounters for technical feasibility. This evaluation yielded a mean modified PQDI-9 score of 42.2 (SD 2.98) out of 45, with the model scoring 4.35 (SD 0.82) out of 5 in the accuracy domain. Following further optimization, version 2 of Sahl AI was evaluated during stage 2, using 55 real-world consultations assessed by 2 independent physician evaluators. Sahl AI achieved mean modified PDQI-9 scores of 42.4 (SD 1.84) out of 45 for Arabic and 37.8 (SD 1.10) out of 40 for English. On the Likert scale, for the accuracy domain, the mean score was 4.53 (SD 0.46) out of 5 for Arabic compared with 4.77 (SD 0.37) out of 5 for English (P=.06). Internal consistency (mean 4.94, SD 0.17) and comprehensibility (mean 4.89, SD 0.21) were the top-rated domains. The targeted survey of a larger cohort of 22 family medicine physicians using Sahl AI showed that most responding physicians agreed that notes were comprehensive and perceived potential time savings and reduced burnout.

Conclusions: Sahl AI, a bilingual Arabic-English medical ambient AI scribe, generates accurate and high-quality notes, and in targeted postimplementation surveys, most physicians agreed that it produced comprehensive notes and perceived potential benefits for time savings and stress or burnout. This provides the first empirical basis for rigorous end-to-end ambient AI scribe evaluation in low-resource languages such as Arabic.

(JMIR Med Inform 2026;14:e83335) doi:10.2196/83335

KEYWORDS

artificial intelligence; AI; natural language processing; speech recognition software; clinical documentation; physician burnout

XSL•FO
RenderX

Introduction

Physicians devote a significant portion of their time to the administrative task of documenting medical encounters with patients [1]. Clinical documentation serves critical functions, including care continuity, quality assurance, and in many jurisdictions, medico-legal protection. It is an essential component for ensuring patient safety and also impacts subsequent tasks, such as submitting insurance claims [2]. The digitization of medical records and the widespread adoption of electronic health record systems have substantially increased the documentation burden on physicians, leading to increased physician burnout [3]. Furthermore, clinical documentation has become a distraction, changing the traditional doctor-patient relationship into a doctor-computer-patient interaction [4].

Clinical documentation is typically completed either during or after patient encounters. Documenting in real time may interrupt the clinical interaction, whereas postvisit documentation may add to physician workload and relies on recall rather than contemporaneous capture. Ambient artificial intelligence (AI) scribes automate clinical note generation and improve the doctor-patient relationship, freeing physicians to focus on personalized care [5,6]. This study evaluates Sahl AI, an AI tool for Arabic and English, using the modified 9-item Physician Documentation Quality Instrument (PDQI-9) tool. This research is unique in its dual-language focus and is particularly relevant in Arabic-speaking countries where clinical encounters are conducted in Arabic, while medical documentation is recorded in English. This supports health care providers who routinely switch between languages during consultations when eliciting patient histories and documenting care.

Arabic is often categorized as a low-resource language in speech technologies due to its complex characteristics. These include its writing system with positional variance and diacritical marks, phonology that features emphatic consonants and vowel length distinctions, and grammar that involves root-based morphology and case-gender agreement. The variability is further compounded by numerous dialects and diglossia, as well as influence from other languages and its significant religious context. This necessitates ongoing translation by the physicians, introducing cognitive strain. Although ambient AI scribes offer a potential solution to mitigate this burden, the effectiveness of bilingual systems in such linguistically complex environments remains insufficiently examined. Therefore, the aim of this study was to develop and assess the feasibility and physician experience of a bilingual Arabic-English medical ambient AI scribe in routine clinical practice.

Methods

Study Design and Objectives

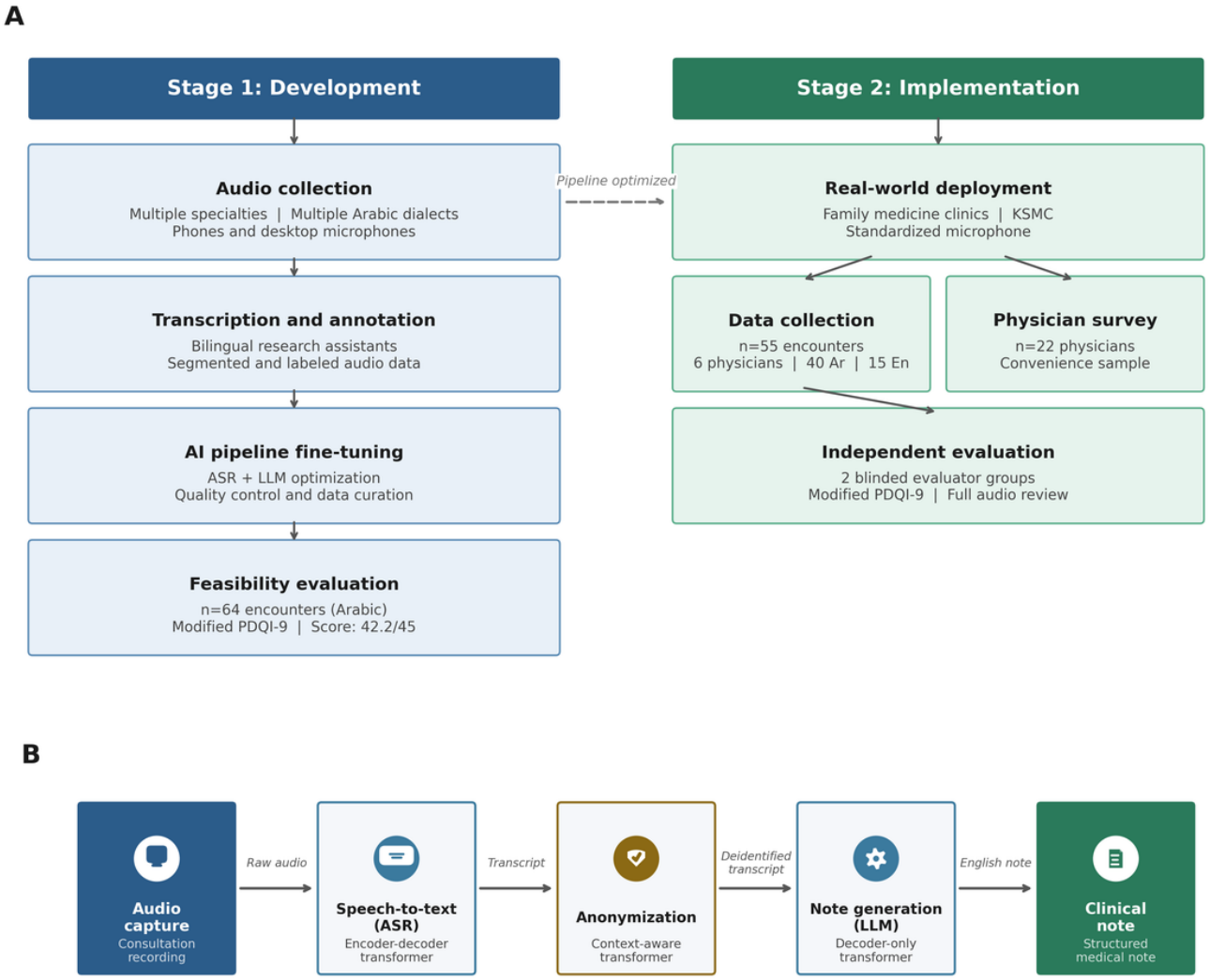
The study was a prospective, single-arm feasibility pilot study evaluating the performance of Sahl AI, a bilingual ambient AI scribe, in routine physician-patient consultations. The primary objectives were to (1) fine-tune an AI pipeline for automated clinical documentation generation and subsequently (2) evaluate the usability of Sahl AI and its impact on the physicians' workflow and the quality of documentation.

Study Setting and Duration

The study was conducted between December 23, 2023, and November 28, 2024, across outpatient clinics, inpatient services, and primary care centers within Riyadh First Health Cluster (Riyadh, Saudi Arabia). The study comprised 2 stages: (1) a development stage, during which data collection and a feasibility evaluation were performed, and (2) an implementation stage, in which the fine-tuned AI pipeline was assessed in a real-world clinical setting (Figure 1A).

XSL•FO
RenderX

Figure 1. (A) Two-stage study design. Stage 1: iterative development including multidialect Arabic audio collection, manual transcription by bilingual research assistants, ASR and LLM fine-tuning, and feasibility evaluation (n=64 encounters; mean 9-item Physician Documentation Quality Instrument [PDQI-9] score: 42.2/45, SD 2.98). Stage 2: real-world implementation at family medicine clinics (King Saud Medical City [KSMC]), with prospective data collection (n=55 encounters; 40 Arabic, 15 English), independent blinded evaluation using the modified PDQI-9, and a parallel physician experience survey (n=22). Dashed arrow indicates pipeline transfer from development to implementation. (B) Artificial intelligence (AI) pipeline architecture: audio capture, encoder-decoder ASR, anonymization via context-aware transformer for personally identifiable information (PII) deidentification, and decoder-only LLM for clinical note generation. ASR: automatic speech recognition; LLM: large language model.



During stage 1, participating specialties included, but were not limited to, general and internal medicine, general surgery, pediatrics, family medicine, obstetrics and gynecology, cardiology, and respiratory medicine. Both new and follow-up patient encounters were included, provided eligibility criteria were met and physicians were able to conveniently record audio conversations during clinical encounters. During stage 2, new patient encounters were sampled during designated family medicine clinic sessions when participating physicians were available. Verbal consent was obtained from patients by the treating physicians and was documented in the clinical notes.

Eligibility Criteria

The inclusion criteria were as follows: patients of all ages and sexes who provided informed consent for the use of Sahl AI during their clinical consultation. For pediatric patients, consent was obtained from a parent or legal guardian.

The exclusion criteria were as follows: patients with cognitive impairment who were unable to provide informed consent.

Collection of Audio Data for Fine-Tuning

During stage 1, audio data were collected to optimize and fine-tune the AI pipeline using Sahl AI's proprietary Data Collection Pipeline. Audio streams from physician-patient conversations were segmented into shorter audio chunks to facilitate accurate labeling. Participating physicians received training on audio recording using various devices and on obtaining informed consent from patients. The recorded audio was securely stored on cloud servers in compliance with the Kingdom of Saudi Arabia's Personal Data Protection Law (PDPL). The audio data were initially retained for 6 months, with the retention period later extended to 1 year following approval of an ethics amendment.

Recordings were collected across multiple clinical specialties during stage 1 using physicians' personal mobile phones as well as desktop microphones, including Anker PowerConf S360 USB Speakerphones and Kaysuda SP300 Bluetooth Speakerphones. A team of bilingual research assistants fluent in Arabic and English reviewed and edited the transcribed audio

XSL•FO
RenderX

segments to ensure high-quality labeled datasets for AI pipeline optimization. The dataset encompassed a wide range of Arabic dialects, including those spoken in Saudi Arabia, Yemen, Egypt, Sudan, the Levant, Kuwait, and North Africa, as well as simplified Arabic spoken by nonnative speakers and diverse English accents. Data scientists subsequently curated the labeled data, excluding poor-quality audio segments identified during review by the research assistants.

The AI pipeline was iteratively tested with internal evaluation metrics before being piloted with physicians during stages 1 and 2 of the study. The pipeline (Figure 1B) comprised multiple components requiring optimization for clinical deployment. In particular, the speech-to-text engine was based on a transformer-based encoder-decoder automatic speech recognition architecture, in which the encoder transforms raw audio signals into high-level semantic representations and the decoder generates the corresponding text transcription. Prior to clinical note generation, transcripts were processed through a deidentification layer that used a context-aware transformer model to detect and redact personally identifiable information. The large language model component was a decoder-only autoregressive transformer trained using causal language modeling.

Fine-tuning focused on improving robustness to spontaneous clinical dialogue, variability in dialects, and background noise commonly encountered in outpatient and inpatient environments. Model training prioritized transcription accuracy for medical conversations. To minimize overfitting and prevent information leakage, training and evaluation datasets were strictly separated.

Feasibility Study

During stage 1, following the fine-tuning and optimization of the AI pipeline, version 1 of Sahl AI was evaluated in a controlled setting to assess feasibility outside of routine clinical practice. This evaluation involved role-plays and mock conversations between physicians, as well as prerecorded simulated physician-patient conversations generated by nonphysicians. These audio recordings were processed by the AI pipeline to generate clinical notes, enabling the assessment of the ambient AI scribe's performance in the Arabic language. Collectively, these activities constituted the feasibility component of the development stage.

Clinical notes were generated in accordance with the documentation requirements of the Central Board for Accreditation of Healthcare Institutions and included standardized headings, such as chief complaint, history of present illness, past medical or surgical history, allergy, medication, social history, family history, physical examination findings, diagnosis, and plan. Participating physicians evaluated the AI-generated clinical notes after completing the role-play scenarios or listening to the prerecorded simulated conversations.

In total, 64 independent assessments were conducted using a modified PDQI-9 (Table 1), a tool that has been previously validated and used in recent evaluations of ambient AI scribes [7,8]. For the purposes of this study, the PDQI-9 tool was adapted to reflect Arabic-language and bilingual audio-to-text workflows. Specifically, the original domains of "up-to-date" and "synthesized" were replaced with "free from hallucination" and "translated accurately to English," respectively. These modified domains have not been formally validated and were used for exploratory assessment only.

XSL•FO
RenderX

Table . Nine-item Physician Documentation Quality Instrument (PDQI-9) tool measuring the quality of the clinical note across 9 dimensions[a].

Domain	Description
Accurate	The note is true. It is free of incorrect information. For example, if the history is about chest pain and the note mentions abdominal pain, then this represents that the note was not accurate.
Thorough	The note is complete and free from omission and documents all of the issues of importance to the patient. For example, if a symptom about visual loss was discussed and was important to the note but not mentioned, then this represents that the note was not thorough.
Useful	The note is extremely relevant, providing valuable information and/or analysis.
Organized	The note is well formed and structured as per CBAHI[b] recommendations in a way that helps the reader understand the patient's clinical course.
Comprehensible	The note is clear, without ambiguity or sections that are difficult to understand.
Succinct	The note is brief, to the point, and without redundancy.
Internally consistent	No part of the note ignores or contradicts any other part. For example, if 1 section of the note mentions that the patient has asthma but another section says no history of asthma, then this is not internally consistent.
Free from hallucination	The note is free of hallucination and only contains information verifiable by the audio conversation. For example, if the note mentions that the patient smokes and drinks alcohol when it was not at all discussed in the note, then this represents hallucination.
Translated accurately to English	The note reflects the Arabic conversation between the physician and the patient accurately in the English language.

[a]Each domain was scored from 1 to 5 (Likert scale), with 1 being "not at all" and 5 being "extremely."

[b]CBAHI: Central Board for Accreditation of Healthcare Institutions.

Real-World Testing

Based on the findings from stage 1, the AI pipeline was further optimized to improve performance prior to real-world deployment. During stage 2, the model was deployed in family medicine clinics for real-world testing involving both new and follow-up patient encounters. From a larger cohort of physicians, 6 physicians who consented to participate received at least 30 minutes of training on the effective use of the software prior to study initiation.

Physician-patient conversations and the corresponding unedited clinical notes generated by Sahl AI were recorded and subsequently evaluated by 2 independent evaluators. Evaluators listened to the complete audio recordings and assessed the associated unedited clinical notes using the modified PDQI-9 tool. To reduce the likelihood of identifying the physician speaker, evaluators were recruited from clinical specialties different from those of the participating physicians; however, complete anonymization could not be guaranteed, as the evaluators were recruited from the same health cluster. Evaluators were blinded to the identities of both patients and physicians. However, as the raw audio was not modified to redact personally identifiable information, complete blinding cannot be assured. Scores from evaluators were averaged prior to calculating the final modified PDQI-9 domain scores.

The evaluation process involved 2 evaluator groups. Evaluator group 1 comprised a single reviewer, while evaluator group 2 consisted of 2 reviewers who divided the assigned clinical notes between them. Each clinical note was reviewed by 2 evaluators—1 from each group. In cases where domain-level scores differed by 2 or more points, both evaluators were asked to rereview the clinical notes and reconcile discrepancies.

To ensure consistency across recordings, a standardized microphone (Kaysuda SP300 Bluetooth Speakerphones) connected to a laptop computer was used for all consultations. Audio recordings were performed by trained research assistants and uploaded directly through the Sahl AI software. Following each consultation, Sahl AI generated a single clinical note immediately using the recorded audio, which was then reviewed in its unedited form by the independent evaluators.

Clinical encounters were categorized as "Arabic" or "English" based on the predominant language of the consultation, as initially determined by research assistants and subsequently confirmed by the evaluators. Code-switched encounters involving substantial use of both languages were classified according to the language comprising the majority (>50%) of clinical content. Regardless of input language, Sahl AI generated all clinical documentation exclusively in English.

Physician Experience

Following real-world deployment in the family medicine department, a targeted physician experience survey was conducted to capture feedback from clinicians with the direct hands-on use of Sahl AI. The survey was completed between 29 July and 28 November 2024 by 22 physicians. Eligibility was limited to physicians who had used the system in at least

XSL•FO
RenderX

3 clinical encounters, ensuring sufficient exposure to evaluate its impact on clinical workflow. Survey items evaluated perceived note comprehensiveness, time savings, ease of use, impact on physician burnout, and impact on clinical workflow using a 5-point Likert scale ranging from strongly agree to strongly disagree. The survey was designed as a descriptive assessment of user experience among active users rather than a population-level evaluation.

Statistical Analysis

During the development stage, individual modified PDQI-9 scores were recorded for each physician's assessment. In contrast, during the implementation stage, the modified PDQI-9 scores were calculated by calculating the mean scores from 2 independent evaluators for each domain and then summing these domain-level means. The overall modified PDQI-9 scores were reported as mean (SD) and were calculated out of a maximum total score of 45 for Arabic conversations and 40 for English conversations, as the "translated accurately to English" domain was not applicable to English consultations. Since the modified PDQI-9 domain scores were ordinal Likert-scale measures and group sizes were unbalanced (40 Arabic vs 15 English consultations), comparisons between Arabic and English groups were conducted using nonparametric Mann-Whitney *U* tests. Differences between evaluator scores were quantified using the mean absolute deviation (MAD), calculated as the average of the absolute differences between paired domain scores. MAD was computed both overall and at the domain level. Descriptive statistics, including means and frequencies, were used to summarize the responses from the physician experience questionnaire.

Ethical Considerations

This study was approved by the institutional review board of King Saud Medical City (protocol H1R1-18-Sep23-01). Verbal informed consent was obtained from all participating patients prior to audio recording and was documented in the medical record. The use of verbal consent was approved by the institutional review board given the minimal-risk nature of the study and the practical limitations of obtaining written consent during routine clinical encounters without disrupting workflow. For pediatric patients, consent was obtained from parents or guardians. All data collection, storage, and analysis procedures were conducted in accordance with confidentiality requirements and complied with the Saudi Data and Artificial Intelligence Authority PDPL [9]. No patient demographic information was collected, and independent evaluators assessed clinical notes without access to any identifying data. Patients did not receive financial compensation for participation, and physicians participated as part of their routine clinical duties.

Results

Feasibility Pilot

During the development stage of the study, feasibility testing was conducted on 64 Arabic-language conversations, with the results summarized in Table 2. The overall mean modified PDQI-9 score across all domains was 42.2 (SD 2.98) out of 45. Clinical notes achieved the highest ratings in the domains of "useful," "organized," and "comprehensible," with mean scores of 4.86 (SD 0.39), 4.87 (SD 0.34), and 4.85 (SD 0.35) out of 5, respectively. The accuracy domain of version 1 of the pipeline was rated 4.35 (SD 0.82) out of 5. Relatively lower scores were observed for the thorough (mean 4.42, SD 0.81) and free from hallucination (mean 4.45, SD 0.85) domains, highlighting areas for further improvement. The identified strengths and limitations informed the subsequent fine-tuning and optimization of the AI pipeline.

Table . Domain-level performance of Sahl AI-generated clinical notes during the development (n=64) and implementation (n=55) stages, assessed using the modified 9-item Physician Documentation Quality Instrument (PDQI-9) tool[a].

Domain	Development stage[b]		Implementation stage[c]	
	Number of cases	Mean (SD)	Number of cases	Mean (SD)
Accurate	64	4.35 (0.82)	55	4.59 (0.45)
Thorough	64	4.42 (0.81)	55	4.55 (0.45)
Useful	64	4.86 (0.39)	55	4.74 (0.33)
Organized	64	4.87 (0.34)	55	4.54 (0.37)
Comprehensible	64	4.85 (0.35)	55	4.89 (0.21)
Succinct	64	4.84 (0.45)	55	4.49 (0.38)
Internally consistent	64	4.73 (0.60)	55	4.94 (0.17)
Free from hallucination	64	4.45 (0.85)	55	4.87 (0.22)
Translated accurately to English	64	4.79 (0.49)	40	4.85 (0.28)

[a]Values represent mean (SD) Likert-scale scores for each documentation quality domain. The overall mean modified PDQI-9 scores are reported separately for Arabic (out of 45) and English-translated (out of 40) conversations.

[b]Average for development stage: 42.2 (SD 2.98) out of 45.

[c]Average for implementation stage: Arabic: 42.4 (SD 1.84) out of 45; English: 37.8 (SD 1.10) out of 40.

XSL•FO
RenderX

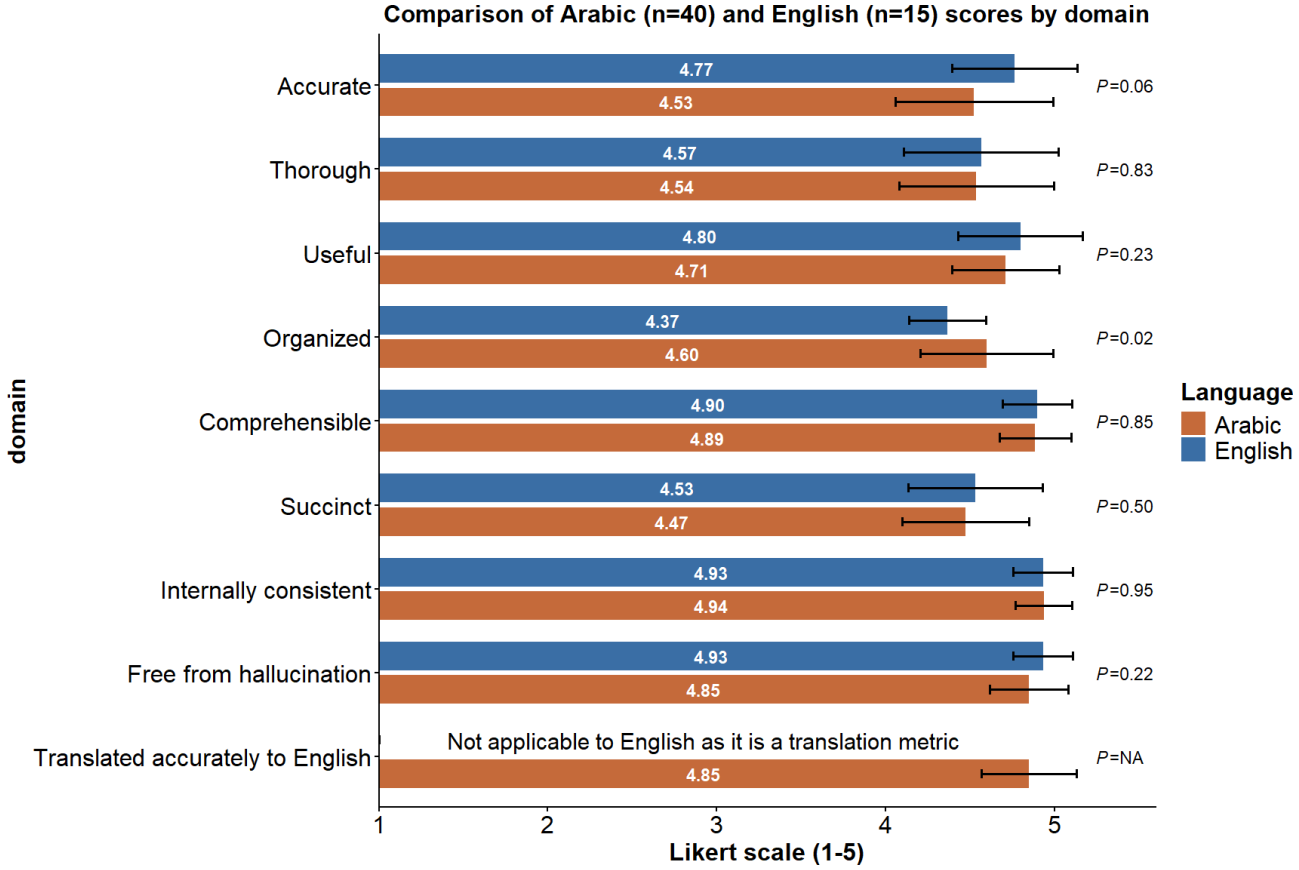
Real-World Implementation

During the implementation stage, real-world testing was conducted in family medicine clinics at King Saud Medical City (Riyadh First Health Cluster). Independent evaluators reviewed 55 new-patient consultations using the modified PDQI-9 tool; 40 consultations involved predominantly Arabic physician-patient conversations, and 15 involved English conversations. The mean modified PDQI-9 score was 42.4 (SD 1.84) out of 45 for Arabic consultations and 37.8 (SD 1.10) out of 40 for English consultations (Table 2).

Across all cases, clinical notes achieved the highest scores in the domains of "internally consistent," "comprehensible," and "free from hallucination" with mean scores of 4.94 (SD 0.17), 4.89 (SD 0.21), and 4.87 (SD 0.22) out of 5, respectively. The mean accuracy domain score for version 2 of the fine-tuned and optimized AI pipeline was 4.59 (SD 0.45) out of 5 across both language groups. Mean scores exceeded 4.5 out of 5 across all domains except for "succinct," which scored 4.49 (SD 0.38), suggesting a tendency toward more verbose clinical notes.

Domain-level scores for Arabic and English encounters were analyzed separately and are presented in Figure 2. A statistically significant difference was observed in the "organized" domain, with higher scores for Arabic compared with English conversations (4.60, SD 0.39 vs 4.37, SD 0.23; P=.02). There was a trend to the English conversations having higher accuracy compared with Arabic conversations (4.77, SD 0.37 vs 4.53, SD 0.46; P=.06). No statistically significant differences were observed across the remaining domains.

Figure 2. Comparison of mean Likert-scale scores for clinical notes generated by Sahl AI, as evaluated by independent reviewers, for Arabic and English conversations across each modified 9-item Physician Documentation Quality Instrument (PDQI-9) domain. Error bars indicate ±1 SD. Between-group differences were assessed using the Mann-Whitney *U* test. NA: not applicable.



Interevaluator Agreement

Interevaluator agreement across modified PDQI-9 domains is presented in Table 3. Overall agreement between evaluator group 1 and evaluator group 2 was high, with a mean absolute deviation of MAD=0.36, indicating close concordance in scoring. Across individual domains, MAD values ranged from 0.13 to 0.66 on a 5-point Likert scale, demonstrating that evaluator ratings typically differed by <1 point. Exact agreement of scores ranged from 34.5% to 87.3%, while agreement within ±1 point on the Likert scale was consistently high across domains (96.4% - 100%).

XSL•FO
RenderX

Table . Interevaluator agreement between evaluator group 1 and evaluator group 2 across modified 9-item Physician Documentation Quality Instrument (PDQI-9) domains[a].

Domain	Number of cases	MAD[b]	Exact agreement (%)	Agreement ±1 (%)	ICC[c]	Cohen kappa
Accurate	55	0.42	60	98.2	0.28	0.28
Thorough	55	0.40	60	100	0.35	0.35
Useful	55	0.35	65.6	100	0.13	0.13
Organized	55	0.60	43.6	96.4	0.12	0.12
Comprehensible	55	0.22	78.2	100	−0.07	−0.06
Succinct	55	0.66	34.5	100	−0.05	−0.05
Internally consistent	55	0.13	87.3	100	0.06	−0.05
Free from hallucination	55	0.26	74.5	100	0.00	0
Translated accurately to English	40	0.20	80	100	0.22	0.22

[a]Agreement is summarized using mean absolute difference, exact agreement (%), and agreement within +1 point on the Likert scale (%), intraclass correlation coefficient (2-way mixed effects, absolute agreement), and weighted Cohen kappa (squared weights).

[b]MAD: mean absolute deviation.

[c]ICC: intraclass correlation coefficient.

The highest interevaluator agreement was observed for the internally consistent (MAD=0.13; exact agreement 87.3%) and translated accurately to English domains (MAD=0.20; exact agreement 80%). Greater variability was observed in more subjective stylistic domains, including succinct (MAD=0.66; exact agreement 34.5%) and organized (MAD=0.60; exact agreement 43.6%). Despite lower exact agreement in these domains, agreement within ±1 point remained high (≥96.4%), supporting overall consistency between evaluators. To supplement the MAD analysis, the intraclass correlation coefficient (ICC; 2-way mixed effects, absolute agreement) and weighted Cohen kappa were calculated for each domain and are reported in Table 3. However, domain-level ICC and kappa values were markedly attenuated by ceiling effects, as 99% of the individual ratings were 4 or 5 on the Likert scale. Under such restricted variance, these metrics can yield paradoxically low or negative values despite high observed concordance; for example, the succinct and internally consistent domains each achieved 100% agreement within ±1 point yet produced negative ICC and kappa values. Given these distributional constraints, the MAD and percentage agreement metrics provide a more clinically meaningful assessment of interevaluator concordance in this setting.

Physician Experience

Physician experience with Sahl AI was evaluated to assess its perceived impact on routine clinical practice. A larger pool of physicians who may not have contributed to the development or implementation stage but used Sahl AI in their clinical practice was surveyed. In total, 22 physicians responded; however, as the total number of invited physicians was not systematically recorded, a formal response rate cannot be calculated, and these results should be interpreted as descriptive feedback from a convenience sample of unknown proportion. The results are summarized in Table 4. Most responding physicians strongly agreed or agreed across all domains that Sahl AI produced comprehensive notes, produced more complete notes compared with current practice, had the potential to save time and reduce physician burnout, was faster, while having a positive impact on their clinical workflow.

XSL•FO
RenderX

Table . Summarization of the physicians' responses to various statements related to the use of Sahl AI and impact on their clinical practice.

Question	Strongly agree, n (%)	Agree, n (%)	Neither agree nor disagree, n (%)	Disagree, n (%)	Strongly disagree, n (%)
Sahl AI produces comprehensive clinical summaries of the consultations	19 (86.4)	3 (13.6)	0 (0)	0 (0)	0 (0)
I feel that Sahl AI can produce more complete notes compared to current documentation practices	14 (63.6)	6 (27.3)	2 (9.1)	0 (0)	0 (0)
Sahl AI was faster and thus has the potential to save time in documenting comprehensive clinical summaries compared to the present practice	19 (86.4)	2 (9.1)	1 (4.5)	0 (0)	0 (0)
I feel that the use of the tool can help reduce physician stress/burnout	15 (68.2)	6 (27.3)	1 (4.5)	0 (0)	0 (0)
It was easy to use the Sahl AI tool	19 (86.4)	2 (9.1)	0 (0)	1 (4.5)	0 (0)
Sahl AI has the potential to positively affect my clinical workflow	16 (72.7)	4 (18.2)	2 (9.1)	0 (0)	0 (0)

Discussion

Principal Findings

This study demonstrates the feasibility and effectiveness of a bilingual Arabic-English medical ambient AI scribe in generating accurate, thorough, and clinically useful documentation in a real-world clinical setting. The study lays a foundation and benchmark for the successful deployment of medical AI technologies for low-resource languages, such as Arabic, and in the bilingual setting. The positive feedback from surveyed physicians regarding the tool's ease of use, potential for time savings and reducing burnout, producing more complete notes and impact on clinical workflow provides preliminary, descriptive evidence of its value in clinical practice. The results are especially relevant to health care systems in Arabic-speaking settings, where improving documentation efficiency and quality remains a critical operational priority. Furthermore, to our knowledge, this study represents the first end-to-end evaluation of a medical ambient AI scribe spanning audio capture to finalized clinical note generation for the Arabic language, extending beyond prior work that has primarily focused on transcript-to-note pipelines.

Comparison to Prior Work

There is a growing body of research evaluating the performance of ambient AI scribes in the English language, but their performance has not been rigorously assessed in other languages, particularly Arabic, which has multiple dialects and is considered a low-resource language in natural language processing due to a dearth of available datasets. This lack of evaluation poses a significant challenge, as the linguistic and cultural nuances of Arabic necessitate tailored solutions for the successful deployment of medical AI technologies in Arabic-speaking regions. Additionally, challenges in recording medical data, such as ensuring data privacy of health data and compliance with laws, such as PDPL, have contributed to this limitation. This study provides a crucial methodology for the collection of data and the training of ambient AI scribes in any language, especially low-resource languages, such as Arabic, which can help address these gaps and pave the way for the successful deployment of these technologies in diverse linguistic and cultural contexts.

Medical ambient AI scribes have been widely adopted in English-speaking health care settings and are generally reported to generate accurate and comprehensive clinical notes. For example, Tierney et al [7] reported a high documentation quality score of 48 out of 50 using a modified PDQI-9 when evaluating English physician-patient transcripts. In contrast, substantially lower scores have been reported in studies evaluating large language models, such as GPT-4, with PDQI-9 scores of 29.7 out of 45 in 1 study and 25 out of 35 using a modified PDQI-9 in another [10,11]. Kernberg et al [11] further noted that the AI-generated notes frequently omitted clinically important information, introduced hallucinated content, or contained factual inaccuracies and that note generation was not consistently reproducible. More recently, an English-language medical ambient AI scribe incorporating audio input in addition to transcripts was evaluated using a modified PDQI-9 tool, achieving a mean score of 41.9 out of 50 [12].

XSL•FO
RenderX

As a reference benchmark, the original PDQI-9 validation study categorized scores of 26.2 out of 45 as "terrible or bad," compared with scores of 36.6 out of 45 as "good or excellent" [8]. Evidence in non-English and low-resource language settings remains limited. A recent study by Lee et al [13] reported a score of 50.2 out of 65 using a modified PDQI-9 tool for synthetic Arabic transcript-to-note generation. Additional efforts have been reported in other non-English contexts, including Spanish-language AI scribe implementations in surgical outpatient settings [14].

In this study, the total mean score was 42.4 (SD 1.84) out of 45 for Arabic conversations and 37.8 (SD 1.10) out of 40 for English conversations. As in prior studies, the assessment instrument was modified to suit the study context. As these modified domains have not been formally validated, the aggregate scores should be interpreted with caution and are not directly comparable to benchmarks established using the original PDQI-9 instrument. However, unlike previous research that primarily assessed note quality based on written transcripts, this study introduces a novel approach by evaluating notes generated directly from original audio recordings of physician-patient consultations. These audio-derived notes were subsequently reviewed by independent evaluators, offering a more authentic and context-rich basis for assessing note quality. This methodology improves existing studies by considering errors from speech-to-text engines, performing an end-to-end evaluation to better reflect real-world situations rather than considering transcripts as ground truth. By evaluating the entire process from audio input to textual output, the approach ensures that the findings are applicable to everyday use cases, enhancing the robustness and reliability of the results. Although the summary of prior work focuses on PDQI-9 and its derivatives for easier comparison, other studies have utilized Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Bidirectional Encoder Representations from Transformers (BERTScore), Crosslingual Optimized Metric for Evaluation of Translation (COMET), and Sheffield Assessment Instrument for Letters (SAIL), among others, which have not been discussed here.

Surveyed physicians perceived that the tool improved note accuracy and quality compared to the clinical practice of typing notes, resulting in more comprehensive and complete documentation compared to current practice. Most responding physicians perceived that the use of Sahl AI can reduce cognitive overload and save time, resulting in a positive overall impact on their clinical workflow. Such findings have been confirmed by other studies, especially showing that the use of ambient AI scribes leads to time saved and reductions in cognitive burden for physicians [10,15-21]. However, a recent report by the Peterson Health Technology Institute concluded that while AI scribes may reduce clinician burnout, it is unclear whether it actually saves time and leads to financial gains [22]. The report concluded that differences in outcomes were likely due to organization-specific metrics, which had variation in how they were defined and measured. Nevertheless, the result of this study is in line with the growing consensus that AI scribes can enhance physician efficiency by alleviating the cognitive burden associated with comprehensive documentation.

Limitations

This study used a single-arm design without a comparator, as comparative evaluation was outside the scope of this work. The absence of baseline comparisons in Arabic or bilingual settings with physician-authored documentation or generic large language models, both in the literature and in this study, limits the ability to draw conclusions regarding comparative effectiveness. Future studies incorporating head-to-head comparisons are required to establish relative performance.

The sample size was limited due to the practical challenges of collecting real-world audio data, which may affect the interpretability and generalizability of the findings. In addition, physicians used Sahl AI as a stand-alone web app that was not integrated into the electronic medical record. Electronic medical record integration is an important factor in simplifying clinical workflows and has been associated with the improved adoption of ambient AI scribes [21]. Real-world testing was conducted exclusively within the family medicine specialty; therefore, the findings may not generalize to other clinical specialties, and further evaluation across diverse clinical domains is warranted. Furthermore, the independent evaluation was conducted by physicians from the same health cluster; although evaluators were blinded to physician and patient identities, full blinding could not be ensured and may have introduced bias, especially as the raw audio was not modified to redact personal identifiable information.

The study did not include a systematic assessment of medical complexity between Arabic and English encounters, which may influence the interpretation of between-group comparisons. Future studies should capture encounter-level characteristics, such as number of medical entities spoken and number of diagnoses, to assess whether any observed language–based differences may be confounded by case complexity. Additionally, encounters were classified as Arabic or English based on the predominant language, without quantifying the degree of code-switching within individual consultations. This may limit the understanding of the model's performance in heavily code-switched versus predominantly monolingual encounters. Separately, a formal safety audit evaluating the potential clinical impact of documentation inaccuracies was not performed. Incorporating such assessments in future work may further strengthen the evidence base. While the training data included multiple Arabic dialects, performance was not systematically evaluated across specific dialect groups, limiting conclusions regarding dialect-level generalizability.

The physician experience survey was based on a convenience sample of clinicians who had used Sahl AI in at least 3 clinical encounters, which may introduce selection bias toward more engaged users. As the survey was distributed through direct outreach during deployment and the total number of physicians invited was not systematically recorded, a formal response rate could not be calculated. Consequently, the survey findings should be interpreted as descriptive user feedback rather than generalizable estimates of physician satisfaction or adoption. As such, the specific percentages reported from the survey should not be interpreted as representative of the broader physician population.

XSL•FO
RenderX

Since the completion of this study, subsequent iterations of the Sahl AI models have been released with improvements in accuracy and processing speed. Further research is needed to evaluate the clinical impact of Arabic medical ambient AI scribes, including objective measures of time savings, documentation quality compared with standard practice, effects on physician-patient interaction, and downstream operational and revenue-cycle outcomes.

Conclusions

This study highlights the potential of medical ambient AI scribe technology to enhance clinical documentation in both Arabic and English clinical environments. Sahl AI demonstrated the ability to generate high-quality clinical notes, with potential benefits including improved documentation quality, time savings, and reduced cognitive burden for physicians. To support broader adoption, future research should address the limitations identified in this study, including evaluation across diverse clinical settings and patient populations, as well as integration with electronic medical record systems. Importantly, this work provides a foundational framework for the development and training of medical ambient AI scribes in low-resource languages, such as Arabic, while also introducing a robust methodology for assessing the quality of AI-generated clinical documentation.

Acknowledgments

We would like to thank all the research assistants involved in labeling the audio data. Generative AI software (ChatGPT [OpenAI; GPT-5.2]) was used to improve the grammar of the text.

Funding

This study received in-kind support from Riyadh First Health Cluster, which funded the research activity of research assistants involved in data collection and labeling of data.

Data Availability

Aggregated modified 9-item Physician Documentation Quality Instrument scores, interevaluator agreement data, and physician survey responses are available upon request from the corresponding author. However, the raw audio recordings were permanently deleted 1 year after collection in accordance with ethics approval requirements and are no longer available.

Authors' Contributions

Conceptualization: ATK, MR, SASA, TNK, UTK, Y Alomran, Y Alsafi
Data curation: MAR, RA, WA, ZB
Formal analysis: UTK
Methodology: ATK, SASA, TNK, UTK, WA, Y Alomran, ZB
Project administration: MR, MAR, SASA
Resources: MR, Y Alomran, SASA
Supervision: ATK, MAR, MR, SASA, TNK, UTK
Validation: TNK, WA, ZB
Visualization: ATK, UTK
Writing – original draft: ATK, SASA, TNK, UTK, WA, ZB
Writing – review & editing: MAR, MR, RA, Y Alomran, Y Alsafi

Conflicts of Interest

UTK, ATK, WA, ZB, MAR, and TNK are employees of, consultants to, or have financial interests in Sahl AI, the company that owns the technology evaluated in this study. The remaining authors declare no competing interests.

References

1. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. J Am Med Inform Assoc 2021 Apr 23;28(5):1057-1061. [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
2. Christino MA, Matson AP, Fischer SA, Reinert SE, Digiovanni CW, Fadale PD. Paperwork versus patient care: a nationwide survey of residents' perceptions of clinical documentation requirements and patient care. J Grad Med Educ 2013 Dec;5(4):600-604. [doi: 10.4300/JGME-D-12-00377.1] [Medline: 24455008]
3. Miao J, Thongprayoon C, Cheungpasitporn W. Should artificial intelligence be used for physician documentation to reduce burnout? Kidney360 2024 May 1;5(5):765-767. [doi: 10.34067/KID.0000000000000430] [Medline: 38523133]
4. Scott D, Purves IN. Triadic relationship between doctor, computer and patient. Interact Comput 1996 Dec;8(4):347-363. [doi: 10.1016/S0953-5438(97)83778-2]

XSL•FO
RenderX

5. Cao DY, Silkey JR, Decker MC, Wanat KA. Artificial intelligence-driven digital scribes in clinical documentation: pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD Int 2024 Jun;15:149-151. [doi: 10.1016/j.jdin.2024.02.009] [Medline: 38571698]
6. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med 2019 Aug;34(8):1626-1630. [doi: 10.1007/s11606-019-05035-1] [Medline: 31090027]
7. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
8. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174. [doi: 10.4338/aci-2011-11-ra-0070] [Medline: 22577483]
9. Personal data protection law. : Saudi Data and Artificial Intelligence Authority (SDAIA); 2023 URL: https://sdaia.gov.sa/en/SDAIA/about/Documents/Personal%20Data%20English%20V2-23April2023-%20Reviewed-.pdf [accessed 2025-06-24]
10. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a digital scribe system on clinical documentation time and quality: usability study. JMIR AI 2024 Sep 23;3:e60020. [doi: 10.2196/60020] [Medline: 39312397]
11. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. J Med Internet Res 2024 Apr 22;26:e54419. [doi: 10.2196/54419] [Medline: 38648636]
12. Palm E, Manikantan A, Mahal H, Belwadi SS, Pepin ME. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. Front Artif Intell 2025;8:1691499. [doi: 10.3389/frai.2025.1691499] [Medline: 41199808]
13. Lee C, Kumar S, Vogt KA, et al. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. Sci Rep 2025;15(1):27619. [doi: 10.1038/s41598-025-10451-x]
14. Gasque RA, Zaietta N, Mollard L, et al. HPB SmartNotes: the impact of artificial intelligence on surgeon workload in the outpatient office. EngMedicine 2025 Dec;2(4):100101. [doi: 10.1016/j.engmed.2025.100101]
15. Stults CD, Deng S, Martinez MC, et al. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Netw Open 2025 May 1;8(5):e258614. [doi: 10.1001/jamanetworkopen.2025.8614] [Medline: 40314951]
16. Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025 Feb 3;8(2):e2460637. [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]
17. Ma SP, Liang AS, Shah SJ, et al. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc 2025 Feb 1;32(2):381-385. [doi: 10.1093/jamia/ocae304] [Medline: 39688515]
18. Shah SJ, Crowell T, Jeong Y, et al. Physician perspectives on ambient AI scribes. JAMA Netw Open 2025 Mar 3;8(3):e251904. [doi: 10.1001/jamanetworkopen.2025.1904] [Medline: 40126477]
19. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J 2024 Sep;11(3):100157. [doi: 10.1016/j.fhj.2024.100157] [Medline: 39371531]
20. Mess SA, Mackey AJ, Yarowsky DE. Artificial intelligence scribe and large language model technology in healthcare documentation: advantages, limitations, and recommendations. Plast Reconstr Surg Glob Open 2025 Jan;13(1):e6450. [doi: 10.1097/GOX.0000000000006450] [Medline: 39823022]
21. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catal Innov Care Deliv 2025 2025 Apr 16;6(5). [doi: 10.1056/CAT.25.0040]
22. Adoption of AI in healthcare delivery systems: early applications and impacts. : Peterson Health Technology Institute; 2025 URL: https://phti.org/wp-content/uploads/sites/3/2025/03/PHTI-Adoption-of-AI-in-Healthcare-Delivery-Systems-Early-Applications-Impacts.pdf [accessed 2026-03-04]

Abbreviations

AI: artificial intelligence
EMR: electronic medical record
ICC: intraclass correlation coefficient
MAD: mean absolute difference
PDPL: Personal Data Privacy Law
PDQI-9: 9-item Physician Documentation Quality Instrument

XSL•FO
RenderX

Edited by A Coristine; submitted 02.Sep.2025; peer-reviewed by A Alghamdi, C Lee, K Pearlman; revised version received 24.Feb.2026; accepted 24.Feb.2026; published 24.Mar.2026.

Please cite as:
Khan UT, Khan AT, Aljaadi W, Alhadlaq R, Baqashmer Z, Alsafi Y, Alomran Y, Al Rusaiyes M, Radif M, Khan TN, Altamimi SAS
A Bilingual Arabic-English Ambient AI Scribe for Clinical Documentation: Prospective Evaluation Study
JMIR Med Inform 2026;14:e83335
URL: https://medinform.jmir.org/2026/1/e83335
doi:10.2196/83335

© Umair Tahir Khan, Ammar Tahir Khan, Waleed Aljaadi, Razan Alhadlaq, Zahran Baqashmer, Yasin Alsafi, Yousef Alomran, Maha Al Rusaiyes, Muaddiyah Radif, Tahir Naeem Khan, Saleh Abdullah Saleh Altamimi. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Large Language Model–Enabled Editing of Patient Audio Interviews From “This Is My Story” Conversations: Comparative Study

Bikram Bains[1*]; Sampath Rapuri[1*]; Edgar Robitaille[1*]; Jonathan Wang[2], BS; Arnav Khera[2], BS; Catalina Gomez[3], BS; Eduardo Reyes[4], MS; Cole Perry[4], BS; Jason Wilson[5], MS; Elizabeth Tracey[5], MS

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States

[2]School of Medicine, Johns Hopkins University, Baltimore, MD, United States

[3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States

[4]Technology Innovation Center, Johns Hopkins Medicine, Baltimore, MD, United States

[5]Division of Spiritual Care and Chaplaincy, Johns Hopkins Medicine, 1800 Orleans St, Baltimore, MD, United States

[*]these authors contributed equally

Corresponding Author:
Elizabeth Tracey, MS
Division of Spiritual Care and Chaplaincy, Johns Hopkins Medicine, 1800 Orleans St, Baltimore, MD, United States

Abstract

Background: This Is My Story (TIMS) was started by Chaplain Elizabeth Tracey to promote a humanistic approach to medicine. Patients in the TIMS program are the subject of a guided conversation in which a chaplain interviews either the patient or their loved one. They are asked four questions to elicit clinically actionable information that has been shown to improve communication between patients and medical providers, strengthening medical providers’ empathy. The original recorded conversation is edited into a condensed audio file approximately 1 minute and 15 seconds in length and placed in the electronic health record where it is easily accessible by all providers caring for the patient.

Objective: TIMS is active at the Johns Hopkins Hospital and has shown value in assisting with provider empathy and communication. It is unique in using audio recordings to accomplish this purpose. As the program expands, there exists a barrier to adoption due to limited time and resources needed to manually edit audio conversations. To address this, we propose an automated solution using a large language model to create meaningful and concise audio summaries.

Methods: We analyzed 24 TIMS audio interviews and created three edited versions of each: (1) expert-edited, (2) artificial intelligence (AI)–edited using a fully automated large language model pipeline, and (3) novice-edited by two medical students trained by the expert. A second expert, blinded to the editor, rated the audio interviews in a randomized order. This expert scored both the audio quality and content quality of each interview on 5-point Likert scales. We quantified transcript similarity to the expert-edited reference using lexical and semantic similarity metrics and identified omitted content relative to that same expert interview.

Results: Audio quality (flow, pacing, clarity) and content quality (coherence, relevance, nuance) were each rated on 5-point Likert scales. Expert-edited interviews received the highest mean ratings for both audio quality (4.84) and content quality (4.83). Novice-edited scored moderately (3.84 audio, 3.63 content), while AI-edited scored slightly lower (3.49 audio, 3.20 content). Novice and AI edits were rated significantly lower than the expert edits (P<.001), but not significantly different from each other. AI and novice-edited interview transcripts had comparable overlap with the expert reference transcript, while qualitative review found frequent omissions of patient identity, actionable insights, and overall context in both the AI and novice-edited interviews. AI editing was fully automated and significantly reduced the editing time compared to both human editors.

Conclusions: An AI-based editing pipeline can generate TIMS audio summaries with comparable content and audio quality to novice human editors with one hour of training. AI significantly reduces editing time and removes the need for manual training; with further validation, it could offer a solution to scale TIMS to a large range of health care settings.

(JMIR Med Inform 2026;14:e80205) doi:10.2196/80205

KEYWORDS

audio recording; communication; This Is My Story; TIMS; distress reduction; empathy; patient interview; provider/patient communication; large language model

XSL•FO
RenderX

Introduction

Recent statistics show that health worker burnout is a widespread issue [1]. A 2022 Centers for Disease Control and Prevention report found that 46% of health workers "often" or "very often" felt burned out, increasing from 32% in 2018. Physicians in the United States also report similarly high burnout rates (56% in 2021, 53% in 2022, and 48% in 2023), with an all-time high physician burnout rate of 63% during the pandemic [2,3]. Some reasons for this burnout include excessive work hours, administrative burdens (such as electronic health record documentation), insufficient support staff, and limited organizational and leadership support [4-8]. These chronic stresses impact both patients and clinicians. For example, Andhavarapu et al [9] mentioned that symptoms of depression, anxiety, and posttraumatic stress disorder were reported in 34% of the health care workers surveyed (while 14% reported severe posttraumatic stress disorder), with the highest prevalence among nursing staff (42.8%) and physicians (25.2%). Similarly, the National Academies' 2019 report found that between 35% and 54% of US nurses and physicians and 45% to 60% of medical students and residents experience substantial burnout symptoms throughout their careers [10].

Empathy can serve as a solution, reducing widespread symptoms of burnout while promoting professional fulfillment and strengthening connection with patients [11-14]. Already, health care organizations have recognized the value of empathy and designed personal and patient-centered interventions within their clinical workflows [15]. For example, the This Is My Story (TIMS) program was developed by Chaplain Elizabeth Tracey at the Johns Hopkins Hospital to bring a more patient-centered and empathetic approach to medicine [16]. Patients who participate in the TIMS program take part in a conversation with a chaplain; if the patient is noncommunicative, a chaplain has a conversation with the patient's loved ones. These conversations are guided by four questions: How do you prefer to be addressed? What brings you joy? What does your medical team need to know to care for you best? What brings you peace?

In the words of Dr Charles Cumming, Director Emeritus of Otolaryngology at Johns Hopkins, *"TIMS is about helping us get back to the proper essence of medicine...it's essential to get to know the patient if we're going to be able to help that patient as best we can"* [17]. TIMS conversations have demonstrated clear benefits for clinical communication and empathy, providing an opportunity for meaningful connection with patients to directly target the emotional aspects of burnout [16,18-20]. Past studies by Tracey et al [21] support the positive outcomes the program has had on patients, their families, and the care team. For example, one previous study reported a 74% increase in staff empathy for patients and a 99% improvement in interactions by patients' loved ones with the medical team. Although it has also been shown to be useful in improving staff empathy and reducing distress by 69%, the process of recording and editing conversations can be labor-intensive [21]. By automating the conversation summarization process, these benefits can be made accessible to a wider range of patients and medical institutions.

In this study, we propose an automated editing pipeline for TIMS interviews using a large language model (LLM) and evaluate whether artificial intelligence (AI)–edited interviews are a viable alternative to manual editing. Because medical students were frequently trained to edit TIMS interviews during the pandemic, they are a reasonable baseline for comparing performance. We designed our analysis around two key hypotheses: (1) that AI-edited interviews maintain similar quality to expert-edited interviews in both audio and content metrics, and (2) that AI-edited interviews can be produced more quickly than interviews produced by expert or novice editors.

Methods

Study Design

We used a within‐subjects, single‐group design in which our reviewer evaluated interviews across three independent editing conditions (expert, AI, novice). Editors were eligible if they had professional experience interpreting patient–clinician audio interviews. Two chaplains from the Johns Hopkins Hospital took part in the study. The novice editors were two medical students who joined the study team from the Johns Hopkins School of Medicine, each having completed an hour-long training session on audio‐editing with an expert editor (Chaplain Elizabeth Tracey). The two novice editors edited 12 randomly assigned audio interviews, mirroring the normal workflow for the TIMS initiative without the AI tool.

Patient Audio Dataset

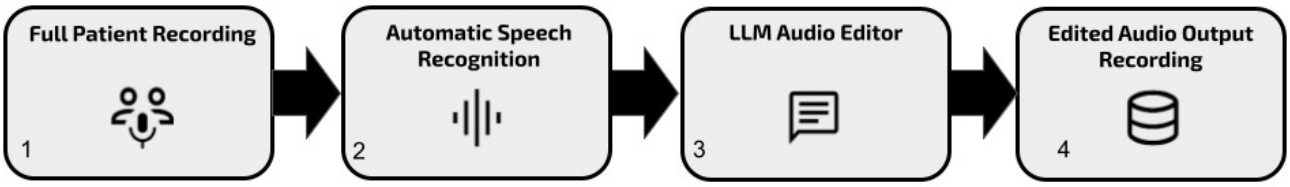
We used a retrospective dataset of audio recordings from 24 patients admitted to the Johns Hopkins Neurosciences Critical Care Unit, a tertiary intensive care unit for patients with diseases of the central or peripheral nervous systems, admitted from departments such as neurosurgery, neurology, and others.

AI Editing Pipeline

Audio recordings were first transcribed using Nvidia's Parakeet-TDT 0.6B v2 automatic speech recognition (ASR) model (Figure 1) [22-24]. ChatGPT-4o [25] processed the transcript using a custom prompt. This prompt asked the model to extract only patient statements that provided essential information about their condition, experiences, feelings, or personal identity. It was also directed to exclude any filler or repetitive content and keep humorous or insightful remarks to preserve patient identity and humanity.

XSL•FO
RenderX

Figure 1. Overview of artificial intelligence–automated workflow for summarizing This Is My Story audio interviews. LLM: large language model.



The model was instructed to include the interviewer's four core questions for the TIMS program while excluding interjections or examples provided by the interviewer. Instructions were given to return a processed transcript that the model estimated would take 1.5 minutes during a TIMS interview. The full model prompt is available in Multimedia Appendix 1. The relevant timestamps identified were then used to splice together the final audio file. Examples of cases in which ASR output might have impacted the LLM output are presented in Table S2 in Multimedia Appendix 2. This process was entirely automated and was run on an NVIDIA GeForce RTX 4090 with 24 GB of RAM.

Survey Design

To evaluate both audio quality and content quality for each edited TIMS interview, we created an online questionnaire for our reviewer to complete using a 5-point Likert scale (1=poor, 5=excellent) for each question. The survey included questions grouped into audio (natural flow, pausing/spacing, transitions, pacing, overall listenability of the interview) and content (conversation flow, speaker/topic tracking, patient representation, understanding of patient characteristics, preparation for interaction for patient providers/care team, nuance of the patient's life, relevance of details) quality domains with all details shown in Textbox 1 below. This same survey was used independently for each edited interview for consistent comparison.

Textbox 1. Summary of domain and survey questions following survey administration and data collection.

Audio quality

1. How natural is the conversation flow?
2. How effective are the pauses and spaces between answers?
3. How smooth are the transitions between questions and answers?
4. How does the pacing of the audio feel overall?
5. Overall, how easy is the interview to listen to?
6. Can you understand the flow of the conversation?

Content quality

7. Can you keep track of the speaker and the topic?
8. Is the patient well represented in this conversation?
9. Do you understand the patient's likes/dislikes (proclivities, idiosyncrasies, etc)?
10. Do you feel prepared to interact with the patient in a meaningful way?
11. How well does the conversation capture the depth and nuance of the patient's experience?
12. Does the conversation convey the patient's emotions effectively?
13. How relevant are the details shared during the conversation for understanding the patient's story?
14. How engaging is the conversation in terms of sustaining your interest in the patient's story?

The survey was created and administered using Google Forms. Content experts (ET, JW, CG) provided qualitative feedback on an initial draft of the questions. This feedback focused on improving the clarity and relevance of each item. The questions were then refined based on this input to better capture the intended domains of audio and content quality. It is important to note that because the survey instrument is novel, its reliability and validity have not yet been formally established. The survey asked questions 1, 5, 10, and 11 to gauge general sentiment for the conversation being rated, and more pointed questions to understand where the audios may differ in terms of score (Textbox 1). Since assessing the flow of conversation relies on both audio and content quality, a question about it was used for both parts of the survey. The individual survey responses are available in Figure S1 in Multimedia Appendix 3.

Before rating each edited version, the reviewer listened to the corresponding raw interview as a baseline reference to better understand the context of the interview. Before rating each interview, the reviewer was asked to first listen to two calibration audios—one edited poorly that should score low across all questions, and an expert-edited interview that would score highly. This allowed us to set audio quality expectations for each extreme on the survey. The survey was completed independently for each condition, with the order of the audio

XSL•FO
RenderX

files randomized by condition for each patient to reduce potential order effects.

Editing durations were recorded automatically for the AI pipeline and self‐reported by each novice editor. Expert editing times were not collected due to the limitations of the retrospective dataset.

Content Analysis

For each edited interview, we generated text transcriptions to study the differences in content among the three types of edited interviews. Text transcriptions were created using the Parakeet transcription model. The novice- and AI-edited interviews were compared to the expert edits, and three members of the study team analyzed differences. Types of errors were identified for both AI- and novice-edited conditions across all samples. The most common types of errors were then formalized and described in the results (Table 1). Content similarity between each condition and expert edits was quantitatively measured using ROUGE-L, ROUGE-1, ROUGE-2, bidirectional encoder representations from transformers (BERT), and METEOR on interview transcripts following studies of medical document summarization [26,27]. All metrics measure the similarity of words between summaries and assign a score from 0 to 1, with the ROUGE scores measuring lexical overlap and other scores [28]. METEOR and the BERT scores were used to assess the semantic overlap. Both factor in semantic similarity between words rather than the exact word choice. METEOR also assigns a penalty for differences in phrasing.

Table . Common omissions and inaccurate portrayals by artificial intelligence and novice editors.

Type of error	Artificial intelligence errors	Novice errors
Omission of actionable patient insights	Failure to include specific interests or hobbies of the patient [talking about favorite musical artists bringing her joy]: “She likes Anita Baker, Regina Belle, and Gladys Knight.”	Failure to include information about the patient’s comfort [informing about her medical condition to better care for the patient]: “She has had eczema since she was about three or four, so her skin has to stay moisturized.”
Omission of patient identity and empathy	Failure to include details relevant to understanding the patient’s background [explaining his occupation and hobbies] : “On the church side, he loves to teach. He is a pastor.”	Failure to mention important characteristics about the patient [claiming that her time at Hopkins has made her more independent and resilient]: “She [patient] worked at Johns Hopkins for over 30 years.”
Omission of emotional background	Failure to include framing details relevant for a patient’s background [talking about what brings the patient joy]: “Me [patient’s husband] … We’ve been married 20 years.”	Failure to include details relevant for a patient’s emotional state and anxiety [explaining how he mainly only trusts his partner for everything]: “[He has] a little bit of a trust issue with the medical field.”
Poor narrative fluency	Prompting questions fail to be edited out of the interview: “Introduce yourself and tell me how you’re related to the patient.”	Filler words before prompting questions fail to be edited out of the interview: “That’s great! So, what brings the patient peace?”

Relationships between audio length and content quality were also explored through simple linear regression of each ROUGE metric on the duration of the original interview (in minutes).

Statistical Analysis

We conducted a Friedman test to compare audio‐quality and content‐quality ratings across conditions, with Bonferroni‐corrected Wilcoxon signed-rank post hoc tests to adjust for multiple comparisons. Editing times were analyzed with an independent‐samples *t* test to test significant differences between the two novice editors. We also examined the relationship between the raw interview length and lexical and semantic score overlap for each editor type using Pearson correlation, testing if each slope differed from zero. We then performed an analysis of covariance with transcript length, editor type, and their interaction term to determine if the slope of the length-overlap relationship differed between AI and novice editors.

Ethical Considerations

Ethical approval was not required for this study as it involved a secondary analysis of anonymized data. The original data collection was conducted under Johns Hopkins institutional review board review and approval of the studies with informed consent obtained from all subjects; the consent allowed for future data use and any participants who declined this future use were not included in this secondary analysis. This study was conducted in accordance with all local, institutional, national, and international regulations on human subject research.

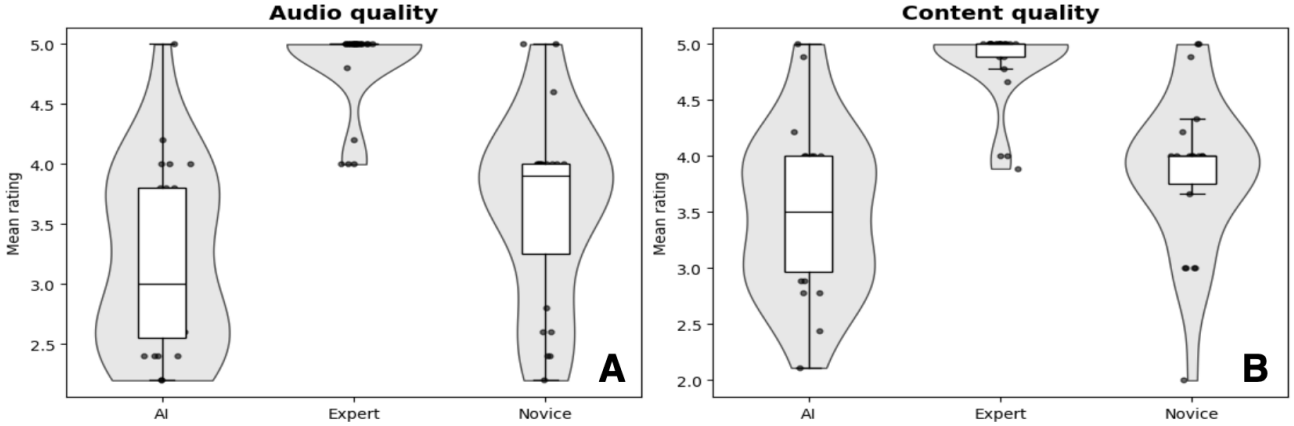
Results

Quantitative Comparison of Editing Quality

Across the three types of editors, the mean audio quality ratings were 3.49 (AI-edited), 3.84 (novice-edited), and 4.84 (expert-edited). Similarly, the mean content quality ratings were 3.20 (AI-edited), 3.63 (novice-edited), and 4.83 (expert-edited). The AI-edited interviews demonstrated a higher variability in the rated content quality compared to the novice-edited interviews (SD 0.73 and SD 0.68, respectively). Both interview types were similarly varied in their audio quality ratings (SD 0.77 and 0.78, respectively). Figure 2 highlights the distributions of content and audio quality ratings across each type of edited interview.

XSL•FO
RenderX

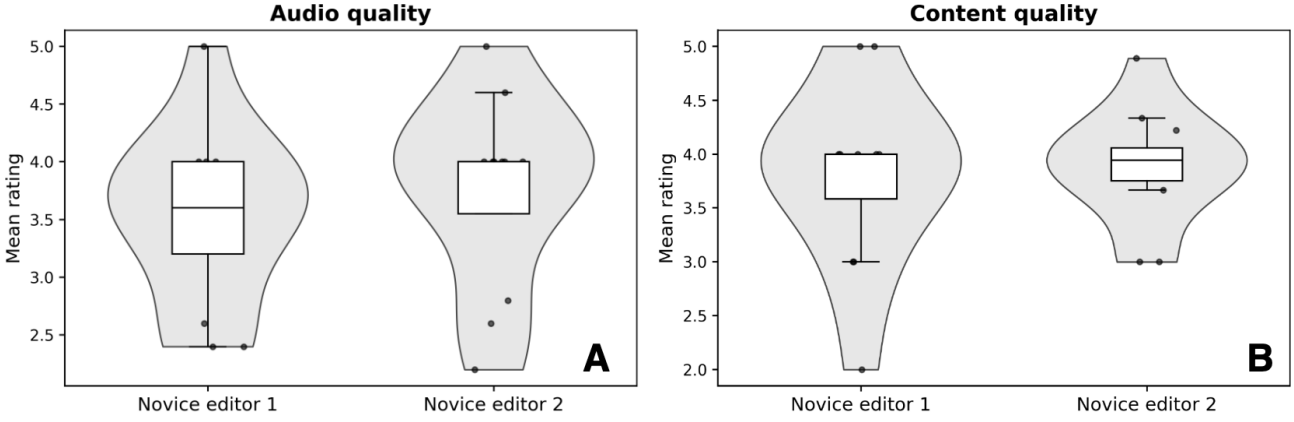
Figure 2. (**A**) A comparison of mean audio quality ratings across the three types of editors aggregated across all 24 audio interviews. (**B**) All editors follow similar trends for the content quality ratings. AI: artificial intelligence.



We observed significant differences in audio and content quality between the novice- and expert-edited interviews (P<.001) as well as between the AI- and expert-edited interviews (P<.001). No significant differences were noted between the AI- and novice-edited interviews for either content quality (P=.31) or audio quality (P=.33). A detailed breakdown of the ratings for each individual survey question across all patient interviews can be found in Figure S1 in Multimedia Appendix 3.

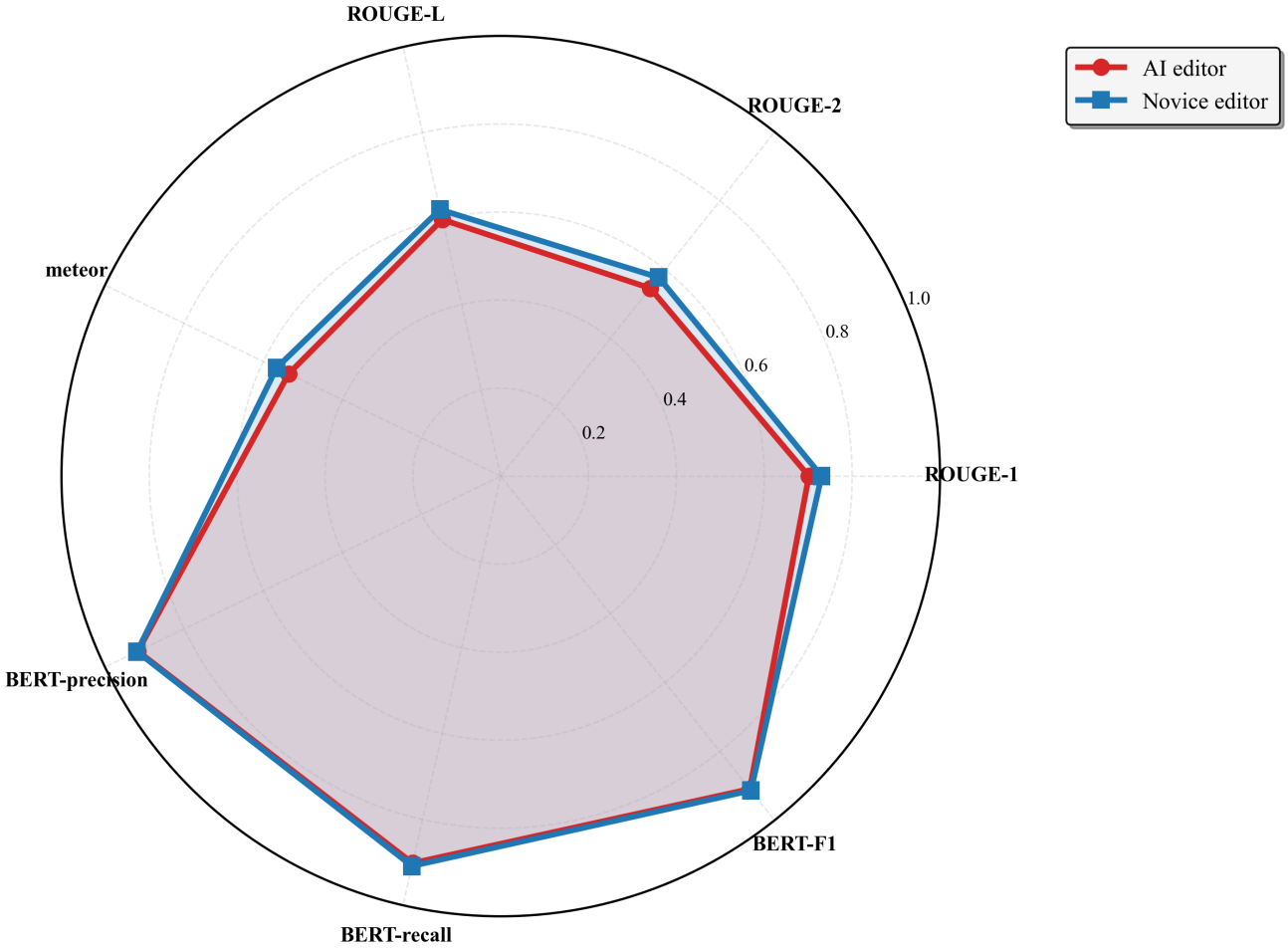
To understand the variability between the novice editors, we broke down the differences in rated audio and content quality in Figure 3. Between the two novice editors, we found that novice editor 1 demonstrated a mean content quality score of 3.81 (SD 0.83) and a mean audio quality score of 3.52 (SD 0.76). The second editor's mean content quality score was measured to be 3.88 (SD 0.52), with a mean audio quality score of 3.75 (SD 0.81). However, neither intragroup difference was significant for both content quality (P>.99) and audio quality (P=.51). Figure 4 shows the mean statistical scores across all audio interviews for both the AI and novice editors.

Figure 3. (**A**) A comparison of mean audio quality ratings between the two novice editors, each of whom edited 12 randomly assigned audio interviews. (**B**) Both editors achieved comparable content quality ratings, but the second novice editor exhibited significantly lower variability. All statistical scores of content similarity highlighted the similarities between the artificial intelligence and novice editors, and we report no statistically significant differences between any metric across both types of editors (P>.05). Detailed scores across each metric are contained in Table S1 in Multimedia Appendix 2.



XSL•FO
RenderX

Figure 4. Mean statistical scores across all 24 audio interviews for both the artificial intelligence and novice editors. AI: artificial intelligence. BERT: bidirectional encoder representations from transformers.



Qualitative Error Analysis

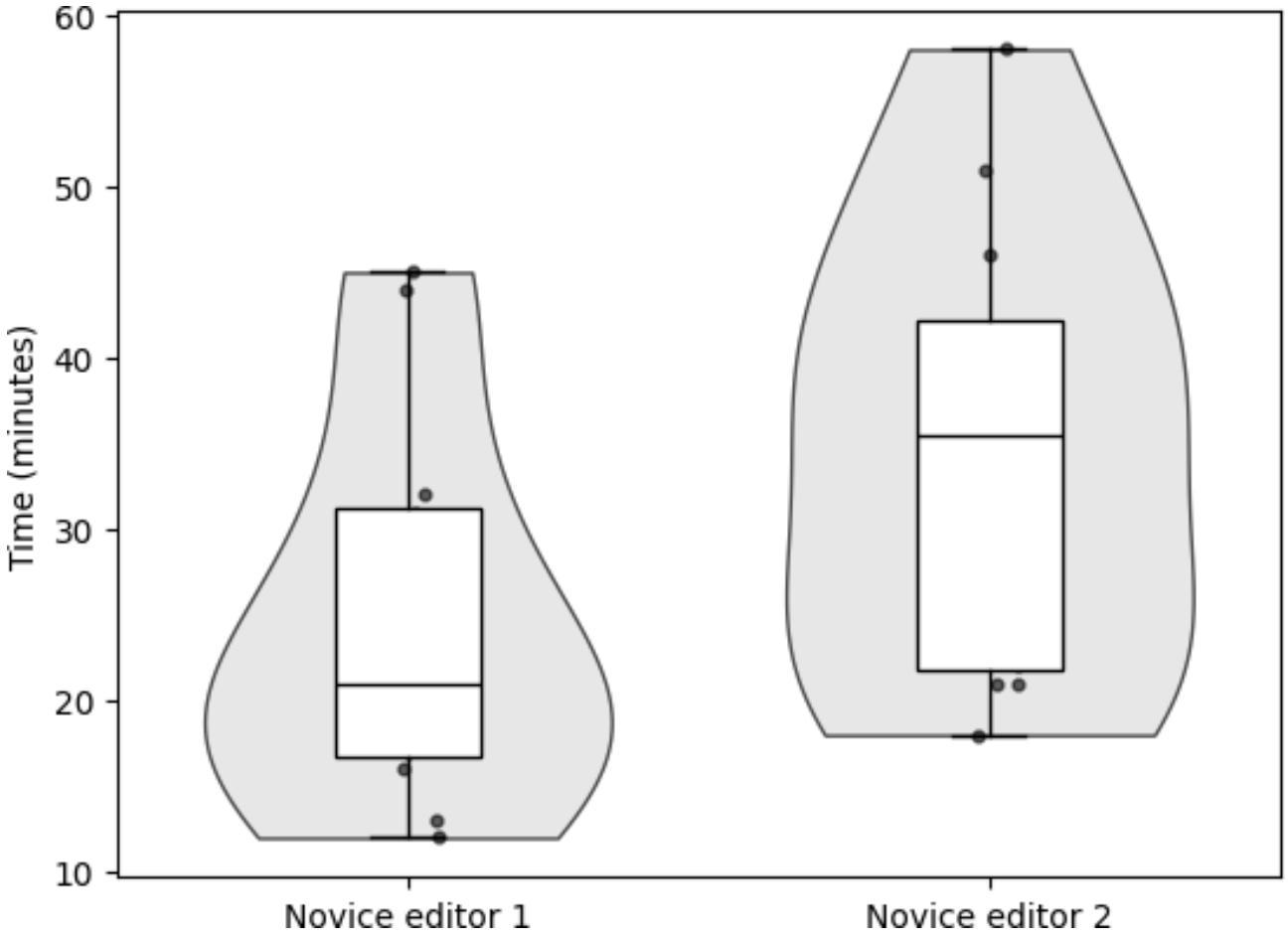
We qualitatively compared the AI- and novice-edited interviews to the expert-edited interviews and found five types of errors repeated across many of the interviews, which are described in Table 1. Many of the errors were similar across the AI- and novice-edited interviews.

Editing Efficiency

On average, each novice editor took 29.54 minutes to edit each interview (SD 12.69 min). However, we report a difference (P=.06) in the time each novice editor takes, with the first editor taking 24.50 minutes (SD 11.18 min) and the second taking 34.58 minutes (SD 13.08 min) to edit each interview. Figure 5 highlights the variability in the time to edit interviews between novice editors. Mean editing times for the expert editor are unavailable as they were not recorded. Based on anecdotal evidence from the expert editor, each audio interview required around 5 to 10 minutes to edit. In contrast to both the expert and novice editors, our automated AI-editing pipeline took less than 10 seconds from ingestion of the raw audio interview to the saving of the edited interview.

XSL•FO
RenderX

Figure 5. Time to edit each interview between each novice editor.

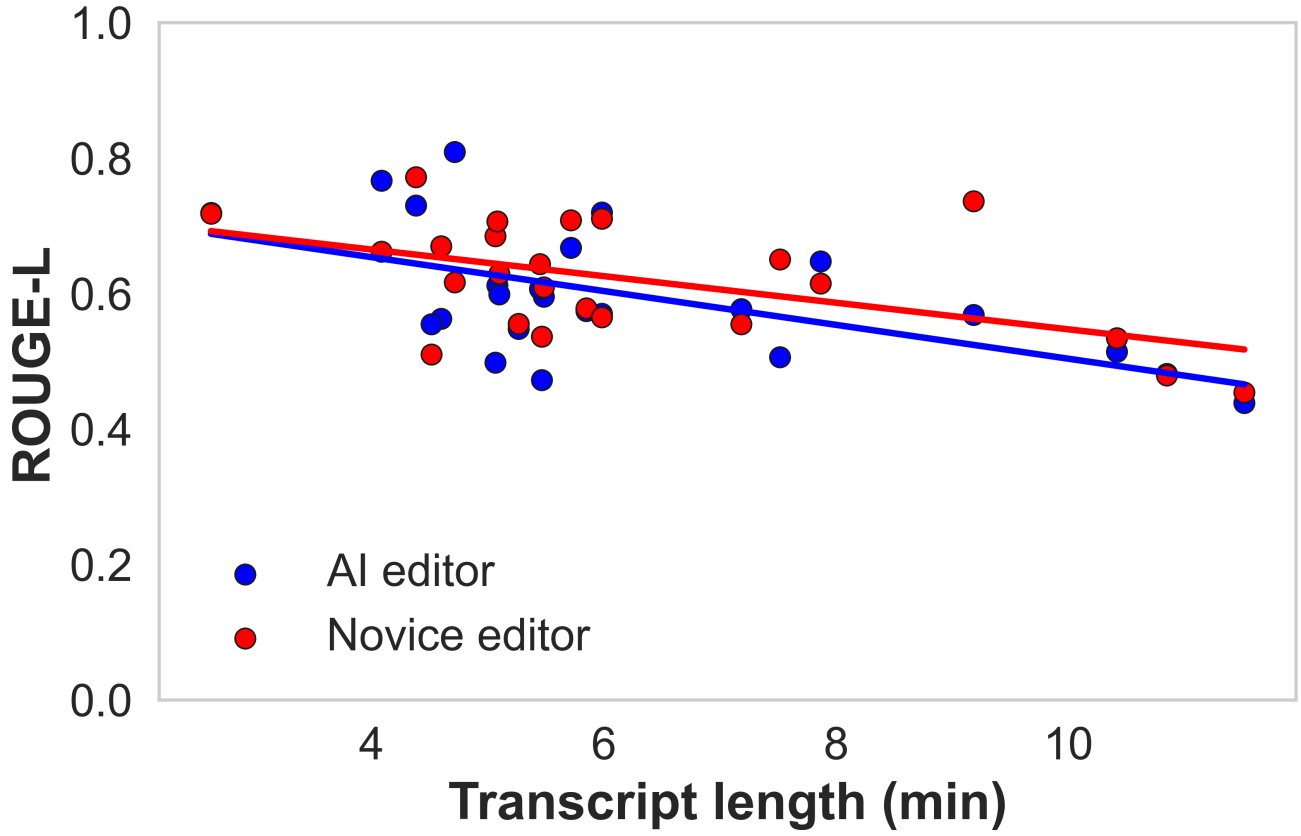


Impact of Interview Length on Editing Quality

We investigated if longer raw audio interviews were associated with changes in the lexical and semantic overlap with the expert reference, as measured by ROUGE-L. For AI-edited interviews, there was a significant negative correlation between transcript length and ROUGE-L scores (r=−0.58, R^2=0.34; Figure 6). This same trend was seen for novice-edited interviews, with a negative correlation (r=−0.52, R^2=0.27). Both slopes were significantly different from zero (P<.05), suggesting longer interviews were associated with lower transcript overlap to expert editor reference for both conditions. Regression plots for ROUGE-1, ROUGE-2, METEOR, and BERT scores showed comparable patterns and are provided in Figure S2 in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 6. Linear regression of the raw transcript length versus ROUGE-L, reflecting the change in lexical similarity to the expert reference with longer interview times.



Discussion

Principal Results

Our study compares the listening experience of patient conversations summarized by an expert editor, a novice editor, and ChatGPT-4o. There was no significant difference in content and audio quality between the AI and novice editors, and both showed varying performance across the samples. Further analysis of the edited transcripts revealed that both of these groups omitted key details. The expert editor had a significantly higher audio and content quality rating than both experimental groups and less variability across samples. Exclusion of natural pauses between phrases and auditory cues to break up conversation led to lower audio quality scores.

Comparison to Prior Work

As burnout remains high among health care workers, TIMS provides an opportunity for meaningful connections with patients to target the emotional aspect of burnout. However, the manpower needed to edit audio is a limiting factor for this program's widespread adoption. Although the audio-editing pipeline introduced was originally developed to expand the TIMS program, these results are also broadly relevant to the use of AI in clinical practice and examine a previously unexplored application of LLMs, specifically editing audio content for a medical context. Previous studies have investigated ChatGPT as a clinical decision-making tool, for taking notes, and analyzing literature [26,28-35]. There have also been studies that demonstrate ChatGPT's capability to elicit empathetic feelings in an emergency setting can even exceed that of clinicians [36-38]. However, the combination of a summarization task to build empathy in a third-party listener has not been investigated. In this study, we aim to evaluate whether an LLM is able to extract emotionally relevant details from a conversation. Additionally, we provide a replicable framework for integrating LLMs in the analysis of patient audio for a broad variety of applications within health care settings. We found that our pipeline faced similar issues raised in previous studies examining ChatGPT's ability to summarize in a medical context. Kernberg et al [39] reported that 58% of structured medical notes from patient-physician interactions omitted important information. A manual analysis of the transcripts revealed that details considered important to a patient's story were also often omitted, highlighting a key disadvantage of LLMs in the literature. ChatGPT also tends to vary widely in the quality of responses across the samples. Although this was a shortcoming observed in the novice editors' performance on audio and content quality (SD 0.68 and SD 0.78), the expert editor's performance was consistent.

In addition to the survey, ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) for the novice and AI editors indicated a high level of word overlap with the expert-edited transcript, with no significant differences between the two groups. High (~0.9) BERT scores (BERT-recall, BERT-precision, BERT-F_1) were also reported, indicating a high degree of semantic overlap that was not necessarily reflected in the ROUGE score. As the length of the audios increased, there was a statistically significant negative correlation between interview length and the ROUGE-L score for both AI and novice editors (P<.05), indicating longer interviews tended to have less lexical overlap

XSL•FO
RenderX

with the expert reference. This implies that there might be a length of audio that may not be as suitable for AI editing that will become more apparent as longer audios are recorded. These results from an established tool align with the insights from the survey, suggesting some level of construct validity for the survey questions.

Limitations

There are key limitations beyond the strengths of this study. First, there is no standardized or validated survey instrument available, so the introduction of a novel survey to assess the impact of each audio on a listener was necessary. However, abstract questions concerning "patient representation" or "nuance" are susceptible to subjective interpretation, a weakness that is amplified by our use of a single blinded reviewer. We attempted to standardize these ratings with calibration audios that were developed, but we cannot exclude the possibility that these subjective quality scores were influenced by rater bias. Despite this concern, the consistently high scores awarded to the expert-edited interviews provide some evidence of the survey's validity, as the rater reliably scored the gold-standard interviews. This survey could be adopted in the future by studies to measure the efficacy of interventions to increase empathy in medicine.

Second, the sample size of the study was also relatively limited, with only 24 samples and 1 recruited rater who was surveyed, which makes the results prone to bias. To build on this work, a larger sample size of patient audio interviews and experienced interview raters should be recruited. Previous volunteers of the program were able to receive iterative feedback on their work over long periods, but the novice editors had approximately 1 hour of training in comparison, so their skills were not as developed. Lastly, we were unable to obtain granular editing time measurements from the expert editor as these were retrospectively edited. However, the AI pipeline's completion time of under 10 seconds represents a multiple-orders-of-magnitude improvement in efficiency against any manual editing process.

Future Directions

We have presented the groundwork for an audio transcription and editing pipeline for humanistic patient conversations. Future work should test newer models as they improve, and others that are currently available besides ChatGPT-4o, with the same pipeline. Other strategies to improve performance include fine-tuning the LLM model, using AI agents to summarize the transcript, testing other ASR models, introducing patient-specific contextual metadata, and further prompt engineering to optimize the output. Error propagation was not formally tracked through the entire editing pipeline, but we have hypothesized an association between ASR errors and the final output quality. Future work should investigate these errors.

Conclusions

We conclude that ChatGPT-4o can create summarized audio files with similar audio and content quality to a novice editor in just a fraction of the time. However, the expert editor outperforms the AI editing pipeline and the novice editors on all metrics. After further validation, this tool could be implemented in the TIMS program to reduce workload and overcome adoption barriers.

Acknowledgments

This research has been generously supported with grants from the John Conley Foundation for Ethics and Philosophy in Medicine. We additionally acknowledge support from the Catalyst Award and the Diversity Innovation Grant from Johns Hopkins University. Lastly, we thank the Johns Hopkins Technology Innovation Center for providing access to a protected health information–compliant version of ChatGPT-4o.

Data Availability

The original interview recordings are not publicly available for privacy protection considerations but are available from the corresponding author on reasonable request. The relevant code is provided here [40]. This code corresponds to the automated audio-editing pipeline that processes raw audio, uses ChatGPT-4o to extract key segments, and stitches those selected segments into an AI-edited audio.

Authors' Contributions

All authors contributed to the conceptualization and study design. SR, BB, and ER curated the data and conducted the analysis. ET and JW managed and supervised the project. All authors reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GPT-4o summarization prompt.
[DOCX File, 16 KB - medinform_v14i1e80205_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Supplementary tables.
[DOC File, 17 KB - medinform_v14i1e80205_app2.doc]

Multimedia Appendix 3
Supplementary figure.
[DOC File, 231 KB - medinform_v14i1e80205_app3.doc]

References

1. Health worker mental health crisis. Centers for Disease Control and Prevention. 2023 Oct 24. URL: https://www.cdc.gov/vitalsigns/health-worker-mental-health/index.html [accessed 2025-07-06]
2. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022 Dec;97(12):2248-2258. [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]
3. Physician burnout statistics 2024: the latest changes and trends in physician burnout by specialty. American Medical Association. 2024 Aug 19. URL: https://www.ama-assn.org/practice-management/physician-health/physician-burnout-statistics-2024-latest-changes-and-trends [accessed 2025-07-06]
4. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018 Jun;283(6):516-529. [doi: 10.1111/joim.12752] [Medline: 29505159]
5. Singh R, Volner K, Marlowe D. Provider burnout. In: StatPearls: StatPearls Publishing. [Medline: 30855914]
6. Belkić K. Toward better prevention of physician burnout: insights from individual participant data using the MD-specific Occupational Stressor Index and organizational interventions. Front Public Health 2025;13:1514706. [doi: 10.3389/fpubh.2025.1514706] [Medline: 40177083]
7. Health worker burnout. Office of the Surgeon General, US Department of Health and Human Services. 2024 May 28. URL: https://www.hhs.gov/surgeongeneral/reports-and-publications/health-worker-burnout/index.html [accessed 2025-07-06]
8. Sipos D, Goyal R, Zapata T. Addressing burnout in the healthcare workforce: current realities and mitigation strategies. Lancet Reg Health Eur 2024 Jul;42:100961. [doi: 10.1016/j.lanepe.2024.100961] [Medline: 39070752]
9. Andhavarapu S, Yardi I, Bzhilyanskaya V, et al. Post-traumatic stress in healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2022 Nov;317:114890. [doi: 10.1016/j.psychres.2022.114890] [Medline: 36260970]
10. National Academies of Sciences, Engineering, and Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being: National Academies Press; 2019. [doi: 10.17226/25521]
11. Empathy: a critical ally in battling physician burnout. American Medical Association. 2019 Feb 5. URL: https://www.ama-assn.org/practice-management/physician-health/empathy-critical-ally-battling-physician-burnout [accessed 2025-07-06]
12. Cairns P, Isham AE, Zachariae R. The association between empathy and burnout in medical students: a systematic review and meta-analysis. BMC Med Educ 2024 Jun 7;24(1):640. [doi: 10.1186/s12909-024-05625-6] [Medline: 38849794]
13. Delgado N, Delgado J, Betancort M, Bonache H, Harris LT. What is the link between different components of empathy and burnout in healthcare professionals? A systematic review and meta-analysis. Psychol Res Behav Manag 2023;16:447-463. [doi: 10.2147/PRBM.S384247] [Medline: 36814637]
14. Wilkinson H, Whittington R, Perry L, Eames C. Examining the relationship between burnout and empathy in healthcare professionals: a systematic review. Burn Res 2017 Sep;6:18-29. [doi: 10.1016/j.burn.2017.06.003] [Medline: 28868237]
15. Nembhard IM, David G, Ezzeddine I, Betts D, Radin J. A systematic review of research on empathy in health care. Health Serv Res 2023 Apr;58(2):250-263. [doi: 10.1111/1475-6773.14016] [Medline: 35765156]
16. Tracey E, Crowe T, Wilson J, Ponnala J, Rodriguez-Hobbs J, Teague P. An introduction to a novel intervention, “This is My Story”, to support interdisciplinary medical teams delivering care to non-communicative patients. J Relig Health 2021 Oct;60(5):3282-3290. [doi: 10.1007/s10943-021-01379-3] [Medline: 34386889]
17. This Is My Story. Johns Hopkins Medicine. 2022 Jun. URL: https://www.hopkinsmedicine.org/news/articles/2022/06/this-is-my-story [accessed 2025-07-06]
18. Tracey E, Wilson J, Im C, Abshire-Saylor M. A brief patient-recorded audio file called TIMS (This Is My Story) improves communication and empathy for healthcare teams in the hospital. J Patient Exp 2024;11:23743735241274015. [doi: 10.1177/23743735241274015] [Medline: 39161418]
19. Wilson J, Tracey E, Ponnala J, Rodriguez-Hobbs J, Crowe T. An ICU expansion of a novel chaplain intervention, “This is My Story,” to support interdisciplinary medical teams delivering care to non-communicative patients in an academic medical center. J Relig Health 2023 Feb;62(1):83-97. [doi: 10.1007/s10943-022-01567-9] [Medline: 35482270]
20. Tracey E, Wilson J, Mathur R, Hager D. Impressions of recording a brief audio file known as a TIMS (This is My Story) file. J Patient Exp 2025;12:23743735251346585. [doi: 10.1177/23743735251346585] [Medline: 40470310]
21. Tracey E, Wilson J, Abshire Saylor M, et al. TIMS: a mixed methods evaluation of the impact of a novel chaplain facilitated recorded interview placed in the medical chart for the medical staff in an ICU during the COVID-19 Pandemic. J Relig Health 2023 Jun;62(3):1532-1545. [doi: 10.1007/s10943-023-01800-z] [Medline: 37014488]

XSL•FO
RenderX

22. Rekesh D, Koluguri NR, Kriman S, et al. Fast conformer with linearly scalable attention for efficient speech recognition. arXiv. Preprint posted online on May 9, 2023. [doi: 10.48550/ARXIV.2305.05084]
23. Xu H, Jia F, Majumdar S, Huang H, Watanabe S, Ginsburg B. Efficient sequence transduction by jointly predicting tokens and durations. arXiv. Preprint posted online on Apr 13, 2023. [doi: 10.48550/ARXIV.2304.06795]
24. NVIDIA/Parakeet TDT 0.6B V2 (En). Hugging Face. 2024. URL: https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2 [accessed 2025-07-06]
25. OpenAI, Hurst A, Lerer A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 22, 2024. [doi: 10.48550/ARXIV.2410.21276]
26. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023 Aug 24;6(1):158. [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
27. Liu Y, Ju S, Wang J. Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences. BMC Med Inform Decis Mak 2024 Mar 14;24(1):75. [doi: 10.1186/s12911-024-02481-8] [Medline: 38486198]
28. Gan RK, Uddin H, Gan AZ, Yew YY, González PA. ChatGPT's performance before and after teaching in mass casualty incident triage. Sci Rep 2023 Nov 21;13(1):20350. [doi: 10.1038/s41598-023-46986-0] [Medline: 37989755]
29. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25(1):e48659. [doi: 10.2196/48659] [Medline: 37606976]
30. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25(1):e48568. [doi: 10.2196/48568] [Medline: 37379067]
31. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
32. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024 Mar 6;15(1):2050. [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
33. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
34. Fraile Navarro D, Coiera E, Hambly TW, et al. Expert evaluation of large language models for clinical dialogue summarization. Sci Rep 2025 Jan 7;15(1):1195. [doi: 10.1038/s41598-024-84850-x] [Medline: 39774141]
35. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
36. Ovsyannikova D, de Mello VO, Inzlicht M. Third-party evaluators perceive AI as more compassionate than expert humans. Commun Psychol 2025 Jan 10;3(1):4. [doi: 10.1038/s44271-024-00182-6] [Medline: 39794410]
37. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
38. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
39. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. J Med Internet Res 2024 Apr 22;26:e54419. [doi: 10.2196/54419] [Medline: 38648636]
40. Rapuris/TIMS_AI_Editing_Pipeline. GitHub. URL: https://github.com/Rapuris/TIMS_AI_Editing_Pipeline [accessed 2025-12-29]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition
BERT: bidirectional encoder representations from transformers
LLM: large language model
TIMS: This Is My Story

XSL•FO
RenderX

Edited by A Coristine; submitted 07.Jul.2025; peer-reviewed by A Kernberg, KH Lin, N Thawinwisan, S Biswas; revised version received 04.Nov.2025; accepted 04.Nov.2025; published 09.Jan.2026.

Please cite as:
Bains B, Rapuri S, Robitaille E, Wang J, Khera A, Gomez C, Reyes E, Perry C, Wilson J, Tracey E
Large Language Model–Enabled Editing of Patient Audio Interviews From "This Is My Story" Conversations: Comparative Study
JMIR Med Inform 2026;14:e80205
URL: https://medinform.jmir.org/2026/1/e80205
doi:10.2196/80205

© Bikram Bains, Sampath Rapuri, Edgar Robitaille, Jonathan Wang, Arnav Khera, Catalina Gomez, Eduardo Reyes, Cole Perry, Jason Wilson, Elizabeth Tracey. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 9.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/6/e18677

(JMIR Med Inform 2026;14:e98661) doi:10.2196/98661

The JMIR Publications Editorial Office is retracting the article:

> *Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process, the integrity of the peer review process, and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. Author Weifeng Fu was contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Fu W. Application of an isolated word speech recognition system in the field of mental health consultation: development and usability study. JMIR Med Inform 2020 Jun 03;8(6):e18677 [FREE Full text] [doi: 10.2196/18677] [Medline: 32384054]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study
JMIR Med Inform 2026;14:e98661
URL: https://medinform.jmir.org/2026/1/e98661
doi:10.2196/98661
PMID:39773461

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/5/e18627/

(JMIR Med Inform 2026;14:e98666) doi:10.2196/98666

The JMIR Publications Editorial Office is retracting the article:

> *Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process and authorship, the integrity of the peer review process, and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. In response to our request for comment, author Yihao Huang indicated that the identified references were cited in a broader methodological context on computational modeling. However, as these references supported statements related to the history and transmission of rabies, the JMIR Publications Editorial Office determined that this response did not fully resolve the concerns regarding the initial use of these references in the published article.

Yihao Huang disagreed with the retraction. Author Mingtao Li did not reply directly and could not be reached regarding the decision to retract the article.

Reference

1. Huang Y, Li M. Application of a mathematical model in determining the spread of the rabies virus: simulation study. JMIR Med Inform 2020 May 27;8(5):e18627 [FREE Full text] [doi: 10.2196/18627] [Medline: 32459185]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study
JMIR Med Inform 2026;14:e98666
URL: https://medinform.jmir.org/2026/1/e98666
doi:10.2196/98666
PMID:40879335

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/5/e18664

(JMIR Med Inform 2026;14:e98667) doi:10.2196/98667

The JMIR Publications Editorial Office is retracting the article:

> *Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process, the integrity of the peer review process, and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. In response to our request for comment, author Yihao Huang indicated that the identified references were cited in a broader methodological context and that multiple references may not have been the most appropriate for statements relating to brucellosis transmission or infection. The JMIR Publications Editorial Office determined that the author's response did not fully resolve the concerns regarding the initial use of these references in the published article.

Yihao Huang disagreed with the decision to retract the article. Author Mingtao Li did not reply directly and could not be reached regarding the decision to retract the article.

Reference

1. Huang Y, Li M. Optimization of precontrol methods and analysis of a dynamic model for brucellosis: model development and validation. JMIR Med Inform 2020 May 27;8(5):e18664 [FREE Full text] [doi: 10.2196/18664] [Medline: 32459180]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation
JMIR Med Inform 2026;14:e98667
URL: https://medinform.jmir.org/2026/1/e98667
doi:10.2196/98667
PMID:

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/5/e18682/

(JMIR Med Inform 2026;14:e98668) doi:10.2196/98668

The JMIR Publications Editorial Office is retracting the article:

> *Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process and the integrity of the peer review process.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. Author Yue Liu was contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Liu Y. Artificial intelligence–based neural network for the diagnosis of diabetes: model development. JMIR Med Inform 2020 May 27;8(5):e18682. [doi: 10.2196/18682]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development
JMIR Med Inform 2026;14:e98668
URL: https://medinform.jmir.org/2026/1/e98668
doi:10.2196/98668
PMID:

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/6/e19202/

(JMIR Med Inform 2026;14:e98669) doi:10.2196/98669

The JMIR Publications Editorial Office is retracting the article:

> *Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. Author Lin Du was contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Du L. Medical emergency resource allocation model in large-scale emergencies based on artificial intelligence: algorithm development. JMIR Med Inform 2020 Jun 25;8(6):e19202 [FREE Full text] [doi: 10.2196/19202] [Medline: 32584262]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development
JMIR Med Inform 2026;14:e98669
URL: https://medinform.jmir.org/2026/1/e98669
doi:10.2196/98669
PMID:

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/7/e19428

(JMIR Med Inform 2026;14:e98670) doi:10.2196/98670

The JMIR Publications Editorial Office is retracting the article:

> *An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process, authorship, and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. The authors were contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Gong L, Zhang X, Li L. An artificial intelligence fusion model for cardiac emergency decision making: application and robustness analysis. JMIR Med Inform 2020 Jul 27;8(7):e19428 [FREE Full text] [doi: 10.2196/19428] [Medline: 32716305]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis
JMIR Med Inform 2026;14:e98670
URL: https://medinform.jmir.org/2026/1/e98670
doi:10.2196/98670
PMID:

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2020/9/e19110/

(JMIR Med Inform 2026;14:e98672) doi:10.2196/98672

The JMIR Publications Editorial Office is retracting the article:

> *Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process and authorship, the source of the animals used in the in vivo experiments, and the source of the study's ethics approval.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. The authors were contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Chen Y, Du J, Sun X, Li Q, Qin M, Xiao Q, et al. Treatment of left ventricular circulation disorder: application of echocardiography information data monitoring. JMIR Med Inform 2020 Sep 16;8(9):e19110 [FREE Full text] [doi: 10.2196/19110] [Medline: 32936076]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring
JMIR Med Inform 2026;14:e98672
URL: https://medinform.jmir.org/2026/1/e98672
doi:10.2196/98672
PMID:17341880

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, ON

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, ON
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2021/4/e19055

(JMIR Med Inform 2026;14:e98673) doi:10.2196/98673

The JMIR Publications Editorial Office is retracting the article:

> *Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process, the integrity of the peer review process, and the relevance of several references.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. The authors were contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Yang Y, Li D. Medical data feature learning based on probability and depth learning mining: model development and validation. JMIR Med Inform 2021 Apr 08;9(4):e19055 [FREE Full text] [doi: 10.2196/19055] [Medline: 33830067]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation
JMIR Med Inform 2026;14:e98673
URL: https://medinform.jmir.org/2026/1/e98673
doi:10.2196/98673
PMID:29875318

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Retraction

Retraction: Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study

JMIR Publications Editorial Office[1]

JMIR Publications, Toronto, Ontario

Corresponding Author:
JMIR Publications
130 Queens Quay East, Suite 1100-1102
Toronto, Ontario
Phone: 1 4165832040
Email: ed-support@jmir.org

Related Article:

Retraction of: https://medinform.jmir.org/2021/5/e21455

(JMIR Med Inform 2026;14:e98674) doi:10.2196/98674

The JMIR Publications Editorial Office is retracting the article:

> *Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study [1]*

This action is being taken due to identified concerns regarding potential manipulation of the submission process and authorship, the integrity of the peer review process, and the source of the study's ethics approval.

The concerns with this publication were identified during an investigation of a series of articles with related concerns. The authors were contacted and offered an opportunity to comment on these findings and the proposed retraction but did not respond to any attempts at communication.

Reference

1. Sheng H, Xu W, Xu B, Song H, Lu D, Ding W, et al. Application of intelligent computer-assisted Taylor 3D external fixation in the treatment of tibiofibular fracture: retrospective case study. JMIR Med Inform 2021 May 14;9(5):e21455 [FREE Full text] [doi: 10.2196/21455] [Medline: 33988516]

Submitted 17.Apr.2026; this is a non–peer-reviewed article;accepted 17.Apr.2026; published 28.Apr.2026.

Please cite as:
JMIR Publications Editorial Office
Retraction: Application of Intelligent Computer-Assisted Taylor 3D External Fixation in the Treatment of Tibiofibular Fracture: Retrospective Case Study
JMIR Med Inform 2026;14:e98674
URL: https://medinform.jmir.org/2026/1/e98674
doi:10.2196/98674
PMID:40027964

© JMIR Publications Editorial Office. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX