Research protocol

DIGITAL HEALTH

Digital Health
Volume 7: 1–9

DOI: 10.1177/20552076211047390
journals.sagepub.com/home/dhj
SAGE

Use and performance of machine learning models for type 2 diabetes prediction in clinical and community care settings: Protocol for a systematic review and meta-analysis of predictive modeling studies

Kushan De Silva[1], Joanne Enticott[1], Christopher Barton[2], Andrew Forbes[3], Sajal Saha[2] and Rujuta Nikam[2]

Abstract

Objective: Machine learning involves the use of algorithms without explicit instructions. Of late, machine learning models have been widely applied for the prediction of type 2 diabetes. However, no evidence synthesis of the performance of these prediction models of type 2 diabetes is available. We aim to identify machine learning prediction models for type 2 diabetes in clinical and community care settings and determine their predictive performance.

Methods: The systematic review of English language machine learning predictive modeling studies in 12 databases will be conducted. Studies predicting type 2 diabetes in predefined clinical or community settings are eligible. Standard CHARMS and TRIPOD guidelines will guide data extraction. Methodological quality will be assessed using a predefined risk of bias assessment tool. The extent of validation will be categorized by Reilly–Evans levels. Primary outcomes include model performance metrics of discrimination ability, calibration, and classification accuracy. Secondary outcomes include candidate predictors, algorithms used, level of validation, and intended use of models. The random-effects meta-analysis of c-indices will be performed to evaluate discrimination abilities. The c-indices will be pooled per prediction model, per model type, and per algorithm. Publication bias will be assessed through funnel plots and regression tests. Sensitivity analysis will be conducted to estimate the effects of study quality and missing data on primary outcome. The sources of heterogeneity will be assessed through meta-regression. Subgroup analyses will be performed for primary outcomes.

Ethics and dissemination: No ethics approval is required, as no primary or personal data are collected. Findings will be disseminated through scientific sessions and peer-reviewed journals.

PROSPERO registration number: CRD42019130886

Keywords

Type 2 diabetes, machine learning, prediction models, meta-analysis, protocol

Submission date: 2 August 2019; Acceptance date: 1 September 2021

[1]Monash Centre for Health Research and Implementation, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Australia
[2]Department of General Practice, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Australia
[3]Biostatistics Unit, Division of Research Methodology, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Australia

Corresponding author:
Kushan De Silva, Monash Centre for Health Research and Implementation, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Locked Bag 29, Level 1, 43-51 Kanooka Grove, Clayton VIC 3168, Australia. **Email:** kushan.ranakombu@monash.edu

Introduction

Machine learning (ML) is a branch of artificial intelligence (AI) involved in the development of algorithms and techniques, which enables computers to learn and gain intelligence based on past experience. It is a computational process in which the system is able to identify and understand input data and consequently apply the acquired information to make decisions and predictions on various phenomena.[1]

One of the main health-related applications of ML is in prediction as well as the diagnosis or prognosis of various biomedical conditions. Diagnostic or prognostic studies have the potential to inform health care administration, clinical decision support, patient monitoring, and interventions.[2] For example, a study that classified diabetes by developing a hybrid intelligence system using ML revealed that it could assist clinicians as a decision support system.[3] Other studies have used ML for the detection of biomedical phenomena, such as the mortality of critically ill patients,[4] image-based skin cancer diagnosis,[5] prognosis of hemorrhage in trauma victims,[6] and tumor diagnosis.[7,8]

Type 2 diabetes (T2DM) has a complex, multi-factorial etiology encompassing interactions of genetic, environmental, and behavioral factors.[9] However, the genetic and non-genetic determinants of T2DM are not fully elucidated.[10] ML studies can be used to examine large health data repositories (coined as "big data") and lead to new knowledge discoveries, such as the unknown determinants of different clinical phenotypes of diabetes, their causal pathways, patterns, and interactions.[11] A systematic review of ML applications revealed that it was predominantly applied for prediction and diagnosis, genetic and environmental determinants, and health care and management of T2DM.[12]

The systematic reviews of prediction models for diabetes to date have exclusively assessed traditional modeling studies.[13–16] To the best of our knowledge, no systematic review of the scope of use and effectiveness of ML-based prediction models for T2DM has been conducted. Nevertheless, ML algorithms have been applied for predicting diabetes,[17–21] and it has been reported that ML methods to address various domains of diabetes research are on the rise.[12]

While prediction models built exclusively on genetic data in laboratory settings, in their nascent stages, might provide little clinical decision support, those applied in clinical and community care settings, incorporating routine clinical information, would be particularly useful for implementing clinical prediction rules and decision-making. Therefore, it is imperative that a synthesis of current scientific evidence on the use and performance of ML-based prediction models of T2DM applied in clinical and community settings is performed. Diabetes is a major public health concern, as it is a global epidemic that entails an enormous disease burden with multiple associated health complications.[22] Timely and effective diagnosis and management of diabetes assisted by ML have the potential to prevent the onset of many of these diabetes-associated complications. In fact, diabetes-associated complications have been classified as potentially preventable hospitalizations in Australia.[23] In this context, this protocol describes a systematic review to identify available ML-based prediction models for T2DM in clinical and community care settings, investigate their predictive performance using meta-analysis, and formulate an evidence synthesis of the extent of use and comparative performance of ML approach for T2DM prediction.

Methods

Preferred reporting items for systematic reviews and meta-analyses protocols (PRISMA-P) checklist (Supplementary Table 1) guided the development of this protocol.[24] Studies published from 2009 will be considered for the review.

Review question

What is the extent of use and the comparative performance of ML models for T2DM prediction in clinical and community settings?

Eligibility criteria

Selection criteria of the studies are outlined using the PICOTS framework in Table 1.

Study design and data sources. Only predictive modeling studies that have explicitly used ML to predict current (diagnostic models) or future (prognostic models) occurrence of T2DM are eligible. Thus, any predictive modeling studies with no explicit ML approach will be excluded. Data sources of these predictive modeling studies could have emanated from any observational or interventional designs. Observational designs may include cross-sectional studies, longitudinal surveys, as well as secondary data sources such as registry-based data. Interventional designs may encompass various randomized controlled trial designs and quasi-experimental studies.

Participants (P). We will include participants in any predictive modeling study reporting ML prediction models for diagnosis or prognosis of T2DM in clinical or community care settings. These participants can be those with diagnosed or undiagnosed T2DM or individuals at high risk being investigated for T2DM. The cohorts used for developing these models generally consist of people with diagnosed or undiagnosed T2DM and individuals perceived as at high risk due to the presence of established risk factors, such as positive family history, obesity, adiposity, race or ethnicity, blood lipid levels, age, and history of prediabetes.

Table 1. Selection criteria of predictive modeling studies in PICOTS format.

	Participants (P)	Intervention (I)	Comparison (C)	Outcomes (O)	Timeframe (T)	Setting (S)	Other limits
Inclusion criteria	Individuals with T2DM Individuals without T2DM being investigated for the condition	ML predictive modeling: supervised, unsupervised, semi-supervised ML, or combinations thereof	Not applicable	Primary: metrics of discrimination ability, calibration, and classification accuracy in T2DM prediction Secondary: candidate predictors, applied algorithms, level of validation, intended use of models	Since 1 January 2009 to date	Clinical care settings, for example, hospitals, long-term-, ambulatory-, acute-care facilities Community care settings, e.g. general practices, community health centers, allied health practices	Language = English
Exclusion criteria	Patients with other clinical phenotypes of diabetes, type 1 diabetes, gestational diabetes Pre-diabetic individuals Diabetic complications	Predictive modeling without an explicit ML approach				Laboratory settings: using only genetic, genomic, or genotype data	

ML: machine learning; T2DM: type 2 diabetes.

Therefore, the proposed ML classifiers will be applicable to discriminate individuals with T2DM from those at high risk of T2DM. No eligibility restrictions will be made on age, gender, ethnicity, and geographic location of participants. Studies focusing only on those with prediabetes or diabetic complications will be excluded.

Interventions (I). Eligible interventions are supervised, unsupervised, semi-supervised ML or any combinations thereof to build prognostic or diagnostic models predicting future or current T2DM, respectively. These models can be at development or validation with or without updating phases. Thus, all three types of studies specified in the CHARMS checklist,[25] that is, prediction model development studies without external validation, prediction model development with external validation on independent data, and external model validation studies with or without model updating, will be included.

Outcome measures (O)

Primary outcomes. The effect of prediction models with respect to the diagnosis or prognosis of T2DM will be measured by reported model performance metrics:

1. Discrimination ability, for example, c-statistic
2. Calibration, for example, Hosmer–Lemeshow statistic
3. Classification measures, for example, sensitivity, specificity, and negative and positive predictive values

Secondary outcomes

1. Candidate predictors
2. Algorithms applied
3. Level of validation: development dataset only (random split of data, resampling methods, e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, and different investigators)
4. Intended use , for example, at the moment of diagnosis of T2DM, in asymptomatic adults to detect undiagnosed T2DM

Studies will be excluded, if the predicted outcome is not binary, categorical, or time-to-event but on a continuous scale with no clear discrimination between the presence and absence of T2DM. Thus, studies describing prediction models that classify patients into arbitrary risk categories (e.g. low risk vs high risk), without providing the personalized estimates of outcome probability are excluded. Moreover, any studies with no reported measures of predictive performance, that is, calibration, discrimination, and classification measures, are excluded.

Time (T). Considering the recent advent of ML as well as the applicability of more recent studies, there will be a restriction on study publishing date. Studies published from 1 January 2009 to the date of search will be the time limit and any done prior to 1 January 2009 will be ineligible.

Settings (S). Studies in both clinical and community settings will be included. Thus, clinical settings such as hospitals, long-term-, ambulatory-, or acute-care facilities as well as community settings, such as general practices, primary care centers, community health centers, allied health practices, and community-based risk factor surveillance surveys, will be eligible. Given the lack of immediate clinical applicability of studies conducted in laboratory settings using only the genomic, genetic, or genotype data, they will be excluded. Since risk factor prevalence as well as participant profiles and characteristics might differ between clinical and community settings, we will report our findings separately for the two settings.

Other limits: language

Only English language articles will be included.

Search strategy

Electronic databases. A uniform search strategy will be developed and applied to following databases: (a) SCOPUS, (b) OVID MEDLINE, (c) MEDLINE In-Process & Other Non-Indexed Citations, (d) EMBASE, (e) Cochrane Library, (f) PsycINFO, (g) CINAHL, (h) Web of Science, (i) Springer, (j) Elsevier, (k) ACM Digital Library, and (l) IEEE Xplore Digital Library. We will also manually search the reference lists of relevant articles retrieved.

Search terms. The search strategy will capture studies that include following PICOTS terms: populations (patient with T2DM, individuals without T2D), intervention (ML), outcomes (predictive performance in terms of T2DM diagnosis or prognosis), time span (1 January 2009 to date), and settings (clinical or community care). Matched terms under each group against possible medical subject headings (MeSH) or keywords as follows will be used in a systematic search through 12 databases.

1. Population terms: diabetes/OR Diabet* OR (hyperglycemia or hyperglycemic) OR (T2DM or diabetes mellitus or late onset diabetes)
2. Intervention terms: ML or deep learning or neural network or support vector machine or classification tree or regression tree or decision tree or random forest or gradient boosting or k-nearest neighbors or supervised learning or unsupervised learning or clustering or PCA or principal component analysis or multifactor dimension reduction or classifier models or data mining or bagging or boosting or naïve Bayes or logistic regression or logistic models or algorithms or

computational modeling or linear modeling or non-linear modeling or ensemble or feature selection
3. Outcome terms: prediction model/OR (predictive model or diagnostic model or prognostic model) OR ROC curve/or (discriminant or c-statistic or area under the curve (AUC) or the area under the receiver operating characteristic curve (AUROC) OR calibration OR validation/or (internal validation or external validation) OR indices OR multivariable OR classification OR models
4. Time limits: 1 January 2009 to date
5. Setting terms: clinical or ambulatory or inpatient or acute or community or primary care or preventive or long-term care

The search strategy detailed above will be developed using the OVID MEDLINE platform with a combination of keywords, wildcards, and truncations and translated to other databases as appropriate with necessary modifications. To increase the relevance of the findings of this review for clinical practice, search will be limited to papers published from 1 January 2009 to date. Resource constraints require that we limit the search to English language papers. We will also search bibliographies of relevant studies for the identification of additional studies.

Hand searching. We will manually search key journals (e.g. Journal of Diabetes Science and Technology, AI in Medicine, JAMA, PLoS One, Expert Systems with Applications, and BMC Medical Research Methodology). If required, direct contact with authors will be undertaken to obtain other relevant articles. Cited original articles in relevant systematic reviews will also be retrieved and analyzed. We will update our literature search using the auto alert system in individual databases before the publication of this review to avoid the exclusion of any recent articles. The search will be re-run before final analysis.

Study selection. All electronically and manually searched records will be merged to remove duplicate citations. Two reviewers (KDS and RN) will independently screen titles and abstracts to identify eligible articles using inclusion and exclusion criteria. After this initial screening, full text articles will be retrieved for all records that have passed the screen, or if exclusion cannot be determined during the screen. Full text articles will then be examined by two independent reviewers (KDS, RN, or SS) against the eligibility criteria. Any discrepancies arising between the reviewers will be resolved by another reviewer (JE or CB). If there is an information gap in a paper and/or a need for further clarification, the author will be contacted to clarify the issue by email. A PRISMA flow diagram will be used to maintain transparency in the article selection process and record remaining studies in each stage of selection with a valid explanation of reasons for exclusion.

Data extraction and management

Data will be extracted from full text papers using a specially developed data extraction form based on the following:

1. Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies (CHARMS) statement[25]
2. **TRIPOD** statement[26]; template for **tr**ansparent reporting of a multivariable prediction model for **i**ndividual **p**rognosis **o**r **d**iagnosis, and
3. Guidelines for developing and reporting ML predictive models in biomedical research by Luo et al.[27]

Data will be extracted by two reviewers. First, a subset of two or three articles will be used as a training set. The training set will be coded by two authors, and others will then review this initial coding. Discrepancies in coding will be resolved during a consensus meeting. The coding scheme will be revised where necessary to ensure the usability and completeness of extraction tool, depending on the training set findings. The revised coding scheme at this point will be checked by a senior researcher of the team (JE, CB, or AF) and modified as required, prior to continuing, to ensure its quality. Thereafter, two reviewers will independently review the remaining articles, and data extraction will be accomplished independently. Any queries about this extraction will be discussed between the two parties and further advice sought from a senior member (JE, CB, or AF) whenever necessary.

We will extract data on: (a) title and abstract, (b) sources of data and study design, (c) type and aims of prediction model (prognostic vs diagnostic), (d) aim(s) of the study, (e) extent of modeling (model development only, model development and external validation, model development, external validation, and updating), (f) target population, (g) outcomes to be predicted, (h) time span of prediction (prognostic models only), (i) intended moment of using the model, (j) participants, (k) candidate predictors, (l) sample size, (m) missing data, (n) ML methods employed, (o) model development, (p) model performance, (q) model evaluation, (r) results, (s) interpretation and discussion (limitations, implications), and (t) other information (supplementary information, funding).

The results will be carefully extracted to make them meta-analyzable. If data presentation is problematic, unclear, missing, or unextractable, authors will be contacted for clarification by email with a response time limit of two weeks. If the author is unresponsive, then they will be classified as uncontactable. We will sub-group studies, as appropriate, based on criteria such as prediction model type, geographic location, and study population.

Assessment of risk of bias and methodological quality

As there are no established checklists specifically designed for the assessment of bias in predictive modeling studies, we will use the classification system created by Van den Boorn et al.,[28] which had been based on TRIPOD statement (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) by Collins et al.[26] Two reviewers (KDS, RN, or SS) will pilot this initial bias assessment tool on a small sample of studies ($n = 5$). Based on the findings of the pilot, a quality assessment of the tool will then be performed by a third author (JE, CB, or AF) and modified as required. This will produce our final quality and bias assessment tool.

Using our final quality and bias assessment tool, the classification of the potential for bias will be done in two stages; each researcher (KDS and CB or JE) will first make notes of potential sources of bias per category separately, and together, they (KDS and CB or JE) will then categorize identified sources of bias. Bias will be determined in areas such as (a) population-related (such as selection bias), (b) predictor-related (such as ill-defined predictors), (c) outcome-related (such as an unclear outcome), (d) sample size-related, (e) missing data-related (such as only complete case analysis), and (f) analysis-related (such as underreporting of statistics). Each study will be categorized as high-, low-, or unclear risk of bias under each of the individual criteria. Based on these categorizations, each study will be given an overall assessment of the low, moderate, or high methodological quality. We will also evaluate reporting criteria (e.g. outcome definition, sample size, and sources of funding) for each of the included studies. The findings of each study's risk of bias assessment will be recorded in a summary table. Low-quality studies will be excluded from meta-analysis.

Analysis of risk of bias

Bias analysis will be as per Van den Boorn et al.[28] and will test three hypotheses.

Hypothesis 1. The higher the impact factor of a journal in which the study was published, the more stringent the internal screening and peer review procedures would be and, hence, the lower the risk of bias.
Hypothesis 2. The higher the impact factor of the journal a prediction model was published in, the better its performance in terms of c-index would be.
Hypothesis 3. The reported c-indices would be larger during model development than during validation due to overfitting. This will be assessed using a one-tailed Wilcoxon signed-rank test.

Hypotheses 1 and 2 will be assessed through the Spearman rank correlation between the journal impact factor[29] (in the year of publication or the closest to publication year available) and the reported c-index as well as between journal impact factor and the potential sources of bias (assessed using the tool for the classification of potential sources of bias developed), respectively. Owing to inequalities of T2DM in different geographical populations,[22] we will examine whether models were constructed and validated with patient cohorts from different continents using the Fisher's exact test.

Analyses will be performed in R statistical software (R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org).[30]

Data synthesis and analysis

Model performance will be described using measures for discriminative ability, calibration, and classification accuracy. Key findings on study design, sources of data, prediction model types, sample size, participant characteristics, aim of the model, methods, presentation of the final prediction model, and outcome measures will be summarized in a tabular format.

Discriminative ability

This is defined as a model's ability to differentiate between those who experience an event and those who do not[31] and is typically quantified by the concordance index (c-index) which has values ranging from 0.5 (no discrimination at all) to 1 (perfect discrimination). It is the generalization of AUROC, a well-known measure of discrimination, and hence diagnostic or prognostic accuracy that would determine its clinical usefulness. The c-indices can be interpreted by the following rule of thumb:

- 0.5 ± 0.6 = no discrimination
- 0.6 ± 0.7 = poor
- 0.7 ± 0.8 = fair
- 0.8 ± 0.9 = good and
- 0.9 ± 1 = excellent discrimination.

These will be reported as described in each study.

Model calibration

Model calibration, in contrast, conveys the goodness of fit, that is, the agreement between observed and average predicted outcomes.[31] Calibration can be displayed visually in a calibration plot and gauged using statistical tests for goodness of fit. These will be reported as described in each study.

Level of validation

The levels of evidence of the discriminatory accuracy of the prediction model as described by Reilly and Evans[32] indicate how extensively a prediction model has been validated and to what extent a model is ready for clinical use.

- Level 1: model development
- Level 2: narrow validation
- Level 3: broader validation
- Level 4: narrow impact analysis
- Level 5: broad impact analysis

Each identified study will be categorized according to Reilly–Evans levels.

Meta-analysis of c-indices

A random-effects meta-analysis of c-indices will be performed to evaluate the discriminative abilities of prediction models using restricted maximum likelihood estimation. The c-indices will be pooled per prediction model, per model type, that is, prognostic versus diagnostic, and per algorithm. When c-statistic has not been reported, it will be estimated from the standard deviation of the linear predictor. When the standard error of the c-statistic has not been reported, it will be either estimated from the confidence interval or approximated from a combination of the reported c-statistic, the total sample size, and the total number of events[33] by adopting a modification of the method proposed by Hanley and McNeil.[34] Logistic transformation as described in Kottas et al.[35] will be applied to all c-index estimates during calculations and then transformed back to ensure that all estimates are bounded by 0 and 1 after pooling. Forest plots specifying various effect sizes, confidence intervals, and summary estimates will be generated. Analyses will be performed using "metamisc,"[36] "metaphor,"[37] and other relevant packages in R statistical software.[30]

Handling missing data

If any missing data exist within studies, the respective authors will be contacted to avoid the inappropriate description of study results and minimize the risk of bias in meta-analysis. Advanced imputation techniques will be used where appropriate.

Assessment of publication bias

Funnel plots will be drawn to visually assess publication bias, and regression tests will be conducted for detecting funnel plot asymmetry caused by the presence of small study effects. Analyses will be performed using the relevant packages in R statistical software (R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org).[30]

Sensitivity analysis

Sensitivity analyses will be conducted to estimate the effects of study quality and missing data on the pooled primary outcome. Two analyses (one including all eligible studies and the other including only studies of high quality) will be performed to determine the effect of study quality. In case of unobtainable data, we will conduct complete case analysis and then perform the sensitivity analysis of the primary outcome (c-indices) to assess the potential impact of missing data on meta-analysis.

Meta-regression and subgroup analysis

To gain insights into the potential sources of between-study heterogeneity in predictive performance, a random-effects meta-regression as recommended by Debray et al.[38] will be performed.

Should enough data be available, we will conduct subgroup analyses for primary outcome. Important subgroup analyses may be performed by: (a) study sub-populations, (b) country settings (high, middle, vs low income), (c) study settings (clinical or community), (d) risk of bias (high, moderate, vs low risk of bias), (e) study designs (observational or interventional), (f) sources of data (randomized trials, case–control, cohort, survey, or registry data), (g) prediction model type (prognostic or diagnostic), and (h) type of algorithm applied (linear, non-linear, and ensemble).

Ethics and dissemination

No formal ethical approval is required, as no primary, personal, and confidential data are being collected in this study. Quality (certainty) of evidence and strength of recommendations will be reported as per Grading of Recommendations Assessment, Development, and Evaluation (GRADE) and PRISMA-P criteria.[39] We will present our findings in Australia and at international conferences in addition to publishing in peer-reviewed journals.

Discussion

This is the first systematic review assessing the use and predictive performance of ML models for T2DM prediction and capturing a wide scope of prediction models. For example, models at varying levels of validation, from those at development stage to those that have been fully externally validated and updated, developed in and aimed at using in different clinical and community settings as well as both diagnostic and prognostic tools are considered. The findings may be reproducible to a broader context due

to the inclusion of a range of studies emanating from both clinical and non-clinical settings. This review will cover a large number of databases as well. The use of only English language articles is a limitation of the review. Poor quality studies and between-study heterogeneity will be of concern, as they may impact on the validity and interpretability findings. Insights into the sources of heterogeneity, however, will be provided by meta-regression and sub-group analyses.

It is anticipated that the findings of this review will be relevant to many stakeholders. First, the review will present a comprehensive overview of features of ML models for T2DM prediction and will highlight any potential gaps in the current literature on this topic. Second, it will produce high-level evidence from peer-reviewed literature on the predictive performance, feasibility, and acceptability of ML prediction models of T2DM. Third, the review could provide information regarding valuable and robust ML models for T2DM prediction, which may guide clinicians on their evidence-based use as a diagnostic tool for the detection of patients with T2DM. Models applicable for community settings may be of use as screening tools. Fourth, the review may be useful for funders to better understand the applications of ML for T2DM prediction, which could assist priority setting in funding allocations. Finally, the findings may support clinicians, researchers, and health policy-makers to design future ML studies to improve T2DM prediction.

Registration and publishing

This protocol is registered on the International Prospective Register of Systematic Reviews (PROSPERO) with registration number CRD42019130886 (https://www.crd.york.ac.uk/PROSPERO/). Important protocol amendments will be documented on PROSPERO. A PRISMA checklist will be used to report the review. The findings of the review will be published in peer-reviewed journals.

Acknowledgments: Authors sincerely thank Anne Young, subject librarian of Faculty of Medicine, Nursing, and Health Sciences, Monash University for her contribution to the development of search strategy.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Kushan De Silva as a PhD student and supported by a scholarship jointly funded by the Australian Government under Research Training Program (RTP) and Monash University via Monash International Tuition Scholarship (MITS). Funders/sponsors had no role in the design of the study protocol, preparation, review, or approval of the manuscript, as well as the decision to submit the manuscript for publication.

Guarantor: KDS

Author contributions: KDS, JE, CB, and SS designed the review concept. KDS, JE, and AF developed study design and literature search strategies. Screening of literature was conducted by KDS, RN, and AF. Study quality risk assessment tools, data extraction tools, data synthesis and meta-analysis, and statistical tests were developed by KDS, JE, and CB. KDS wrote this manuscript and drafted the protocol according to PRISMA-P. The revision of the manuscript was undertaken by all authors.

Ethical Approval: Not applicable, because this article does not contain any studies with human or animal subjects.

Informed Consent: Not applicable, because this article does not contain any studies with human or animal subjects.

Trial registration: Not applicable, because this article does not contain any clinical trials.

ORCID iDs: Kushan De Silva https://orcid.org/0000-0003-0301-0805
Rujuta Nikam https://orcid.org/0000-0002-5584-9306

Supplementary material: Supplemental material for this article is available online.

References

1. Lantz B. *Machine learning with R. 3rd edition.* Birmingham, UK: Packt Publishing Ltd, 2019, pp.1–26.
2. Reddy S, Fox J and Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019; 112: 22–28.
3. Nilashi M, Ibrahim O, Dalvi M, et al. Accuracy improvement for diabetes disease classification: a case on a public medical dataset. *Fuzzy Inform Eng* 2017; 9: 345–357.
4. Yun K, Oh J, Hong TH, et al. Prediction of mortality in surgical intensive care unit patients using machine learning algorithms. *Front Med* 2021; 8: 621861.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–118.
6. Liu NT, Holcomb JB, Wade CE, et al. Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients. *Med Biol Eng Comput* 2014; 52: 193–203.
7. Curioni-Fontecedro A. A new era of oncology through artificial intelligence. *ESMO Open* 2017; 2: e000198.
8. Chen Y, Argentinis JE and Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clin Ther* 2016; 38: 688–701.
9. Hansen T. Type 2 diabetes mellitus--a multifactorial disease. *Ann Univ Mariae Curie Sklodowska Med* 2002; 57: 544–549.

10. Wareham NJ, Franks PW and Harding AH. Establishing the role of gene-environment interactions in the etiology of type 2 diabetes. *Endocrinol Metab Clin North Am* 2002; 31: 553–566.
11. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff* 2014; 33: 1163–1170.
12. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. *Comput Struct Biotechnol J* 2017; 15: 104–116.
13. Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011; 9: 03.
14. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *Br Med J* 2012; 345: e5900.
15. Noble D, Mathur R, Dent T, et al. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011; 343: d7163.
16. Lamain – de Ruiter M, Kwee A, Naaktgeboren CA, et al. Prediction models for the risk of gestational diabetes: a systematic review. *Diagn Progn Res* 2017; 1: 3.
17. Polat K and Güneş S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digit Signal Process* 2007; 17: 702–710.
18. Mani S, Chen Y, Elasy T, et al. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc* 2012; 2012: 606–615.
19. Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 2019; 19: 211.
20. Polat K, Güneş S and Arslan A. A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. *Expert Syst Appl* 2008; 34: 482–487.
21. Wu H, Yang S, Huang Z, et al. Type 2 diabetes mellitus prediction model based on data mining. *Inform Med Unlocked* 2018; 10: 100–107.
22. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet.* 2016; 387: 1513–1530.
23. Australian Institute of Health and Welfare. National healthcare agreement: PI 18-selected potentially preventable hospitalisations, https://meteor.aihw.gov.au/content/index.phtml/itemId/559032 (2015, accessed 23 August 2021).
24. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; 4: 1.
25. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist *PLoS Med* 2014; 11: e1001744.
26. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015; 13: 1.
27. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; 18: e323.
28. Van den Boorn HG, Engelhardt EG, van Kleef J, et al. Prediction models for patients with esophageal or gastric cancer: a systematic review and meta-analysis. *PLoS One* 2018; 13: e0192310.
29. Journal Citation Reports®. Clarivate Analytics 2020, https://jcr.clarivate.com (2020, accessed 23 August 2021).
30. R Core Team. R: a language and environment for statistical computing, https://www.r-project.org/ (2021, accessed 23 August 2021).
31. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361–387.
32. Reilly BM and Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; 144: 201–209.
33. Newcombe RG. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006; 25: 559–573.
34. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.
35. Kottas M, Kuss O and Zapf A. A modified wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. *BMC Med Res Methodol* 2014; 14: 26.
36. Debray T and De Jong V. Metamisc: diagnostic and prognostic meta-analysis, R package version 0.2.4, http://r-forge.r-project.org/projects/metamisc/ (2021, accessed 23 August 2021).
37. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; 36: 1–48.
38. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019; 28: 2768–2786.
39. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008; 336: 924–926.