

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

Machine learning prediction models in primary health care

Hebatullah Abdulazeem^{1*}, Sera Whitelaw², Gunther Schaubberger¹, Stefanie J. Klug¹

¹ Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich (TUM), Munich, Germany

² Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada

*Corresponding author

Dr. Hebatullah Abdulazeem
Chair of Epidemiology
Technical University of Munich
Georg-Brauchle Ring 56
80993 Munich, Germany
Tel: +49892892495
E-mail: hebatullah.abdulazeem@tum.de (HA)

Abstract

Aim: With the rapid advances in technology and data science, machine learning (ML) is being adopted by the health care sector; but there is a lack of literature addressing the health conditions targeted by the ML prediction models within primary health care (PHC). To fill this gap in knowledge, we conducted a systematic review following the PRISMA guidelines to identify the health conditions targeted by ML in PHC.

Methods: We searched the Cochrane Library, Web of Science, PubMed, Elsevier, BioRxiv, Association of Computing Machinery (ACM), and IEEE Xplore databases for studies published from January 1990 to January 2022. We included any primary study addressing ML diagnostic or prognostic predictive models that were supplied completely or partially by real-world PHC data. We performed literature screening, data extraction, and risk of bias assessment. Health conditions were categorized according to international classification of diseases. Extracted data were analyzed quantitatively and qualitatively.

Results: We identified 109 studies investigating 42 health conditions. These studies included 273 ML prediction models supplied by the PHC data of 24.2 million participants from 19 countries. We found that 82% of the studies were retrospective. 76.6% of the studies reported diagnostic predictive ML models. 77% of all reported models aimed for models' development without external validation. Risk of bias assessment revealed that 90.8% of the studies were of high or unclear risk of bias. The most frequently reported health conditions were Alzheimer's disease and diabetes mellitus.

Conclusions: To the best of our knowledge, this is the first review to investigate the extent of the health conditions targeted by the ML prediction models within PHC settings. Our

study provides an important summary on the presently available ML models in PHC, which can be used in further research and implementation efforts.

Introduction

Primary health care (PHC) is considered the gatekeeper, where health education and promotion are provided, non-life-threatening health conditions are diagnosed and treated, and chronic diseases are managed [1]. This form of health maintenance, which aims to provide constant access to high-quality care and comprehensive services, is defined and called for by the WHO global vision for PHC [2]. To achieve these PHC care aims, common health disorders require risk prediction for primary prevention, early diagnosis, follow-up, and timely interventions to avoid diseases exacerbations and complications, all of which are the core practice of PHC [3].

With the high number of patients visiting PHC and the emerge of electronic health records, “Big Data” is generated with subsequent difficulties to be handled by traditional data analytics [4]. Tools that could more accurately predict diseases incidence and progression and offer advice on appropriate treatment could substantially improve the decision-making process. Machine Learning (ML), a subtype of artificial intelligence (AI), provides methods to productively mine this big data, such as predictive models that potentially forecast and predict diseases occurrence and progression [5].

Integrating the PHC medical efforts with the continuously updated technologies constitutes a fusion of numerous disciplines and views aimed at improving the performance of health care regarding patient care and the productivity and efficiency within health care facilities [5,6]. ML models have been developed in health research – most significantly in the last decade - to predict the incidence of diabetes, cancers, and recently COVID-19 pandemic

related illness from health records [7]. A systematic overview of 35 studies published in 2021 investigated the existing literature of AI/ML, but exclusively in relation to World Health Organization indicators [8]. Other literature and scoping reviews examined AI/ML in relation to certain health conditions, such as HIV [9], hypertension [10], and diabetes [11]. Other systematic reviews targeted specific health conditions across multiple health sectors, such as pregnancy care [12], melanoma [13], stroke [14], and diabetes [15]. However, reviews investigating PHC specifically have been fewer [16,17]. It has been reported that research on ML for PHC stands at an early stage of maturity [17]. Similar to ours, a recently published protocol of a systematic review addressing the performance of ML prediction models in multiple different medical fields was published [18]. However, this protocol does not focus on primary care in specific and its search is limited to the years 2018 and 2019. Hence, the current literature is not enough to identify what the diseases targeted by ML prediction models within the real-world primary care are. Furthermore, literature investigating validity and potential impact on health of such models are not abundant. To address this gap, our objective was to encompass the health conditions predicted by using ML models to identify and assess the extent of the body of research within real-world PHC settings.

Methods

We conducted a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [19] and the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) [20]. The protocol for our review was registered on PROSPERO CRD42021264582 [21].

We included primary research articles (peer-reviewed, preprint, or abstract) published in any language. Studies that were published between January 1, 1990, when ML algorithm

with a data-driven approach was first developed [22] and January 4, 2022 were included. Studies that reported real-world exclusive or mixed PHC data for any health condition in ambulatory settings, including referred patients from PHC to other health care facilities, worldwide were included. Studies that reported any prediction ML models within the PHC level that was classified as AI, DL or ML models were included.

Search strategy and selection criteria

A comprehensive and systematic search was performed covering multidisciplinary databases: 1. Cochrane Library, 2. Elsevier (including ScienceDirect, Scopus, and Embase), 3. PubMed, 4. Web of Science (including nine databases), 5. BioRxiv and MedRxiv, 6. Association for Computer Machinery (ACM) Digital Library, and 7. Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library.

To identify potentially relevant studies, we searched literature with the last updated search on January 4, 2022, back to January 1, 1990. The utilized search terms included "machine learning", "artificial intelligence", "deep learning", and "primary health care". Boolean operators and symbols were adapted to each literature database. Hand searches of citations of relevant reviews and a cross-reference check of the retrieved articles was also performed. Conference abstracts and gray literature searches were conducted using the available features of some databases. The full search strategy for all the electronic databases is presented in S1 Appendix. A reference management software (EndNote X9) was used to import references and to remove duplicates.

Literature screening, data collection and statistical analysis

Title and abstract screening for all records were conducted independently by two researchers through the Rayyan platform [23]. Discrepancies were resolved by discussion. All studies that met the eligibility criteria were included in the systematic review.

The data extracted included: meta-data (first author, year, and publisher), source of primary data (country under investigation), datasets used (exclusive PHC data that was generated only within PHC settings or mixed data that was generated within PHC settings in addition to other data sources, such as secondary or tertiary health care), period of data extraction, sample size, and study design, predicted health condition, study objectives (incident diagnostic, prevalent diagnostic or prognostic), aim of model proposal (development without external validation, development with external validation, or external validation without or with update). Data extraction was performed by two authors.

Health conditions extracted were categorized according to international classification of diseases (ICD)-10 version 2019 [24]. Further categorization was based on the ML models' aim. Descriptive statistics (number and percentage of studies) were calculated. Additionally, the overall number of participants was calculated, taking into consideration the potential overlap between the included datasets. This overlap assessment was identified based on similarity of datasets, period of data gathering within each included study, the targeted health condition and the inclusion and exclusion criteria of the participants. The quantitative results were calculated using Microsoft Excel.

Risk of bias and applicability assessment

The 'Prediction model study Risk Of Bias Assessment Tool' (PROBAST) was used to assess the risk of bias and concerns about the applicability of the included studies [25]. The four domains of this tool, which are participants, predictors, outcome, and analysis were addressed. The overall judgement for the risk of bias evaluation and concern of applicability

of the prediction models in PROBAST is ‘low,’ ‘high,’ or ‘unclear.’ In cases when all domains were graded ‘low’ risk of bias, assessment of ‘models developed without external validation’ was downgraded to ‘high’ risk of bias even if all the four domains were of low risk of bias, unless the model’s development was based on an exceptionally large sample size and included some form of internal validation. Results of risk of bias and concern of applicability assessments were presented in a color-coded graph.

Results

Our search strategy yielded 23,045 publications. After duplicate removal, 19,280 publications were screened, of those 167 publications were eligible for the full text screening. A total of 109 publications met our inclusion criteria (Fig 1). A list of the excluded studies with the justification of exclusion is presented in S2 Appendix.

The results of the data extracted in this review are presented in the following paragraphs in the form of geographical and chronological characteristics, studies’ design and the ML models addressed, and the health conditions investigated. Additionally, three tables, Table 1, 2, and 3, are depicting the characteristics of the included studies. Table 1 presents the studies that reported only developing ML prediction models without implication of any external validation of the models developed. Table 2 presents the studies that reported both developing and validating ML prediction models. Whereas Table 3 presents the studies that reported only the validation of previously developed models. In both Table 2 and 3, each row represents different dataset that was used to develop and/or validate the prediction models.

Figure 1 Prisma Flow diagram

Geographical and chronological characteristics

The earliest included study was published in 2002, with the most publications occurring over the past four years. 77.9% (n= 85/109) of the publications were published between 2018 – 2021, (Fig 2). The United States of America (US) and the United Kingdom (UK) were reported in 58.1% of the included publications. While the 109 included publications reported countries 129 times, the US was reported 41 times and the UK 34 times. Other countries were identified but less frequently as depicted in S3 Figure. Usage of exclusive real-world primary health care data as predictors was reported in 77.4% (117 of 151 counts of data sources) across the studies. The remaining 22.6% of the PHC data sources were linked to different data sources, such as health insurance claims, cancer registries, secondary or tertiary health care, or administrative data. In the US, data was obtained mainly from PHC centers. In contrast, the most common source of the UK data was the Clinical Practice Research Datalink (CPRD), which is the largest patients' data registry in the UK [26]. The period of data collection through the studies ranged from 1982 to 2020. The timeframe of patients' data extracted and used to develop and train the ML models among the included studies varied between 2 months and 28 years. Sample sizes used for training and/or validating the models across the included studies ranged from 26 to around 4 million participants. A total number of participants within all the included studies was of 24.2 million. The potential overlaps of the datasets through each publication were investigated using two criteria, which were periods of the data extracted and participant characteristics per study. After identifying the potential overlap, the total number of unique participants was estimated to be 23.7 million.

Figure 2 Number of studies per year of publication until December 31, 2021, in addition to one study included up to January 4, 2022.

Studies' design, objectives, and models

All the included studies were observational in design. Apart from 16 prospective studies, 85.3% (n= 93 of 109) of the studies were retrospective in design, of which 60 studies

were reported as retrospective cohorts. The other reported studies designs were depicted in the supplementary S5 Figure. Regarding the primary objective of the included studies, 76.6% (n= 83 of 109) of the studies were predicting diagnosis of health conditions, either incident (n= 62 of 83) or prevalent (n= 21 of 83). The remaining 23.8% (n= 27 of 109, including one study of two different objectives [27]) predicted prognosis of health conditions, such as remission, improvement, complications, hospitalization, or mortality.

According to CHARMS guidelines, as mentioned earlier, the aim of the studies to use the prediction models can be one of three aims. These aims are model development without external validation, model development with external validation, and external validation of a predeveloped model with or without further model update [20]. The main aim of the included studies was found to be development of prediction models without evaluating the generalizability of the models, i.e., external validations (77%, n= 84 of 109). Another 13 (11.9%) studies developed and externally validated the models and only 12 studies (11%) externally validated previously existing models, but none of these studies reported updating the assessed model.

Within the 109 included publications, 273 models were developed and/or validated. The most frequent used type of ML was the supervised learning 84.2% (n= 230 of 273 models across the included studies). These supervised ML models were identified as follows: random forest (n= 53), logistic regression (n= 42), support vector machine (n= 33), boosting models such as extreme, light, and adaptive boosting (n= 29), decision tree (n= 28), and others such as naïve bias, k-nearest neighbors, and Least Absolute Shrinkage and Selection Operator (LASSO) (n= 45). Reinforcement ML/deep learning techniques, such as neural networks, were reported 36 times (13.1%, of 273 models), cross the studies, either exclusively or in comparison to other supervised ML models. A few studies (n= 3 of 109 studies) developed seven

unsupervised ML models, such as k-means for predicting diseases prognosis through clustering it with the other morbidities [28–30]. A few studies (n= 5) used the natural language processing (NLP) technique as a preparatory step for using the free text clinical notes as (additional) predictors [27,31–34]. A descriptive summary of the types of ML models included is depicted in Fig 3, where Supervised ML models, such as random forest and logistic regression were frequently reported, while the reinforcement and unsupervised ML models were less reported.

Figure 3 Number of models developed and/or validated across studies

A few studies (n=10) compared the performance of the developed ML models to other standard reference techniques that were based on classical statistics, such as classical logistic and Cox regression. In seven studies of them, it was reported that ML models outperformed the classical statistics, providing better insights to discover new associations [29,35–40]. The other studies (n=3) reported either similar [41] or lower performance of ML to classical models.[42][43].

Models' developing attributes, such as features selection and handling of missing data were reported in 68 and 38 studies (of 109), respectively. Models' internal validation using n-fold cross validation and random splitting of the datasets, either one of them or both in the same study, were reported in 90% and 80% of the included studies, respectively. Broader external validation scale was reported in 25 studies in one different setting or more, such as temporal, geographical, or using different population sample validation. On the other hand, models' performance measures of discrimination ability using the area under the receiver operating characteristic curve (AUROC) were reported in 62 studies, where results of these measures range from zero for no discrimination ability to ten for the best ability. One study reported the performance measures using decision analysis curve [116].

Tables 1, 2, and 3 present an overview of the included studies characteristics based on the development and validation stage of the models, grouped according to the ICD-10 classification, and ordered alphabetically within each classification. For each study, study design and the objective of the ML prediction model (incident diagnostic, prevalent diagnostic, or prognostic) were provided. Furthermore, the dataset used in each study was reported based on the national location of the dataset and the health care level source being exclusive if only from a PHC data source or linked if PHC data was reported to be linked to other health care data sources, such as secondary or tertiary health care. Last three columns presented the number of the dataset's population, the timeframe of the data extracted from the dataset used, and the health condition addressed. Nevertheless, in Tables 2 and 3, each study was presented in multiple rows based on the number of the locations used to validate the ML models. An additional panel summary of all the included studies is presented as S5 Appendix.

Table 1 Overview of the included studies reporting ML prediction models developed using primary health care data without conducting external validation (n = 84)

Study	Study design	Objective	Country	Source of data set	Population	Period of extracted data	Health condition
Circulatory System Diseases							
Chen et al (2019) [46]	Retro. nested case control	Incident diagnostic	United States	Exclusive	34,502	05/2000-05/2013	Heart failure
Choi et al (2017) [47]	Retro. case control	Incident diagnostic	United States	Exclusive	32,787	05/2000-05/2013	Heart failure
Du et al (2020) [48]	Retro. cohort	Prognostic	China	Linked	42,676	2010-2018	Hypertension
Farran et al (2013) [49]	Retro. cohort	Incident diagnostic	Kuwait	Linked	270,172	12 years	Any cardiovascular disease
Hill et al (2019) [37]	Retro. cohort	Incident diagnostic	United Kingdom	Exclusive	2,994,837	01/2006-12/2016	Atrial fibrillation
Karapetyan et al (2021) [50]	Retro. cohort	Prognostic	Germany	Exclusive	46,071	02-2020-09/2020	Any cardiovascular disease
Lafreniere et al (2017) [51]	Retro. cohort	Incident diagnostic	Canada	Exclusive	379,027	Not reported	Hypertension
Li et al (2020) [43]	Retro. cohort	Incident diagnostic	United Kingdom	Linked	3,661,932	01/1998-12/2018	Any cardiovascular disease
Lip et al (2021) [52]	Retro. cohort	Prevalent diagnostic	Australia	Linked	926	Not reported	Hypertension
Lorenzoni et al (2019) [53]	Pros. cohort	Prognostic	Italy	Linked	380	2011-2015	Heart failure

Study	Study design	Objective	Country	Source of data set	Population	Period of extracted data	Health condition
Ng et al (2016) [54]	Retro. nested case control	Incident diagnostic	United States	Exclusive	152,095	2003-2010	Heart failure
Nikolaou et al (2021) [55]	Pros. cohort	Prognostic	United Kingdom	Exclusive	6,883	2015-2019	Any cardiovascular disease
Saraju et al (2021) [56]	Retro. cohort	Incident diagnostic	United States	Exclusive	32,192	01/2009-12/2018	Any cardiovascular disease
Selskyy et al (2018) [57]	Retro. case control	Prognostic	Ukraine	Exclusive	63	2011-2012	Hypertension
Shah et al (2019) [27]	Retro. cohort	Prognostic/ Incident diagnostic	United Kingdom	Linked	2,000	Not reported	Myocardial infarction
Solanki et al (2020) [58]	Retro. cohort	Prevalent diagnostic	United States	Exclusive	495	2007-2017	Hypertension
Solares et al (2019) [59]	Pros. cohort	Incident diagnostic	United Kingdom	Exclusive	80,964	entry – 01/2014	Any cardiovascular disease
Ward et al (2020) [60]	Retro. cohort	Incident diagnostic	United States	Exclusive	262,923	01/2009-12/2018	Atherosclerosis
Weng et al (2017) [35]	Pros. cohort	Incident diagnostic	United Kingdom	Exclusive	378,256	01/2005-01/2015	Any cardiovascular disease
Wu et al (2010) [61]	Retro. case control	Incident diagnostic	United States	Exclusive	44,895	01/2003-12/2006	Heart failure
Zhao et al (2020) [34]	Retro. cohort	Incident diagnostic	United States	Linked	4,914	Not reported	Stroke
Digestive System Diseases							
Sáenz Bajo et al (2002) [36]	Retro. cohort	Prevalent diagnostic	Spain	Exclusive	81	01/1999-06/1999	Gastroesophageal reflux
Waljee et al (2018) [62]	Retro. cohort	Prognostic	United States	Linked	20,368	2002-2009	Inflammatory bowel disorders
Endocrine, Metabolic, and Nutritional Diseases							
Akeya et al (2020) [63]	Retro. cohort	Incident diagnostic	United Kingdom	Exclusive	4,027,775	01/1999-06/2019	Familial hypercholesterolemia
Álvarez-Guisasola et al (2010) [42]	Pros. cohort	Incident diagnostic	Spain	Exclusive	2,662	Not reported	Diabetes mellitus
Crutzen et al (2021) [64]	Retro. cohort	Incident diagnostic	The Netherlands	Exclusive	138,767	01/2007-01/2014	Diabetes mellitus
Ding et al (2019) [40]	Retro. case control	Prevalent diagnostic	United States	Exclusive	97,584	1997-2017	Primary Aldosteronism
Dugan et al (2015) [38]	Retro. cohort	Prognostic	United States	Exclusive	7,519	Over 9 years	Obesity
Farran et al (2019) [65]	Retro. cohort	Prognostic	Kuwait	Exclusive	1,837	Over 9 years	Diabetes mellitus
Hammond et al (2019) [66]	Retro. cohort	Prognostic	United States	Exclusive	3,449	01/2008-08/2016	Obesity
Kopitar et al (2020) [67]	Retro. case control	Incident diagnostic	Slovenia	Exclusive	27,050	12/2014-09/2017	Diabetes mellitus
Lethebe et al (2019) [68]	Retro. cohort	Prevalent diagnostic	Canada	Exclusive	1,309	2008-2016	Diabetes mellitus
Looker et al (2015) [69]	Pros. nested case control	Prognostic	United Kingdom	Exclusive	309	12/1998-05/2009	Diabetic nephropathy
Metsker et al (2020) [70]	Retro. cohort	Incident diagnostic	Russia	NR	54,252	07/2009-08/2017	Diabetic polyneuropathy
Metzker et al (2020) [71]	Retro. cohort	Incident diagnostic	Russia	NR	58,462	Not reported	Diabetic polyneuropathy

Study	Study design	Objective	Country	Source of data set	Population	Period of extracted data	Health condition
Nagaraj et al (2019) [72]	Retro. cohort	Prognostic	The Netherlands	Exclusive	11,887	01/2007-12/2013	Diabetes mellitus
Pakhomov et al (2008) [31]	Retro. cohort	Prevalent diagnostic	United States	Exclusive	145	07/2004-09/2004	Diabetic foot
Rumora et al (2021) [73]	Cross sectional	Incident diagnostic	Denmark	Exclusive	97	10/2015-06/2016	Diabetic polyneuropathy
Tseng et al (2021) [33]	Cross sectional	Incident diagnostic	United States	Exclusive	NR	Not reported	Diabetes mellitus
Wang et al (2021) [74]	Retro. cohort	Incident diagnostic	China	Exclusive	1,139	2017-2019	Gestational diabetes
Willimson et al (2020) [75]	Pros. cohort	Incident diagnostic	United States	Linked	866	Not reported	Familial hypercholesterolemia
External Cause of Mortality							
DelPozo-Banos et al (2018) [76]	Retro. case control	Incident diagnostic	United Kingdom	Linked	54,684	2001-2015	Suicidality
Penfold et al (2021) [77]	Retro. cohort	Incident diagnostic	United States	Linked	256,823	Not reported	Suicide
van Mens et al (2020) [78]	Retro. case control	Incident diagnostic	The Netherlands	Exclusive	207,882	2017	Suicidality
Genitourinary System Diseases							
Shih et al (2020) [79]	Retro. cohort	Incident diagnostic	Taiwan	Linked	19,270	01/2015-12/2019	Chronic kidney disease
Zhao et al (2019) [80]	Pros. cohort	Incident diagnostic	United States	Exclusive	61,740	2009–2017	Chronic kidney disease
Mental and Behavioral Diseases							
Alexander et al (2021) [28]	Retro. cohort	Prevalent diagnostic	United Kingdom	Exclusive	7,913	01/1997-06/2016	Alzheimer's disease
Dinga et al (2018) [81]	Pros. cohort	Prognostic	The Netherlands	Linked	804	Not reported	Major depressive disorder
Ford et al (2019) [82]	Retro. case control	Incident diagnostic	United Kingdom	Exclusive	93,120	2000-2012	Alzheimer's disease
Ford et al (2020) [83]	Retro. case control	Incident diagnostic	United Kingdom	Exclusive	95,202	2000-2012	Alzheimer's disease
Ford et al (2021) [84]	Retro. case control	Prevalent diagnostic	United Kingdom	Exclusive	93,426	2000-2012	Alzheimer's disease
Fouladvand et al (2019) [85]	Retro. cohort	Prognostic	United States	Exclusive	3,265	Not reported	Alzheimer's disease
Haun et al (2021) [86]	Cross sectional	Incident diagnostic	Germany	Exclusive	496	Not reported	Anxiety
Jammeh et al (2018) [87]	Retro. Case control	Incident diagnostic	United Kingdom	Exclusive	3,063	06/2010-06/2012	Alzheimer's disease
Jin et al (2019) [88]	Retro. Cohort	Incident diagnostic	United States	Exclusive	923	2010-2013	Major depressive disorder
Kaczmarek et al (2019) [89]	Retro. Case control	Prevalent diagnostic	Canada	Exclusive	890	Not reported	Post-traumatic stress disorder
Ljubic et al (2020) [90]	Retro. Cohort	Incident diagnostic	United States	Exclusive	2,324	Not reported	Alzheimer's disease
Mallo et al (2020) [91]	Retro. case control	Prognostic	Spain	Exclusive	128	2008	Alzheimer's disease
Mar et al (2020) [92]	Retro. case control	Prevalent diagnostic	Spain	Linked	4,003	Not reported	Alzheimer's disease
Pólichlopek et al (2020) [93]	Retro. cohort	Incident diagnostic	The Netherlands	Exclusive	92,621	2007-12/2016	Any mental disorder

Study	Study design	Objective	Country	Source of data set	Population	Period of extracted data	Health condition
Shen et al (2020) [94]	Retro. cohort	Incident diagnostic	China	Exclusive	2,299	2008-2018	Alzheimer's disease
Suárez-Araujo et al (2021) [95]	Retro. case control	Prevalent diagnostic	United States	Exclusive	330	Not reported	Alzheimer's disease
Tsang et al (2021) [96]	Retro. cohort	Prognostic	United Kingdom	Exclusive	592,988	1982-2015	Alzheimer's disease
Zafari et al (2021) [97]	Retro. cohort	Incident diagnostic	Canada	Exclusive	154,118	01/1995-12/2017	Post-traumatic stress disorder
Musculoskeletal and Connective Tissue Diseases							
Emir et al (2014) [98]	Retro. cohort	Incident diagnostic	United States	Exclusive	587,961	2011-2012	Fibromyalgia
Javrik et al (2018) [99]	Pros. cohort	Prognostic	United States	Exclusive	3,971	03/2011-03/2013	Back pain
Kennedy et al (2021) [100]	Retro. case control	Incident diagnostic	United Kingdom	Linked	23,528	Over 6 years	Ankylosing spondylitis
Neoplasms							
Kop et al (2016) [101]	Retro. cohort	Incident diagnostic	The Netherlands	Exclusive	260,000	Not reported	Colorectal cancer
Malhotra et al (2021) [102]	Pros. case control	Incident diagnostic	United Kingdom	Exclusive	56,953	01/2005-06/2009	Pancreatic cancer
Ristanoski et al (2021) [103]	Retro. case control	Incident diagnostic	Australia	Exclusive	683	2016-2017	Lung cancer
Nervous System Diseases							
Cox et al (2016) [104]	Retro. case control	Prevalent diagnostic	United Kingdom	Exclusive	35,510	01/2007-12/2011	Post stroke spasticity
Hrabok et al (2021) [105]	Retro. cohort	Prognostic	United Kingdom	Exclusive	10,499	01/2000-05/2012	Epilepsy
Kwasny et al (2021) [106]	Retro. case control	Incident diagnostic	Germany	Exclusive	3,274	01/2010-12/2017	Progressive supranuclear palsy
Respiratory System Diseases							
Afzal et al (2013) [107]	Retro. cohort	Prevalent diagnostic	The Netherlands	Exclusive	5,032	01/2000-01/2012	Asthma
Doyle et al (2020) [108]	Retro. case control	Incident diagnostic	United Kingdom	Exclusive	112,784	09/2003-09/2017	Nontuberculous mycobacterial lung disease
Kaplan et al (2020) [109]	Retro. cohort	Prevalent diagnostic	United States	Linked	411,563	Not reported	Asthma/obstructive pulmonary disease overlap
Lisspers et al (2021) [110]	Retro. cohort	Prognostic	Sweden	Linked	29,396	01/2000-12/2013	Asthma
Marin-Gomez et al (2021) [111]	Retro. cohort	Incident diagnostic	Spain	Exclusive	7,314	03/04/2020	COVID-19
Nikolaou et al (2021) [112]	Pros. cohort	Prognostic	United Kingdom	Exclusive	13,260	2015-2018	Chronic Obstructive Pulmonary Disease
Pikoula et al (2019) [30]	Retro. cohort	Prognostic	United Kingdom	Linked	30,961	01/1998-01/2016	Chronic Obstructive Pulmonary Disease
Ställberg et al (2021) [113]	Retro. cohort	Prognostic	Sweden	Linked	7,823	01/2000-12/2013	Chronic Obstructive Pulmonary Disease
Stephens et al (2020) [32]	Retro. case control	Incident diagnostic	United States	Exclusive	7,278	2009-2019	Influenza
Trtica-Majnaric et al (2010) [114]	Retro. cohort	Prognostic	Croatia	Exclusive	90	2003-2004	Influenza

Study	Study design	Objective	Country	Source of data set	Population	Period of extracted data	Health condition
Zafari et al (2022) [115]	Retro. cohort	Incident diagnostic	Canada	Exclusive	NR	Not reported	Chronic Obstructive Pulmonary Disease

Table 2 Overview of the included studies reporting ML prediction models developed using primary health care data with conduction of external validation using different datasets (n = 13)

Study	Study design	Objective	Country	Source of dataset	Population	Period of extracted data	Health condition
Endocrine, Metabolic, and Nutritional Diseases							
Hertroijs et al (2018) [116]	Retro. cohort	Prognostic	Netherlands	Exclusive	10,528	01/2006-12/2014	Diabetes mellitus
			Netherlands	Exclusive	3,337	01/2009-12/2013	
Myers et al (2019) [117]	Retro. case control	Incident diagnostic	United States	Exclusive	33,086	09/2013-08/2016	Familial hypercholesterolemia
			United States	Linked	7,805		
			United States	Linked	35,090		
			United States	Linked	8,094		
Perveen et al (2019) [29]	Retro. cohort	Prognostic	Canada	Exclusive	911	08/2003-06/2015	Diabetes mellitus
			Canada	Exclusive	1,970		
Weisman et al (2020) [118]	Retro. cohort	Prevalent diagnostic	Canada	Exclusive	5,402	2010-2017	Diabetes mellitus
			Canada	Linked	29,371		
Mental and Behavioral Diseases							
Amit et al (2021) [119]	Retro. cohort	Prevalent diagnostic	United Kingdom	Exclusive	24,612	2000-2010	Post-partum depression
			United Kingdom	Exclusive	9,193	2010-2017	
			United Kingdom	Exclusive	34,525	2000-2017	
Levy et al (2018) [120]	Pros. cohort	Incident diagnostic	United States	Exclusive	49	Over 9 months	Alzheimer's disease
			United States	Linked	26	Not reported	
Perlis (2013) [121]	Retro. cohort	Prognostic	United States	Exclusive	2,094	1999-2006	Major depressive disorder
			United States	Exclusive	461		

Study	Study design	Objective	Country	Source of dataset	Population	Period of extracted data	Health condition
Raket et al (2020) [45]	Retro. case control	Incident diagnostic	United States	Exclusive	145,720	1990-2018	Psychosis
			United States	Exclusive	4,770		
Musculoskeletal and Connective Tissue Diseases							
Fernandez-Gutierrez et al (2021) [122]	Retro. cohort	Incident diagnostic	United Kingdom	Linked	19,314	2002-2012	Rheumatoid arthritis & Ankylosing spondylitis
			United Kingdom	Linked	1,868		
Jorge et al (2019) [123]	Retro. cohort	Incident diagnostic	United States	Linked	400	Not reported	Systematic lupus erythematosus
			United States	Linked	173	Not reported	
Zhou et al (2017) [124]	Retro. cohort	Incident diagnostic	United Kingdom	Linked	Unclear	10/2013-07/2014	Rheumatoid arthritis
			United Kingdom	Linked	475,580	03/2009-10/2012	
Neoplasms							
Kinar et al (2016) [125]	Retro. cohort	Incident diagnostic	Israel	Exclusive	606,403	01/2003-07/2011	Colorectal cancer
			United Kingdom	Exclusive	30,674	01/2003-05/2012	
Pregnancy, Childbirth, Puerperium							
Suffriyana et al (2020) [126]	Retro. nested case control	Incident diagnostic	Indonesia	Linked	20,975	2015-2016	Preeclampsia
			Indonesia	Linked	1,322	Not reported	
			Indonesia	Linked	904	Not reported	

Table 3 Overview of the included studies reporting previously developed ML prediction models conducting only external validation using primary health care data (n = 12)

Study	Study design	Objective	Country	Source of dataset	Population	Period of extracted data	Health condition
Circulatory System Diseases							
Kostev et al (2021) [127]	Retro. cohort	Incident diagnostic	Germany	Exclusive	11,466	01/2010-12/2018	Stroke
Sekelj et al (2020) [128]	Retro. cohort	Incident diagnostic	United Kingdom	Exclusive	604,135	01/2001-12/2016	Atrial fibrillation
Endocrine, Metabolic, and Nutritional Diseases							
Abramoff et al (2019) [129]	Pros. cohort	Prevalent diagnostic	United States	Exclusive	819	01/2017-07/2017	Diabetic retinopathy
Bhaskaranand et al (2019) [130]	Retro. cohort	Prevalent diagnostic	United States	Exclusive	1,017,001	01/2014-09/2015	Diabetic retinopathy

González-Gonzalo et al (2019) [131]	Retro. case control	Prevalent diagnostic	Spain	Exclusive	288	08/2011-10/2016	Diabetic retinopathy
			Sweden	Exclusive			
			Denmark	Exclusive			
			United States	Linked	4,613	Over 2014	
			United Kingdom	Linked			
Kanagasingam et al (2018) [132]	Pros. cohort	Incident diagnostic	Australia	Exclusive	193	12/2016-05/2017	Diabetic retinopathy
Verbraak et al (2019) [133]	Retro. cohort	Prevalent diagnostic	The Netherlands	Exclusive	1,425	2015	Diabetic retinopathy
Neoplasms							
Birks et al (2017) [134]	Retro. case control	Incident diagnostic	United Kingdom	Exclusive	2,550,119	01/2000-04/2015	Colorectal cancer
Hoogendoorn et al (2016) [135]	Retro. case control	Prevalent diagnostic	The Netherlands	Exclusive	90,000	07/2006-12/2011	Colorectal cancer
Hornbrook et al (2017) [136]	Retro. case control	Incident diagnostic	United States	Linked	17,095	1998-2013	Colorectal cancer
Kinar et al (2017) [137]	Pros. cohort	Incident diagnostic	Israel	Linked	112,584	07/2007-12/2007	Colorectal cancer
Respiratory System Diseases							
Morales et al (2018) [41]	Retro. cohort	Prognostic	United Kingdom	Exclusive	2,044,733	01/2000-04/2014	Chronic Obstructive Pulmonary Disease

Health conditions

Out of the 22 classifications of the ICD-10 version 2019, 11 classifications were addressed in the included studies. Frequently reported classifications were the endocrine, nutritional, and metabolic diseases classification (ICD-10: E) (n= 27 studies of 109, 24.7%), circulatory system diseases (ICD-10: I) (n= 23, 21.1%), and the mental and behavioral disorders classification (ICD-10: F) (n= 22, 20.1%). To a lesser extent, diseases of the respiratory system classifications (ICD-10: J) and neoplasms (ICD-10: C) were addressed in (n=12, 11% and n= 8, 7.3% respectively). 35.9% of the included studies represent other health conditions from the remaining six ICD-10 classifications included. The health conditions addressed are depicted in Tables 1, 2, and 3 and S5 Appendix summary panel.

Endocrine, nutritional, and metabolic diseases (E00-E90)

In 27 studies addressing this classification [29,31,67–75,116,33,117,118,129–133,38,40,42,63–66], populations involved were from 12 countries, mainly the US (41.9%). The studies were published since 2008 with the highest number of studies in 2019 (38.7%). 81% of the included studies reported the development and/or training of the proposed models using exclusive primary health care data of a total number of 4.2 million participants. Data was extracted from different data sources over six months up to over almost 23 years. Four health conditions were identified in this ICD-10 classification, namely diabetes mellitus (E10, E11) with/without complications (n= 21), familial hypercholesterolemia (E78) (n= 3), children obesity (E66) (n= 2), and primary aldosteronism (E26) (n= 1). Incident diagnostic prediction was the most frequently reported outcome (42%). Prevalent diagnostic and prognostic prediction were 32% and 26% respectively. Diabetic retinopathy was the most common complication tackled (n=5 of 21 related diabetes mellitus studies), with using not routine primary health care investigations, such as fundoscopy that is used by the secondary health care. Diabetic foot identification was tackled in only one study using only the free text written by the physicians in the form of clinical notes as a predictor [31]. Two studies investigated prognostic predictive modelling of the short- and long-term levels of HbA1c after insulin treatment [72,116].

Mental and behavioral disorder (F00 – F99)

In 22 studies of this ICD-10 classifications addressing six health conditions [28,45,89–97,119,81,120,121,82–88], the involved population were from eight countries, mainly the US and the UK (n=14). These studies were published since 2013 with the highest number of studies in 2020 (44.4%). Data was extracted from different data sources with varying periods of health records follow up, from one year to almost 28 years. Dementia/Alzheimer's disease (F00) was

addressed in 13 studies, of which one study predicted it within the progression of mental cognitive impairment [85], while another study predicted hospitalization risk [96]. Major depressive disorder (F32) (n= 3) which a study predicted its prognosis within two years and suggested considering the severity of the baseline symptoms for depression prediction [81]. A study claimed to be the first to predict first episode psychosis (F29) and suggested that considerable proportion of the most predictive features were not of a psychiatric nature [45]. A study predicted anxiety (F41) in cancer survivors seeking care in PHC and suggested that fatigue and insomnia were the most important predictors [86]. Lastly, a study used PHC data to predict any mental disorder using different ML modes, claimed that the potentially successful prediction was the best before 180 days of real diagnosis [93].

Circulatory and respiratory health conditions (I00-I99 and J00-J99)

In 35 studies addressing these two ICD-10 classifications, populations involved were from 11 countries, mainly the US and the UK. All the included studies were published since 2010 with the highest number of studies in both groups in 2020 (30.8%). Data was extracted from the different data sources over highly variable period from one month to almost 23 years of longitudinal data.

Six circulatory health conditions were identified in 23 studies [27,34,51–60,35,61,127,128,37,43,46–50]. These conditions were hypertension (I10-I15) (n= 5), heart failure (I50) (n= 5), atrial fibrillation (I48) (n= 2), stroke (I64) (n= 2), atherosclerosis (I70) (n=1), myocardial infarction (I21) (n=1), and any cardiovascular event or disease (n=7). Variable conclusions were reported across these studies. For example, a study reported that variations in the importance of different risk factors depend on the modelling technique [35]. Another study reported that ignoring censoring substantially underestimated risk of cardiovascular disease [43]. Also, systolic blood pressure could be used as a predictor of

hospitalization or mortality [59]. Lastly, predictions levels increase after two years and 4000 patients as requirement to predict incident HF cases, and variability of the best performing model could depend on the method of handling the missed data [53].

Five respiratory health conditions were identified in 12 studies [30,32,114,115,41,107–113]. Chronic obstructive pulmonary disease (COPD) (J40) (n= 5) studies were prognostic predictive studies tackling mortality and hospitalization. Asthma (J45) (n= 2) studies identified known undocumented cases and predicted exacerbation prognosis. COVID-19 (U07) incident cases were predicted within routine PHC visits in one study [111]. With employing results of polychain polymerase reaction (PCR) as predictors, a tree classification approach was reported to be potentially useful in detecting the existence of COVID-19 infection [111]. Contact with a previously infected person was reported as the key factor linked to the development of COVID-19, with recommendation to early detect and isolate the contacts [111].

Other health conditions

Eight studies addressed three neoplasms, namely colorectal cancer (CRC) (C18) (n= 6), lung cancer (C34) (n= 1), and pancreatic cancer (C25) (n= 1). Four studies addressed the same incidence prediction model known as ColonFlag (previously MeScore) to identify CRC cases [125,134,136,137]. Each study predicted incident cases within different time windows before diagnosis; over three months to two years with relative high discrimination ability of the proposed model across the four studies. This ColonFlag model was reported as ‘well-performing’ when used on CRC cases detected at early asymptomatic (often nonanemic), localized stages, as well as when limited to complete blood picture data collected around a year before diagnosis [125].

Three health conditions affecting the nervous system were addressed [104–106], one of which predicted mortality four years before and after diagnosis of epilepsy (G40) with an acceptable performance for identifying those at high risk of early premature mortality [105]. Another study predicted a rare neurodegenerative disease, progressive supranuclear palsy (G23), identified two previously unknown clinical features as predictors associated with the pre-diagnostic stage of this disease [106].

Regarding musculoskeletal and connective tissue disorders [98–100,122–124], back pain (M54) prognosis within PHC settings could be predicted through focusing on patient function-related predictors more than on resolving pain [99]. A study revealed that models can be created using only data from medical records and had prediction values of 70-80% for identifying persons who are at risk of acquiring ankylosing spondylitis (M45) [100]. Two digestive health conditions were addressed in two studies [36,62], which were inflammatory bowel diseases (K50-K52), including Crohn’s disease and ulcerative colitis, and peptic ulcers (K27)/gastroesophageal reflux (K21). Two studies addressed the chronic kidney disease (N18) [28,79], one of them was an incident diagnostic and the other predicted hospitalization and steroid use within 6 months and one year. Three studies tackled suicidality (X60-X84) [76–78], one of which predicted incidence of suicide and reported that PHC records were of little indication of severity [76]. Lastly, one study addressed preeclampsia (O14) with additional reporting of a systematic review of this disease across different health care sectors [126].

Quality assessment

Addressing the included studies using the PROBAST tool revealed that 90.8% (n=99 of 109) of the included studies were of high and unclear risk of bias, as depicted in Fig 5. Analysis domain was the main source of bias, because of underreporting. Additionally, the studies of potential low risk of bias were downgraded from high risk due to the of lack of

external validation of the proposed models (n=20). Only a few studies (n= 11) were reported in accordance with transparent reporting of a multivariable prediction model for individual

Figure 4 Percentage presentation of the results of PROBAST of the two components: risk of bias (4 domains: Participants, predictors, outcome, and analysis) and concern of applicability (3 domains: Participants, predictors, and outcome)

prognosis or diagnosis (TRIPOD) guidelines [138]. Concern of applicability of the addressed models in the PHC of 72 (66%) studies was low. The main source of this concern is the dependence of the predictive models on not-routine PHC data.

Most of the included studies (n= 101 of 109, 92.7%) were published as peer-reviewed publications in biomedical (e.g., PLOS ONE, n= 8) and technical journals (e.g., IEEE, n= 3). Eight studies were preprint and abstracts. National research institutes and universities were the most frequently reported funding support. Most of the studies reported that the funding supporters were not involved in the process or results of the published work. Nevertheless, some studies were supported by industrial companies without clarifying the role of the funding body.

Discussion

ML prediction models could have an immense potential to augment health practice and clinical decision making in various health sectors. Our systematic review provides an outline and summary of the health conditions tackled through ML prediction models using PHC data.

42 health conditions across 109 observational studies were identified. 76.6% of the included studies were diagnostic, while 23.4% were predictive of complications, hospitalization, or morbidity. Alzheimer's disease, diabetes mellitus, heart failure, colorectal cancer, and chronic obstructive pulmonary diseases were the most frequently targeted health

conditions. Less attention was directed to the other reported diseases, such as asthma, children obesity, and dyspepsia.

In the context of PHC, detection and management of evitable and controllable chronic health conditions, such as diabetes mellitus are part of the most vital role of this health care settings [3]. On the other hand, misdiagnosis of diseases can result in abandoned symptoms, ineffective treatment, and preventable deaths [3]. Despite of the early stage of the ML prediction models of such health conditions in PHC [139], primary care setting have gained more attention in many countries [11], similar to our findings. Furthermore, predicting undocumented cases and rare diseases with potential good performance was also reported. Nevertheless, it is suggested to investigating the prediction of other diseases incidence and progression among the health care providers, researchers, and models' developers.

Health conditions diagnostic and prognostic predictions were performed using 273 ML models mainly of supervised learning technique. The models within 77% of the included studies were trained and/or internally validated without evaluating their generalizability. The other 33% of studies present those conducted internal and/or external validation. Despite relatively good performance reported across the included studies, their clinical implication is limited, and further investigations are needed. Furthermore, lack of reporting guidelines usage and overall risk of bias assessment of high to unclear raise concerns about the potential disadvantages of such models.

Technical biases could influence the clinical practice. When a model is trained on historical data, which supports old practice without adaptation to policy changes, then the model reinforce an outmoded practice [140]. Furthermore, due to bias in the training set, change over time, or application of the system in a different population, a mismatch between the data or environment on which the system is trained and that used in operation may result

in an erroneous result [140]. This bias could affect the results of some of the included models being trained with data generated up to 40 years ago. Additionally, lack of reporting the different health systems prevents the proper estimation of the applicability of external validation results. Hence, it is recommended to properly report the development and performance measures attributes of the models under progress in the present and the future. Additionally, it is encouraged to validate the proposed models within different geographical and temporal settings with proper reporting of the main up-to-date criteria of the health system addressed as well.

The main source of the extracted participants' data was exclusive primary health care data. Various other sources of data were linked to the PHC data, such as secondary and tertiary health care. These health data represented a total of 24.2 million participants mainly within PHC settings. A large majority of the models' development and/or validation was conducted in the US and the UK (58.1%) with a noticeable rise since 2018.

Despite the fact that big data generated through the health records is a strong fit for ML tools, the coding system itself does not universally follow the same criteria for diseases [141]. Furthermore, PHC has no standardized definition globally with a wide variability of the services provided. Hence, different health system and terminology of diseases and symptoms across the world could limit the consistency of the models' performance [141]. Additionally, uncoded free-text clinical notes and the lack of proper coding, such as using ('race' and 'ethnicity') and ('suicide' and 'suicide attempts') to be documented as a single input, affect the predictive power of the models [142]. Other drawbacks reported, similar to our findings, were underrepresentation of healthy persons and retrospective temporal dimension of the extracted predictors [142]. Therefore, routine care data collected according to a documentation system might not fully match the proposed questions with the models' developed. Additionally,

misclassification bias and incomplete health records represented a major limitation, as reported in the included studies. Even with proper classification, certain diseases require confirmatory diagnosis using higher care services, such as magnetic resonance imaging (MRI) [143], which is missing from PHC. Therefore, it would be advisable that models' developers propose solutions for the digital documentation systems, when possible, based on the addressed health condition to overcome the limitations faced with discussing these solutions' benefits and applicability. With that approach, more evidence-based literature would be available for the stakeholders to implement further enhancements.

On the other hand, the unequal distribution of papers across countries could be related partially to the low publication rate in the low-income countries or lack of proper big data documentation systems. However, this justification does not clarify the reason of the unequal distribution of publication among the middle- and high-income countries. Hence, the transition from using the conventional medical records to integrating the predictive models in PHC is far from simple and necessitates specialized processing techniques. Furthermore, solid technical infrastructure as well as strong academic and governmental support are essential for promoting and supporting long-term and broad-reaching PHC data gathering efforts [142,144]).

Lastly, based on the high variability of the structure and reporting styles identified across the studies, i.e., medical versus engineering point of view and style, it is recommended to augment the participation of health professionals through the development process of the health related predictive models to critically evaluate, assess, adopt, and challenge the validation of the models within practices, given the increasing popularity of digitally connected health systems [5]. Furthermore, ML engineers must be aware of the unintended implications of their algorithms and ensure that they are created with the global and local communities in mind [145]. Hence, it is advisable to obtain an efficient cooperation between ML developers

and the health care professionals to provide new insights for tackling the potential biases. Additionally, it is suggested to integrate the basic understanding of ML concepts and techniques among the under- and post-graduate education programs.

Strength and limitations

Our review was conducted following a predesigned comprehensive protocol [21]. We identified the health conditions targeted within primary care settings as an encompassing of literature and identifying the gaps needed to be tackled. However, the main limitation of our review's quality of evidence, first, is the reviewing of observational studies that mostly lacked external validation of the proposed models. Second, regarding our search strategy, some studies could have been missed if they exclusively used 'big data', 'statistical modelling', 'statistical learning' or similar terms instead of our search string as noted in [146]. Third, limiting our scope to the clinical health conditions resulted in excluding other conditions that could be reported and predicted within the PHC, such as domestic violence and drug abuse [3]. Fourth, guiding our work using ICD-10 might lead to excluding potentially relevant studies, such as a study that used frailty as a medical syndrome [147]. Lastly, we neither extract thoroughly the performance measures of each study nor conduct a meta-analysis, because of the broad heterogeneity across studies. In the future, we plan to update our review - considering this noticeable rise of the PHC ML studies – while also modifying our methodology to reduce the identified limitations. Additionally, we plan to use the new under-progress specific ML guidelines TRIPOD-AI and PROBAST-AI when published to strengthen quality and reporting of our findings [148].

In conclusion, ML prediction models within PHC is gaining traction. Further studies are needed, especially those with prospective designs and more representative samples. Working among multi-discipline teams to tackle ML in primary care increases the trust of the

models and their implementations with further consideration of improving quality of development and reporting of the ML predictive models. More research is required to continue to fill the gaps in knowledge surrounding the emergence of PHC data.

Acknowledgment

Dr. Marcos André Gonçalves, PhD and his colleague Bruna Zanotto, MSc. provided their feedback on the project's primary draft. Dr. phil. Luana Fiengo Tanaka retrieved the inaccessible studies.

References

1. Aoki M. Editorial: Science and roles of general medicine. *Japanese J Natl Med Serv.* 2001;55: 111–114. doi:10.11261/iryō1946.55.111
2. Troncoso EL. The Greatest Challenge to Using AI/ML for Primary Health Care: Mindset or Datasets? *Front Artif Intell.* 2020;3: 53. doi:10.3389/frai.2020.00053
3. Hashim MJ. A definition of family medicine and general practice. *J Coll Physicians Surg Pakistan.* 2018;28: 76–77. doi:10.29271/jcpsp.2018.01.76
4. Cao L. Data science: A comprehensive overview. *ACM Comput Surv.* 2018;50: 1–42. doi:10.1145/3076253
5. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemy C, Terry AL, et al. Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges. *Yearb Med Inform.* 2019;28: 41–46. doi:10.1055/s-0039-1677901
6. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to

- systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460. doi:10.1136/bmj.i6460
7. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput Sci*. 2021;2: 160. doi:10.1007/s42979-021-00592-x
 8. Do Nascimento IJB, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Goncalves MA, et al. Impact of big data analytics on people's health: Overview of systematic reviews and recommendations for future studies. *J Med Internet Res*. 2021;23: e27275. doi:10.2196/27275
 9. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. *Curr HIV/AIDS Rep*. 2020;17: 171–179. doi:10.1007/s11904-020-00490-6
 10. Amaratunga D, Cabrera J, Sargsyan D, Kostis JB, Zinonos S, Kostis WJ. Uses and opportunities for machine learning in hypertension research. *Int J Cardiol Hypertens*. 2020;5: 100027. doi:10.1016/j.ijchy.2020.100027
 11. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J*. 2017;15: 104–116. doi:10.1016/j.csbj.2016.12.005
 12. Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, Yeh TY, et al. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis. *JMIR Med Informatics*. 2020;8: e16503. doi:10.2196/16503
 13. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy

- and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. *Br J Dermatol.* 2009;161: 591–604. doi:10.1111/j.1365-2133.2009.09093.x
14. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One.* 2020;15: e0234722. doi:10.1371/journal.pone.0234722
 15. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. *J Med Internet Res.* 2018;20: e10775. doi:10.2196/10775
 16. Rahimi SA, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: Systematic scoping review and critical appraisal. *Journal of Medical Internet Research J Med Internet Res*; Sep 1, 2021. doi:10.2196/29839
 17. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: A scoping review. *Ann Fam Med.* 2020;18: 250–258. doi:10.1370/afm.2518
 18. Andaur Navarro CL, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open.* 2020;10: e038832. doi:10.1136/bmjopen-2020-038832
 19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ.* British Medical Journal Publishing Group; 2021. doi:10.1136/bmj.n71
 20. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et

- al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med.* 2014;11: e1001744. doi:10.1371/journal.pmed.1001744
21. Abdulazeem H, Whitelaw S, Schauburger G, Klug S. Development and Performance of Prediction Machine Learning Models supplied by Real-World Primary Health Care Data: A Systematic Review and Meta-analysis. In: PROSPERO 2021 CRD42021264582 [Internet]. 2021. Available: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021264582
 22. Schapire RE. The Strength of Weak Learnability. *Mach Learn.* 1990;5: 197–227. doi:10.1023/A:1022648800760
 23. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev.* 2016;5: 210. doi:10.1186/s13643-016-0384-4
 24. World Health Organization. ICD-10 Version:2019. In: International Classification of Diseases [Internet]. 2019 [cited 1 Sep 2021]. Available: <https://icd.who.int/browse10/2019/en#/XIV>
 25. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med.* 2019;170: W1–W33. doi:10.7326/M18-1377
 26. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015;44: 827–836. doi:10.1093/ije/dyv098

27. Shah AD, Bailey E, Williams T, Denaxas S, Dobson R, Hemingway H. Natural language processing for disease phenotyping in UK primary care records for research: A pilot study in myocardial infarction and death. *J Biomed Semantics*. 2019;10. doi:10.1186/s13326-019-0214-4
28. Alexander N, Alexander DC, Barkhof F, Denaxas S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med Inform Decis Mak*. 2021;21. doi:10.1186/s12911-021-01693-6
29. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. *Sci Rep*. 2019;9: 13805. doi:10.1038/s41598-019-49563-6
30. Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. *BMC Med Inform Decis Mak*. 2019;19: 86. doi:10.1186/s12911-019-0805-0
31. Pakhomov SVS, Hanson PL, Bjornsen SS, Smith SA. Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning. *J Am Med Informatics Assoc*. 2008;15: 198–202. doi:10.1197/jamia.M2585
32. Stephens KA, Au MA, Yetisgen M, Lutz B, Suchsland MZ, Ebell MH, et al. Leveraging UMLS-driven NLP to enhance identification of influenza predictors derived from electronic medical record data. In: *BioRxiv [preprint] [Internet]*. 2020 [cited 4 Jan 2022]. doi:10.1101/2020.04.24.058982
33. Tseng E, Schwartz JL, Rouhizadeh M, Maruthur NM. Analysis of Primary Care

- Provider Electronic Health Record Notes for Discussions of Prediabetes Using Natural Language Processing Methods. *J Gen Intern Med.* 2021;35: S11–S12. doi:10.1007/s11606-020-06400-1
34. Zhao Y, Fu S, Bielinski SJ, Decker P, Chamberlain AM, Roger VL, et al. Abstract P259: Using Natural Language Processing and Machine Learning to Identify Incident Stroke From Electronic Health Records. *Circulation.* 2020;141. doi:10.1161/circ.141.suppl_1.p259
35. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One.* 2017;12. doi:10.1371/journal.pone.0174944
36. Sáenz Bajo N, Barrios Rueda E, Conde Gómez M, Domínguez Macías I, López Carabaño A, Méndez Díez C. Use of neural networks in medicine: concerning dyspeptic pathology. *Aten Primaria.* 2002;30: 99–102. doi:10.1016/s0212-6567(02)78978-6
37. Hill NR, Ayoubkhani D, McEwan P, Sugrue DM, Farooqui U, Lister S, et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS One.* 2019;14: e0224582. doi:10.1371/journal.pone.0224582
38. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform.* 2015;6: 506–520. doi:10.4338/ACI-2015-03-RA-0036
39. Barons MJ, Parsons N, Griffiths F, Thorogood M. A comparison of artificial neural network, latent class analysis and logistic regression for determining which patients benefit from a cognitive behavioural approach to treatment for non-specific low back pain. 2013 IEEE Symposium on Computational Intelligence in Healthcare and e-health

- (CICARE). University of Warwick, Coventry CV4 7 AL, United Kingdom: IEEE; 2013. pp. 7–12. doi:10.1109/CICARE.2013.6583061
40. Ding X, Ajmal I, Trerotola OSc, Fraker D, Cohen J, Wachtel H, et al. EHR-based modeling specifically identifies patients with primary aldosteronism. In: *Circulation* [Internet]. 2019 [cited 22 Sep 2021]. Available: <https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=emed20&AN=630921513>
 41. Morales DR, Flynn R, Zhang J, Trucco E, Quint JK, Zutis K. External validation of ADO, DOSE, COTE and CODEX at predicting death in primary care patients with COPD using standard and machine learning approaches. *Respir Med*. 2018;138: 150–155. doi:10.1016/j.rmed.2018.04.003
 42. Álvarez-Guisasola F, Conget I, Franch J, Mata M, Mediavilla JJ, Sarria A, et al. Adding questions about cardiovascular risk factors improve the ability of the ADA questionnaire to identify unknown diabetic patients in Spain. *Diabetologia*. 2010;26: 347–352. doi:10.1016/S1134-3230(10)65008-9
 43. Li Y, Sperrin M, Ashcroft DM, Van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: Longitudinal cohort study using cardiovascular disease as exemplar. *BMJ*. 2020;371: m3919. doi:10.1136/bmj.m3919
 44. Ngufor C, Caraballo PJ, O’Byrne TJ, Chen D, Shah ND, Pruinelli L, et al. Development and Validation of a Risk Stratification Model Using Disease Severity Hierarchy for Mortality or Major Cardiovascular Event. *JAMA Netw Open*. 2020;3. doi:10.1001/jamanetworkopen.2020.8270

45. Raket LL, Jaskolowski J, Kinon BJ, Brasen JC, Jönsson L, Wehnert A, et al. Dynamic ElecTronic hEalth reCord deTectioN (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digit Heal.* 2020;2: e229–e239. doi:10.1016/S2589-7500(20)30024-8
46. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: Implications for temporal modeling with respect to time before diagnosis, data density, data quantity, and data type. *Circ Cardiovasc Qual Outcomes.* 2019;12: e005114. doi:10.1161/CIRCOUTCOMES.118.005114
47. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Informatics Assoc.* 2017;24: 361–370. doi:10.1093/jamia/ocw112
48. Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, et al. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation. *JMIR Med Informatics.* 2020;8: e17257. doi:10.2196/17257
49. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open.* 2013;3. doi:10.1136/bmjopen-2012-002457
50. Karapetyan S, Schneider A, Linde K, Donnachie E, Hapfelmeier A. SARS-CoV-2 infection and cardiovascular or pulmonary complications in ambulatory care: A risk assessment based on routine data. *PLoS One.* 2021;16: e0258914.

doi:10.1371/journal.pone.0258914

51. LaFreniere D, Zulkernine F, Barber D, Martin K. Using machine learning to predict hypertension from a clinical dataset. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE; 2016. pp. 1–7. doi:10.1109/SSCI.2016.7849886
52. Lip S, McCallum L, Reddy S, Chandrasekaran N, Tule S, Bhaskar RK, et al. Machine Learning Based Models for Predicting White-Coat and Masked Patterns of Blood Pressure. *J Hypertens*. 2021;39: e69. doi:10.1097/01.hjh.0000745092.07595.a5
53. Lorenzoni G, Sabato SS, Lanera C, Bottigliengo D, Minto C, Ocagli H, et al. Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. *J Clin Med*. 2019/08/28. 2019;8. doi:10.3390/jcm8091298
54. Ng K, Steinhubl SR, Defilippi C, Dey S, Stewart WF. Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time before Diagnosis, Data Diversity, Data Quantity, and Data Density. *Circ Cardiovasc Qual Outcomes*. 2016;9: 649–658. doi:10.1161/CIRCOUTCOMES.116.002797
55. Nikolaou V, Massaro S, Garn W, Fakhimi M, Stergioulas L, Price D. The cardiovascular phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying machine learning to the prediction of cardiovascular comorbidities. *Respir Med*. 2021/07/15. 2021;186: 106528. doi:10.1016/j.rmed.2021.106528
56. Sarraju A, Ward A, Chung S, Li J, Scheinker D, Rodríguez F. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Hear*. 2021;8: e001802. doi:10.1136/openhrt-2021-001802
57. Selskyy P, Vakulenko D, Televiak A, Veresiuk T. On an algorithm for decision-making

- for the optimization of disease prediction at the primary health care level using neural network clustering. *Fam Med Prim Care Rev.* 2018;20: 171–175. doi:10.5114/fmpcr.2018.76463
58. Solanki P, Ajmal I, Ding X, Cohen J, Cohen D, Herman D. Abstract P185: Using Electronic Health Records To Identify Patients With Apparent Treatment Resistant Hypertension. *Hypertension.* 2020;76. doi:10.1161/hyp.76.suppl_1.p185
 59. Ayala Solares JR, Canoy D, Raimondi FED, Zhu Y, Hassaine A, Salimi-Khorshidi G, et al. Long-Term Exposure to Elevated Systolic Blood Pressure in Predicting Incident Cardiovascular Disease: Evidence From Large-Scale Routine Electronic Health Records. *J Am Heart Assoc.* 2019;8. doi:10.1161/JAHA.119.012129
 60. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digit Med.* 2020;3: 125. doi:10.1038/s41746-020-00331-1
 61. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Med Care.* 2010;48: S106–S113. doi:10.1097/MLR.0b013e3181de9e17
 62. Waljee AK, Lipson R, Wiitala WL, Zhang Y, Liu B, Zhu J, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm Bowel Dis.* 2018;24: 45–53. doi:10.1093/ibd/izx007
 63. Akyea RK, Qureshi N, Kai J, Weng SF. Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care. *NPJ Digit Med.* 2020;3: 142. doi:10.1038/s41746-020-00349-5

64. Crutzen S, Belur Nagaraj S, Taxis K, Denig P. Identifying patients at increased risk of hypoglycaemia in primary care: Development of a machine learning-based screening tool. *Diabetes Metab Res Rev.* 2021;37: e3426. doi:10.1002/dmrr.3426
65. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A, Thanaraj TA. Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. *Front Endocrinol (Lausanne).* 2019;10. doi:10.3389/fendo.2019.00624
66. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, et al. Predicting childhood obesity using electronic health records and publicly available data. *PLoS One.* 2019;14: e0215571. doi:10.1371/journal.pone.0215571
67. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep.* 2020;10: 11981. doi:10.1038/s41598-020-68771-z
68. Lethebe BC, Williamson T, Garies S, McBrien K, Leduc C, Butalia S, et al. Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. *C open.* 2019;7: E246–E251. doi:10.9778/cmajo.20180142
69. Looker HC, Colombo M, Hess S, Brosnan MJ, Farran B, Dalton RN, et al. Biomarkers of rapid chronic kidney disease progression in type 2 diabetes. *Kidney Int.* 2015;88: 888–896. doi:10.1038/ki.2015.199
70. Metsker O, Magoev K, Yanishevskiy S, Yakovlev A, Kopanitsa G, Zvartau N. Identification of diabetes risk factors in chronic cardiovascular patients. *Stud Health Technol Inform.* 2020;273: 136–141. doi:10.3233/SHTI200628

71. Metzker O, Magoev K, Yanishevskiy S, Yakovlev A, Kopanitsa G. Risk factors for chronic diabetes patients. *Stud Health Technol Inform.* 2020;270: 1379–1380. doi:10.3233/SHTI200451
72. Nagaraj SB, Sidorenkov G, van Boven JFM, Denig P. Predicting short- and long-term glycosylated haemoglobin response after insulin initiation in patients with type 2 diabetes mellitus using machine-learning algorithms. *Diabetes, Obes Metab.* 2019;21: 2704–2711. doi:10.1111/dom.13860
73. Rumora AE, Guo K, Alakwaa FM, Andersen ST, Reynolds EL, Jørgensen ME, et al. Plasma lipid metabolites associate with diabetic polyneuropathy in a cohort with type 2 diabetes. *Ann Clin Transl Neurol.* 2021;8: 1292–1307. doi:10.1002/acn3.51367
74. Wang J, Lv B, Chen X, Pan Y, Chen K, Zhang Y, et al. An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres. *BMC Pregnancy Childbirth.* 2021;21: 814. doi:10.1186/s12884-021-04295-2
75. Williamson L, Wojcik C, Taunton M, McElheran K, Howard W, Staszak D, et al. Finding Undiagnosed Patients With Familial Hypercholesterolemia in Primary Care Using electronic Health Records. *J Am Coll Cardiol.* 2020;75: 3502. doi:10.1016/s0735-1097(20)34129-2
76. DelPozo-Banos M, John A, Petkov N, Berridge DM, Southern K, Loyd KL, et al. Using neural networks with routine health records to identify suicide risk: Feasibility study. *JMIR Ment Heal.* 2018;5: e10144. doi:10.2196/10144
77. Penfold RB, Johnson E, Shortreed SM, Ziebell RA, Lynch FL, Clarke GN, et al. Predicting suicide attempts and suicide deaths among adolescents following outpatient

- visits. *J Affect Disord.* 2021;294: 39–47. doi:10.1016/j.jad.2021.06.057
78. van Mens K, Elzinga E, Nielen M, Lokkerbol J, Poortvliet R, Donker G, et al. Applying machine learning on health record data from general practitioners to predict suicidality. *Internet Interv.* 2020;21: 100337. doi:10.1016/j.invent.2020.100337
79. Shih CC, Lu CJ, Chen G Den, Chang CC. Risk prediction for early chronic kidney disease: Results from an adult health examination program of 19,270 individuals. *Int J Environ Res Public Health.* 2020;17: 1–11. doi:10.3390/ijerph17144973
80. Zhao J, Gu S, McDermaid A. Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Math Biosci.* 2019;310: 24–30. doi:10.1016/j.mbs.2019.02.001
81. Dinga R, Marquand AF, Veltman DJ, Beekman ATF, Schoevers RA, van Hemert AM, et al. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. *Transl Psychiatry.* 2018;8: 241. doi:10.1038/s41398-018-0289-1
82. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee 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.* 2019;19: 248. doi:10.1186/s12911-019-0991-9
83. Ford E, Starlinger J, Rooney P, Oliver S, Banerjee S, van Marwijk H, et al. Could dementia be detected from UK primary care patients' records by simple automated methods earlier than by the treating physician? A retrospective case-control study. *Wellcome Open Res.* 2020;5: 120. doi:10.12688/wellcomeopenres.15903.1

84. Ford E, Sheppard J, Oliver S, Rooney P, Banerjee S, Cassell JA. Automated detection of patients with dementia whose symptoms have been identified in primary care but have no formal diagnosis: A retrospective case-control study using electronic primary care records. *BMJ Open*. 2021;11: e039248. doi:10.1136/bmjopen-2020-039248
85. Fouladvand S, Mielke MM, Vassilaki M, St. Sauver J, Petersen RC, Sohn S. Deep Learning Prediction of Mild Cognitive Impairment using Electronic Health Records. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019. pp. 799–806. doi:10.1109/BIBM47256.2019.8982955
86. Haun MW, Simon L, Sklenarova H, Zimmermann-Schlegel V, Friederich HC, Hartmann M. Predicting anxiety in cancer survivors presenting to primary care – A machine learning approach accounting for physical comorbidity. *Cancer Med*. 2021;10: 5001–5016. doi:10.1002/cam4.4048
87. Jammeh EA, Carroll CB, Pearson Stephen W, Escudero J, Anastasiou A, Zhao P, et al. Machine-learning based identification of undiagnosed dementia in primary care: A feasibility study. *BJGP Open*. 2018;2: bjgpopen18X101589-bjgpopen18X101589. doi:10.3399/bjgpopen18X101589
88. Jin H, Wu S. Use of patient-reported data to match depression screening intervals with depression risk profiles in primary care patients with diabetes: Development and validation of prediction models for major depression. *JMIR Form Res*. 2019;3: e13610–e13610. doi:10.2196/13610
89. Kaczmarek E, Salgo A, Zafari H, Kosowan L, Singer A, Zulkernine F. Diagnosing PTSD using electronic medical records from Canadian primary care data. *ACM International Conference Proceeding Series*. School of Computing, Queen’s University,

Kingston, Canada; 2019. pp. 23–29. doi:10.1145/3362966.3362982

90. Ljubic B, Roychoudhury S, Cao XH, Pavlovski M, Obradovic S, Nair R, et al. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. *Comput Methods Programs Biomed.* 2020;197: 105765. doi:10.1016/j.cmpb.2020.105765
91. Mallo SC, Valladares-Rodriguez S, Facal D, Lojo-Seoane C, Fernández-Iglesias MJ, Pereiro AX. Neuropsychiatric symptoms as predictors of conversion from MCI to dementia: A machine learning approach. *Int Psychogeriatrics.* 2020;32: 381–392. doi:10.1017/S1041610219001030
92. Mar J, Gorostiza A, Ibarondo O, Cernuda C, Arrospide A, Iruin A, et al. Validation of Random Forest Machine Learning Models to Predict Dementia-Related Neuropsychiatric Symptoms in Real-World Data. *J Alzheimer's Dis.* 2020;77: 855–864. doi:10.3233/JAD-200345
93. Pólchłopek O, Koning NR, Büchner FL, Crone MR, Numans ME, Hoogendoorn M. Quantitative and temporal approach to utilising electronic medical records from general practices in mental health prediction. *Comput Biol Med.* 2020;125. doi:10.1016/j.combiomed.2020.103973
94. Shen X, Wang G, Rick Yiu-Cho Kwan, Choi KS. Using dual neural network architecture to detect the risk of dementia with community health data: Algorithm development and validation study. *JMIR Med Informatics.* 2020;8: e19870. doi:10.2196/19870
95. Suárez-Araujo CP, García Báez P, Cabrera-León Y, Prochazka A, Rodríguez Espinosa N, Fernández Viadero C, et al. A Real-Time Clinical Decision Support System, for Mild Cognitive Impairment Detection, Based on a Hybrid Neural Architecture. *Bangyal WH,*

- editor. *Comput Math Methods Med.* 2021;2021: 1–9. doi:10.1155/2021/5545297
96. Tsang G, Zhou SM, Xie X. Modeling Large Sparse Data for Feature Selection: Hospital Admission Predictions of the Dementia Patients Using Primary Care Electronic Health Records. *IEEE J Transl Eng Heal Med.* 2021;9. doi:10.1109/JTEHM.2020.3040236
 97. Zafari H, Kosowan L, Zulkernine F, Signer A. Diagnosing post-traumatic stress disorder using electronic medical record data. *Health Informatics J.* 2021;27. doi:10.1177/14604582211053259
 98. Emir B, Mardekian J, Masters ET, Clair A, Kuhn M, Silverman SL. Predictive modeling of a fibromyalgia diagnosis: Increasing the accuracy using real world data. Meeting: 2014 ACR/ARHP Annual Meeting. ACR; 2014.
 99. Jarvik JG, Gold LS, Tan K, Friedly JL, Nedeljkovic SS, Comstock BA, et al. Long-term outcomes of a large, prospective observational cohort of older adults with back pain. *Spine J.* 2018;18: 1540–1551. doi:10.1016/j.spinee.2018.01.018
 100. Kennedy J, Kennedy N, Cooksey R, Choy E, Siebert S, Rahman M, et al. Predicting a diagnosis of ankylosing spondylitis using primary care health records – a machine learning approach. *medRxiv.* 2021; 2021.04.22.21255659. doi:10.1101/2021.04.22.21255659
 101. Kop R, Hoogendoorn M, Teije A ten, Büchner FL, Slottje P, Moons LMG, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Comput Biol Med.* 2016;76: 30–38. doi:10.1016/j.combiomed.2016.06.019
 102. Malhotra A, Rachet B, Bonaventure A, Pereira SP, Woods LM. Can we screen for

- pancreatic cancer? Identifying a sub-population of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. *PLoS One*. 2021;16: e0251876–e0251876. doi:10.1371/journal.pone.0251876
103. Ristanoski G, Emery J, Gutierrez JM, McCarthy D, Aickelin U. Primary Care Datasets for Early Lung Cancer Detection: An AI Led Approach. *Lecture Notes in Computer Science*. AIME; 2021. pp. 83–92. doi:10.1007/978-3-030-77211-6_9
 104. Cox AP, Raluy M, Wang M, Bakheit AMO, Moore AP, Dinet J, et al. Predictive analysis for identifying post stroke spasticity patients in UK primary care data. *Pharmacoepidemiol Drug Saf*. 2014;23: 422–423.
 105. Hrabok M, Engbers JDT, Wiebe S, Sajobi TT, Subota A, Almohawes A, et al. Primary care electronic medical records can be used to predict risk and identify potentially modifiable factors for early and late death in adult onset epilepsy. *Epilepsia*. 2021;62: 51–60. doi:10.1111/epi.16738
 106. Kwasny MJ, Oleske DM, Zamudio J, Diegidio R, Höglinger GU. Clinical Features Observed in General Practice Associated With the Subsequent Diagnosis of Progressive Supranuclear Palsy. *Front Neurol*. 2021;12: 637176. doi:10.3389/fneur.2021.637176
 107. Afzal Z, Engelkes M, Verhamme KMC, Janssens HM, Sturkenboom MCJM, Kors JA, et al. Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. *Pharmacoepidemiol Drug Saf*. 2013;22: 826–833. doi:10.1002/pds.3438
 108. Doyle OM, van der Laan R, Obradovic M, McMahon P, Daniels F, Pitcher A, et al. Identification of potentially undiagnosed patients with nontuberculous mycobacterial lung disease using machine learning applied to primary care data in the UK. *Eur Respir*

J. 2020;56: 2000045. doi:10.1183/13993003.00045-2020

109. Kaplan A, Cao H, Fitzgerald JM, Yang E, Iannotti N, Kocks JWH, et al. Asthma/COPD Differentiation Classification (AC/DC): Machine Learning to Aid Physicians in Diagnosing Asthma, COPD and Asthma-COPD Overlap (ACO). D22 COMORBIDITIES IN PEOPLE WITH COPD. American Thoracic Society; 2020. p. A6285. doi:10.1164/ajrccm-conference.2020.201.1_MeetingAbstracts.A6285
110. Lisspers K, Ställberg B, Larsson K, Janson C, Müller M, Łuczko M, et al. Developing a short-term prediction model for asthma exacerbations from Swedish primary care patients' data using machine learning - Based on the ARCTIC study. *Respir Med.* 2021;185: 106483. doi:10.1016/j.rmed.2021.106483
111. Marin-Gomez FX, Fàbregas-Escurriola M, Seguí FL, Pérez EH, Camps MB, Peña JM, et al. Assessing the likelihood of contracting COVID-19 disease based on a predictive tree model: A retrospective cohort study. *PLoS One.* 2021;16: e0247995. doi:10.1371/journal.pone.0247995
112. Nikolaou V, Massaro S, Garn W, Fakhimi M, Stergioulas L, Price DB. Fast decliner phenotype of chronic obstructive pulmonary disease (COPD): Applying machine learning for predicting lung function loss. *BMJ Open Respir Res.* 2021;8. doi:10.1136/bmjresp-2021-000980
113. Ställberg B, Lisspers K, Larsson K, Janson C, Müller M, Łuczko M, et al. Predicting hospitalization due to copd exacerbations in swedish primary care patients using machine learning – based on the arctic study. *Int J COPD.* 2021;16: 677–688. doi:10.2147/COPD.S293099
114. Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. Prediction of influenza

- vaccination outcome by neural networks and logistic regression. *J Biomed Inform.* 2010;43: 774–781. doi:10.1016/j.jbi.2010.04.011
115. Zafari H, Langlois S, Zulkernine F, Kosowan L, Singer A. AI in predicting COPD in the Canadian population. *BioSystems.* 2022;211: 104585. doi:10.1016/j.biosystems.2021.104585
 116. Hertroijs DFL, Elissen AMJ, Brouwers MCGJ, Schaper NC, Köhler S, Popa MC, et al. A risk score including body mass index, glycated haemoglobin and triglycerides predicts future glycaemic control in people with type 2 diabetes. *Diabetes, Obes Metab.* 2018;20: 681–688. doi:10.1111/dom.13148
 117. Myers KD, Knowles JW, Staszak D, Shapiro MD, Howard W, Yadava M, et al. Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data. *Lancet Digit Heal.* 2019;1: e393–e402. doi:10.1016/S2589-7500(19)30150-5
 118. Weisman A, Tu K, Young J, Kumar M, Austin PC, Jaakkimainen L, et al. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. *BMJ Open Diabetes Res Care.* 2020;8. doi:10.1136/bmjdr-2020-001224
 119. Amit G, Girshovitz I, Marcus K, Zhang Y, Pathak J, Bar V, et al. Estimation of postpartum depression risk from electronic health records using machine learning. *BMC Pregnancy Childbirth.* 2021;21. doi:10.1186/s12884-021-04087-8
 120. Boaz L, Samuel G, Elena T, Nurit H, Brianna W, Rand W, et al. Machine Learning Detection of Cognitive Impairment in Primary Care. *Alzheimers Dis Dement.* 2017;1:

S111. doi:10.36959/734/372

121. Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry*. 2013;74: 7–14. doi:10.1016/j.biopsych.2012.12.007
122. Fernández-Gutiérrez F, Kennedy JI, Cooksey R, Atkinson M, Choy E, Brophy S, et al. Mining primary care electronic health records for automatic disease phenotyping: A transparent machine learning framework. *Diagnostics*. 2021;11. doi:10.3390/diagnostics11101908
123. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. *Semin Arthritis Rheum*. 2019;49: 84–90. doi:10.1016/j.semarthrit.2019.01.002
124. Zhou S-M, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis. *PLoS One*. 2016;11: e0154515. doi:10.1371/journal.pone.0154515
125. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: A binational retrospective study. *J Am Med Informatics Assoc*. 2016;23: 879–890. doi:10.1093/jamia/ocv195
126. Sufriyana H, Wu YW, Su ECY. Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. *EBioMedicine*. 2020;54: 102710.

doi:10.1016/j.ebiom.2020.102710

127. Kostev K, Wu T, Wang Y, Chaudhuri K, Tanislav C. Predicting the risk of stroke in patients with late-onset epilepsy: A machine learning approach. *Epilepsy Behav.* 2021;122: 108211. doi:10.1016/j.yebeh.2021.108211
128. Sekelj S, Sandler B, Johnston E, Pollock KG, Hill NR, Gordon J, et al. Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study. *Eur J Prev Cardiol.* 2021;28: 598–605. doi:10.1177/2047487320942338
129. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med.* 2019/07/16. 2018;1: 39. doi:10.1038/s41746-018-0040-6
130. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda SR, et al. The value of automated diabetic retinopathy screening with the EyeArt system: A study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther.* 2019;21: 635–643. doi:10.1089/dia.2019.0164
131. González-Gonzalo C, Sánchez-Gutiérrez V, Hernández-Martínez P, Contreras I, Lechanteur YT, Domanian A, et al. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. *Acta Ophthalmol.* 2019;98: 368–377. doi:10.1111/aos.14306
132. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of Artificial Intelligence-Based Grading of Diabetic Retinopathy in Primary Care. *JAMA Netw open.* 2018;1: e182665. doi:10.1001/jamanetworkopen.2018.2665

133. Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care*. 2019;42: 651–656. doi:10.2337/dc18-0148
134. 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*. 2017;6: 2453–2460. doi:10.1002/cam4.1183
135. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artif Intell Med*. 2016;69: 53–61. doi:10.1016/j.artmed.2016.03.003
136. Hornbrook MC, Goshen R, Choman E, O’Keeffe-Rosetti M, Kinar Y, Liles EG, et al. Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Dig Dis Sci*. 2017;62: 2719–2727. doi:10.1007/s10620-017-4722-8
137. Kinar Y, Akiva P, Choman E, Kariv R, Shalev V, Levin B, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. *PLoS One*. 2017;12: e0171759. doi:10.1371/journal.pone.0171759
138. Collins GS, Reitsma JB, Altman DG, Moons KGMM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med*. 2015;162: 55–63. doi:10.7326/M14-0697
139. Daines L, McLean S, Buelo A, Lewis S, Sheikh A, Pinnock H. Systematic review of clinical prediction models to support the diagnosis of asthma in primary care. *NPJ Prim care Respir Med*. 2019;29: 19. doi:10.1038/s41533-019-0132-z

140. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019;28: 231–237. doi:10.1136/bmjqs-2018-008370
141. Nickel B, Barratt A, Copp T, Moynihan R, McCaffery K. Words do matter: a systematic review on how different terminology for the same condition influences management preferences. *BMJ Open.* 2017;7: e014129. doi:10.1136/BMJOPEN-2016-014129
142. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science American Medical Informatics Association; 2020 pp. 191–200.*
143. Kaneko H, Umakoshi H, Ogata M, Wada N, Iwahashi N, Fukumoto T, et al. Machine learning based models for prediction of subtype diagnosis of primary aldosteronism using blood test. *Sci Rep.* 2021;11: 9140. doi:10.1038/s41598-021-88712-8
144. Gentil M-L, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: A systematic review of the literature. *BMC Med Inform Decis Mak.* 2017;17. doi:10.1186/s12911-017-0538-x
145. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17: 195. doi:10.1186/s12916-019-1426-2
146. Bakker L, Aarts J, Uyl-de Groot C, Redekop W, Groot CUD, Redekop W. Economic evaluations of big data analytics for clinical decision-making: A scoping review. *J Am Med Informatics Assoc.* 2020;27: 1466–1475. doi:10.1093/jamia/ocaa102

147. Williamson T, Aponte-Hao S, Mele B, Lethebe BC, Leduc C, Thandi M, et al. Developing and validating a primary care EMR-based frailty definition using machine learning. *Int J Popul Data Sci.* 2020;5: 1344. doi:10.23889/IJPDS.V5I1.1344
148. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, 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.* 2021;11: e048008. doi:10.1136/bmjopen-2020-048008

Supporting information

S1 Appendix Search strategy

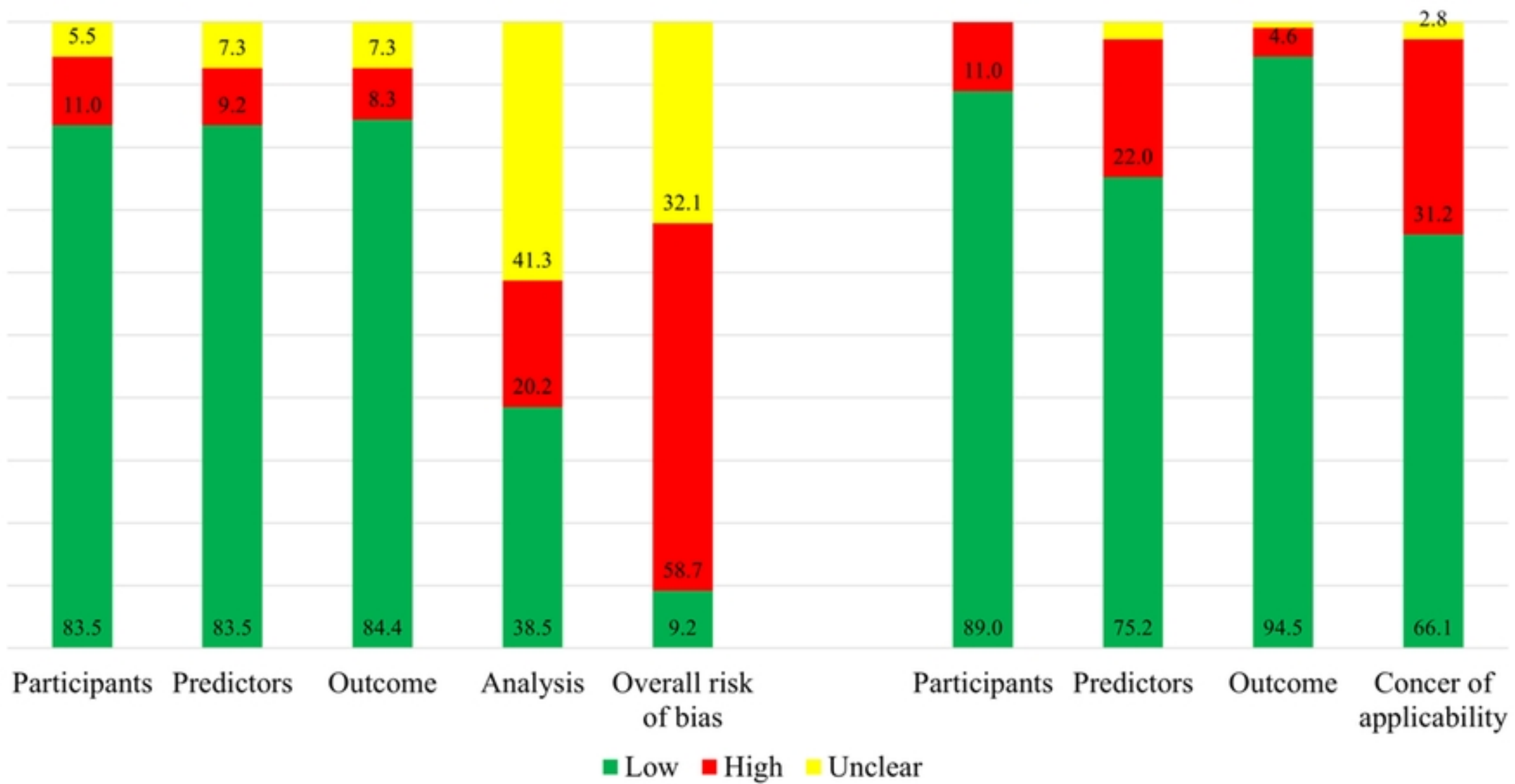
S2 List of excluded studies with reasons (n = 58)

S3 Figure Countries under study with total number of studies per country

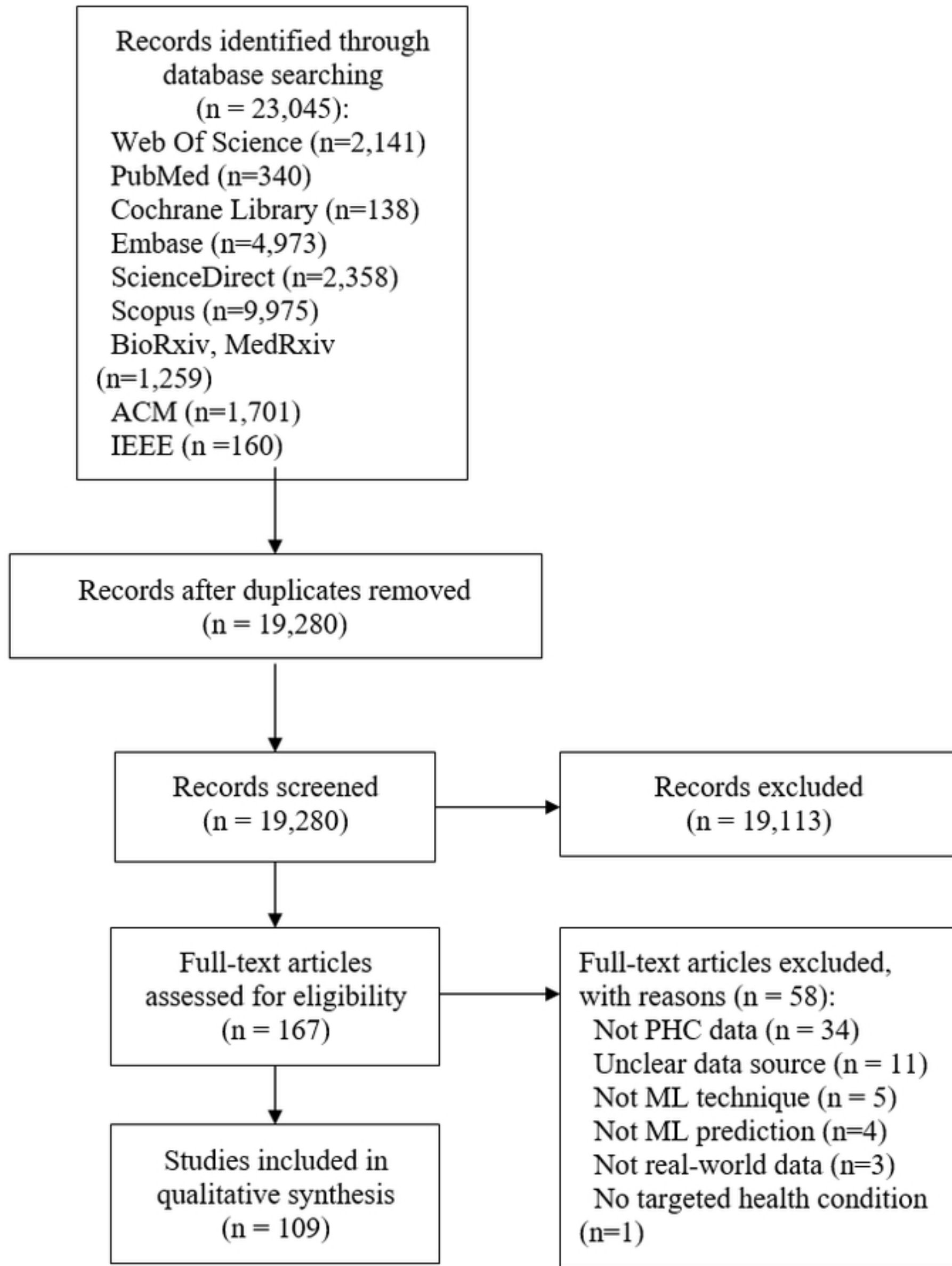
S4 Figure of studies designs

S5 Appendix Panel of the included studies' (n =109) characteristics

S6 Prisma Checklist

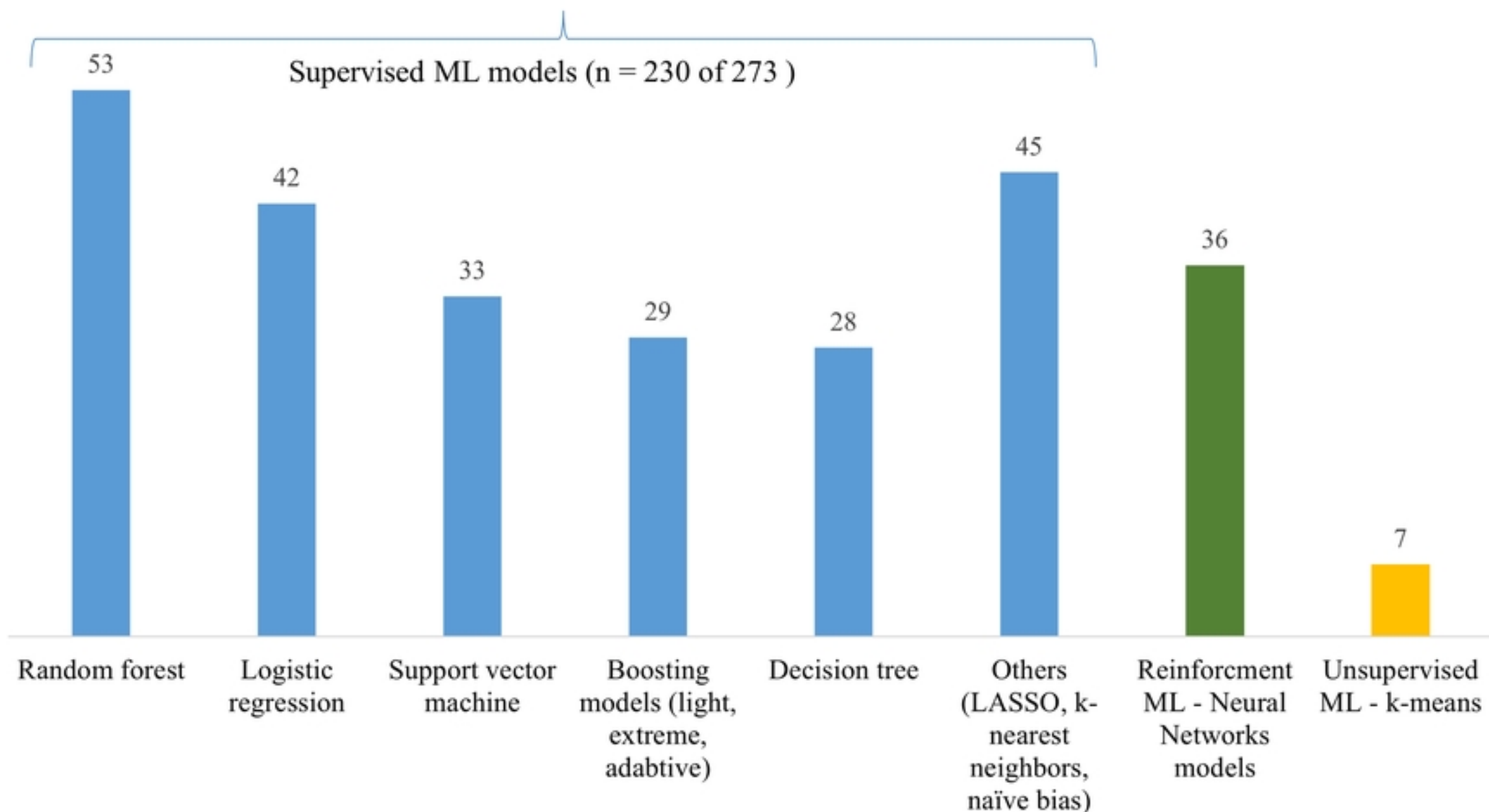


Results of PROBABAST (Risk of bias assessment)



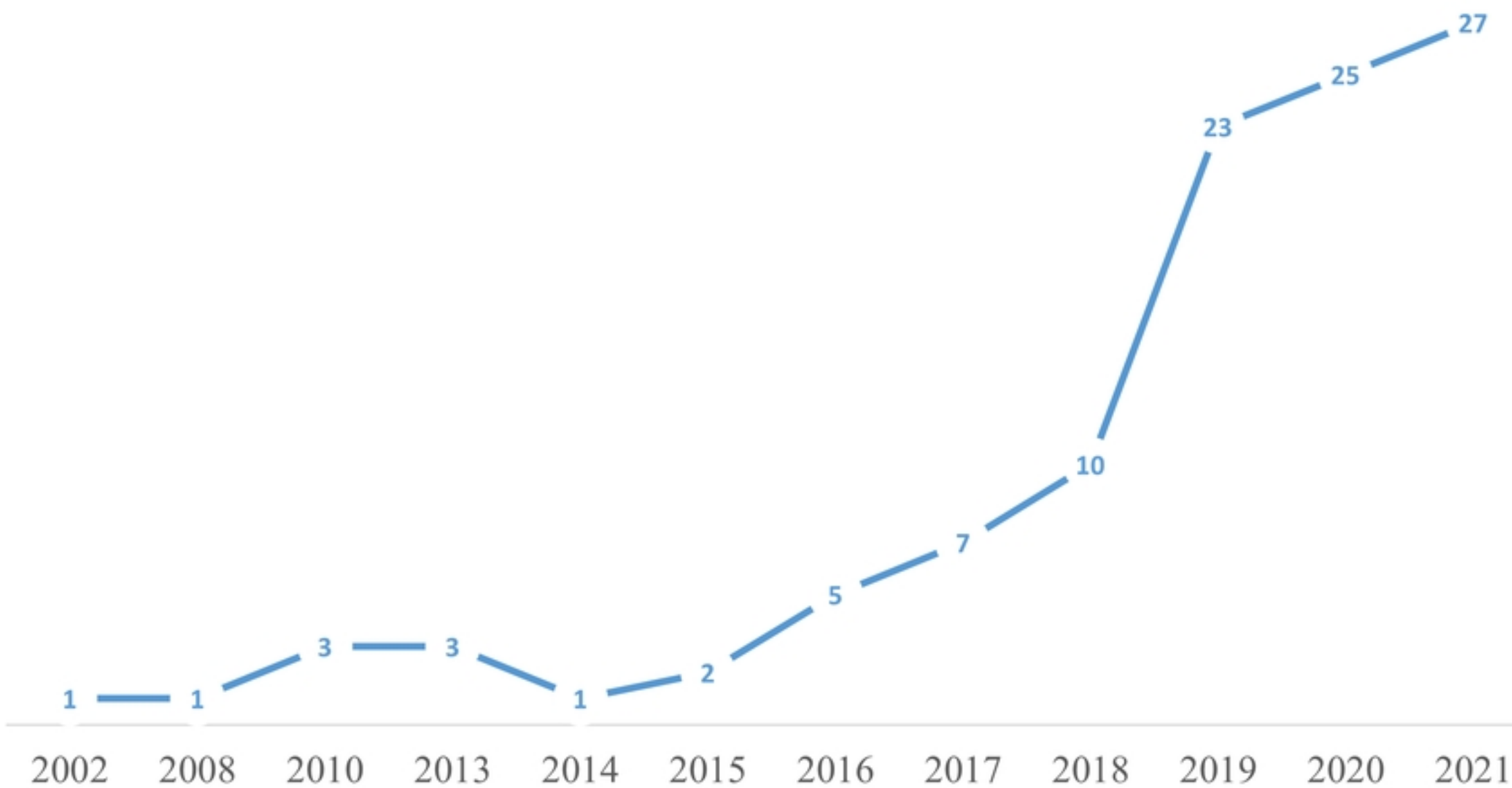
PRISMA flow chart

Number of models developed and/or validated across the included studies
(n = 273 models in 109 studies)



Frequency of ML prediction models techniques reported

NUMBER OF STUDIES PER PUBLICATION YEAR



Number of included studies per publication year