Review

Using Large Language Models for Chronic Disease Management Tasks: Scoping Review

Henry Mukalazi Serugunda[1], MSc, MBA; Ouyang Jianquan[2], PhD; Hasifah Kasujja Namatovu[3], PhD; Paul Ssemaluulu[4], PhD; Nasser Kimbugwe[5], PhD; Christopher Garimoi Orach[6], PhD; Peter Waiswa[7], PhD

[1]Department of Information Technology, School of Computing and Informatics Technology, Makerere University, Kampala, Uganda
[2]School of Computer Science and School of Cyberspace, Xiangtan University, Yuhu District, Xiangtan, Hunan, China
[3]Department of Information Systems, School of Computing and Informatics Technology, Makerere University, Kampala, Uganda
[4]Department of Computer Science, Faculty of Computing and Library Science, Kabale University, Kabale, Uganda
[5]Department of Networks, School of Computing and Informatics Technology, Makerere University, Kampala, Uganda
[6]Department of Community Health and Behavioral Sciences, School of Public Health, College of Health Sciences, Makerere University, Kampala, Uganda
[7]Department of Health Policy Planning and Management, School of Public Health, College of Health Sciences, Makerere University, Kampala, Uganda

Corresponding Author:

Ouyang Jianquan, PhD
School of Computer Science and School of Cyberspace
Xiangtan University
Engineering Building, 2nd Floor
Yuhu District, Xiangtan, Hunan 411105
China
Phone: 86 73158292718 ext 186
Email: oyjq@xtu.edu.cn

Abstract

Background: Chronic diseases present significant challenges in health care, requiring effective management to reduce morbidity and mortality. While digital technologies like wearable devices and mobile applications have been widely adopted, large language models (LLMs) such as ChatGPT are emerging as promising technologies with the potential to enhance chronic disease management. However, the scope of their current applications in chronic disease management and associated challenges remains underexplored.

Objective: This scoping review investigates LLM applications in chronic disease management, identifies challenges, and proposes actionable recommendations.

Methods: A systematic search for English-language primary studies on LLM use in chronic disease management was conducted across PubMed, IEEE Xplore, Scopus, and Google Scholar to identify articles published between January 1, 2023, and January 15, 2025. Of the 605 screened records, 29 studies met the inclusion criteria. Data on study objectives, LLMs used, health care settings, study designs, users, disease management tasks, and challenges were extracted and thematically analyzed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews guidelines.

Results: LLMs were primarily used for patient-centered tasks, including patient education and information provision (18/29, 62%) of studies, diagnosis and treatment (6/29, 21%), self-management and disease monitoring (8/29, 28%), and emotional support and therapeutic conversations (4/29, 14%). Practitioner-centered tasks included clinical decision support (8/29, 28%) and medical predictions (6/29, 21%). Challenges identified include inaccurate and inconsistent LLM responses (18/29, 62%), limited datasets (6/29, 21%), computational and technical (6/29, 21%), usability and accessibility (9/29, 31%), LLM evaluation (5/29, 17%), and legal, ethical, privacy, and regulatory (10/29, 35%). While models like ChatGPT, Llama, and Bard demonstrated use in diabetes management and mental health support, performance issues were evident across studies and use cases.

Conclusions: LLMs show promising potential for enhancing chronic disease management across patient and practitioner-centered tasks. However, challenges related to accuracy, data scarcity, usability, and ethical concerns must be addressed to ensure patient safety and equitable use. Future studies should prioritize the integration of LLMs with low-resource platforms,

wearable and mobile technologies, developing culturally and age-appropriate interfaces, and establishing robust regulatory and evaluation frameworks to support safe, effective, and inclusive use in health care.

JMIR Med Inform 2025;13:e66905; doi: 10.2196/66905

Keywords: chronic diseases; disease management; artificial intelligence in health care; large language models; natural language processing; NLP; generative pre-trained transformer; GPT

Introduction

Chronic diseases, such as diabetes, heart disease, asthma, lung disease, depression, hypertension, Alzheimer disorder, and cancer, are a significant global burden on health care systems [1-3]. These conditions often lead to long-term health issues and have profound physical, psychological, and social impacts on patients [1,2]. Chronic diseases demand continuous, personalized care, often resource-intensive and difficult to scale [1]. Therefore, disease management, which encompasses screenings, regular check-ups, monitoring, coordination of treatment, medication adherence, lifestyle modifications, and patient education, is crucial for improving patient outcomes, enhancing quality of life, and reducing the overall burden on health care systems [1]. However, the resource-intensive nature of personalized, continuous care often makes it inaccessible to many patients, particularly in underserved populations where limited access to health care professionals and resources creates significant barriers to effective disease management [3].

In recent years, the use of digital technologies such as wearable devices [4], mobile apps [5], and chatbots [6,7] has grown significantly in the management of chronic diseases. These have mainly been used for health care tasks, including patient education, symptom monitoring, and medication management [4-7]. The recent emergence of large language models (LLMs) such as GPT, Palm, Llama, and LaMDA [8-11] has demonstrated growing potential in health care applications. These models have been applied in health care for tasks such as personalized treatment recommendations, medical diagnosis, medical record summarization, and interpretation of clinical data to support clinical decision-making and disease management [12-16].

Chronic disease management requires continuous monitoring, patient education, treatment coordination, and personalized care strategies [1]. Recent advancements in LLMs have introduced new possibilities for improving these tasks. For instance, ChatGPT has been explored in providing personalized health advice, enhancing patient engagement, and supporting symptom monitoring [12,13]. In diabetes management, GPT-based models have been investigated for interpreting continuous glucose monitoring data, providing personalized lifestyle recommendations [17], and assessing individualized risk profiles for complications such as retinopathy [17]. Beyond diabetes, LLMs such as LLaMA and GPT have been investigated for mental health support [18], blood pressure measurement using wearable bio signals [19], management of sickle cell anemia [20], and dissemination of information on inflammatory bowel diseases to patients and health care professionals [21].

Despite these applications, several challenges affect the effectiveness of LLMs in chronic disease management, including inaccurate responses, limited and biased datasets, and ethical concerns [17,22,23]. These issues raise concerns regarding the accuracy, reliability, and clinical applicability of LLM-generated recommendations [24]. Given these challenges, a comprehensive review is essential to assess the current applications, identify key limitations, and propose strategies to enhance the effectiveness and safety of LLMs in chronic disease care. While existing reviews explore LLMs in general health care, this scoping review uniquely focuses on their role in chronic disease management. It synthesizes evidence across patient- and practitioner-centered applications, domain-specific challenges, and provides actionable recommendations. Specifically, this review aims to evaluate the current applications of LLM in chronic disease management tasks, identify key challenges, and provide actionable recommendations to address identified challenges.

Methods

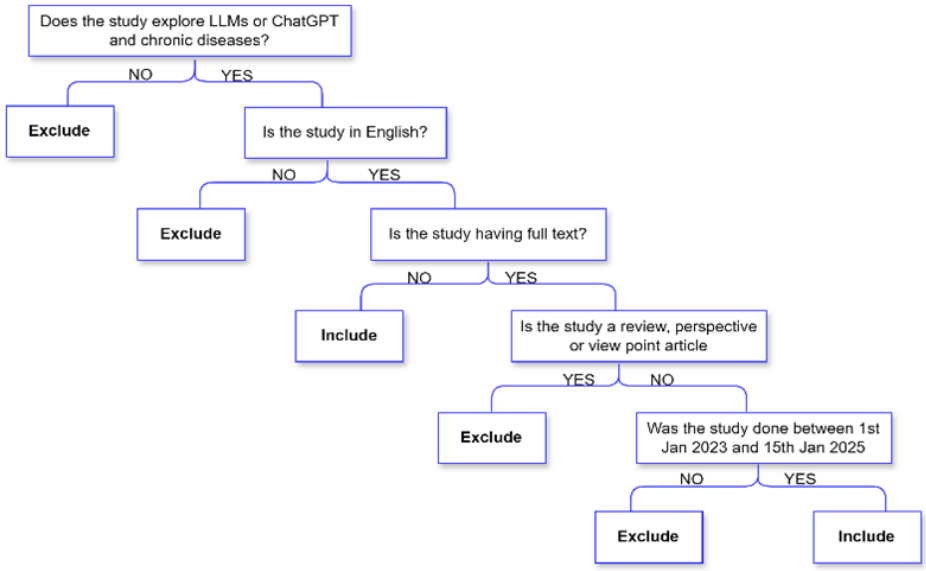
Search Strategy and Information Sources

This scoping review explored the use of LLMs in chronic disease management following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [25]. A comprehensive search was conducted in PubMed, Scopus, IEEE Xplore, and Google Scholar, selected for their coverage of peer-reviewed medical, technical, and AI-related research. Google Scholar was included to capture a broad range of academic publications, including preprints and conference papers that may not be covered by traditional databases. Search terms included combinations of “Large language models,” “LLMs,” “ChatGPT,” “chronic diseases,” and “chronic disease management.” The search targeted both published and unpublished English-language articles from January 1, 2023, to January 15, 2025, ensuring a focus on recent advancements in LLM applications in health care.

Article Selection

Studies were included if they focused on applications of LLM in chronic disease management, provided full-text access, were published in English, and appeared between January 2023 and January 15, 2025. Exclusion criteria eliminated nonprimary research (reviews, editorials, viewpoints, and commentaries), abstracts without full text, non-English publications, and articles outside the date range. To capture emerging research and potentially studies, reputable non-peer–reviewed preprints from established repositories such as arXiv and medRxiv were included. Figure 1 illustrates the eligibility screening process with a decision tree.

Figure 1. Decision tree for assessing article eligibility. The exclusion of certain publication types was necessary to ensure the review focused on primary research and empirical studies that directly address the application of large language models in chronic disease management.



Data Extraction and Synthesis

The data from selected studies were extracted into a structured form that captured study objectives, the LLMs used, health care settings, study designs, disease management tasks, identified challenges, evaluation methods, and target users. These extracted data points were selected to ensure a structured and objective analysis aligned with the study's aims. Study objectives provided insight into the intended applications of LLMs in chronic disease management, while health care settings contextualized their use across clinical and patient-centered environments. Study design and evaluation methods were included to assess methodological rigor and the validity of findings. In addition, data on LLM models, disease management tasks, and key challenges enabled a systematic evaluation of current applications, limitations, and areas for future research.

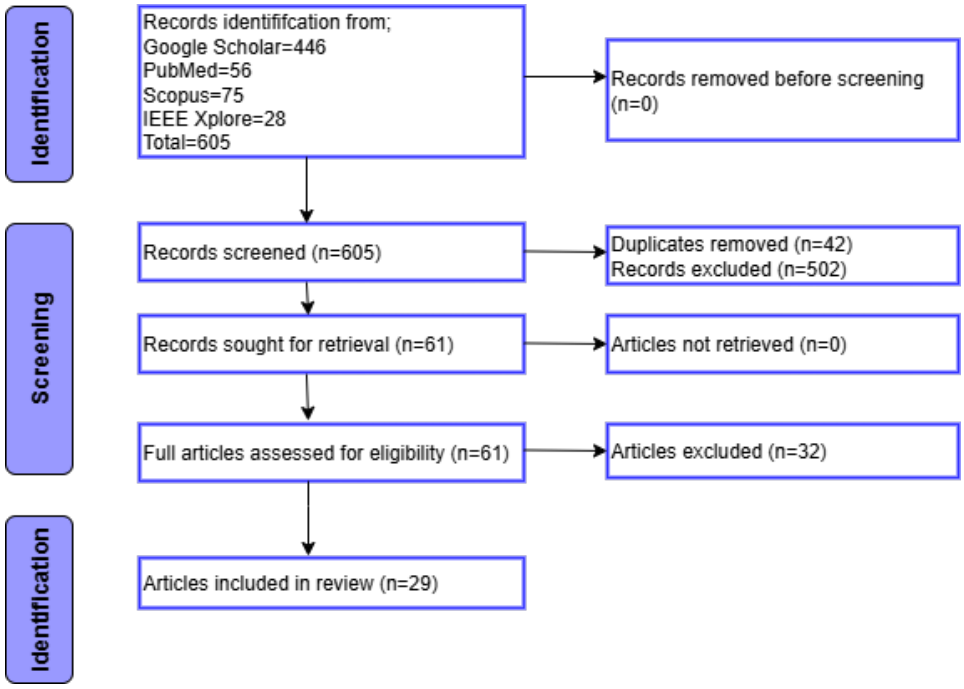
The Mixed Methods Appraisal Tool (MMAT, version 2018) [26] was used to perform a formal methodological quality assessment. The MMAT was selected due to its flexibility in evaluating diverse study designs included in the review. Two reviewers independently appraised each study using the 5 criteria relevant to its methodological design, with discrepancies resolved through discussion and reached consensus on final ratings. Consistent with scoping review methodology, the studies were not excluded based on quality, but results inform the interpretation of findings [25]. The findings from the included studies were synthesized and presented in alignment with the study objectives. A thematic analysis approach was used to categorize qualitative insights, grouping findings into patient-centered tasks, practitioner-centered tasks, and challenges. Discrepancies in data interpretation were resolved through consensus among the reviewers. Reference management and citation generation were conducted using Mendeley.

Results

Included Studies

The PRISMA flow diagram (Figure 2) outlines each stage of the study selection process. The PRISMA-ScR framework was followed to ensure transparency and reproducibility, incorporating detailed search strategies and clearly defined exclusion criteria. A total of 446 records were identified from Google Scholar, 75 from Scopus, 56 from PubMed, and 28 from IEEE Xplore. After removing duplicates, 242 unique records underwent title and abstract screening, resulting in 61 articles for full-text review. Following the application of eligibility criteria, 29 articles were included in the final analysis.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.



Characteristics of Included Articles

Table 1 provides the characteristics of the 29 articles included in this scoping review. The articles were categorized based on publication type, with the majority of the studies being journal articles (n=18), followed by conference papers (n=8), and preprints (n=3). The studies used a variety of research designs, including experimental (n=10), qualitative research (n=3), comparative studies (n=4), cross-sectional study (n=2), case study (n=1), observational (n=3), retrospective cohort designs (n=3), mixed methods (n=1), pilot study (n=1), and prospective study (n=1). Table 1 provides more details on the characteristics of the reviewed studies.

Table 1. Study characteristics (n=29).

Study	Country	Article category	Study objective	Health care setting	Study design	Evaluation
Montagna et al [6]	Italy	Conference paper	To design and implement a system architecture for a chatbot-based home blood pressure monitoring solution.	Home care setting	Experimental (system design and prototype development)	Human evaluation
Yang et al [16]	China	Preprint	To explore the application of a fine-tuned model-based outpatient treatment support system for treating patients with diabetes and to evaluate its effectiveness and potential value.	Clinical (West China Hospital) home care	Experimental (fine-tuning)	Human evaluation and automated evaluation metrics
Raghu et al [17]	India	Journal article	To evaluate the ability of ChatGPT to predict the diabetic retinopathy risk.	—[a]	Comparative study	Human evaluation
Song et al [18]	The Republic of Korea	Preprint	To investigate the experiences of individuals using LLM[b] chatbots for mental health support.	Korea Advanced Institute of Science and Technology	Qualitative study	Human evaluation
Liu et al [19]	China	Conference paper	Explore the use of LLMs for cuffless blood pressure measurement using wearable bio signals	Home care settings	Experimental (cuffless blood pressure measurement using LLMs)	Human evaluation

Study	Country	Article category	Study objective	Health care setting	Study design	Evaluation
Ogundare et al [20]	Nigeria	Conference paper	To investigate the potential of LLMs in ambulatory devices for sickle cell anemia management.	Home care setting	Case study	Automated evaluation metrics
Cankurtaran et al [21]	Turkey	Journal article	To evaluate the performance of ChatGPT within the context of inflammatory bowel disease.	—	Cross-sectional study	Human evaluation
Wang et al [22]	China	Conference paper	To enhance the diagnosis and treatment of depression.	Clinical and homecare settings	Experimental (pre-training and fine-tuning)	Human evaluation
Abdullahi et al [23]	Germany	Journal article	To explore the potential of three popular Large Language Models in medical education to enhance the diagnosis of rare and complex diseases.	Home care setting	Qualitative study	Human evaluation and automated evaluation metrics
Al Anezi [24]	Saudi Arabia	Journal article	To analyze the use of ChatGPT as a virtual health coach for chronic disease management.	Home care setting	Quasi-experimental design	Human evaluation
Athavale et al [27]	United States	Journal article	To assess whether chatbots could assist with answering patient questions and electronic health record inbox management	Clinical (Division of Vascular Surgery, Stanford University School of Medicine in Palo Alto)	Experimental (chatbot assistance in chronic venous disease management)	Human evaluation
Soto-Chávez et al [28]	Colombia	Journal article	To evaluate the reliability and readability of Spanish chronic disease information presented to ChatGPT	—	Cross-sectional study	Human evaluation
Abbas et al [29]	Pakistan	Journal article	To assess the predictive accuracy of ChatGPT-assisted machine learning models for various chronic diseases.	Clinical (Tertiary hospital)	Observational study	Automated evaluation metrics
Anderson et al [30]	United States	Conference paper	The study aims to discover and rank novel relationships between various aspects of this condition.	—	Experimental	Human evaluation and automated evaluation metrics
Ding et al [31]	Taiwan	Conference paper	To develop and evaluate Large Language Multimodal Models that integrate clinical notes and laboratory test results for predicting the risk of chronic diseases, particularly type 2 diabetes mellitus	Clinical (Eastern Memorial Hospital in Taiwan	Retrospective cohort study	Automated evaluation metrics
Jairoun et al [32]	Malaysia	Journal article	To investigate the benefits and risks associated with the application of ChatGPT in managing diabetes and metabolic illnesses	—	Qualitative study	Human evaluation
Mondal and Naskar [33]	India	Journal article	To evaluate GPT-4's competency in reviewing diabetic patient management plans	General medical setting	Comparative study	Human evaluation and automated evaluation metrics

Study	Country	Article category	Study objective	Health care setting	Study design	Evaluation
			compared to expert reviews.			
Liu et al [34]	N/S[a]	Journal article	To leverage LLMs and multi-prompt engineering for chronic disease management, specifically for detecting mental disorders through user-generated textual content.	Online platforms	Experimental (few-shot learning)	Automated evaluation metrics
Liao et al [35]	Taiwan	Conference paper	To develop an EHR-based chronic disease prediction platform using LLMs for diabetes, heart disease, and hypertension.	Clinical (Far Eastern Memorial Hospital, Taiwan)	Retrospective cohort study	Automated evaluation metrics
Ding et al [36]	Taiwan	Journal article	Predict new-onset type 2 diabetes using large language multimodal models with EHR data	Clinical (Far Eastern Memorial Hospital, Taiwan)	Retrospective cohort study	Automated evaluation metrics
Dao et al [37]	Ireland and Singapore	Conference paper	Design and evaluate an AI chatbot system using GPT-3.5 for proactive diabetes prevention	Community-based setting	Experimental (AI design and evaluation)	Automated evaluation metrics
Khan [38]	United States	Journal article	Assess the efficacy of ChatGPT in facilitating self-management strategies for diabetic patients.	Outpatient diabetes care	Observational study	Human evaluation
Mondal et al [39]	India	Journal article	To evaluate the effectiveness of ChatGPT, an LLM, in providing answers to queries related to lifestyle-related diseases or disorders	Clinical and academic settings	Observational study	Human evaluation
Young et al [40]	United States	Journal article	Assess LLMs' capacity to deliver age-appropriate explanations of chronic pediatric conditions to enhance patient understanding.	Clinical (Boston Children's Hospital)	Pilot study	Human evaluation
Li et al [41]	China	Journal article	Develop DeepDR-LLM, an integrated AI system for primary diabetes care and diabetic retinopathy (DR) screening	Low-resource primary care settings	Experimental	Human evaluation and automated evaluation metrics
Ying et al [42]	China	Preprint	To evaluate the feasibility and utility of ChatGPT in diabetes education using retrospective and real-world patient questions.	Outpatient setting	Mixed methods	Human evaluation
Li et al [43]	China	Journal article	To evaluate the performance of LLMs in diabetes-related queries and their potential to assist in diabetes training for primary care physicians	Primary diabetes care, endocrinology, and diabetes management.	Prospective study	Human evaluation and automated evaluation metrics
Hussain and Grundy [44]	Australia	Journal article	Evaluate the responses of ChatGPT models to queries from diabetes patients, assessing their accuracy, biases, and	Home care setting	Comparative study	Human evaluation

Study	Country	Article category	Study objective	Health care setting	Study design	Evaluation
			limitations in providing self-management advice.			
Wang et al [45]	China	Journal article	To evaluate the potential of the RISE framework to improve LLMs' performance in accurately and safely responding to diabetes-related inquiries.	Home care setting	Comparative study	Human evaluation

[a]Not available.
[b] LLM: large language model.

As shown in Table 1, most studies were conducted in China (7/29, 24%) and the United States (4/29, 14%), with limited representation of low-resource settings. Experimental designs were predominant (10/29, 35%), and nearly half of the studies (14/29, 48%) focused on diabetes management. The studies primarily focused on adult populations (28/29, 96%), with only 1 study (1/29, 4%) specifically addressing pediatric applications. Human evaluation was the most common evaluation method (16/29, 55%), followed by automated evaluation metrics (10/29, 35%) used in prototype evaluation, with some studies using both approaches (3/29, 10%). Studies were carried out in diverse health care settings, with home care settings (10/29, 35%) and clinical settings (9/29, 31%) being the most common. This diversity in study characteristics reflects the broad application of LLMs across various health care contexts for chronic disease management.

Methodological Quality Assessment

Using the Mixed Methods Appraisal Tool [26], the studies were categorized as quantitative descriptive studies (19/29, 66%), comprizing observational, cross-sectional, case study, system development, prototype evaluation, and comparative analyses. Quantitative nonrandomized studies (3/29, 10%) included retrospective cohort and quasi-experimental designs, qualitative studies (5/29, 17%), and 2 studies with mixed methods design (2/29, 7%). Of the 29 studies, 18 studies (62.1%) were classified as high quality (meeting ≥4 of 5 criteria), 9 studies (31.1%) as moderate quality (meeting 2-3 criteria), and 2 studies (6.9%) as low quality (meeting ≤1 criterion). The most common methodological limitations identified across studies included inadequate sampling strategy descriptions, limited participant demographic reporting, use of synthetic clinical data, and lack of external validation. Detailed quality appraisal results for each study are provided in Multimedia Appendix 1. Table 2 provides an overview of the LLM types, users, tasks, and challenges identified across the included studies.

Table 2. Large language model tasks and challenges (N=29).

Study	LLM[a]	Users	Disease management tasks	Challenges
Montagna et al [6]	GPT-3	Individuals with hypertension	Patient engagement blood pressure monitoring and management.	• The incorrect way patients measure their blood pressure
Yang et al [16]	ChatGLM-6B	Patients diagnosed with diabetes	Treatment recommendations, suggesting appropriate laboratory tests, and medication	• Inadequate understanding of complex medical records • The small size of the training data
Raghu et al [17]	ChatGPT	Practitioner (ophthalmologist)	Patient education, medical reports: diagnoses and predictions	• Incorrect information, privacy and protection of patient data
Song et al [18]	ChatGPT Llama	Individuals who have used LLM chatbots for mental health support	• Providing emotional support, engaging in therapeutic conversations, and offering recommendations tailored to individual contexts. • Addressing specific stressors or challenges faced by individuals.	• Cultural Misalignments: Participants noted that recommendations from LLM chatbots often felt like they were translated from stereotypical American responses. • Linguistic Biases: Participants often felt compelled to use English when interacting with LLM chatbots.

Study	LLM[a]	Users	Disease management tasks	Challenges
Liu et al [19]	Gemma-7B, Mistral-7B, Yi-6B, MedAlpaca-7B, LLaMA2-7B, LLaMA3-8B, Qwen2-7B, PalmyraMed-20B, PMCLLaMA13B, OpenBioLLM-8B	Patients with hypertension	Cardiovascular disease management via cuffless blood pressure measurement.	• Therapeutic Misalignment. • Dataset imbalances diminish accuracy • Privacy concerns in real-world deployment • Need for calibration to mitigate individual variability
Ogundare et al [20]	Unspecified	Sickle cell patients and clinicians	Assessing anemia severity in real-time, predicting time to vaso-occlusive episodes, and communicating with emergency personnel.	• Creation of a reliable non-invasive tool for angiogenic level assessment, development of a biophysics model, and practical considerations of LLM communication with emergency personnel
Cankurtaran et al [21]	ChatGPT	Inflammatory bowel disease patients and health care professionals.	• Tailored responses, educational resources • monitoring and follow-up, patient empowerment • Decision support	• Insufficient responses • Limited scope of knowledge (up-to-date information)
Wang et al [22]	LLaMA-7B, ChatGLM-6B Alpaca.	Individuals with depression.	Diagnosis and treatment of depression	• Absence of pretraining data sets on depression • Hallucination problem • Evaluation methodologies emphasize predictive performance and lack • Quantification of the impact on patient treatment
Abdullahi et al [23]	Bard ChatGPT 3.5 and GPT-4	• Physicians • Medical Students Resident Nurses • Nurse Practitioners	• Clinical Decision Support • Medical Education • Disease Diagnosis	• Inconsistency in Responses, LLMs do not always explicitly indicate their level of uncertainty due to Limited Scope, Sample Size, and knowledge. ChatGPT-3.5 and GPT-4 were limited to health care data available up to 2021 • LLMs may generate different responses for the same prompt
Al Anezi [24]	ChatGPT	Outpatients with chronic diseases	• Providing information about patient conditions, treatment plans, and medication schedules. Reminders for medication intake, appointments, or lifestyle adjustments. • Assisting in behavior change efforts by providing evidence-based strategies, personalized goal-setting techniques, and	• Limited physical examination, Lack of human connection and empathy • Complexity of individual cases • Privacy and security concerns • Legal and ethical challenges, language and cultural barriers, technical limitations, diagnostic limitations, and lack of reliability and trust

Study	LLM[a]	Users	Disease management tasks	Challenges
			reminders for healthy habits. • Identifying barriers to behavior change and exploring solutions to overcome them. • Monitoring blood pressure, blood glucose levels, or weight and providing feedback based on shared data.	• Ineffectiveness in emergencies
Athavale et al [27]	ChatGPT 4.0	Patient	Answered administrative and non-complex medical questions well, and electronic health record inbox management. Answering complex medical questions	• Hallucinations • Need for extensive supervised training by subject experts • No regulatory approval
Soto-Chávez et al [28]	ChatGPT	Patients with chronic diseases using the Spanish language	Evaluating the reliability and readability of ChatGPT-generated patient information on chronic diseases in Spanish.	• ChatGPT was trained in English, which affects the accuracy of responses in Spanish • Lower reliability on chronic diseases like heart failure and chronic kidney disease
Abbas et al [29]	GPT-3.5	Machine Learning engineers, clinical researchers	Chronic disease prediction	• lack of longitudinal data • limited generalizability
Anderson et al [30]	GPT (Generative Pre-trained Transformer)	Practitioners	Discover and rank novel relationships between various aspects of chronic lower back pain.	• The GPT-based approach took around half an hour to process approximately 500 pairs, making it computationally intensive. • Achieving strong agreement among human evaluators
Ding et al [31]	MedAlpaca	• Patients with early diabetes • Patients with multiclass chronic diseases	Early prediction of diabetes Prediction of multiclass chronic diseases	• Lower positive rates when using only laboratory blood values • Missing tests for most patients when using only laboratory blood values. • Integrating multimodal data from clinical notes and laboratory test results • Difficulty in model explainability for early disease prediction
Jairoun et al [32]	ChatGPT	diabetes and metabolic illnesses, endocrinologists and diabetologists	Patient support and education Tailored treatment	• Diagnostic mistakes • Patient data security and privacy • Limitations on generalizability • Integration difficulties and workflow errors, and Compliance with laws and regulations • Absence of empathy and human contact

Study	LLM[a]	Users	Disease management tasks	Challenges
Mondal et al [33]	GPT-4	Health care professionals	Reviewing and evaluating diabetes management plans for guideline adherence	• GPT-4's difficulties in handling complex clinical judgments, such as medication adjustments and treatment modifications in varied clinical scenarios
Liu et al [34]	GPT-2 and T5	Patients with mental disorders	Detection of mental disorders (depression, anorexia, pathological gambling, self-harm) through user-generated textual content.	• Need for personalized prompts to capture individual user characteristics. • Integration of medical knowledge into prompts for accurate detection. • Handling noisy and lengthy user-generated content. Few-shot learning with minimal labeled data
Liao et al [35]	BERT BiomedBERT Flan-T5-large-770M GPT-2	Physicians, health care providers	Prediction of chronic diseases (diabetes, heart disease, hypertension) using EHR data.	• Difficulty in classifying diseases with lower positive rates (eg, hypertension) • Need for interpretability in model predictions • Integration of multimodal data (clinical notes and blood test results) for accurate predictions
Ding et al [36]	BERT, Roberta, BiomedBERT, Flan-T5, GPT-2	Researchers, health care professionals	Predict new-onset T2DM, early detection, and risk assessment	• Handling multimodal data, missing values, and model interpretability
Dao et al [37]	GPT-3.5	Individuals at risk of diabetes or with prediabetes	Instant Q&A and advice Personalized reminders Data analysis for tailored guidance Health resource aggregation Emotional support	• Engagement barriers in prevention programs (eg, transportation, personal responsibilities) • Lack of research on AI in diabetes prevention • Need for reliable, context-aware AI responses
Khan [38]	ChatGPT	Diabetic patients	Real-time education and support Blood glucose monitoring guidance Medication adherence advice Lifestyle/diet recommendations Emergency detection	• Inaccuracies in medical information (eg, insulin storage guidelines, trial data mix-ups) • Lack of emotional support/ empathy • Limited to pre-2021 knowledge • Difficulty distinguishing medical terminologies • Low adoption among older adults
Mondal et al [39]	ChatGPT-4	Patients and health care professionals	Answering patient questions (causes, symptoms, treatment, diet) Providing information on managing Crohn's disease (CD) and ulcerative colitis (UC). Addressing professional queries (classification, diagnosis,	• Insufficient elaboration on medical agents and surgical indications • Inadequate information for patients. ChatGPT provided different answers to the

Study	LLM[a]	Users	Disease management tasks	Challenges
			disease activity, prognostic markers, complications)	same question across sessions • Lower reliability/usefulness scores for patient-directed questions compared to professional-focused ones. • Outdated information
Young et al [40]	GPT-4 Gemini 1.0 Ultra	Pediatric patients, health care providers, and caregivers	Generating explanations for chronic conditions	• Age-appropriateness discrepancies between models (GPT-4 versus Gemini) • Lack of direct feedback from pediatric patients; reliance on clinician evaluations.
li et al [41]	LLaMA	Primary care physicians	Individualized diabetes management recommendations	• Underdiagnosis and poor primary diabetes management
Ying et al [42]	GPT-3.5	Physicians, laypersons, and type 2 diabetes patients	Diabetes education and personalized Q&A support	• Lower real-world performance, variability by prompt, trust, and safety concerns
Li et al [43]	ChatGPT-3.5 ChatGPT-4.0 Google Bard MedGPT LlaMA2-7B	Researchers Primary care physicians	Answering diabetes-related exam questions and assisting in diabetes training.	• Poor performance in both Chinese and English diabetes-related questions • LLMs may provide misleading explanations and difficulty with multiple-choice and case analysis questions
Hussain and Grundy [44]	ChatGPT-3.5 ChatGPT-4	Diabetes patients and health care providers	Patient education, treatment recommendations, insulin management, dietary advice	• Inaccuracies in medical advice • Lack of personalization • Failure to recognize regional variations • Incorrect assumptions about blood glucose units limitations in addressing complex patient histories
Wang et al [45]	GPT-4 Anthropic Claude 2 Google Bard	• Clinicians • Diabetes patients	Responding to diabetes-related inquiries and providing accurate and comprehensive information for diabetes self-management.	• Lack of specialized medical knowledge in commercially available LLMs • Susceptibility to generating inaccurate or misleading information • Need for real-time, domain-specific knowledge to improve accuracy and reliability • Ensuring responses are safe, accurate, and understandable for patients

[a] LLM: large language model.

As shown in Table 2, GPT models were the most commonly used (14/29, 48%), followed by LLaMA variants (5/29, 17%), the Bard model (3/29, 10%), and BERT-based models (2/29, 7%). LLMs were primarily used for patient education and information provision (18/29, 62%), with most studies targeting patients (18/29, 62%) rather than health

care providers (11/29, 38%). Inaccurate and inconsistencies in responses (18/29, 62 %) were the most frequently reported challenge across studies.

Objective 1: The Tasks in Chronic Disease Management Performed by LLMs

Our literature synthesis revealed that LLMs have significant potential to improve various chronic disease management tasks. The tasks identified have been broadly categorized into patient-centered tasks and practitioner-centered tasks.

Patient-Centered Tasks

Patient Education and Information Provision

Eighteen studies (n=18) delved into the use of LLMs in providing health information to enhance patient health literacy [16,17,21,23,24,27,28,32,33,37-41]. Key applications included using ChatGPT to provide personalized guidance on diabetes management [16,38,39,42], educational content for diabetic retinopathy and inflammatory bowel disease patients [17,21], generating age-appropriate explanations of chronic pediatric conditions [40], and supporting physician training in diabetes management [43]. In addition, LLMs supported treatment adherence through tools like ChatGPT and GPT-3.5, offering tailored medication reminders, appointment scheduling, and strategies for behavior change [24,37]. For Spanish-speaking populations, ChatGPT was evaluated for reliability and readability of chronic disease information [28], while multiple ChatGPT versions were assessed for regional variations in diabetes education quality [44].

Diagnosis and Treatment

Six studies (n=6) examined the role of LLMs in assisting with diagnosis and treatment recommendations. These studies explored the potential of LLMs to suggest appropriate laboratory tests, generating differential diagnoses and medication options tailored to the individual patient's condition [16,22,32,34]. Notable applications included enhancing depression diagnosis and treatment through fine-tuning of models like LLaMA-7B and ChatGLM-6B [22], supporting the diagnosis of rare and complex diseases [23], and detecting mental disorder patterns through analysis of user-generated content [34]. Furthermore, integrating AI-driven diagnostic and treatment capabilities with diabetes management systems showed particular promise in low-resource primary care settings [41].

Self-Management and Disease Monitoring

Eight studies (n=8) addressed using LLMs for self-management and disease monitoring. These studies explored how LLMs provide guidance on managing chronic conditions, promote patient engagement, and support home disease monitoring [6,18,20,21,24,37,38,42]. Key applications included developing chatbot architectures for home blood pressure monitoring [6], creating cuffless blood pressure measurement systems using wearable biosignals [19], and assessing real-time disease severity in sickle cell anemia [20]. LLMs also demonstrated value in detecting emergencies such as hypoglycemic episodes in diabetic patients and guiding appropriate actions [38]. Additional applications encompassed monitoring tools for inflammatory bowel disease management [21], personalized reminders for diabetes prevention [37], and comprehensive health parameter tracking with feedback based on patient-shared data [24]. These implementations highlight the potential of LLMs to enhance patient self-management through continuous monitoring and timely intervention guidance.

Emotional Support and Therapeutic Conversations

Four studies (n=4) explored the role of LLMs in providing emotional support and engaging in therapeutic conversations for patients managing chronic diseases. The review identified several key applications, including investigating LLM chatbots for mental health support with tailored recommendations addressing specific stressors [18], evaluating ChatGPT as a virtual health coach identifying barriers to behavior change [24], assessing GPT-3.5's emotional support capabilities in proactive diabetes prevention [37], and examining ChatGPT's ability to provide coping strategies for diabetic patients [38].

Practitioner-Centered Tasks

Clinical Decision Support

Eight studies (n=8) investigated the use of LLMs for clinical decision support. The review identified several key applications, including generating personalized medical reports with treatment options and diagnostic procedures for conditions like diabetic retinopathy [17] and inflammatory bowel disease [21], assessing LLMs' diagnostic accuracy compared with human experts in rare and complex diseases [23], and exploring potential use for electronic health record inbox management [27]. Other applications included using GPT to discover and rank novel relationships between aspects of chronic lower back pain [30], evaluating diabetes management plans using GPT-4 [33], disease classification and prognosis [39], and evaluating LLMs' competency in answering diabetes-related exam questions for physician training [43].

Medical Predictions

Six studies (n=6) explored the predictive capabilities of LLMs in chronic disease management. The review identified several key applications, including predicting diabetic retinopathy risk [17], developing ChatGPT-assisted machine learning models for chronic disease classification [29], and integrating multimodal data from electronic health records and laboratory tests to predict new-onset type 2 diabetes [31,36]. Additional applications included creating an EHR-based prediction platform for diabetes, heart disease, and hypertension [35] and implementing integrated AI systems for diabetes risk assessment in primary care settings [41].

LLMs in chronic disease management are predominantly utilized for patient education and information provision, accounting for (18/29) of reported applications.

Self-management and disease monitoring and clinical decision support each account for 28% (8/29) of applications. Diagnosis and treatment tasks, along with medical predictions, both constitute 21% (6/29) of applications, while emotional support and therapeutic conversations account for 14% (4/29). Percentages exceed 100% due to thematic overlaps where individual studies addressed multiple tasks.

LLMs in chronic disease management are predominantly utilized for patient education and information provision, accounting for (18/29) of reported applications. Self-management and disease monitoring and clinical decision support each account for 28% (8/29) of applications. Diagnosis and treatment tasks, along with medical predictions, both constitute 21% (6/29) of applications, while emotional support and therapeutic conversations account for 14% (4/29). Percentages exceed 100% due to thematic overlaps where individual studies addressed multiple tasks.

Objective 2: Challenges Associated With Using LLMs for Chronic Disease Management

The challenges identified and presented in Table 2 were categorized as follows.

Inaccurate and Inconsistencies in Responses

Eighteen studies (n=18) highlighted issues with hallucinations, diagnostic errors, and unreliable outputs [6,16,17,21-23,27-29,32-34,38,39,42-45]. The review identified several key challenges, including hallucinations where models like ChatGPT and LLaMA generate reasonable but factually incorrect information [22,45], diagnostic errors in conditions ranging from depression to inflammatory bowel disease [21,32,44], and inconsistent responses to identical prompts without indicating uncertainty levels [23,39,42]. In addition, models demonstrated limited understanding of complex medical records [16], struggled with regional variations in medical practice [44], provided insufficient elaboration on medical treatments [39], and showed difficulty distinguishing medical terminologies [38]. Further challenges included lower reliability in the Spanish language for specific chronic conditions like heart failure and chronic kidney disease [28] as well as limited generalizability due to restricted training populations [29,32]. These inaccuracies stem from erroneous input data, such as incomplete or incorrect test results [6,21,22,27,31].

Limited Datasets and Knowledge

Six studies (n=6) identified challenges related to limited datasets and knowledge cutoffs in LLM applications for chronic disease management. The review highlighted several key limitations, including scarcity of disease-specific datasets [16,22], dataset imbalances affecting predictions for conditions like hypertension [19], and knowledge limitations that restrict LLM awareness of the current medical guidelines [21,23,39].

Computational and Technical Challenges

Six studies (n=6) highlight significant computational and technical challenges in deploying LLMs for chronic disease management. The review identified several key limitations, including resource-intensive processing that results in prolonged training time in resource-constrained environments [30]. In addition, technical challenges include integrating multimodal data from clinical notes and laboratory results [31,36], ensuring model explainability for early disease prediction [35], and handling noisy user-generated content in mental health applications [31,34]. Further challenges involve difficulties in integrating LLMs into clinical workflows [32] and managing complex clinical judgments, such as medication adjustments and treatment modifications [33].

Usability and Accessibility Concerns

Nine studies (n=9) identified usability and accessibility concerns surrounding LLMs in chronic disease management tasks. Notably, the restriction to textual inputs limits use for tasks involving multimodal diagnostic tasks [28,40] language and cultural misalignments [18,28], while age-inappropriate outputs pose challenges for pediatric care [40]. In addition, poor interpretability of model predictions for early disease prediction and risk assessment [31,35,36]. Additional challenges included a lack of empathy and ineffectiveness in emergencies [24,29], digital literacy gaps restricting adoption among older adults [29,38]. Furthermore, these studies also noted how insufficient transparency in model decision-making processes hindered trust and clinical acceptance [35,36].

LLM Evaluation

Five studies (n=5) noted challenges involving LLM evaluation. Notable challenges identified included that automated evaluation metrics primarily focus on predictive performance and fail to assess the impact on patient treatment outcomes [22], difficulties in achieving consensus among human evaluators when assessing LLM outputs [30], discrepancies between model performance in test environments versus real-world applications [42], difficulties in consistently evaluating language models across different diabetes-related tasks [43], and significant variations in age-appropriateness scoring between different LLM platforms [40].

Legal, Ethical, Privacy, and Regulatory Concerns

Ten studies (n=10) identified legal, ethical, privacy, and regulatory challenges of using LLMs in chronic disease management. The review highlighted several critical concerns, including privacy and data security vulnerabilities [17,19,24,27,28,32], absence of regulatory approval and standardized guidelines [27], and compliance issues with health care laws across different jurisdictions [27,32]. In addition, language and cultural barriers posed additional challenges, particularly for non-English speakers [18,28], while bias and equity issues stemming from limited training data diversity raised concerns about health care disparities [24,32,42]. Studies also noted ethical challenges around accountability for errors [24], lack of transparency

in decision-making [22,27], and limitations in addressing complex ethical dilemmas in clinical care [44,45].

The most prevalent challenge identified was inaccurate and inconsistent responses, reported in 62% (18/29) of studies. Legal, ethical, privacy, and regulatory concerns followed, appearing in 35% (10/29) of studies. Usability and accessibility issues were noted in 31% (9/29) of studies. Computational and technical limitations, as well as dataset and knowledge constraints, were each reported in 21% (6/29) of studies. Additionally, 17% (5/29) of studies highlighted limitations in evaluation methodologies.

Discussion

Principal Findings

This scoping review presents 3 significant findings from the 29 included studies (n=29): (1) LLMs are mostly used for both patient-centered (18/29, 62%) and practitioner-centered (11/29, 38%) tasks, with patient education and information provision emerging as the major application (18/29, 62%); (2) despite promising applications, significant challenges still exist, particularly regarding LLM response accuracy (18/29, 62% of studies), ethical concerns (10/29, 35%), and usability issues (9/29, 31%); and (3) methodological quality varies considerably across studies, with journal articles demonstrating higher quality (13/18, 72%) compared with conference papers (3/8, 38%) and preprints (1/3, 33%). These findings highlight both the considerable promise and significant limitations of current LLM applications in chronic disease management, which are examined in detail below.

Chronic Disease Management Tasks

Chronic disease management is an approach to managing chronic illnesses involving screenings, regular check-ups, monitoring, coordination of treatment, medication adherence, lifestyle modifications, and patient education [1]. The findings from this scoping review reveal an increasing interest in leveraging LLMs like ChatGPT to support both patient-centered and practitioner-centered tasks [6,16,22-24,30,31].

Patient-Centered Tasks

The majority of the studies (18/29, 62%) focused on patient-centered tasks, reflecting the emphasis on patient active engagement in chronic disease management [3,46]. Patients' active engagement enables them to monitor their symptoms, disease progress, weight, and adverse drug effects, and adhere to medication and visits [3,46]. This review found that LLMs support various patient-centered applications, including patient education and information provision, disease monitoring and self-management, emotional support and therapeutic conversations, and diagnosis and treatment assistance.

Patient education and information provision emerged as the most prominent application (18/29, 62%), with LLMs providing health information about conditions, treatment plans, and medication schedules [16,17,21,23,24,27,28,32,33,37-41]. With the right health information, individuals with chronic diseases can easily self-manage their conditions. Diagnosis and treatment applications accounted for (6/29, 21%). Studies experimented using LLMs to suggest laboratory tests, for diagnosis, and for medication generation tailored to individual patient conditions [16,22,23,32,34,41]. However, using LLMs for diagnostic and treatment has been criticized due to concerns about hallucinations and misinterpretations of clinical guidelines [22,23], highlighting the need for continued research to ensure patient safety. Therefore, LLM usage should not replace health practitioners but instead serve as complementary tools.

Self-management and disease monitoring applications (8/29, 28%) demonstrated how LLMs can facilitate home-based monitoring of various physiological parameters [6,19-21,24,37,38,42]. Recent studies have highlighted that patient engagement with health monitoring technologies is crucial for improving health outcomes in chronic disease management [47]. Studies have also shown that wearable technologies integrated with LLMs provide real-time patient-centered health data that can better inform self-management decision-making [48]. However, ensuring consistent long-term engagement is still a challenge.

Emotional support and therapeutic conversations (4/29, 14%) represented an emerging application area [18,24,37,38]. Studies showed that LLMs can provide psychological support through tailored recommendations addressing specific stressors [18], identifying barriers to behavior change [24], and offering coping strategies for patients with diabetes [38]. Emotional support is increasingly recognized as essential in chronic disease management [49], which helps patients overcome psychological barriers to treatment adherence and lifestyle modifications.

Practitioner-Centered Tasks

Practitioner-centered tasks (11/29, 38%) mainly revolved around clinical decision support and medical predictions. Clinical decision support applications (8/29, 28%) provided health care practitioners with actionable information to enhance decision-making. Studies demonstrated that LLMs can generate personalized medical reports, generate treatment recommendations, and support diagnostic processes that assist health care specialists in making informed decisions [17,21]. However, while LLMs enhance diagnostic efficiency, concerns regarding inconsistent outputs pose barriers to clinical adoption. Medical prediction applications account for (7/29, 24%) of LLM use in chronic disease management, showing strong potential for early disease detection and risk stratification. By integrating structured and unstructured clinical data, such as lab results, clinical notes, and imaging, LLMs enable more comprehensive and accurate predictive models compared with traditional methods [17,29,31,34-36].

Notably, real-time risk assessment tools, like the ambulatory device developed for sickle cell anemia management [20], demonstrate how LLMs can predict complications before symptoms appear. However, challenges relating to medical accuracy still limit their seamless integration into clinical workflows [6,16,17,21-23,27]. A significant emerging trend is the development of retrieval-augmented generation

(RAG) frameworks that enhance prediction accuracy by dynamically incorporating up-to-date medical knowledge [22,35,50-53]. Future advancements should focus on seamless integration of LLMs with existing medical systems, wearable health technologies, and mobile health applications to improve interoperability and trustworthiness.

Methodological Quality and Strength of Evidence

The methodological quality assessment revealed notable trends that should be considered while interpreting the results of this scoping review. High-quality studies (18/29, 62%) were not uniformly distributed across LLM application tasks, studies examining medical prediction applications [29,31,35,36] and patient education and information provision [28,32,38,39] had significantly stronger methodological rigor. Studies investigating emotional support showed mixed quality, with some high-quality qualitative research [18,24], a quantitative descriptive study [37] alongside a moderate-quality mixed study [38]. Notably, self-management and disease monitoring studies showed significant methodological heterogeneity, with quality ratings ranging from low to high.

Journal articles demonstrated a substantially higher proportion of high-quality studies (13/18, 72%) compared with conference papers (3/8, 38%) and preprints (1/3, 33%), suggesting that peer-review processes enhance methodological quality. The identified common limitations, including inadequate sampling strategies, limited participant demographic reporting, and insufficient methodological transparency, were more prevalent in lower-tier publication sources, including conference papers and preprints. Quantitative descriptive studies, especially those focused on system design and prototype testing (10/29, 34.5%), showed mixed quality ratings ranging from low to high, with a common limitation being the use of synthetic data, lack of clinical validation, as they commonly prioritized technical utility. These methodological patterns significantly influence the reliability of the findings.

Challenges and Corresponding Recommendations

This section discusses the key challenges identified in the review and presents corresponding recommendations to mitigate these issues. Each challenge is followed by practical solutions to enhance the applicability of LLMs in chronic disease management.

Inaccurate Data and Inconsistencies in Responses

Inaccurate and inconsistent responses emerged as a significant challenge (18/29, 62%), primarily due to poor data quality, inherent biases in training datasets, and the limited scope of knowledge constrained by the model's training cutoff [6,22,23,31]. Given that LLM performance is intrinsically linked to data quality, flawed datasets inevitably propagate errors in outputs, a manifestation of the "garbage in, garbage out" principle [54-56]. The issue of biases in AI models has gathered significant attention in recent literature [54-56], prompting possible migration strategies [57-59]. These limitations carry critical implications for health care, as erroneous LLM outputs may lead to incorrect clinical decisions, posing significant risks to patient safety [8,12]. Hence, researchers are exploring technical solutions such as advancing domain-specific fine-tuning techniques [60,61], leveraging retrieval-augmented generation (RAG) frameworks [22,50-52], and refining outputs through reinforcement learning (RL) and prompt engineering techniques [62-64]. In addition, implementing expert validation protocols has emerged as a crucial safeguard to ensure adherence to evidence-based practice.

Limited Datasets

The scarcity of high-quality datasets for chronic diseases accounted for (6/29, 21%). Studies highlighted limitations and narrow coverage of publicly accessible clinical training datasets due to data privacy and institutional restrictions [16,19,21-23,39]. Given that experimental studies often require substantial model training datasets, the absence of adequate data poses a significant challenge to the effectiveness and success of these studies. Therefore, to address this gap, synthetic datasets and data augmentation techniques have been explored by studies [15,65]. However, these methods risk amplifying pre-existing biases in source data [54-56]. Therefore, dataset validation is essential to ensure quality, collaborative partnerships with health care institutions to access real-world clinical datasets and knowledge distillation techniques, where smaller models can be trained on outputs from larger, clinically validated models, reducing dependency on raw data volume, can be explored [66].

Computational Resources and Technical Challenges

The computational demands of training LLMs for health care applications remain a significant challenge (6/29, 21%). Consequently, low computing resources approaches, such as quantization, parameter-efficient fine-tuning (PEFT) techniques like Low-Rank Adaptation (LORA), Quantized LoRA (QLORA), Weight-Decomposed Low-Rank Adaptation, and REFT [67-73], are evolving and are popularly used in fine-tuning LLMs in low-resource computing environments. In addition, adapter-based tuning methods provide lightweight alternatives by injecting trainable adapter layers into frozen pretrained models, enabling task-specific fine-tuning without updating the entire model parameters [74,75]. Building on these advances, the development of lightweight LLMs optimized for mobile devices presents a promising direction for extending AI-based chronic disease management to resource-constrained settings [76].

Usability and Accessibility Concerns

Usability and accessibility concerns accounted for (9/29, 31%), including issues with text-only interfaces for some LLMs, cultural misalignments, and outputs ill-suited for pediatric or elderly populations [18,28,40]. While studies highlighted text-only interfaces as a critical limitation of LLMs in health care [28,40], recent advances in multimodal

architectures have addressed this gap. These models now integrate and interpret diverse data modalities, including medical images, audio, and structured documents, while generating composite textual and visual outputs [31,77-79]. In addition, having dynamic and simplified interfaces to accommodate low-digital-literacy users, pediatric users, and for different cultural and language settings could further improve LLM usability and adaptation.

Legal, Ethical Issues, and Regulatory Issues

Legal, ethical, and regulatory concerns (10/29, 35%) remain a key issue in LLM adoption in health care. Studies identified data privacy, biases, misinformation, responsibility, and accountability for LLM-generated content as key concerns. Although several AI frameworks have been proposed, the absence of standardized guidelines and regulatory approvals creates significant gaps [27,80-82]. This regulatory vacuum risks inconsistent model development and validation practices, unaddressed ethical dilemmas regarding accountability for AI-generated recommendations, and potential mismatches between rapidly evolving LLM capabilities and static health care regulations [27,83]. Addressing these ethical concerns requires standard guidelines and rules to ensure responsible use in health care settings[84]. Therefore, future efforts must prioritize the development of a comprehensive regulatory framework.

LLM Evaluation

Evaluation gaps accounted for (5/29, 17%), reflecting critical shortcomings in current methodologies for assessing LLM performance in clinical contexts. The existing automated evaluation metrics mainly focus on predictive performance using metrics such as accuracy and F_1-scores that lack medical and treatment knowledge [22,85]. These metrics may produce misleading conclusions if not appropriately selected [85]. Furthermore, human evaluation, although valuable, introduces subjectivity and interrater variability, yet the absence of a standardized LLM evaluation framework makes attaining consensus among human raters challenging [30,85]. A promising direction involves a hybrid evaluation approach that integrates human expert reviews with automated metrics. Future efforts should prioritize the development of standardized LLM evaluation frameworks tailored to health care settings. Table 3 summarizes the challenges associated with applying LLMs in chronic disease management with proposed recommendations.

Table 3. Key challenges and recommendations.

Challenge	Key observations	Recommendation
Inaccurate and inconsistent responses	Hallucinations, inconsistent responses, outdated knowledge	Adopting RAG frameworks, fine-tuning, and prompt engineering. RL, expert validation of LLM-generated recommendations
Limited datasets	Scarcity of datasets, missing data, and dataset imbalances.	Use synthetic data, data augmentation, partnerships with health care institutions, knowledge distillation
Computational demands	High computational demands	Adopt PEFT (LORA and QLORA), quantization use lightweight models for mobile devices
Ethical and privacy concerns	Privacy concerns, language and cultural barriers, and lack of regulatory oversight	Develop a regulatory framework
Usability issues	Restriction to textual inputs, lack of empathy, ineffectiveness in emergencies Age-appropriate interaction	Use multimodal LLMs. Dynamic interfaces to accommodate low-digital-literacy users, age-appropriate interaction modes customize to different cultural settings Integrate with wearable devices and mobile health apps
Evaluation challenges	Predictive performance metrics, subjectivity, and interrater variability	Develop a standardized LLM evaluation framework for health care.

Limitations of the Study

Although quality assessment was conducted using MMAT, studies were not excluded based on their methodological rigor. As a result, including moderate and low-quality studies may have influenced the reliability and consistency of the reviewer's findings. The varied methodological designs across studies may have affected the interpretation of results and conclusions drawn. Furthermore, the review was limited to only English-language publications, which may have introduced language bias and limited the generalizability of our findings, particularly in contexts where LLMs are being adapted to local languages or integrated into culturally specific health care practices. The exclusion of databases such as Embase and Web of Science may have limited the comprehensiveness of the search. Future research can consider broader database coverage and non-English sources to enhance diversity and scope.

Implications for Practice and Future Research

This scoping review reveals several critical implications for the integration of LLMs in chronic disease management. First, it is essential to address the accuracy issues identified in 18/29, 62% of studies. This calls for both technical solutions (domain-specific fine-tuning, reinforcement learning (RL), and retrieval-augmented generation (RAG) frameworks

[61,64]) and nontechnical solutions (expert validation and collaborative partnerships with health care institutions to access and use real-world clinical data). Second, enhancing accessibility across diverse patient populations requires developing culturally adapted LLM interfaces, implementing age-appropriate interaction modes [40], and integrating with low-resource platforms such as SMS-based systems and lightweight mobile apps for various populations [76]. Third, robust governance frameworks must be established to address the ethical, legal, and privacy concerns noted in 10/29, 35% of studies to ensure regulatory compliance.

Finally, future research should focus on multimodal LLMs that can synthesize diverse data inputs from wearable devices and mobile health applications for holistic patient monitoring [19,31,35] and develop resource-efficient deployment techniques. The predominance of diabetes-focused studies (14/29, 48%) highlights a potential research gap in addressing other prevalent chronic conditions. Similarly, there is a research gap in age-inclusive LLM design, given the overwhelming focus on adult populations (28/29, 96%) and the lack of pediatric studies. Addressing these gaps would enhance the clinical relevance and equitable application of LLMs across the full spectrum of chronic disease management.

Conclusion

This scoping review highlights the growing potential of LLMs in supporting chronic disease management through patient education, diagnosis and treatment, emotional support, self-management support, decision support, and prediction tasks. While LLMs offer promising capabilities, their effective integration into health care still requires addressing key challenges related to accuracy, accessibility, usability, and ethical and privacy concerns. Future research should focus on integrating LLMs with mobile and wearable technologies, creating culturally and age-appropriate interfaces, and exploring integration with low-resource platforms. Addressing these research gaps will ensure equitable and safe use of LLMs across diverse health care settings.

Acknowledgments

The authors acknowledge the valuable inputs of the reviewers and the editor, and the support of Mr. Bazekketa Datson of Ndejje University, Faculty of Science and Computing, for illustrating the figures used in this scoping review. The authors did not receive any funding for this review study. During the preparation of this work, the authors used Grammarly to improve the readability and language of the paper. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Authors' Contributions

SHM contributed to conceptualization, methodology, data extraction and analysis, writing-original draft, and writing-review and editing. OJ managed conceptualization, methodology, data extraction and analysis, and writing-review and editing. HKN handled data extraction and analysis and review and editing. CGO, PS, PW, and NK contributed to review and critical input.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed methodological quality appraisal results for each included study using the Mixed Methods Appraisal tool (MMAT).
[DOCX File (Microsoft Word File), 79 KB-Multimedia Appendix 1]

Checklist 1

PRISMA-ScR checklist.
[PDF File (Adobe File), 154 KB-Checklist 1]

References

1. Bardhan I, Chen H, Karahanna E. Connecting systems, data, and people: a multidisciplinary research roadmap for chronic disease management. MIS Q. Mar 1, 2020;44(1):185-200. [doi: 10.25300/MISQ/2020/14644]
2. Hacker K. The burden of chronic disease. Mayo Clinic Proceedings: Innovations, Quality & Outcomes. Feb 2024;8(1):112-119. [doi: 10.1016/j.mayocpiqo.2023.08.005]
3. Holmen H, Larsen MH, Sallinen MH, et al. Working with patients suffering from chronic diseases can be a balancing act for health care professionals - a meta-synthesis of qualitative studies. BMC Health Serv Res. Feb 10, 2020;20(1):98. [doi: 10.1186/s12913-019-4826-2] [Medline: 32039723]
4. Xie Y, Lu L, Gao F, et al. Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: a new paradigm in smart healthcare. CURR MED SCI. Dec 2021;41(6):1123-1133. [doi: 10.1007/s11596-021-2485-0] [Medline: 34950987]
5. Jiménez-Muñoz L, Gutiérrez-Rojas L, Porras-Segovia A, Courtet P, Baca-García E. Mobile applications for the management of chronic physical conditions: a systematic review. Intern Med J. Jan 2022;52(1):21-29. [doi: 10.1111/imj.15081] [Medline: 33012045]

6. Montagna S, Ferretti S, Klopfenstein LC, Florio A, Pengo MF. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. GoodIT '23: Proceedings of the 2023 ACM Conference on Information Technology for Social Good. Sep 6, 2023:205-212. [doi: 10.1145/3582515.3609536]
7. Haque A, Chowdhury M, Soliman H. Transforming chronic disease management with chatbots: key use cases for personalized and cost-effective care. Presented at: 2023 Sixth International Symposium on Computer, Consumer and Control (IS3C); Jun 30 to Jul 3, 2023:367-370; Taichung, Taiwan. 2023.[doi: 10.1109/IS3C57901.2023.00104]
8. Reddy S. Evaluating large language models for use in healthcare: a framework for translational value assessment. Informatics in Medicine Unlocked. 2023;41(May):101304. [doi: 10.1016/j.imu.2023.101304]
9. Huang X, Lian J, Lei Y, Yao J, Lian D, Xie X. Recommender AI agent: integrating large language models for interactive recommendations. arXiv. Preprint posted online on Jan 30, 2024. URL: https://arxiv.org/abs/2308.16505 [Accessed 2025-08-24]
10. Snoswell CL, Snoswell AJ, Kelly JT, Caffery LJ, Smith AC. Artificial intelligence: augmenting telehealth with large language models. J Telemed Telecare. Jan 2025;31(1):150-154. [doi: 10.1177/1357633X231169055] [Medline: 37041736]
11. Khan MA, Mohammad N. A review on large language models: architectures, applications, taxonomies, open issues and challenges. TechRxiv. Preprint posted online on Sep 27, 2023. [doi: 10.36227/techrxiv.24171183.v1]
12. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci. Aug 2023;2(4):255-263. [doi: 10.1002/hcs2.61] [Medline: 38939520]
13. Sharma P, Parasa S. ChatGPT and large language models in gastroenterology. Nat Rev Gastroenterol Hepatol. Aug 2023;20(8):481-482. [doi: 10.1038/s41575-023-00799-8] [Medline: 37253794]
14. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus. May 2023;15(5):e39305. [doi: 10.7759/cureus.39305] [Medline: 37378099]
15. Latif A, Kim J. Evaluation and analysis of large language models for clinical text augmentation and generation. IEEE Access. Apr 2024;12:1-1. [doi: 10.1109/ACCESS.2024.3384496]
16. Yang H, li jiaxi, liu siru, Liu J. Exploring the potential of large language models in personalized diabetes treatment strategies. In Review. Preprint posted online on 2023. [doi: 10.21203/rs.3.rs-3995740/v1]
17. Raghu K, S T, S Devishamani C, M S, Rajalakshmi R, Raman R. The utility of ChatGPT in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis [response to letter]. Clin Ophthalmol. 2024;18:313-314. [doi: 10.2147/OPTH.S461186] [Medline: 38317795]
18. Song I, Pendse SR, Kumar N. The typing cure: experiences with large language model chatbots for mental health support. arXiv. Preprint posted online on May 9, 2025. [doi: https://arxiv.org/abs/2401.14362]
19. Liu Z, Chen C, Cao J, et al. Large language models for cuffless blood pressure measurement from wearable biosignals. BCB '24: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 1-11. [doi: 10.1145/3698587.3701447]
20. Ogundare O, Sofolahan S. Large language models in ambulatory devices for home health diagnostics: a case study of sickle cell anemia management. arXiv. Preprint posted online on May 5, 2023. [doi: 10.1007/978-3-031-40971-4_42]
21. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. Cureus. Oct 2023;15(10):e46736. [doi: 10.7759/cureus.46736] [Medline: 38022227]
22. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. Presented at: 2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS); Aug 12-13, 2023:532-536; Dali, China. [doi: 10.1109/CCIS59572.2023.10263217]
23. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ. Feb 13, 2024;10:e51391. [doi: 10.2196/51391] [Medline: 38349725]
24. Al-Anezi FM. Exploring the use of ChatGPT as a virtual health coach for chronic disease management. Learn Health Syst. Jul 2024;8(3):e10406. [doi: 10.1002/lrh2.10406] [Medline: 39036525]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI. 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
27. Athavale A, Baier J, Ross E, Fukaya E. The potential of chatbots in chronic venous disease patient management. JVS Vasc Insights. 2023;1:100019. [doi: 10.1016/j.jvsvi.2023.100019] [Medline: 37701430]

28. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Health. 2024;10:20552076231224603. [doi: 10.1177/20552076231224603] [Medline: 38188865]
29. Abbas S, Iftikhar M, Shah MM, Khan SJ. ChatGPT-assisted machine learning for chronic disease classification and prediction: a developmental and validation study. Cureus. Dec 2024;16(12):e75851. [doi: 10.7759/cureus.75851] [Medline: 39822450]
30. Anderson P, Lin D, Davidson J, et al. Bridging domains in chronic lower back pain: large language models and ontology-driven strategies for knowledge graph construction. bioRxiv. Preprint posted online on Mar 14, 2024. [doi: 10.1101/2024.03.11.584505]
31. Ding JE, Thao PNM, Peng WC, et al. Large language multimodal models for 5-year chronic disease cohort prediction using EHR data. arXiv. Preprint posted online on Aug 29, 2024. URL: https://arxiv.org/abs/2403.04785
32. Jairoun AA, Al-Hemyari SS, Shahwan M, Al-Qirim T, Shahwan M. Benefit-risk assessment of ChatGPT applications in the field of diabetes and metabolic illnesses: a qualitative study. Clin Med Insights Endocrinol Diabetes. 2024;17:11795514241235514. [doi: 10.1177/11795514241235514] [Medline: 38495947]
33. Mondal A, Naskar A. Artificial intelligence in diabetes care: evaluating GPT-4's competency in reviewing diabetic patient management plan in comparison to expert review. Endocrinology (including Diabetes Mellitus and Metabolic Disease). 2024. [doi: 10.1101/2024.04.12.24305732]
34. Liu H, Zhang W, Xie J, et al. Few-shot learning for chronic disease management: leveraging large language models and multi-prompt engineering with medical knowledge injection. Proceedings of the 58th Hawaii International Conference on System Sciences. 2025. [doi: 10.24251/HICSS.2025.084] [Medline: 38681743]
35. Liao C, Kuo WT, Hu IH, et al. EHR-based mobile and web platform for chronic disease risk prediction using large language multimodal models. CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Oct 21, 2024:5244-5248. [doi: 10.1145/3627673.3679227]
36. Ding JE, Thao PNM, Peng WC, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. Sci Rep. Sep 6, 2024;14(1):20774. [doi: 10.1038/s41598-024-71020-2] [Medline: 39237580]
37. Dao D, Teo JYC, Wang W, Nguyen HD. LLM-powered multimodal AI conversations for diabetes prevention. Presented at: ICMR '24: International Conference on Multimedia Retrieval; 1-6; Phuket, Thailand. Jun 10, 2024.URL: https://dl.acm.org/doi/proceedings/10.1145/3643479 [Accessed 2025-09-02]
38. Khan M. Assessing the efficacy of ChatGPT in facilitating self-management strategies among diabetic patients section. European Chemical Bulletin. 2023;12(10):10490-10502. URL: https://www.researchgate.net/publication/373420033_Assessing_the_Efficacy_of_ChatGPT_in_Facilitating_Self-Management_Strategies_among_Diabetic_Patients_Section_A-Research_paper_10490_Eur [Accessed 2025-09-02]
39. Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in answering queries related to lifestyle-related diseases and disorders. Cureus. Nov 2023. [doi: 10.7759/cureus.48296]
40. Young CC, Enichen E, Rao A, et al. Pilot study of large language models as an age-appropriate explanatory tool for chronic pediatric conditions. medRxiv. Preprint posted online on Aug 7, 2024. [doi: 10.1101/2024.08.06.24311544]
41. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. Oct 2024;30(10):2886-2896. [doi: 10.1038/s41591-024-03139-8] [Medline: 39030266]
42. Ying Y, Wang Y, Yuan S, et al. Exploration of ChatGPT application in diabetes education based on multi-dataset. medRxiv. Preprint posted online on Sep 27, 2023. [doi: 10.1101/2023.09.27.23296144]
43. Li H, Jiang Z, Guan Z, et al. Large language models for diabetes training: a prospective study. Sci Bull Sci Found Philipp. Mar 2025;70(6):934-942. [doi: 10.1016/j.scib.2025.01.034]
44. Hussain W, Grundy J. Advice for diabetes self-management by ChatGPT models: challenges and recommendations. arXiv. Preprint posted online on Jan 14, 2025. [doi: https://arxiv.org/abs/2501.07931]
45. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res. Nov 8, 2024;26:e58041. [doi: 10.2196/58041] [Medline: 39046096]
46. Eaton C, Vallejo N, McDonald X, et al. User engagement with mHealth interventions to promote treatment adherence and self-management in people with chronic health conditions: systematic review. J Med Internet Res. Sep 24, 2024;26:e50508. [doi: 10.2196/50508] [Medline: 39316431]
47. Cheikh-Moussa K, Mira JJ, Orozco-Beltran D. Improving engagement among patients with chronic cardiometabolic conditions using mHealth: critical review of reviews. JMIR Mhealth Uhealth. Apr 8, 2020;8(4):e15446. [doi: 10.2196/15446] [Medline: 32267239]
48. Mattison G, Canfell O, Forrester D, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res. Jul 1, 2022;24(7):e36690. [doi: 10.2196/36690] [Medline: 35776492]

49. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. Sep 1, 2018;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
50. Lewis P, Perez E, Pictus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. Preprint posted online on Apr 12, 2021. [doi: 10.48550/arXiv.2005.11401]
51. Zakka C, Chaurasia A, Shad R, et al. Almanac: retrieval-augmented language models for clinical medicine. Res Sq. May 2, 2023:rs.3.rs-2883198. [doi: 10.21203/rs.3.rs-2883198/v1] [Medline: 37205549]
52. Wang Y, Ma X, Chen W. Augmenting black-box llms with medical textbooks for clinical question answering. arXiv. Preprint posted online on Feb 23, 2025. [doi: 10.48550/arXiv.2309.02233]
53. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics ACL 2024. 2024:6233-6251. [doi: 10.18653/v1/2024.findings-acl.372]
54. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. Jun 2023;5(6):e333-e335. [doi: 10.1016/S2589-7500(23)00083-3] [Medline: 37120418]
55. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. J Data and Information Quality. Jun 30, 2023;15(2):1-21. [doi: 10.1145/3597307]
56. Liang PP, Wu C, Morency LP, Salakhutdinov R. Towards understanding and mitigating social biases in language models. arXiv. Preprint posted online on Jun 24, 2021. [doi: 10.48550/arXiv.2106.13219] [Medline: 34545335]
57. Verma S, Ernst M, Just R. Removing biased data to improve fairness and accuracy. arXiv. Preprint posted online on Feb 5, 2021. [doi: 10.48550/arXiv.2102.03054]
58. Kim E, Lee J, Choo J. BiaSwap: removing dataset bias with bias-tailored swapping augmentation. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021:14972-14981; Montreal, QC, Canada. 2021.[doi: 10.1109/ICCV48922.2021.01472]
59. Ernst JS, Marton S, Brinkmann J, et al. Bias mitigation for large language models using adversarial learning. Presented at: CEUR Workshop on Fairness and Bias in AI; Krakow, Poland. 2023.URL: https://ceur-ws.org/Vol-3523/paper11.pdf [Accessed 2025-08-24]
60. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. Cureus. Jun 2023;15(6):e40895. [doi: 10.7759/cureus.40895] [Medline: 37492832]
61. Alghanmi I, Espinosa-Anke L, Schockaert S. Self-supervised intermediate fine-tuning of biomedical language models for interpreting patient case descriptions. Proceedings of the 29th International Conference on Computational Linguistics. 2022. URL: https://aclanthology.org/2022.coling-1.123/ [Accessed 2025-09-02]
62. Heston TF, Khun C. Prompt engineering in medical education. IME. 2023;2(3):198-205. [doi: 10.3390/ime2030019]
63. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. Oct 4, 2023;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
64. Wang J, Shi E, Yu S, et al. Prompt engineering for healthcare: methodologies and applications. arXiv. Preprint posted online on Apr 23, 2024. [doi: 10.48550/arXiv.2304.14670]
65. Sarker S, Qian L, Dong X. Medical data augmentation via chatgpt: a case study on medication identification and medication event classification. arXiv. Preprint posted online on Jun 10, 2023. [doi: https://arxiv.org/abs/2306.07297]
66. Yang C, Zhu Y, Lu W, et al. Survey on knowledge distillation for large language models: methods, evaluation, and application. ACM Trans Intell Syst Technol. 2024. [doi: 10.1145/3699518]
67. Liao B, Meng Y, Monz C. Parameter-efficient fine-tuning without introducing new latency. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1. 4242-4260; 2023.[doi: 10.18653/v1/2023.acl-long.233]
68. Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell. 2023;5(3):220-235. [doi: 10.1038/s42256-023-00626-4]
69. Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. arXiv. Preprint posted online on Aug 26, 2022. [doi: 10.48550/arXiv.2205.05638]
70. Hu E, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on Oct 16, 2021. [doi: 10.48550/arXiv.2106.09685]
71. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. arXiv. Preprint posted online on May 23, 2023. [doi: 10.48550/arXiv.2305.14314]
72. Liu SY, Wang CY, Yin H, et al. DoRA: weight-decomposed low-rank adaptation. arXiv. Preprint posted online on Jul 9, 2024. [doi: 10.48550/arXiv.2402.09353]
73. Wu Z, Arora A, Wang Z, et al. ReFT: representation finetuning for language models. arXiv. Preprint posted online on May 22, 2024. [doi: 10.48550/arXiv.2404.03592]

74. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. arXiv. Preprint posted online on Jun 13, 2019. [doi: 10.48550/arXiv.1902.00751]
75. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1). Preprint posted online on Jan 2021. [doi: 10.18653/v1/2021.acl-long.353]
76. Wang X, Dang T, Kostakos V, Jia H. Efficient and personalized mobile health event prediction via small language models. Presented at: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom '24). 2353-2358; Dec 4, 2024.[doi: 10.1145/3636534.3698123]
77. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res. Nov 2, 2023;25:e52865. [doi: 10.2196/52865] [Medline: 37917126]
78. Belyaeva A, Cosentino J, Hormozdiari F, et al. Multimodal llms for health grounded in individual-specific data. arXiv. Preprint posted online on Jul 20, 2023. [doi: 10.48550/arXiv.2307.09018]
79. Gemini Team Google. Gemini: a family of highly capable multimodal models. arXiv. Preprint posted online on May 9, 2025. [doi: 10.48550/arXiv.2312.11805]
80. Ali H, Qadir J, Alam T, Househ M, Shah Z. ChatGPT and large language models in healthcare: opportunities and risks. Presented at: IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)). 2023.[doi: 10.1109/AIBThings58340.2023.10291020]
81. Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access. 2024;12:31014-31035. [doi: 10.1109/ACCESS.2024.3369912]
82. Goirand M, Austin E, Clay-Williams R. Implementing ethics in healthcare AI-based applications: a scoping review. Sci Eng Ethics. Sep 3, 2021;27(5):5. [doi: 10.1007/s11948-021-00336-3] [Medline: 34480239]
83. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 6, 2023;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
84. Lareyre F, Raffort J. Ethical concerns regarding the use of large language models in healthcare. EJVES Vasc Forum. 2024;61:1. [doi: 10.1016/j.ejvsvf.2023.10.003] [Medline: 38025830]
85. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform. Oct 2021;28(1):1-7. [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]

Abbreviations

LLM: large language model
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RAG: retrieval augmented generation
RL: reinforcement learning

Edited by Alexandre Castonguay; peer-reviewed by Ali Jafarizadeh, Soroosh Tayebi Arasteh; submitted 26.09.2024; final revised version received 06.05.2025; accepted 23.05.2025; published 29.09.2025

Please cite as:
Serugunda HM, Jianquan O, Kasujja Namatovu H, Ssemaluulu P, Kimbugwe N, Garimoi Orach C, Waiswa P
Using Large Language Models for Chronic Disease Management Tasks: Scoping Review
JMIR Med Inform 2025;13:e66905
URL: https://medinform.jmir.org/2025/1/e66905
doi: 10.2196/66905