JMIR Publications
The Leading eHealth Publisher

Diagnostic prediction models for primary care, based on artificial intelligence and electronic health records: a systematic review

Liesbeth Hunik, Asma Chaabouni, Twan van Laarhoven, Tim C olde Hartman, Ralph TH Leijenaar, Jochen WL Cals, Annemarie A Uijen, Henk J Schers

Submitted to: Journal of Medical Internet Research
on: June 03, 2024

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript **5**
Supplementary Files **31**
Multimedia Appendixes **32**
Multimedia Appendix 1 ..... **32**
CONSORT (or other) checklists **33**
CONSORT (or other) checklist 0 ..... **33**

Preprint
JMIR Publications

Diagnostic prediction models for primary care, based on artificial intelligence and electronic health records: a systematic review

Liesbeth Hunik[1] MD, MSc; Asma Chaabouni[1] MD, MSc; Twan van Laarhoven[2] PhD; Tim C olde Hartman[1] PhD, MD; Ralph TH Leijenaar[3] PhD; Jochen WL 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 NL
[2]Institute for Computing and Information Science, Radboud University Nijmegen NL
[3]Department of Family Medicine, Care and Public Health Research Institute, Maastricht University Maastricht NL

Corresponding Author:
Liesbeth Hunik MD, MSc
Department of primary and community care, Research Institute for Medical Innovation, Radboudumc
Geert Grooteplein Zuid 21
NIjmegen
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 (EHR) 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: To provide an overview of diagnostic prediction models based on AI and EHR in primary care and to evaluate the content of each model, 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 primary care EHRs and developing or testing a diagnostic prediction model for health conditions. Two independent reviewers used a standardised 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 one chronic healthcare condition (n=11). From the 15 papers, 13 described a study that developed a diagnostic prediction model and 2 described a study that externally validated and tested the model in a primary care 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, high risk of bias was found in 9 papers. Biases covered an unjustified small sample size (n=5), not excluding predictors from the outcome definition (n=2) and the inappropriate evaluation of the performance measures (n=2). Unclear risk of bias was found in 6 papers, as no information was provided on the handling of missing data (n=10) and no results were reported from the multivariate analysis (n=9). Applicability was unclear in 10 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 primary care focused on one chronic condition. Only two papers tested the model in a primary care 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 primary care.

(JMIR Preprints 03/06/2024:62862)
DOI: https://doi.org/10.2196/preprints.62862

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Diagnostic prediction models for primary care, based on artificial intelligence and electronic health records: a systematic review

Liesbeth Hunik[1], MSc, MD (ORCID-ID: 0000-0003-4791-8823)
Asma Chaabouni[1], MSc, MD (ORCID-ID: 0000-0001-5731-9993)
Twan van Laarhoven[2], MSc, PhD (ORCID-ID: 0000-0001-7597-0579)
Tim C olde Hartman[1], MD, PhD (ORCID-ID: 0000-0003-2078-1206)
Ralph TH Leijenaar[3], MSc, PhD (ORCID-ID: 0000-0001-8642-9657)
Jochen WL Cals[3], MD, PhD, professor of general practice (ORCID-ID: 0000-0001-9550-5674)
Annemarie A Uijen[1], MD, PhD (ORCID-ID: 0000-0002-7703-6250)
Henk J Schers[1], MD, PhD, professor of general practice (ORCID-ID: 0000-0002-9362-9451)

[1]Department of primary and community care, Research Institute for Medical Innovation, Radboudumc, Nijmegen, 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, Netherlands

Corresponding author

Liesbeth Hunik (ORCID-ID: 0000-0003-4791-8823)
Liesbeth.hunik@radboudumc.nl
0031 24 365 56 38

Funding

This study was funded by ZonMw file number: 839150005

Word count

3457 words

Keywords

Primary care, electronic health records, artificial intelligence, diagnostic decision support

Preprint JMIR Publications

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 (EHR) 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

To provide an overview of diagnostic prediction models based on AI and EHR in primary care and to evaluate the content of each model, 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 primary care EHRs and developing or testing a diagnostic prediction model for health conditions. Two independent reviewers used a standardised 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 one chronic healthcare condition (n=11). From the 15 papers, 13 described a study that developed a diagnostic prediction model and 2 described a study that externally validated and tested the model in a primary care 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, high risk of bias was found in 9 papers. Biases covered an unjustified small sample size (n=5), not excluding predictors from the outcome definition (n=2) and the inappropriate evaluation of the performance measures (n=2). Unclear risk of bias was found in 6 papers, as no information was provided on the handling of missing data (n=10) and no results were reported from the multivariate analysis (n=9). Applicability was unclear in 10 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 primary care focused on one chronic condition. Only two papers tested the model in a primary care 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 primary care.

Preprint JMIR Publications

Introduction

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].

Clinical prediction models used to be built on data from large databases, such as data collected for research purposes, claims data or data from electronic health records (EHR) [9,10]. EHR data consists of structured data, which is data in standardized format, and unstructured data, which is free text data. Primary care (PC) EHR data provides 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 [11]. For that reason, AI has the potential to transform EHR data into a valuable tool for predicting diagnosis in daily primary care practice.

Reviews on the value of AI in PC are scarce and previous research had different aims. For example, Kueper et al. provided an overview of diagnostic prediction models based on AI in PC [12]. 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 [13] or focused on diagnostic and prognostic models about a specific health care condition [14]. 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 EHR in PC was largely lacking in current literature. Therefore, we systematically reviewed the literature in order to provide an overview of the studies conducted and critically evaluate the way these AI-based diagnostic prediction models were developed and validated.

Methods

We performed a systematic review according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [15]. The protocol for this study was registered in PROSPERO (nr: CRD42022320002).

Search strategy

Our search was adapted from the search strategy developed by Kueper et al. [12]. 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 part 1). EHR was not part of the search strategy, because literature suggests that we might miss important studies when including EHR or related terms in the search strategy [11]. 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 28th, 2023. We focused on intervention and observational studies. We excluded systematic reviews, meta-analyses, case studies,

editorials, protocols, and conference posters/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.

Study selection

Four inclusion criteria were used to select the papers. First, *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. Second, *diagnostic prediction model*; models had to predict a health condition applicable during a GPs consultation. Prediction models that identified a disease in a database, rather than predicting a disease for an individual, were excluded. Third, *AI*; this included all machine learning and deep learning techniques. We directed our focus to data driven prediction models without using medical images as input data. Fourth, *EHR-based data;* EHR data had to be used for the development or validation of the model. EHR was defined as PC data from EHRs, medical records or clinical notes. See Multimedia Appendix part 2 for the full screening guidance.

Title and abstract screening was done in management software Rayyan (rayyan.ai) by two 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, resolved by a third reviewer (AU). Backward citation searching was conducted on the included papers and finished on November 7th, 2023.

Data extraction

Data extraction of included papers was done by two independent reviewers (LH and AC). They used a standardised data extraction form adapted from the CHARMS checklist (CHecklist for critical

Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) [16]. 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/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 signalling questions divided into 4 domains (participants, predictors, outcome, and analysis) [17,18]. Overall judgement (i.e., low, unclear, or high) of risk of bias is based on the 4 domains. If one domain is considered to have a high risk of bias, the overall judgement is scored as a high risk of bias. If at least one domain is considered to have an unclear risk of bias (without a domain with high risk of bias), the overall judgement is scored as unclear risk of bias. Applicability concern was rated based on 3 domains (participants, predictors, and outcome) and an overall judgement of applicability (i.e., low, unclear, or high) was also given with the same approach as the risk of bias scoring. Applicability evaluation depends on the review question [17], we translated applicability assessment as usability of the diagnostic prediction model in a PC setting. Conflicts in data extraction between the two reviewers were resolved by discussion and if no consensus was reached, were resolved by a third reviewer (TvL).

Preprint JMIR Publications

Results

Description of included studies

We retrieved 10,657 records using our search strategy. After duplicate removal, we conducted title and abstract screening on 7,146 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 part 3. The data used in these 34 papers were collected from different sources, including secondary care data sets (n = 17), questionnaires (n = 4) and the knowledge of different health care providers (n = 5).

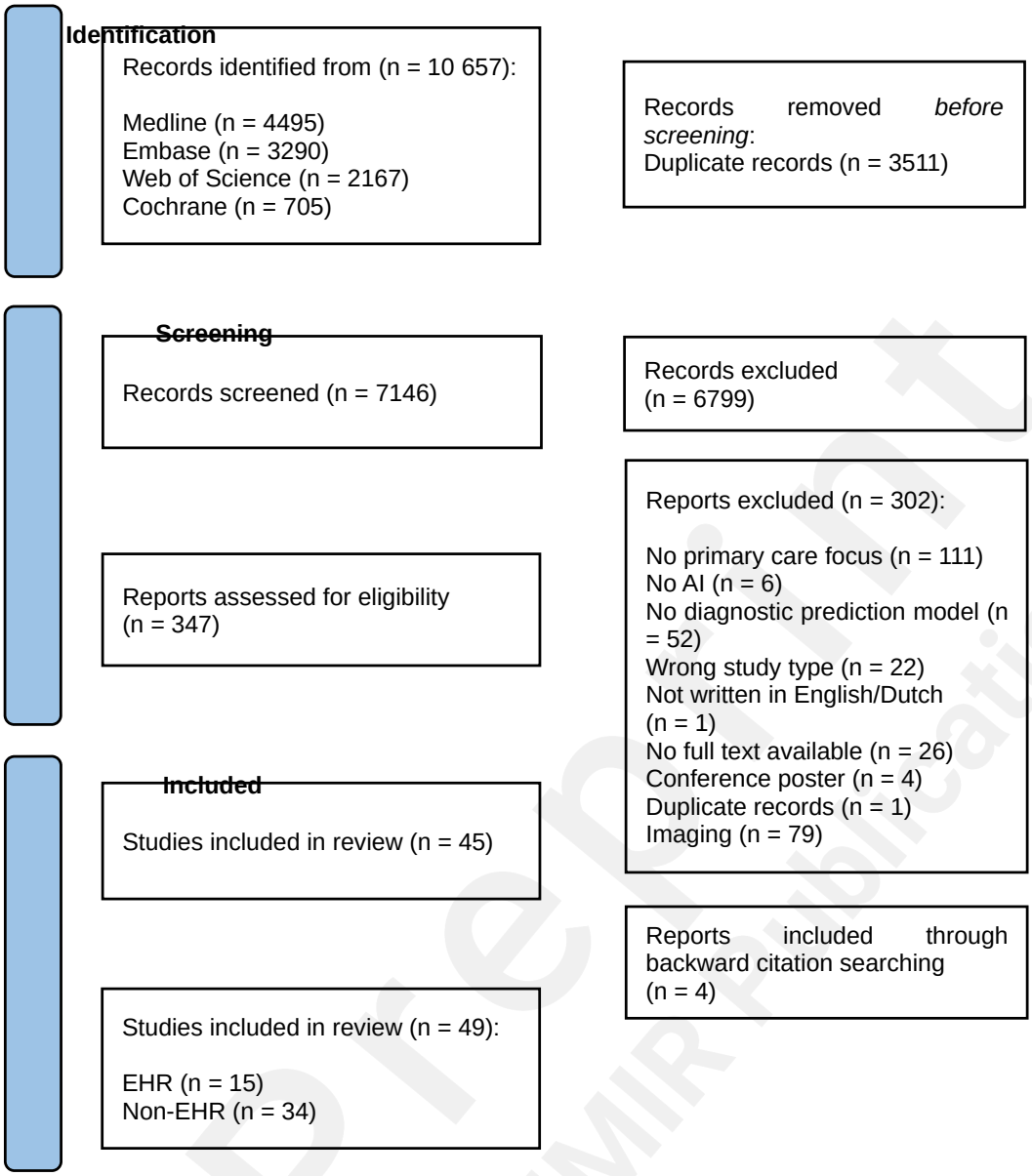
Overview of the EHR-based papers

Of the 15 EHR papers, 13 included the development of a prediction model [19-31]. In Table 1, the data extraction per paper can be found. The included EHR papers covered various outcomes, mostly chronic conditions (n = 11) [20-29,32]. The most frequent predicted outcomes were dementia (n = 3) [20,21,24], and asthma and/or Chronic Obstructive Pulmonary Disease (COPD) (n = 3) [22,27,32]. Other study outcomes are shown in Table 1. All included papers used predictors registered in EHR. Predictors included findings from clinical examination (n = 6) [20,26,27,29,32,33], laboratory results (n = 5) [22,23,26,29,33] and medication (n = 4) [20,22,25,30]. All models used structured data.

Two papers externally validated and tested a prediction model in a PC setting [32,33]. One paper had a prospective approach and tested the diagnostic performance of a prediction model for Asthma and COPD [32]. Ten papers were published after 2020 [19,22-25,27,28,30,32,33]. Most data sources used in the studies originated from Europe (n = 8) [19-23,25,28,32], followed by North America (n = 5) [24,26,29,30,32,33] and Asia (n = 2) [27,31].

Preprint JMIR Publications

Figure 1. PRISMA flow chart of study selection.



AI technique

All of the included studies performed at least one 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).

Table 1. Extracted information from EHR papers.

Author, year	Country	Study design and study type	Outcome	Dataset information	Predictors	AI technique
Barnes, 2020 [24]	USA	Retrospective cohort study Developmental	Dementia	Kaiser Permanente Washington 4 330 participants	Demographics, diagnosis, vital signs, healthcare utilization, medication	LR
Briggs, 2022 [23]	UK	Nested case control study Developmental	Oesophago-gastric cancer	General Practice Research Database 40 348 participants (7471 cases and 32 877 controls)	Demographics, symptoms, laboratory (lab) results	RF, SVM, LR, NB, XGBoost
Dhanda, 2023 [33]	USA	Retrospective cohort study Developmental + Validation	Urine culture result/urine tract infection (UTI)	Data for development from emergency department. Data for external validation from outpatient family medicine department. 80 859 participants (80 387 development, 472 external validation)	Demographics, urine analysis, vital signs, symptoms, history of UTI, higher risk clinical features	XGBoost, RF, NN
Dros, 2022 [25]	Netherlands	Nested case control study Developmental	Primary Sjögren's Syndrome (pSS)	Nivel Primary Care Database (PCD) 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)	Demographics, diagnosis, medication, health utilization	LR, RF

Ellertsson, 2021 [19]	Iceland	Retrospective cohort study Developmental	Common clinical headaches	Data from 15 primary Health Care of the Capital Area clinics Unknown number of participants, 800 clinical notes	Headache symptoms, sex, age, family history	RF
Ford, 2019 [20]	UK	Nested case control study Developmental	Dementia	Clinical Practice Research Datalink data 93 120 participants (46 560 cases, 46 560 controls)	Symptoms of physical or cognitive frailty, medical history, healthcare utilization, ethnicity, family history of dementia, intoxications, BMI, blood pressure, psychological diagnoses and treatment	RF, NB, SVM, NN
Jammeh, 2018 [21]	UK	Case control study Developmental	Dementia	NHS Devon with 18 GP surgeries 3063 participants (850 cases, 2213 controls)	Demographics, long-term conditions, and consultations	LR, RF, NB, SVM
Kocks, 2023 [32]	Netherlands	Prospective observational study Validation	Asthma, COPD or Asthma-COPD overlap	Nivel PCD 116 cases, tested on 180 specialists from 9 countries (external validation)	Symptoms, BMI, spirometry scores, smoking, diagnosis of chronic/allergic rhinitis, age at onset of respiratory disease	Multinominal LR
LaFreniere, 2016 [26]	Canada	Case control study, nested is unclear Developmental	Hypertension	Canadian Primary Care Sentinel Surveillance Network (CPCSSN) 379 027 participants (185 371 cases, 193 656 controls)	Demographics, BMI, blood pressure, lab	NN
Lin, 2023 [27]	China	Retrospective cohort study	COPD	Public health data of Chinese residents	Demographics, smoking, BMI, chronic cough, shortness of	18 methods, including:

		Developmental		1875 participants	breath, biofuel use, and family history. Based on questionnaire for COPD.	Decision tree, LR, discriminant analysis (linear and quadratic), SVM, gradient boosting classifiers, NN, gaussian process classifier, KNN, NB
Mariani, 2021 [22]	Netherlands	Retrospective cohort study Developmental	Asthma and COPD	Data from Dutch primary care laboratory in Groningen 3659 participants	Demographics, symptoms, diagnosis, medication, lab results, referrals, spirometry results	SVM, RF, KNN
Nemlander, 2023 [28]	Sweden	Nested case control study Developmental	Non-metastatic colorectal cancer	VEGA regional administrative healthcare database 2681 participants (542 cases, 2139 controls)	NMCRC stage, number of GP consultations, diagnosis codes	Stochastic gradient boosting
Perveen, 2016 [29]	Canada	Retrospective cohort study Developmental	Diabetes mellitus	CPCSSN 4678 participants (377 cases, 4301 controls)	Demographics blood pressure, lab results	Decision tree, bagging, ADAboost
Singh, 2021 [30]	USA	Retrospective cohort study Developmental	Anterior segment vision threatening disease (asVTD)	EHR of the University of Michigan 2942 participants (133 cases, 2809 controls)	Demographics, history of eye problems, symptoms, medication	Elastic net LR
Su, 2019 [31]	China	Retrospective cohort study	Top 100 diagnoses	National Hospital Ambulatory Medical Care	Demographics, symptoms, past medical history	NN

		Developmental	(within general diagnoses)	Survey and the National Ambulatory Medical Care Survey Unknown number of participants, top 100 diagnosis selected from 2 000 000 records		

Abbreviations: LR: logistic regression, SVM: support vector machine, RF: random forest, KNN: K-nearest neighbours, NN: neural network, NB: naïve Bayes, XGBoost: extreme gradient boosting

Preprint JMIR Publications

Quality assessment

Risk of bias

None of the studies assessed by the PROBAST tool had a low risk of bias. We found high risk of bias in 9 studies and an unclear risk of bias in 6 studies (Table 2). In Multimedia Appendix part 4, the full assessment of the PROBAST tool can be found.

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 (n = 5) [19,25,27,30,32] and irrelevant model performance measures that were used to evaluate the model (n = 2) [29,32]. The main reasons for an unclear risk of bias in the analysis domain were lack of clarity on how missing data were handled (n = 10) [19-21,23,24,28-31,33], and on how the predictors and their assigned weights in the final model correspond to results from the reported multivariate analysis (n = 9) [19,21,22,26-31]. Even though measures of calibration are not part of the signalling questions of the PROBAST, we noticed that only 4 papers [23,24,30,33] 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 (n = 1) [32], and not excluding the predictors from the outcome definition (n = 2). For example, Perveen et al. included fasting glucose levels to predict diabetes [29] and Kocks et al. included spirometry findings to predict Asthma and COPD [32]. The two 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 (n = 9) [21,25-31,33], and the lack of clarity on the outcome definition (n = 7) [21-23,25,27,29,31].

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 two studies as both studies excluded participants at high risk of the outcome [28,33].

Another reason for high risk of bias was a non-appropriate data source was used in one study [21] because the authors described the study as a case control-study although the study was not nested as recommended in the PROBAST guidelines [17,18].

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 (n = 4) [19,21,23,31].

Table 2. Risk of bias per domain using the PROBAST tool.

	Participants	**Predictors**	**Outcome**	**Analysis**	**Overall**
Barnes [24]	Low	Low	Low	Unclear	Unclear
Briggs [23]	Low	Unclear	Unclear	Unclear	Unclear
Dhanda [33]	High	Low	Unclear	Unclear	High
Dros [25]	Low	Low	Unclear	High	High
Ellertsson[19]	Low	Unclear	Low	High	High
Ford [20]	Unclear	Low	Low	Unclear	Unclear
Jammeh[21]	High	Unclear	Unclear	Unclear	High
Kocks [32]	Low	Low	High	High	High
LaFreniere [26]	Unclear	Low	Unclear	Unclear	Unclear
Lin [27]	Unclear	Low	Unclear	High	High
Mariani, [22]	Unclear	Low	Unclear	Unclear	Unclear
Nemlander [28]	High	Low	Unclear	Unclear	High
Perveen [29]	Low	Unclear	High	High	High
Singh [30]	Low	Low	Unclear	High	High
Su [31]	Unclear	Unclear	Unclear	Unclear	Unclear
	Total • Low: 7 • Unclear: 5 • High: 3	**Total** • Low: 10 • Unclear: 5 • High: 0	**Total** • Low: 3 • Unclear: 10 • High: 2	**Total** • Low: 0 • Unclear: 9 • High: 6	**Total** • Low: 0 • Unclear: 6 • High: 9

Applicability

Overall, we found an unclear concern for applicability in 10 papers and a low concern for applicability in 5 papers (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 (n = 8) [21,26-31,33].

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 (n = 5) [19,21,23,29,31]. For example, one paper lacked information on how notes were annotated before they were used as predictors in the model [19].

The unclear concern for applicability in the participants domain was mainly due to lack of information on inclusion and exclusion criteria (n = 3) [26,27,31].

Table 3. Applicability per domain using the PROBAST tool.

	Participants	**Predictors**	**Outcome**	**Overall**
Barnes [24]	Low	Low	Low	Low
Briggs [23]	Low	Unclear	Unclear	Unclear
Dhanda [33]	Low	Low	Unclear	Unclear
Dros [25]	Low	Low	Low	Low
Ellertsson[19]	Low	Unclear	Low	Unclear
Ford [20]	Low	Low	Low	Low
Jammeh[21]	Low	Unclear	Unclear	Unclear
Kocks [32]	Low	Low	Low	Low
LaFreniere [26]	Unclear	Low	Unclear	Unclear
Lin [27]	Unclear	Low	Unclear	Unclear
Mariani, [22]	Low	Low	Low	Low
Nemlander [28]	Low	Low	Unclear	Unclear
Perveen [29]	Low	Unclear	Unclear	Unclear
Singh [30]	Low	Low	Unclear	Unclear
Su [31]	Unclear	Unclear	Unclear	Unclear
	Total Low: 12 Unclear: 3 High: 0	**Total** Low: 10 Unclear: 5 High: 0	**Total** Low: 6 Unclear: 9 High: 0	**Total** Low: 5 Unclear: 10 High: 0

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 one chronic condition. Merely two papers tested the model in a PC setting. All of the included studies performed at least one 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 two reviews evaluated the risk of bias in clinical prediction models on a wide range of diseases in PC studies [13,14]. Most of the included studies in these reviews showed a high to unclear risk of bias, which is in line with our findings [13,14]. However, there appear to be differences in grading compared to Abdulazeem et al. [14]. 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 Rahimi et al. did not report details on the reasons they coded sub-domains as high or unclear risk of bias, for which reason we are unable to make a formal comparison with our results [13].

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 [34,35]. These similarities include the unjustified small sample size in EHRs studies, inappropriate evaluation in the performance measures and

flaws in handling of missing data [34-36].

The most used Al 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 [14,34,36-38].

In general, studies analysing EHRs are subject to a high risk of bias, because these data are collected for clinical rather than research purposes [17]. Hence, clinical prediction models developed on EHR are more difficult to reproduce and generalize given the heterogeneity of coding systems and database infrastructures [14]. In line with models analysed in previous studies [13,14,36], most of the clinical prediction models were not externally validated. Most of the studies developed in PC were performed in developed countries and may not have taken into account regional/global differences in the availability of certain predictors [12-14]. For example, some predictors may not be easy to obtain in PC settings in less developed countries (e.g. spirometry results for the prediction of asthma/COPD). Together, all these factors limit 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 EHR in the search strategy [11]. A second strength is that the findings on the risk of bias were carefully assessed by two independent reviewers 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 [35], 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 EHR.

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 machine leaning and deep learning techniques. Given the lack of a widely accepted definition of AI, other reviews use other criteria for AI or machine learning [34,36,38]. 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 GPs 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. Even though the PROBAST guidelines are highly detailed and reliable in evaluating clinical prediction models [39], PROBAST has been criticised for being less specific and less applicable for AI-based models compared to 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 model, as stated in a number of standardized frameworks designed for prediction model developers [40,41]. We found that only two 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 [12,14]. However, in general, chronic conditions are not known to be difficult to diagnose in PC. Two examples from our included papers are the diagnosis hypertension

predicted on the variable high blood pressure [26], and the diagnosis diabetes predicted on the variable high glucose levels [29]. 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 one 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 is important in developing clinical prediction models with a higher success rate of clinical implementation. We recommend involving relevant stakeholders in early stages of the development of a new model.

To improve the methodology in future studies, our findings suggests 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 [11] 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 [42] and PROBAST guidelines, in communicating their findings [3], particularly now that the TRIPOD-AI extension is released [43]. 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 [44]. However, we only found two papers that performed external validation.

Conclusion

AI based prediction models using EHR data are not ready yet 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 primary care, 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.

Preprint JMIR Publications

Acknowledgements

We would like to thank Lori van den Hurk for her help with the screening process.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

AI: artificial intelligence
EHR: electronic health records
GP: general practitioner
PC: primary care
PROBAST: Prediction model Risk Of Bias ASsessment Tool
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

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
2. Moons KG, 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. doi:10.1136/bmj.b375
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
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
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
6. Mistry P. Artificial intelligence in primary care. *Br J Gen Pract*. Sep 2019;69(686):422-423. doi:10.3399/bjgp19X705137
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
8. Lin S. A Clinician's Guide to Artificial Intelligence (AI): Why and How Primary Care Should Lead the Health Care AI Revolution. *The Journal of the American Board of Family Medicine*. 2022;35(1):175-184. doi:10.3122/jabfm.2022.01.210226
9. 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
10. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and

Clinical Medicine. *N Engl J Med*. Sep 29 2016;375(13):1216-9. doi:10.1056/NEJMp1606181
11. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. 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
12. 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
13. Abbasgholizadeh Rahimi S, Legare 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
14. 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
15. 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
16. Moons KG, de Groot JA, 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
17. 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
18. 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
19. 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
20. 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
21. 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
22. Mariani S, Lahr MMH, Metting E, Vargiu E, Zambonelli F. Developing an ML pipeline for asthma and COPD: The case of a Dutch primary care service. *Int J Intell Syst*. Nov 2021;36(11):6763-6790. doi:10.1002/int.22568
23. 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)doi:10.3390/cancers14205023
24. 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
25. Dros JT, Bos I, Bennis FC, et al. Detection of primary Sjogren'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

26. LaFreniere D, Zulkernine F, Barber D, Martin K. Using Machine Learning to Predict Hypertension from a Clinical Dataset. *Proceedings of 2016 Ieee Symposium Series on Computational Intelligence (Ssci)*. 2016;doi:10.1109/SSCI.2016.7849886
27. Lin XS, 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
28. 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;182:100-106. doi:10.1016/j.ejca.2023.01.011
29. 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
30. 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
31. Su GX, Wen JH, Zhu ZW, et al. An Approach of Integrating Domain Knowledge into Data-Driven Diagnostic Model. *Stud Health Technol*. 2019;264:1594-1595. doi:10.3233/Shti190551
32. 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 e3. doi:10.1016/j.jaip.2023.01.017
33. 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*. Jan-Feb 2023;21(1):11-18. doi:10.1370/afm.2902
34. Andaur Navarro C, Damen JA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281. doi:10.1136/bmj.n2281
35. 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
36. 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
37. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259
38. 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
39. Moons KG, Altman DG, Reitsma JB, Collins GS, Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Development I. 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):303-5. doi:10.1097/PAP.0000000000000072
40. Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning:

users' guides to the medical literature. *Jama*. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489
41. 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)doi:10.1136/bmjhci-2023-100749
42. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol*. Feb 2015;68(2):134-43 doi: 10.1016/j.jclinepi.2014.11.010. doi:10.1016/j.jclinepi.2014.11.010
43. 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
44. 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

Preprint JMIR Publications

Supplementary Files

Preprint JMIR Publications

Multimedia Appendixes

Search strategy, screening guidance, table with all included papers, probast checklist, references of appendix.
URL: http://asset.jmir.pub/assets/e8849f8758434a1949081b7fa16b3907.docx

Preprint JMIR Publications

CONSORT (or other) checklists

PRISMA checklist.
URL: http://asset.jmir.pub/assets/20f5c9880f392604457f77f41f86ad75.pdf

Preprint JMIR Publications