Review

Diagnostic Prediction Models for Primary Care, Based on AI and Electronic Health Records: Systematic Review

Liesbeth Hunik[1], MSc, MD; Asma Chaabouni[1], MSc, MD; Twan van Laarhoven[2], PhD; Tim C Olde Hartman[1], PhD, MD; Ralph T H Leijenaar[3], PhD; Jochen W L Cals[3], Prof Dr; Annemarie A Uijen[1], PhD, MD; Henk J Schers[1], Prof Dr

[1]Department of Primary and Community Care, Research Institute for Medical Innovation, Radboudumc, Nijmegen, The Netherlands
[2]Institute for Computing and Information Science, Radboud University, Nijmegen, The Netherlands
[3]Department of Family Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

Corresponding Author:

Liesbeth Hunik, MSc, MD
Department of Primary and Community Care
Research Institute for Medical Innovation, Radboudumc
Geert Grooteplein Zuid 21
Nijmegen 6525 GA
The Netherlands
Phone: 31 243618181
Email: liesbeth.hunik@radboudumc.nl

Abstract

Background: Artificial intelligence (AI)–based diagnostic prediction models could aid primary care (PC) in decision-making for faster and more accurate diagnoses. AI has the potential to transform electronic health records (EHRs) data into valuable diagnostic prediction models. Different prediction models based on EHR have been developed. However, there are currently no systematic reviews that evaluate AI-based diagnostic prediction models for PC using EHR data.

Objective: This study aims to evaluate the content of diagnostic prediction models based on AI and EHRs in PC, including risk of bias and applicability.

Methods: This systematic review was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. MEDLINE, Embase, Web of Science, and Cochrane were searched. We included observational and intervention studies using AI and PC EHRs and developing or testing a diagnostic prediction model for health conditions. Two independent reviewers (LH and AC) used a standardized data extraction form. Risk of bias and applicability were assessed using PROBAST (Prediction Model Risk of Bias Assessment Tool).

Results: From 10,657 retrieved records, a total of 15 papers were selected. Most EHR papers focused on 1 chronic health care condition (n=11, 73%). From the 15 papers, 13 (87%) described a study that developed a diagnostic prediction model and 2 (13%) described a study that externally validated and tested the model in a PC setting. Studies used a variety of AI techniques. The predictors used to develop the model were all registered in the EHR. We found no papers with a low risk of bias, and high risk of bias was found in 9 (60%) papers. Biases covered an unjustified small sample size, not excluding predictors from the outcome definition, and the inappropriate evaluation of the performance measures. The risk of bias was unclear in 6 papers, as no information was provided on the handling of missing data and no results were reported from the multivariate analysis. Applicability was unclear in 10 (67%) papers, mainly due to lack of clarity in reporting the time interval between outcomes and predictors.

Conclusions: Most AI-based diagnostic prediction models based on EHR data in PC focused on 1 chronic condition. Only 2 papers tested the model in a PC setting. The lack of sufficiently described methods led to a high risk of bias. Our findings highlight that the currently available diagnostic prediction models are not yet ready for clinical implementation in PC.

Trial Registration: PROSPERO CRD42022320002; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022320002

JMIR Med Inform 2025;13:e62862; doi: 10.2196/62862

Keywords: primary care; electronic health records; artificial intelligence; EHR; AI; systematic review; decision-making; AI-based diagnostic; applicability; assessment tool

Introduction

Background

The diagnostic process is a core task of general practitioners (GPs). However, making a diagnosis may be a challenging task given the diversity, complexity, and early presentation of symptoms. Clinical prediction models are intended to improve the diagnostic process [1]. These models can support the health care provider by predicting serious illness [2]. In the last years, the interest in artificial intelligence (AI) techniques for the development of prediction models has been growing [3,4]. AI-based prediction models could aid in decision-making for faster and more accurate diagnoses, with more diagnostic efficiency that can benefit patients' health [5-8]. Examples are prediction tools that can predict colorectal cancer in patients [9,10].

Clinical prediction models used to be built on data from large databases, such as data collected for research purposes, claim data, or data from electronic health records (EHRs) [11,12]. EHR data consist of structured data, which are data in standardized format, and unstructured data, which are free-text data. Primary care (PC) EHR data provide extensive and longitudinal data from a patient's health trajectory and changes over time. AI might prove to be a valuable method to extract clinically useful and actionable insight from this vast and complex source of patient data [13]. For that reason, AI has the potential to transform EHR data into a valuable tool for predicting diagnosis in daily PC practice.

Reviews on the value of AI in PC are scarce, and previous research had different aims. For example, Kueper et al [14] provided an overview of diagnostic prediction models based on AI in PC. However, the authors did not assess the quality of these diagnostic prediction models. Other research in this field explored AI systems in community-based primary health care [15] or focused on different machine learning (ML)–based diagnostic and prognostic models that predicted a health care condition [16]. As AI has the potential to support and improve the diagnostic process, high-quality and validated prediction models are crucial in order to ensure patient safety after clinical implementation. Although a variety of prediction models for PC have been developed, to our knowledge, there are currently no systematic reviews on AI-based diagnostic prediction models for PC using EHR data.

Objective

Evaluation of the content and quality assessment of AI-based diagnostic prediction models using EHRs in PC was largely lacking in current literature. Therefore, we systematically reviewed the literature in order to critically evaluate the content of these AI-based diagnostic prediction models, including risk of bias and applicability.

Methods

Study Design

We performed a systematic review according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [17] (the PRISMA checklist is provided in Checklist 1). The protocol for this study was registered in PROSPERO (nr: CRD42022320002). The research team included stakeholders such as practicing GPs, researchers, methodologists, and AI experts in the design, analysis, and reporting of the study.

Search Strategy and Study Selection

Our search was adapted from the search strategy developed by Kueper et al [14]. It combines two concepts including a wide range of different terms used to describe the concepts: (1) artificial intelligence and (2) primary care (for full search strategy, see Multimedia Appendix 1, part 1 [18-67]). EHRs were not part of the search strategy, because literature suggests that we might miss important studies when including EHRs or related terms in the search strategy [13]. We searched in the following databases: MEDLINE, Embase, Web of Science, and Cochrane. There were no restrictions concerning the publication date. The last search update was conducted on August 28, 2023. We focused on intervention and observational studies. We excluded systematic reviews, meta-analyses, case studies, editorials, protocols, and conference posters or abstracts. Full text had to be available to be selected for screening. The literature had to be written in English or Dutch. Duplicate publications were removed with EndNote 20.

Inclusion and Exclusion Criteria

Four inclusion criteria were used to select the papers: (1) primary care focus: this included PC data, models that were tested in a PC setting, or PC had to be specifically mentioned in the aim of the study; (2) diagnostic prediction model: models had to predict a health condition applicable during a GP's consultation; prediction models that identified a disease in a database, rather than predicting a disease for an individual, were excluded; (3) AI: this included all ML and deep learning techniques; we directed our focus to data-driven prediction models without using medical images as input data; and (4) EHR-based data: EHR data had to be used for the development or validation of the model. EHRs were defined as PC data from EHRs, medical records, or clinical notes. See Multimedia Appendix 1, part 2 [18-67], for the full screening guidance.

Title and abstract screening was done in management software Rayyan (rayyan.ai) by 2 independent reviewers (LH and LvdH). Conflicts were resolved by a third reviewer (AU). Full-text screening was done by the same independent reviewers. Conflicts were resolved by discussion, and if no consensus was reached, they were resolved by a third

reviewer (AU). Backward citation searching was conducted on the included papers and finished on November 7, 2023.

Data Extraction and Quality Assessment

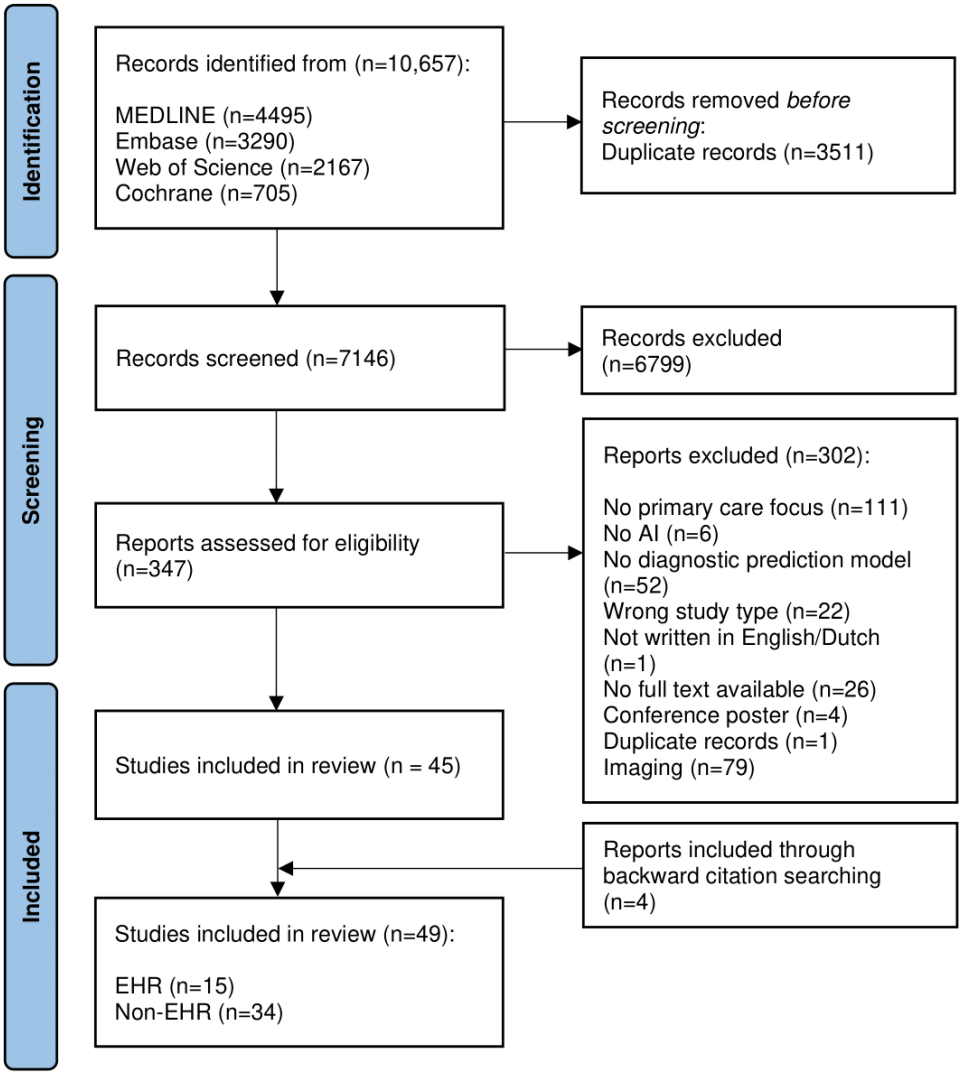
Data extraction of included papers was done by 2 independent reviewers (LH and AC). They used a standardized data extraction form adapted from the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) [68]. Basic information was extracted from all papers. The extraction of more detailed information was focused on EHR-based papers. For all papers (EHR and non-EHR papers), we extracted general information (first author, year of publication, title, data source, and country of data source), study design (retrospective or prospective), and outcome (predicted health condition). For the EHR papers, we additionally extracted dataset information (name of the dataset and sample size: number of participants used for model training, testing, or validation), AI technique, and predictors (the potentially used predictors used to develop the model). Risk of bias and applicability were assessed using PROBAST (Prediction Model Risk of Bias Assessment Tool). This tool includes 20 signaling questions divided into 4 domains (participants, predictors, outcome, and analysis) [69,70]. Overall judgment (ie, low, unclear, or high) of risk of bias is based on the 4 domains. If 1 domain is considered to have a high risk of bias, the overall judgment is scored as a high risk of bias. If at least 1 domain is considered to have an unclear risk of bias (without a domain with high risk of bias), the overall judgment is scored as unclear risk of bias. Applicability concern was rated based on 3 domains (participants, predictors, and outcome) and an overall judgment of applicability (ie, low, unclear, or high) was also given with the same approach as the risk-of-bias scoring. Applicability evaluation depends on the review question [69], and we translated applicability assessment as usability of the diagnostic prediction model in a PC setting. Conflicts in data extraction between the 2 reviewers (LH and AC) were resolved by discussion, and if no consensus was reached, they were resolved by a third reviewer (TvL).

Results

Description of Included Studies

We retrieved 10,657 records using our search strategy. After duplicate removal, we conducted title and abstract screening on 7146 records. A total of 347 records met the eligibility criteria for full-text screening. After full-text screening, 45 records were included. Backward citation searching yielded an additional 4 papers. A total of 49 papers were thus included in the review (Figure 1). Of the included papers, we identified 15 EHR papers and 34 non-EHR papers. A detailed description of the 34 non-EHR papers can be found in Multimedia Appendix 1, part 3 [18-67]. The data used in these 34 papers were collected from different sources, including secondary care datasets (n=17), questionnaires (n=4), and the knowledge of different health care providers (n=5).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. AI: artificial intelligence; EHR: electronic health record.



Overview of the EHR-Based Papers

Of the 15 EHR papers, 13 (87%) included the development of a prediction model [18-30]. In Table 1, the data extraction per paper can be found. The included EHR papers covered various outcomes, mostly chronic conditions (11/15, 73%) [19-28,31]. The most frequent predicted outcomes were dementia (3/15, 20%) [19,20,23], asthma (3/15, 20%), or chronic obstructive pulmonary disease (COPD) (3/15, 20%) [21,26,31]. Other study outcomes are shown in Table 1. All included papers used predictors registered in EHRs. Predictors included findings from clinical examination (n=6) [19,25,26,28,31,32], laboratory results (n=5) [21,22,25,28,32], and medication (n=4) [19,21,24,29]. All models used structured data.

Two papers externally validated and tested a prediction model in a PC setting [31,32]. One paper had a prospective approach and tested the diagnostic performance of a prediction model for asthma and COPD [31]. Ten papers (10/15, 67%) were published after 2020 [18,21-24,26,27,29,31,32]. Most data sources used in the studies originated from Europe (8/15, 53%) [18-22,24,27,31], followed by North America (5/15, 33%) [23,25,28,29,31,32] and Asia (2/15, 13%) [26,30].

Table 1. Extracted information from electronic health record papers.

Author, year	Country	Study type	Study design	Outcome	Dataset	Participants and inclusion criteria	Predictors	AI technique
Barnes et al (2020) [23]	United States	Developmental	Retrospective cohort study	To identify patients at high risk of unrecognized dementia	Data from Kaiser Permanente Washington	4330 participants aged at least 65 years, community member with no dementia	Demographics, diagnosis, vital signs, health care usage, medication	LR[a]
Briggs et al (2022) [22]	United Kingdom	Developmental	Nested case-control study	To predict risk of esophagogastric cancer	Data from General Practice Research Database	40,348 participants with esophagus or gastric cancer (7471 cases and 32,877 matched controls) diagnosed after 2000 (aged ≥40 years)	Demographics, symptoms, laboratory results	RF[b], SVM[c], LR, NB[d], XGBoost[e]
Dhanda et al (2023) [32]	United States	Developmental + Validation	Retrospective cohort study	To predict urine culture result without microscopy data to predict urinary tract infection	Data from emergency department (developmental phase). Data from primary care outpatient family medicine department at University of Kansas Medical Center (external validation)	80,859 participants (80,387 development, 472 external validation) with an ordered urinalysis and urine culture	Demographics, urine analysis, vital signs, symptoms, history of urinary tract infection, higher risk of clinical features	XGBoost, RF, NN[f]
Dros et al (2022) [24]	Netherlands	Developmental	Nested case-control study	To identify primary Sjögren syndrome	Data from Nivel Primary Care Database linked with Diagnosis Related Groups Information System dataset	930,590 participants (1411 cases, 1411 controls for training phase and all of the 929,179 controls for testing phase), with primary Sjögren syndrome from 2017	Demographics, diagnosis, medication, health usage	LR, RF
Ellertsson et al (2021) [18]	Iceland	Developmental	Retrospective cohort study	To diagnose common clinical headaches (cluster headache, migraine [with or without aura], tension headache)	Data from 15 primary Health Care of the Capital Area clinics	Unknown number of participants, 800 clinical notes from patients with 4 headache diagnoses from 2006 to 2020	Headache symptoms, sex, age, family history	RF
Ford et al (2019) [19]	United Kingdom	Developmental	Nested case-control study	To detect dementia	Data from Clinical Practice Research Datalink data	93,120 participants (46,560 cases with a dementia diagnosis code between 2000 and 2012, 46,560 controls)	Symptoms of physical or cognitive frailty, medical history, health care usage, ethnicity, family history of dementia, intoxications, BMI, blood pressure, psychological diagnoses, and treatment	RF, NB, SVM, NN
Jammeh et al (2018) [20]	United Kingdom	Developmental	Case-control study	To identify undiagnosed dementia	NHS Devon dataset with 18 participating GP[g] surgeries	3063 participants (850 cases with a dementia diagnosis code, 2213 controls)	Demographics, long-term conditions, and consultations	LR, RF, NB, SVM
Kocks et al (2023) [31]	Netherlands	Validation	Prospective observational study	To diagnose asthma, COPD[h], or asthma-COPD overlap	Data from Nivel Primary Care Database	116 cases, tested on 180 specialists from 9 countries (external validation) from patients aged ≥40	Symptoms, BMI, spirometry scores, smoking, diagnosis of chronic or allergic rhinitis, age at	Multinomial LR

Author, year	Country	Study type	Study design	Outcome	Dataset	Participants and inclusion criteria	Predictors	AI technique
						years, with complete data file	onset of respiratory disease	
LaFreniere et al (2016) [25]	Canada	Developmental	Case-control study, nested is unclear	To predict hypertension	EHR data from Canadian Primary Care Sentinel Surveillance Network	379,027 participants (185,371 cases with hypertension, 193,656 controls with no hypertension and with no 8 specific chronic conditions)	Demographics, BMI, blood pressure, laboratory results	NN
Lin et al (2023) [26]	China	Developmental	Retrospective cohort study	To identify COPD	Public health data from EHRs and electronic medical records of Chinese residents	1875 participants with lung symptoms or chronic lung disease	Demographics, smoking, BMI, chronic cough, shortness of breath, biofuel use, and family history. Based on the questionnaire for COPD	18 methods, including: Decision tree, LR, discriminant analysis (linear and quadratic), SVM, gradient boosting classifiers, NN, Gaussian process classifier, KNN[i], NB
Mariani et al (2021) [21]	Netherlands	Developmental	Retrospective cohort study	To diagnose asthma and COPD or asthma-COPD overlap	Data from Dutch primary care laboratory in Groningen	3659 participants with asthma or COPD from 2007 to 2017	Demographics, symptoms, diagnosis, medication, laboratory results, referrals, spirometry results	SVM, RF, KNN
Nemlander et al (2023) [27]	Sweden	Developmental	Nested case-control study	To identify nonmetastatic colorectal cancer	Regional administrative health care database from Västra Götaland Region	2681 participants (542 cases with a cancer diagnosis, 2139 controls)	Nonmetastatic colorectal cancer stage, number of GP consultations, diagnosis codes	Stochastic gradient boosting
Perveen et al (2016) [28]	Canada	Developmental	Retrospective cohort study	To classify diabetes mellitus in 3 adult age groups	EHR data from Canadian Primary Care Sentinel Surveillance Network	4678 participants (377 cases of diabetes, 4301 controls with no diabetes) with all documented risk factors	Demographics blood pressure, laboratory results	Decision tree, bagging, ADAboost
Singh et al (2022) [29]	United States	Developmental	Retrospective cohort study	To predict anterior segment vision-threatening disease (asVTD)	EHRs of the University of Michigan	2942 participants with anterior segment eye complaint (133 cases with asVTD, 2809 controls) with PC notes with ophthalmologist visit	Demographics, history of eye problems, symptoms, medication	Elastic net LR
Su et al (2019) [30]	China	Developmen-tal	Retrospec-tive cohort study	Top 100 diagnoses (within general diagnoses)	National Hospital Ambulatory Medical Care Survey and the National Ambulatory Medical Care Survey	Unknown number of participants, top 100 diagnosis selected from 2,000,000 records	Demographics, symptoms, past medical history	NN

[a]LR: logistic regression.
[b]RF: random forest.
[c]SVM: support vector machine.
[d]NB: naïve Bayes.
[e]XGBoost: extreme gradient boosting.
[f]NN: neural network.
[g]GP: general practitioner.
[h]COPD: chronic obstructive pulmonary disease.
[i]KNN: K-nearest neighbors.

AI Technique

All of the included studies performed at least 1 supervised AI technique (Table 1). The most used AI techniques were random forest (9 papers), logistic regression (7 papers), support vector machines (5 papers), boosting algorithms (5 papers), neural networks (5 papers), and naïve Bayes (4 papers).

Quality Assessment: Risk of Bias

None of the studies assessed by the PROBAST tool had a low risk of bias. We found a high risk of bias in 9 studies (9/15, 60%) and an unclear risk of bias in 6 studies (6/15, 40%; Table 2). In Multimedia Appendix 1, part 4 [18-67], the full assessment of the PROBAST tool can be found.

Table 2. Risk of bias per domain using the Prediction model Risk Of Bias ASsessment Tool.

	Participants	Predictors	Outcome	Analysis	Overall
Barnes [23]	Low	Low	Low	Unclear	Unclear
Briggs et al [22]	Low	Unclear	Unclear	Unclear	Unclear
Dhanda et al [32]	High	Low	Unclear	Unclear	High
Dros et al [24]	Low	Low	Unclear	High	High
Ellertsson et al [18]	Low	Unclear	Low	High	High
Ford et al [19]	Unclear	Low	Low	Unclear	Unclear
Jammeh et al [20]	High	Unclear	Unclear	Unclear	High
Kocks et al [31]	Low	Low	High	High	High
LaFreniere et al [25]	Unclear	Low	Unclear	Unclear	Unclear
Lin et al [26]	Unclear	Low	Unclear	High	High
Mariani et al [21]	Unclear	Low	Unclear	Unclear	Unclear
Nemlander et al [27]	High	Low	Unclear	Unclear	High
Perveen et al [28]	Low	Unclear	High	High	High
Singh et al [29]	Low	Low	Unclear	High	High
Su et al [30]	Unclear	Unclear	Unclear	Unclear	Unclear

The most significant source of bias was found in the analysis domain. The main reasons for the high risk of bias in this domain were the insufficient number of participants with the outcome (5/15, 33%) [18,24,26,29,31] and irrelevant model performance measures that were used to evaluate the model (2/15, 13%) [28,31]. The main reasons for an unclear risk of bias in the analysis domain were lack of clarity on how missing data were handled (10/15, 67%) [18-20,22,23,27-30,32], and on how the predictors and their assigned weights in the final model correspond to results from the reported multivariate analysis (9/15, 60%) [18,20,21,25-30]. Although measures of calibration are not part of the signaling questions of the PROBAST, we noticed that only 4 papers (4/15, 27%) [22,23,29,32] used calibration to assess the performance of the model.

The second significant source of bias was found in the outcome domain. The main reasons for the high risk of bias in this domain were the determination of the predictors with a prior knowledge of the outcome (1/15, 7%) [31] and not excluding the predictors from the outcome definition (2/15, 13%). For example, Perveen et al [28] included fasting glucose levels to predict diabetes and Kocks et al [31] included spirometry findings to predict asthma and COPD. The 2 main reasons for an unclear risk of bias in the outcome domain were lack of clarity on the time interval between the outcome and the predictors (9/15, 60%) [20,24-30,32] and the lack of clarity on the outcome definition (7/15, 47%) [20-22,24,26,28,30].

The third domain with risk of bias was the participants domain. A high risk of bias in the participants domain was found because inclusion and exclusion criteria were not appropriate in 2 studies (2/15, 13%) as both studies excluded participants at high risk of the outcome [27,32]. Another reason for the high risk of bias was a nonappropriate data source that was used in 1 study [20] because the authors described the study as a case-control study although the study was not nested as recommended in the PROBAST guidelines [69,70]. The predictors domain was the domain with the lowest risk of bias. The lack of clarity that resulted in an unclear risk of bias covered mainly insufficient information on whether the predictors were defined and assessed in a similar way for all participants (4/15, 27%) [18,20,22,30].

Applicability

Overall, we found an unclear concern for applicability in 10 papers (10/15, 67%) and a low concern for applicability in 5 papers (5/15, 33%; Table 3). The unclear concern for applicability to our research question was mainly noticed in the outcome domain due to a lack of clarity in reporting the time interval between the outcomes and predictors (8/15, 53%) [20,25-30,32].

Table 3. Applicability per domain using the Prediction model Risk Of Bias Assessment Tool.

	Participants	Predictors	Outcome	Overall
Barnes et al [23]	Low	Low	Low	Low
Briggs et al [22]	Low	Unclear	Unclear	Unclear
Dhanda et al [32]	Low	Low	Unclear	Unclear
Dros et al [24]	Low	Low	Low	Low
Ellertsson et al [18]	Low	Unclear	Low	Unclear
Ford et al [19]	Low	Low	Low	Low
Jammeh et al [20]	Low	Unclear	Unclear	Unclear
Kocks et al [31]	Low	Low	Low	Low
LaFreniere et al [25]	Unclear	Low	Unclear	Unclear
Lin et al [26]	Unclear	Low	Unclear	Unclear
Mariani et al [21]	Low	Low	Low	Low
Nemlander et al [27]	Low	Low	Unclear	Unclear
Perveen et al [28]	Low	Unclear	Unclear	Unclear
Singh et al [29]	Low	Low	Unclear	Unclear
Su et al [30]	Unclear	Unclear	Unclear	Unclear

In the predictors domain we also found an unclear concern for applicability due to the lack of clarity in the definition of the included predictors (5/15, 33%) [18,20,22,28,30]. For example, 1 paper lacked information on how notes were annotated before they were used as predictors in the model [18]. The unclear concern for applicability in the participants' domain was mainly due to lack of information on inclusion and exclusion criteria (3/15, 20%) [25,26,30].

Discussion

Principal Results

We systematically reviewed the literature for studies about AI-based diagnostic prediction models for PC. These models were developed with different data sources, such as questionnaire data, secondary care data, or EHR data. Only 15 out of 49 models were developed using data from EHRs. Most of the models using EHR data focused on just 1 chronic condition. Merely 2 papers tested the model in a PC setting. All of the included studies performed at least 1 supervised AI technique, most often with random forest or logistic regression. Evaluation with the PROBAST guidelines showed an unclear to high risk of bias for all EHR papers. In most of the papers, we found unclear concerns about the applicability to our research question.

Comparison With Prior Work

To the best of our knowledge, only 2 reviews evaluated the risk of bias in clinical prediction models on a wide range of diseases in PC studies [15,16]. Most of the included studies in these reviews showed a high to unclear risk of bias, which is in line with our findings [15,16]. However, there appear to be differences in grading compared with Abdulazeem et al [16]. They considered incomplete reporting and the absence of external validation a high risk of bias, whereas in our systematic review, these points were considered as an unclear risk of bias and no risk of bias, respectively. The study by Abbasgholizadeh et al [15] did not report details on the reasons they coded subdomains as high or unclear risk of bias, for which reason we are unable to make a formal comparison with our results.

Systematic reviews evaluating AI-based clinical prediction models in other medical fields have followed the same grading criteria as we did and found similar flaws in the analysis domain as we did in our systematic review [71,72]. These similarities include the unjustified small sample size in EHR studies, inappropriate evaluation in the performance measures, and flaws in handling of missing data [71-73].

The most used AI techniques were random forest, logistic regression, support vector machines, boosting algorithms, and neural networks. In previous systematic reviews, random forest and support vector machines are also more often found as most used methodology [16,71,73-75]. This might be explained by the well-described strong performance and ease of interpretability of random forests and support vector machines, particularly when working with lower-quality structured data. Most PC EHRs are primarily used for clinical purposes, with secondary purposes for research [69]. Thus, the challenges associated with using such EHRs to develop prediction models have been widely documented and include missing values and inconsistencies in data entry [13]. These challenges are inherent to the data and should be addressed at the preprocessing stage. We did not find papers that used generative AI methods (such as large language models). Our retrieved papers developed or validated tools based only on structured data (numbers or codes such as laboratory results, vital signs, and diagnosis codes from International Classification of Primary Care or *ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*]) rather than unstructured data or written text, where large language models work well on. Literature found it valuable for the performance of the model to use unstructured data together with structured data for prognostic prediction models [76,77].

We think that future studies about diagnostic prediction tools will increasingly use generative AI methods, although it is still difficult to integrate them into clinical workflows [78].

In general, studies analyzing EHRs are subject to a high risk of bias, because these data are collected for clinical rather than research purposes [69]. Hence, clinical prediction models developed on EHRs are more difficult to reproduce and generalize, given the heterogeneity of coding systems and database infrastructures [16]. In line with models analyzed in previous studies [15,16,73], most of the clinical prediction models were not externally validated. Most of the studies developed in PC were performed in high-income countries and may not have taken into account regional or global differences in the availability of certain predictors [14-16]. For example, some predictors may not be easy to obtain in PC settings in low-income countries (eg, spirometry results for the prediction of asthma or COPD). Furthermore, the lack of stratified analyses in most studies implies that we cannot draw conclusions about how diagnostic models perform across different equity groups. Together, all these factors limit the generalizability of the clinical prediction models.

Strengths and Limitations

The main strength of the study is the extensive search strategy with no date limit in a large and diverse range of studies on AI prediction models in PC. Not including "EHR" in the search strategy added rigor to our study as a recent review suggests that important papers could have been missed when we included EHRs in the search strategy [13]. A second strength is that the findings on the risk of bias were carefully assessed by 2 independent reviewers (LH and AC) with experience in clinical PC, and the conflicts were discussed with other experts in the field of PC and AI. Unlike previous systematic reviews that found a high proportion of studies with a high concern of applicability to the research question [72], we noticed no high concern for applicability in any study. We believe that the findings shared in our review are highly reliable in highlighting the current situation of AI studies in PC using EHRs.

The main limitation of this study is the broad definition of the terminology for the search strategy, which may have prevented us from capturing all relevant studies. For example, we included all studies that used ML and deep learning techniques. Given the lack of a widely accepted definition of AI, other reviews use other criteria for AI or ML [71,73,75]. Similarly, given our definition of diagnostic prediction models, we considered a diagnostic prediction model to be a model that predicts a health condition during a GP's consultation. As a result, multiple prediction models that identified a disease in a database were excluded. The second limitation is the use of the PROBAST guidelines to determine the risk of bias and applicability in evaluating AI prediction models. Although the PROBAST guidelines are highly detailed and reliable in evaluating clinical prediction models [33], PROBAST has been criticized for being less specific and less applicable for AI-based models than traditional statistical methods. Considering this criticism, a protocol on the extension of PROBAST into PROBAST-Artificial Intelligence (PROBAST-AI) has been published with the aim to develop a PROBAST-AI tool to better support evaluation of prediction model studies that applied AI [3]. The PROBAST-AI tool has not yet been published.

Future Research and Practical Implications

The relevance of the applicability of prediction models in clinical practice should be the priority when developing clinical prediction models, as stated in a number of standardized frameworks designed for prediction model developers [79,80]. We found that only 2 models were tested in PC settings. Moreover, most studies included in this review predict chronic conditions. This is also seen in previous reviews evaluating clinical prediction models in PC [14,16]. However, in general, chronic conditions are not known to be difficult to diagnose in PC. Two examples from our included papers are the diagnosis of hypertension predicted on the variable high blood pressure [25] and the diagnosis of diabetes predicted on the variable high glucose levels [28]. These predictions might not be as useful in clinical practice, even if the model performance metrics are excellent. Nevertheless, chronic conditions are highly prevalent in PC and for conditions that are influenced by several and complex factors, prediction models may facilitate the diagnostic process for the GP. As most tools focused on predicting 1 condition, GPs would have to use many prediction tools side by side to predict the correct diagnosis in daily practice. All these findings highlight that involving more practicing GPs and asking what they need are important in developing clinical prediction models with a higher success rate of clinical implementation. We recommend involving relevant stakeholders in the early stages of the development of a new model.

To improve the methodology in future studies, our findings suggest that a special focus is required on reporting areas such as methods for internal validation, appropriate inclusion of participants, and a proper sample size calculation. A high risk of bias mainly found in the analysis and outcome domains should be alarming as this questions the methodology of the included papers. We found an unclear risk of bias and unclear concern for applicability in more than half of the included studies, mainly related to poor reporting, for example, about missing data. Missing data is known as a large challenge for EHR data [13], and extra attention should therefore be paid to reporting this. Researchers can benefit from the use of the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement [81] and PROBAST guidelines in communicating their findings [3], particularly now that the TRIPOD-AI extension is released [82]. To enhance the applicability of the prediction model, we highlight the importance of clear reporting on the time interval between predictors and outcome, a clear definition of the outcome and predictors, and a clear description of the inclusion and exclusion criteria. Differences in recording between EHRs might lower the performance of the model in the external validation step, and external validation is a crucial step for

generalizable and reliable models [76]. However, we found only 2 papers that performed external validation.

Conclusions

AI-based prediction models using EHR data are not yet ready for implementation into PC daily practice. The number of studies found was limited, and reproducibility and generalizability were insufficient. For a diagnostic prediction model to be used in PC, it is important that GPs and relevant stakeholders are involved in the development, that the model is externally validated, and that it is appropriately recorded.

Acknowledgments

This study was funded by ZonMw file number: 839150005. The authors would like to thank Lori van den Hurk for her help with the screening process.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy, screening guidance, table with all included papers, PROBAST (Prediction Model Risk of Bias Assessment Tool) checklist, and references of appendix.
[DOCX File (Microsoft Word File), 290 KB-Multimedia Appendix 1]

Checklist 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe File), 70 KB-Checklist 1]

References

1. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. Apr 2021;132:142-145. [doi: 10.1016/j.jclinepi.2021.01.009] [Medline: 33775387]
2. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. Feb 23, 2009;338:b375. [doi: 10.1136/bmj.b375] [Medline: 19237405]
3. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. Jul 9, 2021;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
4. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. Jun 14, 2017;38(23):1805-1814. [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
5. Liyanage H, Liaw ST, Jonnagaddala J, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform. Aug 2019;28(1):41-46. [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
6. Mistry P. Artificial intelligence in primary care. Br J Gen Pract. Sep 2019;69(686):422-423. [doi: 10.3399/bjgp19X705137] [Medline: 31467001]
7. Summerton N, Cansdale M. Artificial intelligence and diagnosis in general practice. Br J Gen Pract. Jul 2019;69(684):324-325. [doi: 10.3399/bjgp19X704165] [Medline: 31249070]
8. Lin S. A clinician's guide to artificial intelligence (AI): Why and how primary care should lead the health care AI revolution. J Am Board Fam Med. 2022;35(1):175-184. [doi: 10.3122/jabfm.2022.01.210226] [Medline: 35039425]
9. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med. Oct 2017;6(10):2453-2460. [doi: 10.1002/cam4.1183] [Medline: 28941187]
10. Burnett B, Zhou SM, Brophy S, et al. Machine learning in colorectal cancer risk prediction from routinely collected data: a review. Diagnostics (Basel). Jan 13, 2023;13(2):301. [doi: 10.3390/diagnostics13020301] [Medline: 36673111]
11. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. BMJ Open. Oct 27, 2020;10(10):e037860. [doi: 10.1136/bmjopen-2020-037860] [Medline: 33109649]
12. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. Sep 29, 2016;375(13):1216-1219. [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
13. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. Jan 2017;24(1):198-208. [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
14. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. Ann Fam Med. May 2020;18(3):250-258. [doi: 10.1370/afm.2518] [Medline: 32393561]

15. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res. Sep 3, 2021;23(9):e29839. [doi: 10.2196/29839] [Medline: 34477556]
16. Abdulazeem H, Whitelaw S, Schauberger G, Klug SJ. A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. PLoS One. 2023;18(9):e0274276. [doi: 10.1371/journal.pone.0274276] [Medline: 37682909]
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Ellertsson S, Loftsson H, Sigurdsson EL. Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy. Scand J Prim Health Care. Dec 2021;39(4):448-458. [doi: 10.1080/02813432.2021.1973255] [Medline: 34585629]
19. Ford E, Rooney P, Oliver S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Med Inform Decis Mak. Dec 2, 2019;19(1):248. [doi: 10.1186/s12911-019-0991-9] [Medline: 31791325]
20. Jammeh EA, Carroll CB, Pearson SW, et al. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open. Jul 2018;2(2):bjgpopen18X101589. [doi: 10.3399/bjgpopen18X101589] [Medline: 30564722]
21. Mariani S, Metting E, Lahr MMH, Vargiu E, Zambonelli F. Developing an ML pipeline for asthma and COPD: the case of a Dutch primary care service. Int J of Intelligent Sys. Nov 2021;36(11):6763-6790. [doi: 10.1002/int.22568]
22. Briggs E, de Kamps M, Hamilton W, Johnson O, McInerney CD, Neal RD. Machine learning for risk prediction of oesophago-gastric cancer in primary care: comparison with existing risk-assessment tools. Cancers (Basel). Oct 14, 2022;14(20):5023. [doi: 10.3390/cancers14205023] [Medline: 36291807]
23. Barnes DE, Zhou J, Walker RL, et al. Development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. J Am Geriatr Soc. Jan 2020;68(1):103-111. [doi: 10.1111/jgs.16182] [Medline: 31612463]
24. Dros JT, Bos I, Bennis FC, et al. Detection of primary Sjögren's syndrome in primary care: developing a classification model with the use of routine healthcare data and machine learning. BMC Prim Care. Aug 9, 2022;23(1):199. [doi: 10.1186/s12875-022-01804-w] [Medline: 35945489]
25. LaFreniere D, Zulkernine F, Barber D, Martin K. Using machine learning to predict hypertension from a clinical dataset. Presented at: 2016 IEEE Symposium Series on Computational Intelligence (SSCI; Dec 6-9, 2016; Athens, Greece. [doi: 10.1109/SSCI.2016.7849886]
26. Lin X, Lei Y, Chen J, et al. A case-finding clinical decision support system to identify subjects with chronic obstructive pulmonary disease based on public health data. Tsinghua Sci Technol. Jun 2023;28(3):525-540. [doi: 10.26599/TST.2022.9010010]
27. Nemlander E, Ewing M, Abedi E, et al. A machine learning tool for identifying non-metastatic colorectal cancer in primary care. Eur J Cancer. Mar 2023;182(100-6):100-106. [doi: 10.1016/j.ejca.2023.01.011] [Medline: 36758474]
28. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016;82:115-121. [doi: 10.1016/j.procs.2016.04.016]
29. Singh K, Thibodeau A, Niziol LM, et al. Development and validation of a model to predict anterior segment vision-threatening eye disease using primary care clinical notes. Cornea. Aug 1, 2022;41(8):974-980. [doi: 10.1097/ICO.0000000000002877] [Medline: 34620768]
30. Su G, Wen J, Zhu Z, et al. An approach of integrating domain knowledge into data-driven diagnostic model. Stud Health Technol Inform. Aug 21, 2019;264:1594-1595. [doi: 10.3233/SHTI190551] [Medline: 31438248]
31. Kocks JWH, Cao H, Holzhauer B, et al. Diagnostic performance of a machine learning algorithm (asthma/chronic obstructive pulmonary disease [COPD] differentiation classification) tool versus primary care physicians and pulmonologists in asthma, COPD, and asthma/COPD overlap. J Allergy Clin Immunol Pract. May 2023;11(5):1463-1474. [doi: 10.1016/j.jaip.2023.01.017] [Medline: 36716998]
32. Dhanda G, Asham M, Shanks D, et al. Adaptation and external validation of pathogenic urine culture prediction in primary care using machine learning. Ann Fam Med. 2023;21(1):11-18. [doi: 10.1370/afm.2902] [Medline: 36690486]
33. Moons KGM, Altman DG, Reitsma JB, Collins GS, Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Development Initiative. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. Adv Anat Pathol. Sep 2015;22(5):303-305. [doi: 10.1097/PAP.0000000000000072] [Medline: 26262512]
34. Ahmed MM, Sayed AM, El Abd D, et al. Diagnosis of coronavirus disease 2019 and the potential role of deep learning: insights from the experience of Cairo University Hospitals. J Int Med Res. Jul 2022;50(7). [doi: 10.1177/03000605221109392]

35. Ahmed MM, Sayed AM, Khafagy GM, et al. Accuracy of the traditional COVID-19 phone triaging system and phone triage-driven deep learning model. J Prim Care Community Health. 2022;13(21501319221113544):21501319221113544. [doi: 10.1177/21501319221113544] [Medline: 35869692]
36. Basta M, John Simos N, Zioga M, et al. Personalized screening and risk profiles for mild cognitive impairment via a machine learning framework: implications for general practice. Int J Med Inform. Feb 2023;170(104966):104966. [doi: 10.1016/j.ijmedinf.2022.104966] [Medline: 36542901]
37. Blanes-Vidal V, Lindvig KP, Thiele M, Nadimi ES, Krag A. Artificial intelligence outperforms standard blood-based scores in identifying liver fibrosis patients in primary care. Sci Rep. Feb 21, 2022;12(1):2914. [doi: 10.1038/s41598-022-06998-8] [Medline: 35190650]
38. Braido F, Santus P, Corsico AG, et al. Chronic obstructive lung disease "expert system": validation of a predictive tool for assisting diagnosis. Int J Chron Obstruct Pulmon Dis. 2018;13(1747-53):1747-1753. [doi: 10.2147/COPD.S165533] [Medline: 29881264]
39. Brooks GJ, Ashton RE, Pethybridge RJ. DERMIS: a computer system for assisting primary-care physicians with dermatological diagnosis. Br J Dermatol. Dec 1992;127(6):614-619. [doi: 10.1111/j.1365-2133.1992.tb14875.x] [Medline: 1476920]
40. Cruz-Gutiérrez V, Posada-Zamora MA, Sánchez-López A. An efficient expert system for diabetes with a Bayesian inference engine. Adv Soft Comput, Micai. 2016;10062:54-64. [doi: 10.1007/978-3-319-62428-0_5]
41. Dong W, Tse TYE, Mak LI, et al. Non-laboratory-based risk assessment model for case detection of diabetes mellitus and pre-diabetes in primary care. J Diabetes Investig. Aug 2022;13(8):1374-1386. [doi: 10.1111/jdi.13790] [Medline: 35293149]
42. Exarchos TP, Rigas G, Bibas A, et al. Mining balance disorders' data for the development of diagnostic decision support systems. Comput Biol Med. Oct 1, 2016;77(240-8):240-248. [doi: 10.1016/j.compbiomed.2016.08.016] [Medline: 27619194]
43. Faris H, Habib M, Faris M, Elayan H, Alomari A. An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine. Inform Med Unlocked. 2021;23:100513. [doi: 10.1016/j.imu.2021.100513]
44. Farmer N, Schilstra MJ. A knowledge-based diagnostic clinical decision support system for musculoskeletal disorders of the shoulder for use in a primary care setting. Shoulder Elbow. Apr 2012;4(2):141-151. [doi: 10.1111/j.1758-5740.2011.00165.x]
45. Farmer N. An update and further testing of a knowledge-based diagnostic clinical decision support system for musculoskeletal disorders of the shoulder for use in a primary care setting. J Eval Clin Pract. Oct 2014;20(5):589-595. [doi: 10.1111/jep.12153] [Medline: 24828447]
46. Grill E, Groezinger M, Feil K, Strupp M. Developing and implementing diagnostic prediction models for vestibular diseases in primary care. Stud Health Technol Inform. 2016;228(735-9):735-739. [doi: 10.3233/978-1-61499-678-1-735] [Medline: 27577483]
47. Harabor V, Mogos R, Nechita A, et al. Machine learning approaches for the prediction of hepatitis B and C seropositivity. Int J Environ Res Public Health. Jan 29, 2023;20(3):2380. [doi: 10.3390/ijerph20032380] [Medline: 36767747]
48. Heckerling PS, Canaris GJ, Flach SD, Tape TG, Wigton RS, Gerber BS. Predictors of urinary tract infection based on artificial neural networks and genetic algorithms. Int J Med Inform. Apr 2007;76(4):289-296. [doi: 10.1016/j.ijmedinf.2006.01.005] [Medline: 16469531]
49. Hejlesen OK, Olesen KG, Dessau R, Beltoft I, Trangeled M. Decision support for diagnosis of Lyme disease. Stud Health Technol Inform. 2005;116(205-10):205-210. [Medline: 16160260]
50. Koch Nogueira PC, Venson AH, de Carvalho MFC, Konstantyner T, Sesso R. Symptoms for early diagnosis of chronic kidney disease in children—a machine learning-based score. Eur J Pediatr. Aug 2023;182(8):3631-3637. [doi: 10.1007/s00431-023-05032-x] [Medline: 37233777]
51. Liu X, Zhang W, Zhang Q, et al. Development and validation of a machine learning-augmented algorithm for diabetes screening in community and primary care settings: a population-based study. Front Endocrinol. 2022;13:1043919. [doi: 10.3389/fendo.2022.1043919]
52. Maizels M, Wolfe WJ. An expert system for headache diagnosis: the Computerized Headache Assessment tool (CHAT). Headache. Jan 2008;48(1):72-78. [doi: 10.1111/j.1526-4610.2007.00918.x] [Medline: 17868352]
53. Pasic A, Pasic L, Pasic A. The artificial intelligence based diagnostic assistant—AIDA. Presented at: 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON); Jun 14-16, 2022:114-119; Palermo, Italy. [doi: 10.1109/MELECON53508.2022.9843070]

54. Rahimi SA, Kolahdoozi M, Mitra A, et al. Quantum-inspired interpretable ai-empowered decision support system for detection of early-stage rheumatoid arthritis in primary care using scarce dataset. Mathematics. Feb 2022;10(3):496. [doi: 10.3390/math10030496]
55. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv. Preprint posted online on Jun 27, 2018. [doi: 10.48550/arXiv.1806.10698]
56. Salmeron JL, Rahimi SA, Navali AM, Sadeghpour A. Medical diagnosis of Rheumatoid Arthritis using data driven PSO–FCM with scarce datasets. Neurocomputing. Apr 2017;232:104-112. [doi: 10.1016/j.neucom.2016.09.113]
57. Sanaeifar A, Eslami S, Ahadi M, Kahani M, Vakili Arki H. DxGenerator: An improved differential diagnosis generator for primary care based on MetaMap and semantic reasoning. Methods Inf Med. Dec 2022;61(5-06):174-184. [doi: 10.1055/a-1905-5639] [Medline: 35858654]
58. Shen EX, Lord A, Doecke JD, et al. A validated risk stratification tool for detecting high-risk small bowel Crohn's disease. Aliment Pharmacol Ther. Jan 2020;51(2):281-290. [doi: 10.1111/apt.15550] [Medline: 31769537]
59. Suárez-Araujo CP, García Báez P, Cabrera-León Y, et al. A real-time clinical decision support system, for mild cognitive impairment detection, based on a hybrid neural architecture. Comput Math Methods Med. 2021;2021(5545297):5545297. [doi: 10.1155/2021/5545297] [Medline: 34257699]
60. Tsoi KKF. Application of artificial intelligence on a symptom diagnostic platform for telemedicine a pilot case study. Presented at: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC); Oct 6-9, 2019:806-813; Bari, Italy. URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8906183 [Accessed 2025-08-12] [doi: 10.1109/SMC.2019.8914229]
61. Velickovski F, Ceccaroni L, Roca J, et al. Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. J Transl Med. Nov 28, 2014;12(Suppl 2):25471545. [doi: 10.1186/1479-5876-12-S2-S9] [Medline: 25471545]
62. Velu SR, Ravi V, Tabianan K. Data mining in predicting liver patients using classification model. Health Technol. Nov 2022;12(6):1211-1235. [doi: 10.1007/s12553-022-00713-3]
63. Xiao T, Wang C, Yang M, et al. Use of virus genotypes in machine learning diagnostic prediction models for cervical cancer in women with high-risk human papillomavirus infection. JAMA Netw Open. Aug 1, 2023;6(8):e2326890. [doi: 10.1001/jamanetworkopen.2023.26890] [Medline: 37531108]
64. Yoshihara A, Yoshimura Noh J, Inoue K, et al. Prediction model of Graves' disease in general clinical practice based on complete blood count and biochemistry profile. Endocr J. Sep 28, 2022;69(9):1091-1100. [doi: 10.1507/endocrj.EJ21-0741] [Medline: 35387949]
65. Yu C, Peng YY, Liu L, Wang X, Xiao Q. Leukemia can be effectively early predicted in routine physical examination with the assistance of machine learning models. J Healthc Eng. 2022;2022:8641194. [doi: 10.1155/2022/8641194] [Medline: 36465253]
66. Zardab M, Balarajah V, Banerjee A, et al. Differentiating ductal adenocarcinoma of the pancreas from benign conditions using routine health records: a prospective case-control study. Cancers (Basel). Jan 2023;15(1):280. [doi: 10.3390/cancers15010280]
67. Zhang H, Yin M, Liu Q, et al. Machine and deep learning-based clinical characteristics and laboratory markers for the prediction of sarcopenia. Chin Med J. 2023;136(8):967-973. [doi: 10.1097/CM9.0000000000002633]
68. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. Oct 2014;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
69. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. Jan 1, 2019;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
70. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. Jan 1, 2019;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
71. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. Oct 20, 2021;375:n2281. [doi: 10.1136/bmj.n2281] [Medline: 34670780]
72. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review. Acad Emerg Med. Feb 2021;28(2):184-196. [doi: 10.1111/acem.14190] [Medline: 33277724]
73. Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. Feb 2023;154:8-22. [doi: 10.1016/j.jclinepi.2022.11.015] [Medline: 36436815]

74. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. Apr 4, 2019;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
75. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. Jun 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
76. Seinen TM, Fridgeirsson EA, Ioannou S, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc. Jun 14, 2022;29(7):1292-1302. [doi: 10.1093/jamia/ocac058] [Medline: 35475536]
77. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak. Oct 29, 2020;20(1):280. [doi: 10.1186/s12911-020-01297-6] [Medline: 33121479]
78. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. Sep 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
79. Liu Y, Chen PH, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA. Nov 12, 2019;322(18):1806-1816. [doi: 10.1001/jama.2019.16489] [Medline: 31714992]
80. Sujan M, Smith-Frazer C, Malamateniou C, et al. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Inform. Jun 2023;30(1):e100749. [doi: 10.1136/bmjhci-2023-100749] [Medline: 37364922]
81. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol. Feb 2015;68(2):134-143. [doi: 10.1016/j.jclinepi.2014.11.010] [Medline: 25579640]
82. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]

Abbreviations

AI: artificial intelligence
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
COPD: chronic obstructive pulmonary disease
EHR: electronic health record
GP: general practitioner
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
ML: machine learning
PC: primary care
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROBAST-AI: Prediction Model Risk of Bias Assessment Tool-Artificial Intelligence
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by Arriel Benis; peer-reviewed by Brendan Delaney, Dillon Chrimes, Douglas Manuel; submitted 03.06.2024; final revised version received 06.06.2025; accepted 09.06.2025; published 22.08.2025

Please cite as:
Hunik L, Chaabouni A, van Laarhoven T, Olde Hartman TC, Leijenaar RTH, Cals JWL, Uijen AA, Schers HJ
Diagnostic Prediction Models for Primary Care, Based on AI and Electronic Health Records: Systematic Review
JMIR Med Inform 2025;13:e62862
URL: https://medinform.jmir.org/2025/1/e62862
doi: 10.2196/62862