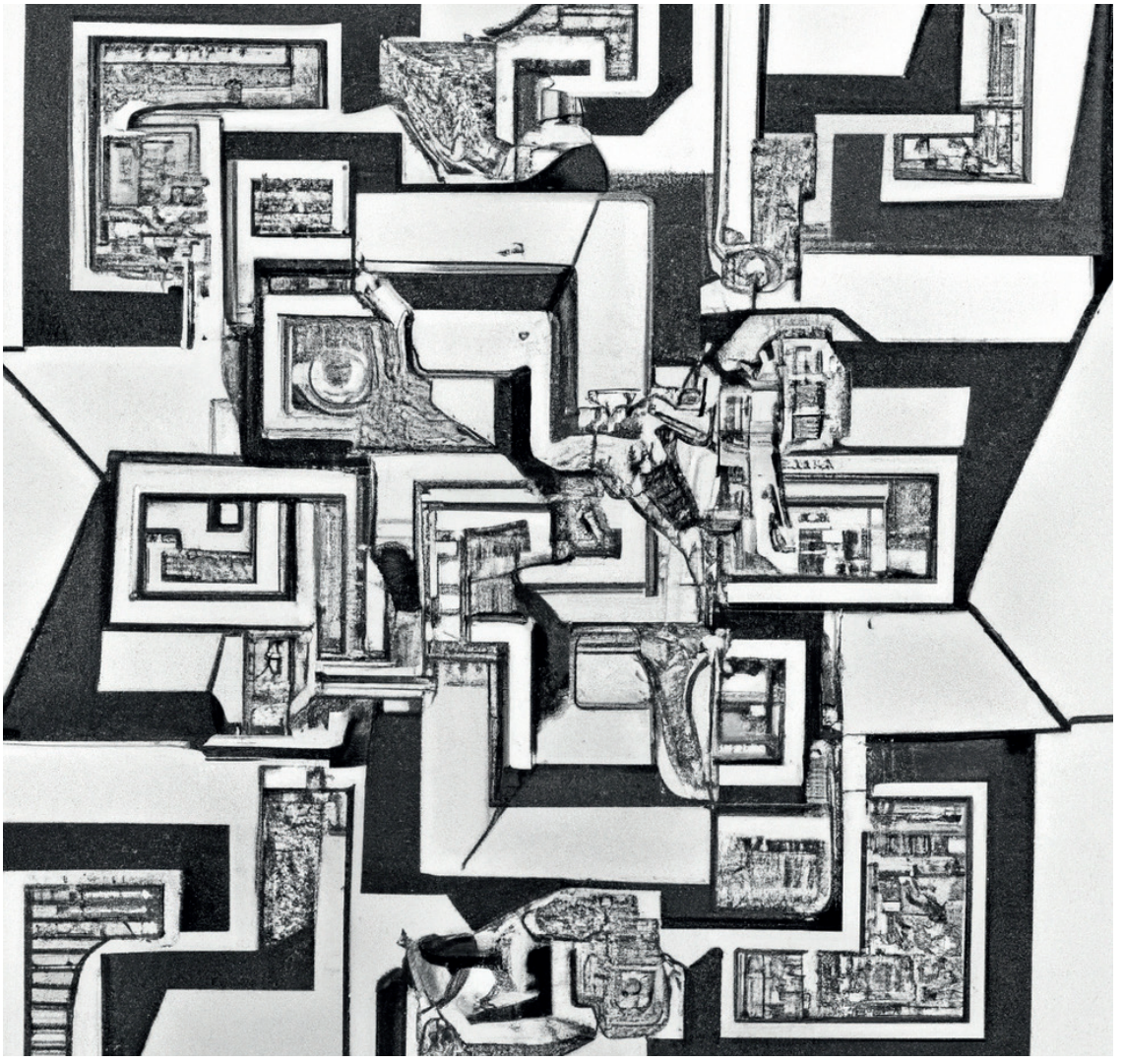
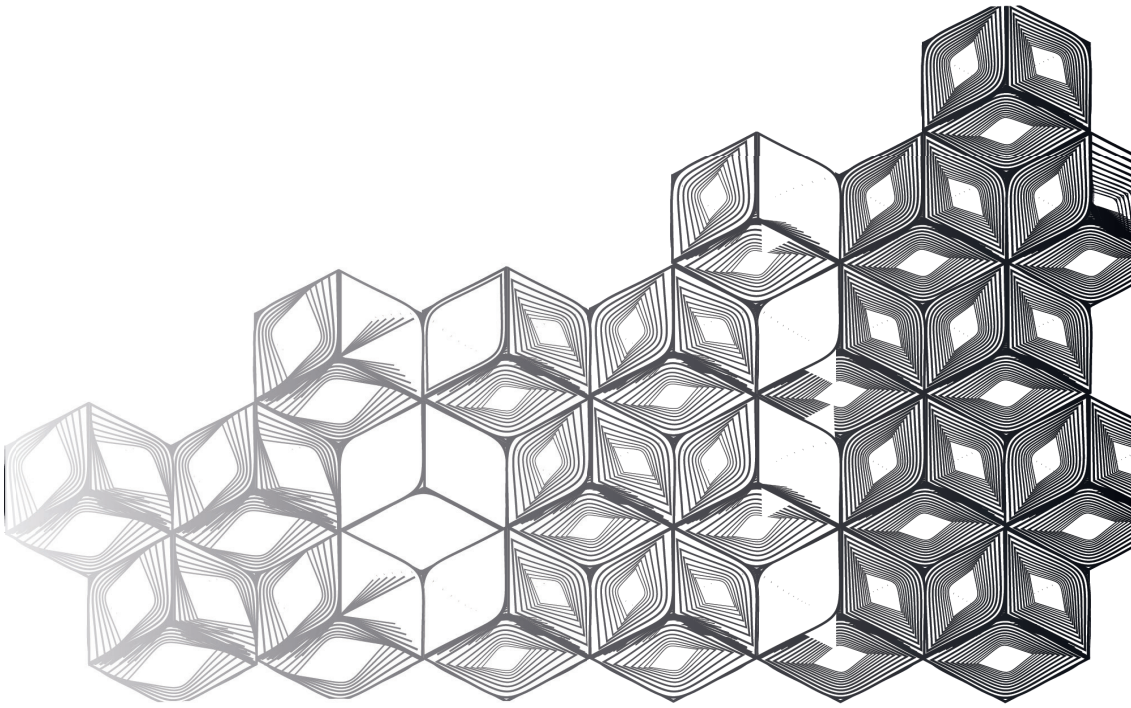
QUALITY OF MACHINE LEARNING PREDICTION MODELS IN HEALTHCARE



CONSTANZA LOURDES ANDAUR NAVARRO



En recuerdo a mi padre,
César Andaur Bastías (1951-2020).
Nos vemos al final del camino...

QUALITY OF MACHINE LEARNING PREDICTION MODELS IN HEALTHCARE

Constanza Lourdes Andaur Navarro

Quality of Machine Learning Prediction Models in Healthcare

ISBN 978-94-6483-091-0
Author Constanza Lourdes Andaur Navarro
Cover Design DALL•E
Lay-out Constanza Lourdes Andaur Navarro, Utrecht, The Netherlands
Printed by Ridderprint | www.ridderprint.nl
Financial support by the Julius Center for Health Sciences and Primary care for the publication of this thesis is gratefully acknowledged.

Quality of machine learning prediction models in healthcare

Kwaliteit van voorspellingsmodellen op basis van machine learning in de gezondheidszorg
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 16 mei 2023 des ochtends te 10.15 uur

door

Constanza Lourdes Andaur Navarro

geboren op 25 oktober 1988
te Providencia, Chili

Promotoren:

Prof. dr. K.G.M. Moons
Prof. dr. L. Hooft

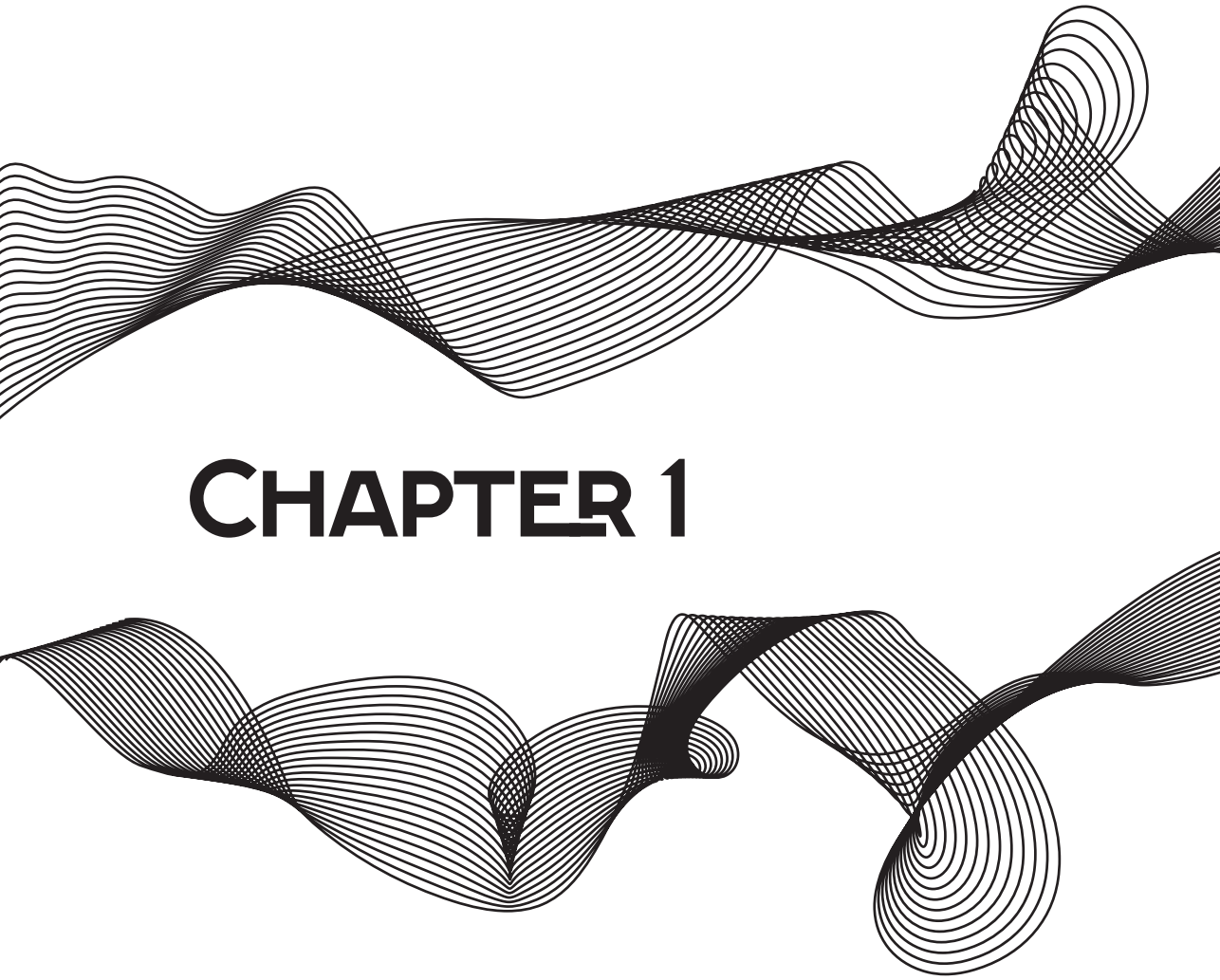
Copromotor:

Dr. J.A.A.G. Damen

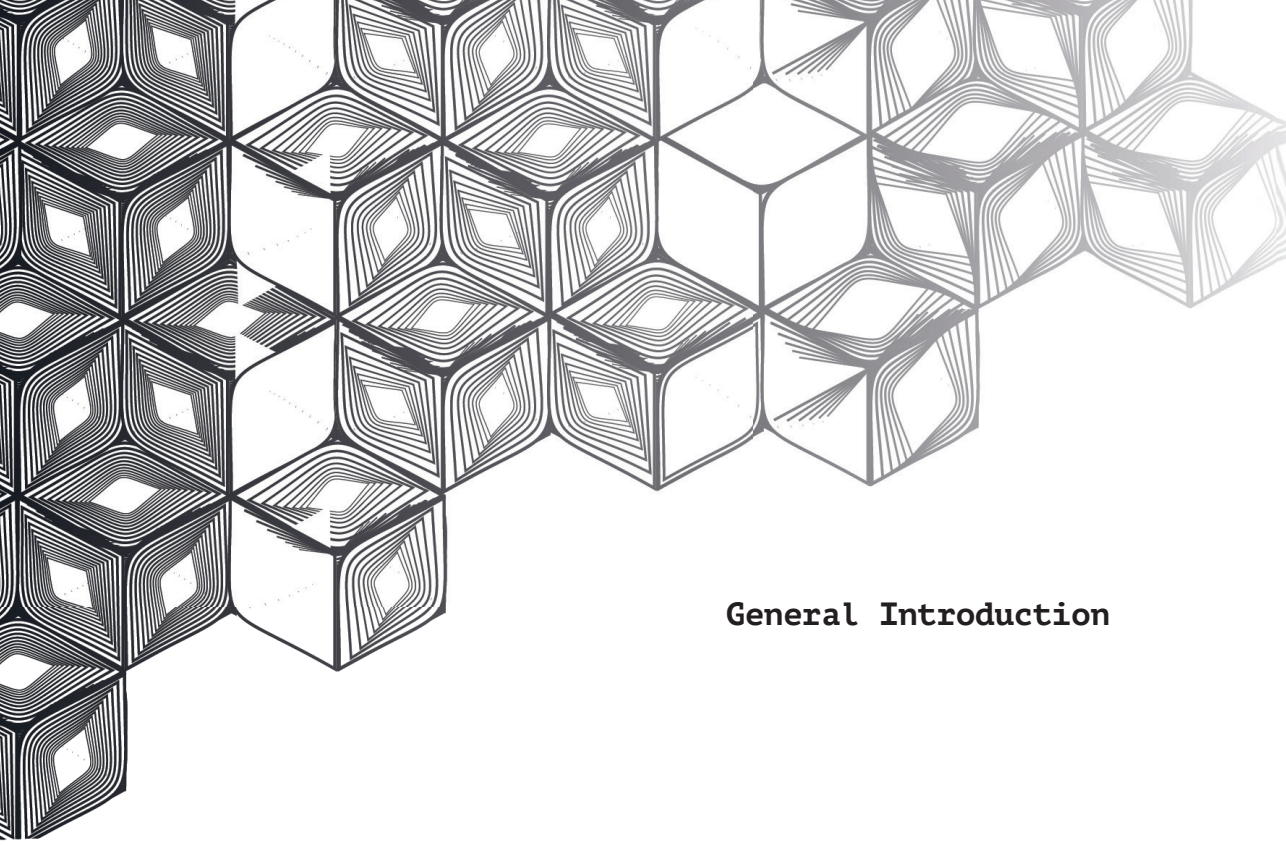
Beoordelingscommissie:

Prof. dr. M.C.J.M. Sturkenboom
Prof. dr. F.E. Scheepers (voorzitter)
Prof. dr. M.J.N.L. Benders
Prof. dr. P.G.M. Van Der Heijden
Prof. dr. M.G.W. Dijkgraaf

Table of contents



CHAPTER 1



General Introduction

What is a clinical prediction model?

A clinical prediction model is a combination of multiple variables (called predictors) to estimate the individualized probability of having a specific outcome (diagnostic models) or developing a particular outcome within a specific period (prognostic models).[1–3] Outcomes can be specific events, such as death and complications, but they can also be quantities, such as blood pressure, or changes in pain or quality of life. Broadly, the path before implementing a prediction model in clinical practice involves model development, validation, and assessment of their potential impact on daily practice (figure 1).[4] The presentation of clinical prediction models ranges from simple risk scores to real-time clinical decision support tools.[5] Well known examples used daily in clinical practice are the APGAR score to predict how well a new born will do outside the mother's womb and the Framingham score to predict the risk of heart disease within the next 10 years.[6]

During risk model development, predictors are combined to obtain a single absolute risk estimate.[7] Predictors can be patient and disease characteristics, medical imaging, test results, biosensors, genome sequencing, and even insurance claims. Traditionally, most risk prediction models are fitted using regression techniques, either logistic regression for short-term discrete outcomes or time-to-event regression for long-term outcomes. Before any prediction model is adopted in clinical practice, it is necessary to show that it provides estimates that are valid outside the specific data set wherein it was developed. Validation studies provide evidence on how well a model performs in a different study population, setting, time, or even in a different health domain.[8] Finally, model implementation studies provide insight on whether the use of a model ultimately improves decision making and behavior, and subsequently health-related outcomes.[8]

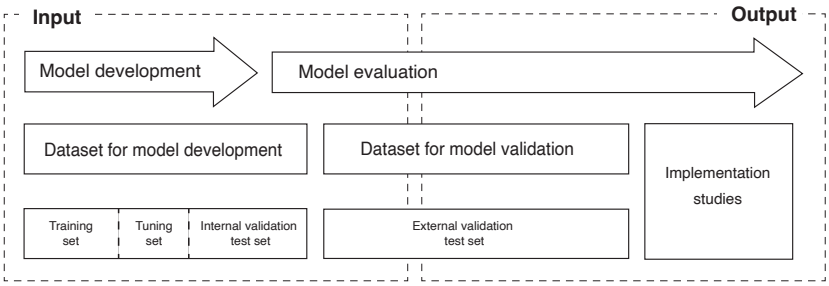


Figure 1. Paths to implement a prediction model in clinical practice

Why are prediction models important?

Healthcare professionals intuitively combine and judge an individual's information to estimate the probability of having or developing a certain outcome. Prediction models aim to support healthcare professionals by providing an objective and formal identification of individuals at, e.g., a high or low risk of a particular outcome and thus, aim to improve shared decision-making on subsequent management or changes in behavior, and consequently improve health-related outcomes of the individual. Prediction models can be used, e.g., to assist in primary prevention of healthy individuals, foresee recurrent events in patients, personalize treatments, and support learning healthcare systems. Furthermore, prediction models largely support stratified or personalized medicine, where decisions regarding prevention, monitoring or treatment choices are informed by an individual's profile and therefore, constitute an important asset across several pathways of the health(care) process.[9]

Worldwide, there is an increased demand for more efficient and sustainable healthcare. Technologies based on artificial intelligence (AI) and machine learning (ML) might have an important role in reducing diagnostic and treatment mistakes, avoid wasting of resources, improve workflows, and reduce inequities. Developed using vast amounts of data and computational power, AI-based prediction models might potentially provide more accurate predictions than the current healthcare system can.

What is Artificial Intelligence and Machine learning?

AI-based technologies with healthcare applications have grown in popularity over the past years. Artificial intelligence is widely defined as a branch of computer science that aims to developed machines capable of replicating tasks that would typically require human intelligence (figure 2). Machine learning, a branch of AI, has received attention as potentially promising modelling technique to speed up the process of diagnosis and improve outcome prognostication. Studies have demonstrated that AI-based prediction models can achieve expert-level diagnosis and prognosis in multiple disease contexts.[10,11] These models or algorithms have the potential to recognize complex patterns in data (i.e. pixels in an image, genomics) and automatically build flexible prediction models.[12]

Machine learning can be mainly sub-classified into supervised and unsupervised learning.[13] Supervised machine learning refers to algorithms in which a model is fit on a range of predictors with a known outcome, similar to the prevailing prediction models based on regression techniques. Once the model is developed, it will be capable of predicting the outcome when applied to new data. A variety of supervised machine learning methods are available, including

Artificial Neural Networks (ANN), Support Vector Machine (SVM), Classification and Regression Trees (CART), Random Forest (RF), and many more.[14–17] In contrast, unsupervised machine learning refers to methods in which a model is fit without a known outcome. These methods are thus used to find undefined patterns or clusters within a dataset.[18] Examples are principal component analysis (PCA) and factor analysis (FA). Unsupervised learning will not be further discussed in this thesis.

In biomedical literature, there is increased interest on the potential of machine learning for health(care). The application of different approaches for prediction model development (i.e. training) and validation (i.e. testing) have created the sense of a new "culture" beyond traditional modelling approaches, such as regression techniques.[19] Although both approaches share similarities and are rooted in statistics, machine learning remains for many users a 'black box'.[20,21]

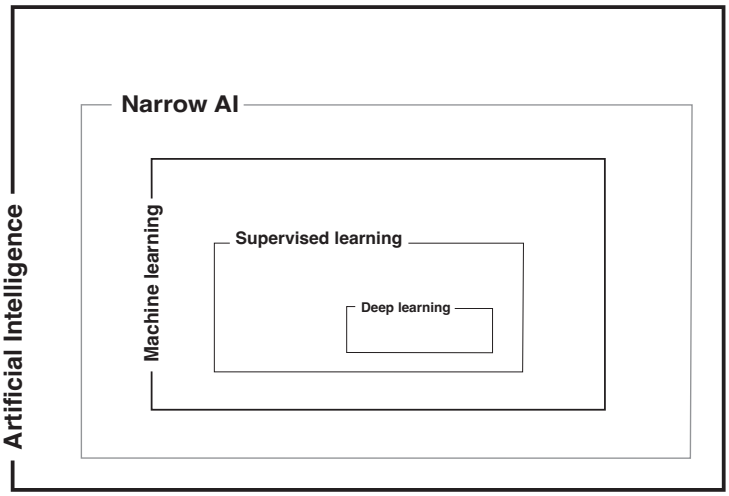


Figure 2. Concepts related to artificial intelligence

What is the current quality of prediction model studies?

Systematic reviews evaluating the methodological conduct and reporting of studies on prediction models developed with traditional regression techniques have all consistently concluded that these studies are plagued with deficiencies in study design, inadequate methodology, and poor reporting.[22] For example, prediction model studies typically fail to properly address missing values or, more surprisingly, do not share their developed models in a way that these can be replicated by independent researchers nor used by healthcare professionals.[23] Moreover, shortcomings in the design and statistical methods can make any study finding vulnerable to overinterpretation or its conclusions and implications overstated. Authors may,

for example, use exaggerated language or choose a particular statistical analysis to shape the impression their results will produce in readers.[24] These practices, deliberately or not, are known as 'spin'.[25,26] While there is clear evidence of the poor quality of studies on prediction models developed using regression techniques, it is yet unclear what is the quality of studies on prediction models developed using supervised machine learning.

Since 2019, PROBAST — Prediction model Risk of Bias ASsessment Tool, has been available to support the assessment of the methodological quality of studies that report either on the development, validation, or update of prediction models, regardless of the clinical domain, predictors, outcomes, or modelling technique used.[27,28] Similarly, TRIPOD – Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis, has been available as a consensus-based reporting guideline aiming to increase the completeness of reporting, and ultimately facilitate critical appraisal and further (independent) validations of prediction models.[29,30]

The number of studies on prediction models is increasing exponentially.[31,32] For the prediction of COVID-19 and its consequences, more than 200 models were published within one year alone.[33] The appetite for AI-based prediction algorithms with healthcare applications may outpace the development of rigorous evidence to support such groundbreaking innovations. Accordingly, potential or actual limitations may be overlooked leading to the premature adoption of poorly developed AI-based prediction models in daily health(care) and prevention.[34]

Researchers, healthcare professionals, and decision-makers frequently need to identify robust prediction models with actual value for health(care). They need to judge whether reported findings are justified in the context of methodological quality, and to which extent the findings of such studies are reproducible and generalizable. However, poor scientific practices and inadequate scrutiny and reporting of the study design and methods make this process challenging, leading to unclear evidence of its value and utility for health(care). Tailored tools to assess the quality and reporting of AI-based prediction models are therefore essential and urgently needed to ensure reliable, fast, and valuable application.[35]

What is the purpose of this thesis?

Due to the relative novelty of applying machine learning to develop prediction models for health(care), there is little information on the quality of published studies. Whether studies behind such promising claims can easily be validated by other researchers and whether they are likely to perform well on new individuals and settings has not been examined in detail. Furthermore, tools to support the quality assessment of these AI-based prediction models are currently missing.[35] Therefore,

the purpose of this thesis is to explore the current status of studies on clinical prediction models that used supervised machine learning as modelling technique in all medical domains. The original studies presented in this thesis are embedded in an umbrella review project aiming to unwrap the strengths and deficiencies, as well as to identify gaps for further methodological research on AI-based prediction models and support the development of quality assessment tools for such studies.

Poorly reported studies hinder its proper appraisal, replication, and ultimately, usefulness. From chapter 2 onwards, we assess the completeness of reporting and overinterpretation of findings in studies on AI-based prediction models in all medical domains. In **chapter 2**, we present the peer reviewed protocol of the umbrella review project. In **chapter 3**, we evaluated the adherence to the reporting guideline TRIPOD. Accurate representation of study findings is crucial to preserve public trust. The words used to describe the findings of a study could affect perceptions of the usefulness and transportability of prediction models, regardless of the modelling approach. In **chapter 4**, we look at 'spin' practices, that is more favorable reporting practices than justified by the results that might mislead readers in studies on prediction models using supervised machine learning. Given the subjective nature of spin evaluation, several challenges arose during its systematic assessment. We introduce a tailored framework to evaluate spin in prediction model studies in **chapter 5**.

Poorly conducted studies hinder a prediction model's validation and transportability into "real- world" health(care) settings. From chapter 6 onwards, we provide an overview of how AI-based models are currently built, and we critically appraised their methodological quality and risk of bias. In **chapter 6,** we describe in detail the study design and modelling choices in AI-based prediction models across medical specialties. In **chapter 7**, we provide further insights on the methodological conduct of AI-based models in oncology. Using PROBAST as benchmark, we evaluate the methodological quality and risk of bias of prediction models develop using supervised machine learning in **chapter 8**. In **chapter 9**, we compared the predictive performance of models built under different machine learning techniques.

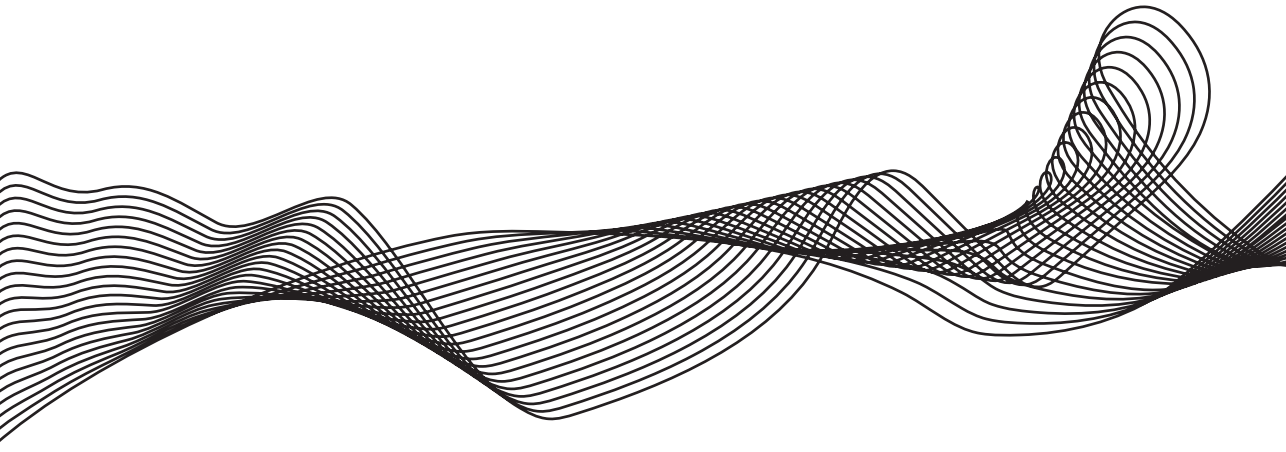
Few quality assessment tools for prediction models developed using machine learning are currently available. From chapter 10 onwards, we describe the initiative to extent TRIPOD and PROBAST to cover studies on prediction models developed using artificial intelligence. We introduce in **chapter 10** the peer-reviewed protocol for a large-scale international project aimed at the development of both tools: TRIPOD-AI and PROBAST-AI. In **chapter 11**, the protocol for PROBAST-AI is described in detail.

In **chapter 12**, we provide the conclusions of this thesis, and we end with a general discussion focusing on the first advances of PROBAST-AI.

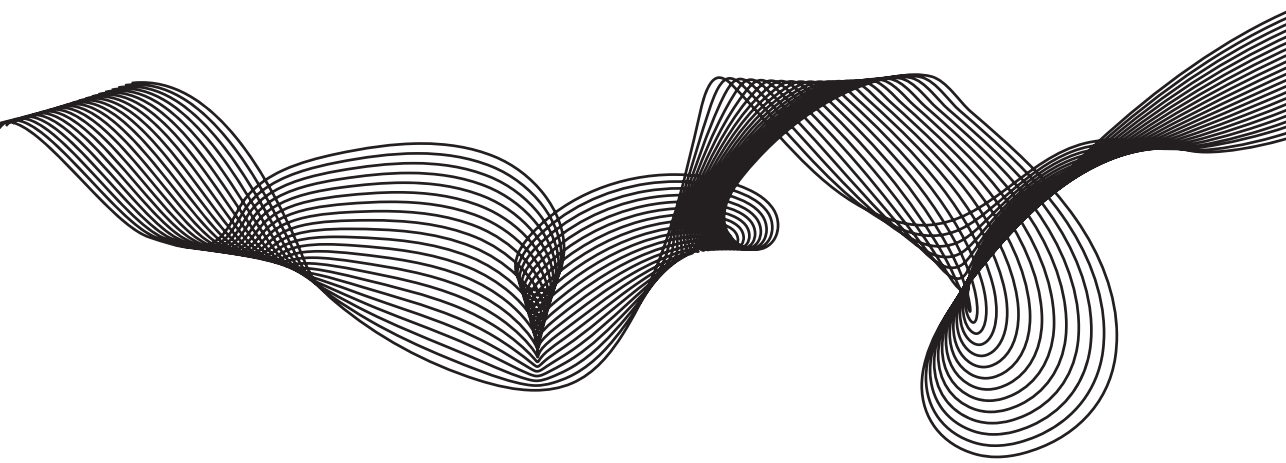
REFERENCES

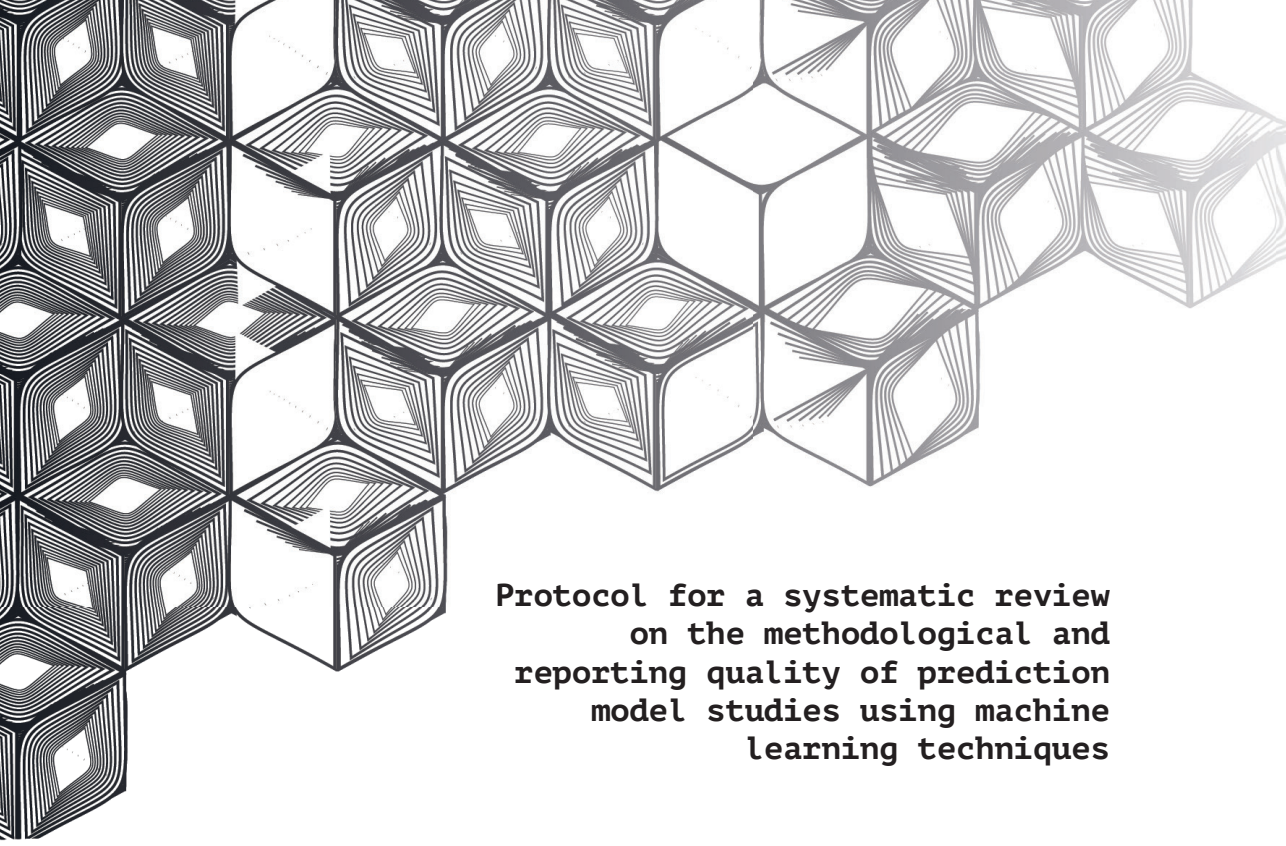
1. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2). doi:10.1371/journal.pmed.1001381
2. Riley, Richard D; van der Windt, Danielle; Croft, Peter; Moons KGM. Prognosis Research in Health Care: Concepts, Methods, and Impact. Oxford University Press; 2019. doi:10.1093/med/9780198796619.001.0001
3. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132(April):142-145. doi:10.1016/j.jclinepi.2021.01.009
4. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? BMJ. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
5. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ. 2019;365. doi:10.1136/bmj.l737
6. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-1847. doi:10.1161/01.CIR.97.18.1837
7. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-690. doi:10.1136/heartjnl-2011-301246
8. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247
9. Hingorani AD, Van Der Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ. 2013;346. doi:10.1136/bmj.e5793
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118. doi:10.1038/nature21056
11. 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. 2018;1(1). doi:10.1038/s41746-018-0040-6
12. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA - J Am Med Assoc. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391
13. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
14. Cortes C, Vapnik V. Support-Vector Networks. Machine. 1995;20(5):273-297. doi:10.1109/64.163674
15. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324
16. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115-133. doi:https://doi.org/10.1007/BF02478259
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. 1st editio.; 1984.
18. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):1-18. doi:10.1186/s12874-019-0681-4
19. Breiman L. Statistical Modeling: The Two Cultures. Stat Sci. 2001;16(3):199-231.
20. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189
21. Handelman GS, Kok HK, Chandra R V., et al. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. AJR Am J Roentgenol. 2019;212(1):38-43. doi:10.2214/AJR.18.20224
22. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. BMC

Med Res Methodol. 2014;14(1):40. doi:10.1186/1471-2288-14-40
23. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Informatics Assoc. 2022;00(0):1-7. doi:10.1093/jamia/ocac002
24. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Misreporting of Diagnostic Accuracy Studies : Evidence of "Spin." Radiology. 2013;267(2):581-588. doi:10.1148/radiol.12120527/-/DC1
25. Fletcher RH, Black B. Spin in scientific writing: Scientific mischief and Legal Jeopardy. Med Law. 2007;26(3):511-525.
26. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-2619. doi:10.1073/PNAS.1710755115
27. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med. 2019;170(1):W1-W33. doi:10.7326/M18-1377
28. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-58. doi:10.7326/M18-1376
29. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55. doi:10.7326/M14-0697
30. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73. doi:10.7326/M14-0698
31. Jong Y de, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, Diepen M van. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). Nephrology. Published online July 8, 2021. doi:10.1111/NEP.13913
32. Faes L, Sim DA, Van Smeden M, Held U, Bossuyt PM, Bachmann LM. Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? Front Digit Heal. 2022;1:833912. doi:10.3389/fdgth.2022.833912
33. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369. doi:10.1136/bmj.m1328
34. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. npj Digit Med. 2022;5(1). doi:10.1038/s41746-021-00549-7
35. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. Published online 2019. doi:10.1016/S01406736(19)302351



CHAPTER 2





Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques

Constanza L Andaur Navarro
Johanna AA Damen
Toshihido Takada
Steven WJ Nijman
Paula Dhiman
Jie Ma
Gary S Collins
Ram Bajpai
Richard D Riley
Karl GM Moons
Lotty Hooft

BMJ Open 2020; 10(11):e038832

ABSTRACT

Introduction. Studies addressing the development and/or validation of diagnostic and prognostic prediction models are abundant in most clinical domains. Systematic reviews have shown that the methodological and reporting quality of prediction model studies is suboptimal. Due to the increasing availability of larger, routinely collected and complex medical data, and the rising application of Artificial Intelligence (AI) or Machine Learning (ML) techniques, the number of prediction model studies is expected to increase even further. Prediction models developed using AI or ML techniques are often labeled as a "black box" and little is known about their methodological and reporting quality. Therefore, this comprehensive systematic review aims to evaluate the reporting quality, the methodological conduct, and the risk of bias of prediction model studies that applied ML techniques for model development and/or validation.

Methods and Analysis. A search will be performed in PubMed to identify studies developing and/or validating prediction models using any ML methodology and across all medical fields. Studies will be included if they were published between January 2018 and December 2019, predict patient-related outcomes, use any study design or data source, and available in English. Screening of search results and data extraction from included articles will be performed by two independent reviewers. The primary outcomes of this systematic review are: (1) the adherence of ML based prediction model studies to the Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD), and (2) the risk of bias in such studies as assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). A narrative synthesis will be conducted for all included studies. Findings will be stratified by study type, medical field, and prevalent machine learning methods, and will inform necessary extensions or updates of TRIPOD and PROBAST to better address prediction model studies that used AI or ML techniques.

Ethics and dissemination. Ethical approval is not required for this study because only available published data will be analyzed. Findings will be disseminated through peer-reviewed publications and scientific conferences.

Systematic review registration: PROSPERO, CRD42019161764.

ARTICLE SUMMARY

Strengths and Limitation of this study

1. This protocol increases transparency to the methods and definitions that we used in our review and are applied to develop prediction model studies using Artificial Intelligence or Machine Learning.
2. The systematic review will provide an overview and critical appraisal of the methodological and reporting quality, and risk of bias of prediction model studies using Machine Learning.
3. The findings of the review will provide the needed evidence for the development of tailored methodological and reporting guidelines for prediction model studies based on Machine Learning techniques.
4. We will build a sensitivity search strategy by using terms related to Machine Learning techniques, as well, as conventional prediction techniques.
5. Language restriction to English might exclude additional studies published in other languages.

INTRODUCTION

Clinical prediction models aim to estimate the individualized probability that a particular outcome, e.g. condition or disease, is present (diagnostic models) or whether a specific outcome will occur in the future (prognostic models).[1-4] Studies addressing the development, validation, and updating of prediction models are abundant in most clinical domains. For example, in cardiovascular disease, more than 350 prediction models have been developed and only a few have been validated.[5] Moreover, systematic reviews have shown that, within different medical domains, the methodological and reporting quality of prediction model studies is suboptimal.[6-10] Due to the increasing availability of larger, routinely collected and complex medical data, and the rising application of Artificial Intelligence (AI) or Machine Learning (ML) techniques for clinical prediction, the number of prediction model studies is expected to increase even further.

Machine Learning can be described as techniques that directly and automatically learn from data without being explicitly programmed for that task, and often without any prior assumption.[11,12,13] Thus ML relies on patterns and inferences from the data itself. A perceived advantage of ML over conventional statistical techniques is its ability to analyze "big", non-linear and high dimensional data, and thus its ability to model complex associations and scenarios. Due to the novelty, diversity, flexibility, and complexity of ML techniques, ML based prediction model studies are often considered as uninterpretable for many users. Inadequate reporting of, e.g. data sources, study design, modeling processes, number of predictors, and other data assumptions, makes prediction models developed with ML techniques published in medical journals difficult to interpret and to be validated by other researchers, creating barriers to their use in daily clinical practice.

Complete reporting is essential to judge the validity of any prediction model as it facilitates: study replication, independent validation of the prediction model, risk of bias assessments, interpretation of the results, meta-analysis of prediction models, and the judgment of the value and applicability of such model in real clinical settings for individualized predictions.[14] While complete reporting reveals the strengths and limitations of a prediction model, it also enhances the use and implementation of prediction model in clinical practice. The "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)" statement has been available since 2015, providing a checklist of 22 items considered essential for informative reporting of diagnostic or prognostic prediction model studies.[15,16] Similarly, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) was published in 2019 to guide the critical appraisal of prediction model studies.[17,18] PROBAST provides signaling questions to facilitate both the applicability and risk of bias assessment of prediction model studies across four domains: participants,

predictors, outcome, and analysis. This assessment can only be correctly implemented if prediction model studies are properly reported. Although TRIPOD and PROBAST both covered all types of prediction modelling studies, including those using ML techniques, their focus was on regression-based modeling. The challenges and necessity for reporting and quality assessment guidelines in the AI/ML field has been addressed by several authors and this has led to initiatives such as, CONSORT-AI (for randomized controlled trials), and SPIRIT-AI (for clinical trial protocols). Similarly, for prediction model studies using ML, TRIPOD-ML and PROBAST-ML have been announced.[19-21]

To improve the quality, transparency, and usability of ML based prediction models in medicine, it is important to explore the current use and reporting of ML techniques in prediction model studies, to evaluate the methodological conduct and risk of bias using PROBAST, and assess the adherence to TRIPOD by performing a comprehensive systematic review.[3,15-18, 22]

Study aim

- The primary aim of this systematic review is to evaluate the reporting and the methodological conduct of studies reporting on prediction models developed with supervised ML techniques, across all medical fields. Specific objectives are to:
- Evaluate the reporting quality of prediction models developed using ML techniques based on TRIPOD.
- Assess the methodological quality and the risks of bias in prediction model development or validation studies using ML techniques based on PROBAST.
- Identify key and emerging concepts for the development of tailored adaptations or extensions of both TRIPOD and PROBAST.

METHODS

Our systematic review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) on 19 December 2019 (CRD42019161764). This protocol was prepared using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA-P) 2015 statement.[23]

Eligibility criteria

Articles will be eligible for this review when describing primary studies on the development and/or validation of a multivariable diagnostic or prognostic prediction model with at least 2 predictors, using any supervised ML methodology within all medical fields, and published between January 2018 and December 2019. This last inclusion criterion is to obtain the most contemporary sample of articles that would reflect the current practices of applied methods in the ML prediction model field to date. We will include studies with any study design and data source, all patient-related health outcomes, all outcome formats, and restricted to humans only. Further details about inclusion criteria are given in Table 1.

Table 1 Definition of inclusion criteria

Inclusion criteria	Definition
Any study design	Articles that report the development and/or validation of a prediction model based on experimental studies or observational studies. This includes randomized controlled trials, prospective and, retrospective cohort, case-control studies, and case-cohort studies.
Using at least 2 predictors for risk prediction	Articles that report the development and/or validation of a prediction model using at least 2 predictors. Articles that use imaging or speech parameters as structured data plus other predictors such as e.g. clinical, demographics, histological and genetic risk scores features will be included.
Any data sources	Articles that report the development and/or validation of a prediction model using any structured data source, e.g. electronic medical records, claims data, and individual patient data meta-analysis data.
Any supervised ML technique	Articles that report the use of any ML technique for development and/or validation of a prediction model. We will consider as a ML technique, a statistical technique based on advanced computational capacity and lower human intervention. More specifically, we will focus on supervised ML technique.
Patient health-related outcomes	Articles that report the development and/or validation of a prediction model whose main outcome is on an individual patient level. We will include articles assessing diagnosis, prognosis, and health services performance, such as length of stay or triage assessment.
All outcome measures format	Articles that report the development and/or validation of a prediction model whose main outcome has one of the following formats: continuous, binary, ordinal, multinomial, and time-to-event.

Articles will be excluded from this review when reporting models that make predictions for enhancing the reading of images or signals (rather than for prediction of health outcomes in individuals) or, use only genetic or molecular markers as candidate predictors. Furthermore, prognostic factor studies, secondary research, conference abstracts, and studies for which no full-text is available will also be

excluded. The search will be restricted to articles available in English only. Further details about exclusion criteria are given in Table 2.

Table 2 Definition of exclusion criteria

Exclusion criteria	Definition
Images or signal studies	Articles that report the development and/or validation of a prediction model for enhancing the reading of images, pathological samples, or signals. The purpose of this articles is to improve the accuracy of an instrument rather than providing a clinical outcome.
Only genetic and/or molecular predictors	Articles that report the development and/or validation of a prediction model using only genetic and/or molecular candidate predictors. These articles are often based on high-dimensional data and unsupervised ML techniques.
Prognostic factors studies	Articles that report the identification of prognostic factors associated with a clinical outcome in an individual.
Secondary research	Articles that report narrative reviews, systematic reviews about prediction model studies in a specific medical field. Guidelines, expert's opinions. and letters to the editor will also be excluded.
Conference abstract	Articles that report the development and/or validation of a prediction model presented in a conference. Such articles, by definition, do not report all the information required for assessment.
Full-text not available	Articles that report the development and/or validation of a prediction model for which full-text is not accessible online.

Information sources

A literature search will be systematically applied in one major public-available electronic medical literature databases (PubMed) from 01 January 2018 to 31 December 2019.

Search strategy

The search strategy was built using keywords including ML-related terms (i.e. 'supervised learning', 'support vector machine', 'neural network'), prediction-related terms[24] (i.e. 'risk', 'prognosis'), and several performance measures for prediction modelling (i.e. 'AUC', 'O:E ratio'). For search refinement, we selected 30 articles aligned with our inclusion/exclusion criteria to create a "golden bullet" set. This set was analyzed using SWIFT-Reviewer to obtain the most frequent words in the included articles by topic modelling.[25] In MedlinerRanker, the analysis of the included and excluded golden bullets articles allowed us to obtain the most discriminative words to be considered in the search strategy.[26] The final search strategy is presented in supplemental file 1.

Study records

Data management

Study record information including title and abstract from the searched online database will be imported into Endnote Citation Manager and Rayyan systematic

review software.[27] These platforms will track and backup all activities when authors conduct the literature review process. Once eligible studies are identified, full-text articles will be downloaded for full-text screening and data extraction. Data items (below) will be extracted from the final included studies for review using Research Data Capture (REDCap) software.[28]

Selection process

Two researchers, from a combination of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD) will independently screen the titles and abstracts to identify eligible studies according to the eligibility criteria. Two independent researchers, from the combination of the previous seven reviewers, will review the full text for potentially eligible articles; one researcher (CLAN) will screen all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) will collectively screen a portion of the same articles for agreement. Disagreements between reviewers will be solved by consensus or consultation with a third investigator, if necessary (JAAD). The study flow will be presented in a PRISMA flowchart.[29]

Data collection process

We will perform a double data extraction for all included articles. Two reviewers will independently extract data from each article using a standardized data extraction form. One researcher (CLAN) will extract data from all articles and six other researchers (TT, SWJN, PD, JM, RB, JAAD) will collectively extract data from the same articles. The data extraction form will be piloted on five papers and amended if necessary. Disagreements in data extraction will be discussed between the two reviewers, and adjudicated by a third reviewer (KGMM, GSC, RDR or LH), if necessary. The authors of the articles will be contacted for further information and clarification if needed. Data and records will be maintained by the lead investigator (CLAN) and stored on a shared secure platform for access by all investigators (REDCap).

Data items

Data to be extracted will be informed by TRIPOD using the TRIPOD data extraction adherence guidance, PROBAST and the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS).[15-18,22,30] Additional items specifically relating to ML techniques for prediction model purposes, will also be extracted.

Extracted data will include study design for the development and validation of the model, outcomes to be predicted, setting, the intended use of the prediction model, study population, data source, patient characteristics, total study sample size, number of individuals with the outcome, number of predictors (candidate and final), internal validation type, predictive performance measures (discrimination

and calibration), number of models developed, and the details of the machine learning technique used to develop each model (e.g. technique, pre-processing, data cleaning, optimization algorithm, predictors selection, penalization techniques, hyperparameters, code, and data availability, etc.). This form will contain instructions for the reviewers on how to assess the models presented in the articles. For example, the number of models developed will be based on how many ML techniques were used, including if several hyperparameters are tuned. We will set a limit to the number of models for data extraction to 10. The number of predictors will be counted based on what is reported in the article and/or supplemental file. If not stated, the number of predictors will be reported as unclear. The final data extraction form is presented in supplemental file 2.

Outcomes

The primary outcomes of this systematic review are the adherence to the TRIPOD reporting guideline and the risk of bias assessed using PROBAST.[17,18,22]

Assessment of risk of bias

The risk of bias of individual studies is one of our outcomes of interest and will be assessed using PROBAST.[17,18]

Data synthesis

We will conduct a narrative synthesis of the extracted data. Data will be summarized using descriptive statistics and visual plots. Numbers and percentages will be used to describe categorical data about the reporting, methodological conduct, and risks of bias of the studies. The distribution of continuous data, such as sample size and the number of predictors, will be assessed and described using mean and standard deviation for normally distributed data and using median and 25th and 75th percentiles for non-normally distributed data. The risk of bias assessment will be summarized and graphically presented for each PROBAST domain and as an overall risk of bias judgement. Results will be stratified by study type (development with internal validation and/or external validation), medical field, and prevalent ML techniques.

Meta-bias(es)

Meta-bias will not be investigated in this study.

Confidence in cumulative evidence

The strength of the body of evidence will not be assessed in this study.

Amendments

Protocol amendments will be listed and made available on the PROSPERO registration. The date, description, and rationale will be given for each amendment.

ETHICS AND DISSEMINATION

Ethical approval is not required for this study because only available published data will be analyzed. The findings of this systematic review will be published in an open-access journal to ensure access for all stakeholders and disseminated in various scientific conferences.

Patient and public involvement

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

DISCUSSION

The use of ML has been increasingly recognized as a powerful tool to improve healthcare by enabling related professionals to make decisions based on the increasingly available and diverse sources of (bio)medical data. Particularly, ML based prediction algorithms are considered the key to unlock the increasingly available data sources, are intended to better inform real-time clinical decisions, support early-warning systems, and provide superhuman imaging diagnostics.[31] However, published research about this topic rarely provides adequate information about the final predictive model, and its estimates and performance. Even more scarce is research where the prediction model is accessible for patients and healthcare professionals alike. Hence, ML based prediction model studies are often seen as uninterpretable. This aspect of ML techniques is problematic especially in medical diagnosis and prognosis, hampering the judgement of quality, clinical acceptance, and implementation.

At present, there is a limited number of systematic reviews regarding the reporting and methodological quality of ML based prediction model studies and their risks of bias.[32,33,34] In this systematic review, we will review across all medical fields, the current use of ML techniques in prediction model development, validation and updating studies, the methodological conduct and risks of bias of using PROBAST, and the adherence to the reporting guideline for such studies using TRIPOD. Particularly, we will assess the extent to which risks of bias and reporting of ML based prediction model studies match the current recommendations from TRIPOD and PROBAST,[22] and the implications of these results to update or extend them to TRIPOD-ML and PROBAST-ML.

So far, our findings should be considered within limitations. Machine Learning is a recently developed concept and without a clear scope yet. Therefore, a sensitive search strategy is hard to build, which may result in a large number of abstracts to screen at initial stages. Additionally, we are only able to include articles in English, which will underrepresent research available in other languages.

Acknowledgements
The authors would like to thank and acknowledge the support of René Spijker, information specialist.

Authors Contributions
The study concept and design were conceived by CLAN, JD, KGMM, LH, PD, GSC, and RDR. CLAN, JD, TT, SN, PD, JM and RB will conduct article screening and data extraction. CLAN will perform data analysis. All authors drafted this manuscript, revised it for important content, and have provided the final approval of this version. CLAN, the corresponding author, is the guarantor of the review.

Funding
GSC is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is funded by the NIHR Oxford BRC. The funder was not involved in the development of this protocol.

Competing interests
None declared.

Provenance and peer review
Not commissioned; externally peer-reviewed.

Data sharing statement
The final dataset that supports the findings of this study will be available from the corresponding author upon reasonable request.

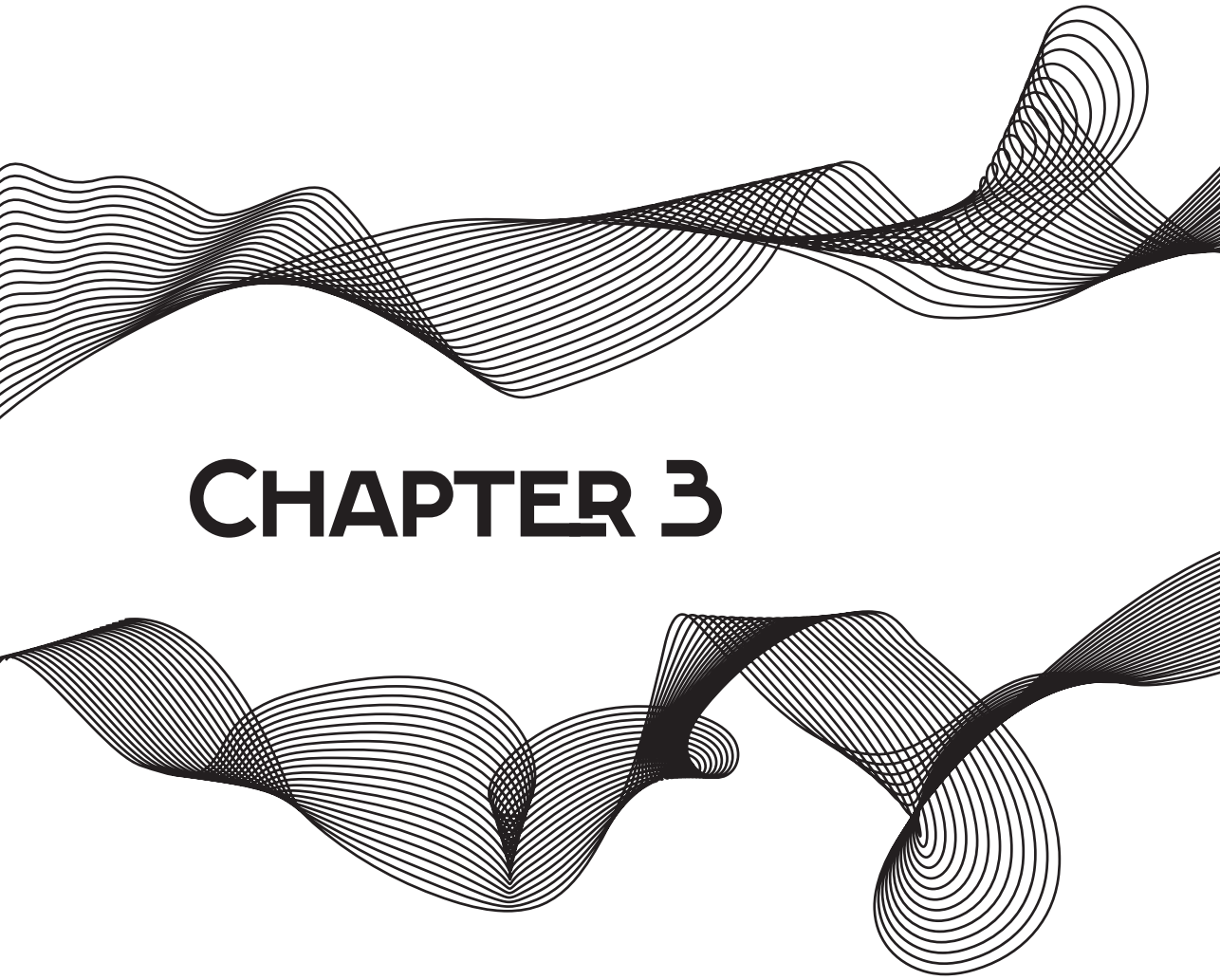
Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1136/bmjopen-2020-038832

Abbreviations
AI: Artificial Intelligence; ML: Machine Learning; TRIPOD: Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis; PROBAST: Prediction model Risk Of Bias Assessment Tool.

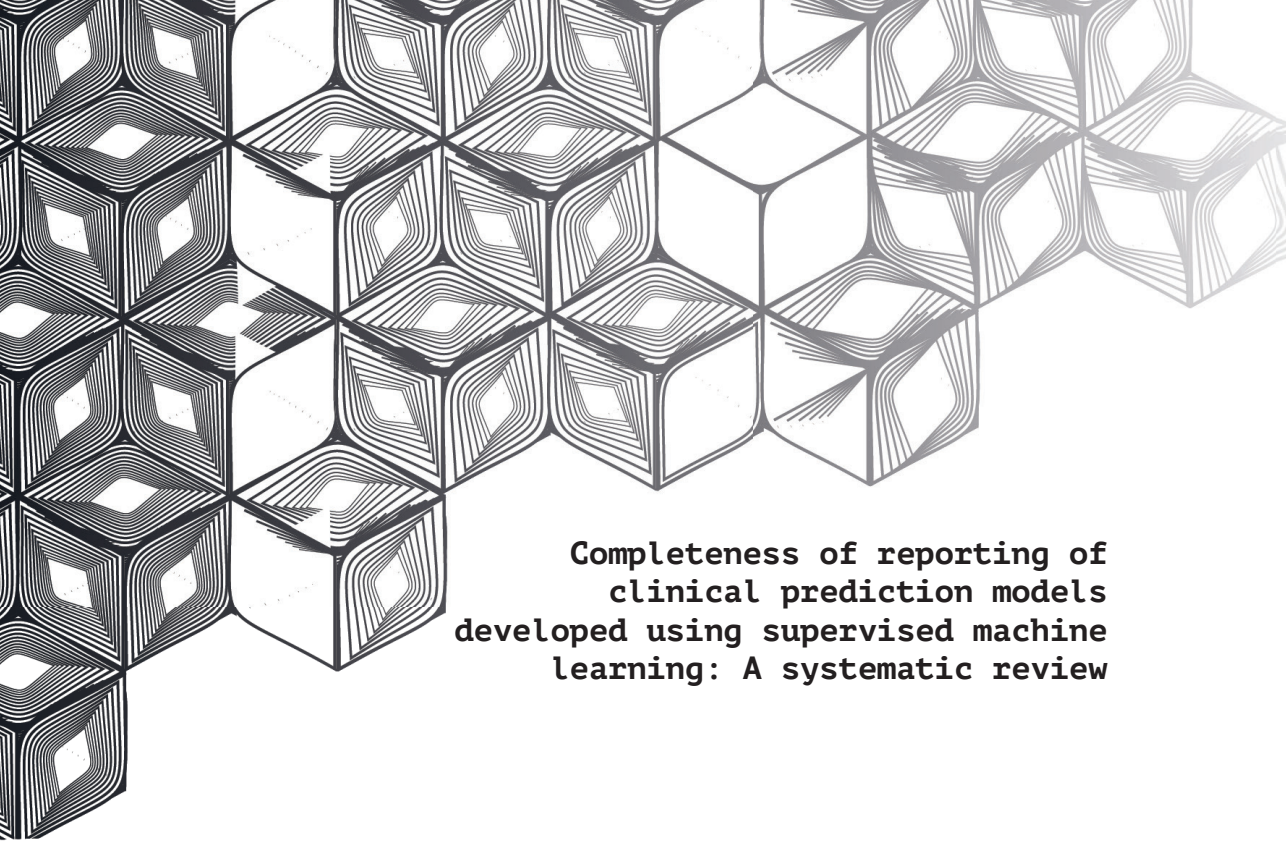
REFERENCES

1. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: What, why, and how? BMJ 2009; doi:10.1136/bmj.b375.
2. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med 2013; doi:10.1371/journal.pmed.1001381.
3. Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Healthcare: Concepts, Methods, and Impact. First ed. Oxford University Press; 2019.
4. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Second ed. Cham, Switzerland: Springer; 2019
5. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. BMJ 2016; doi:10.1136/bmj.i2416.
6. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. BMC Med 2018; doi:10.1186/s12916-018-1099-2.
7. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014; doi:10.1186/1471-2288-14-40.
8. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. PLoS Med 2012; doi:10.1371/journal.pmed.1001221
9. Collins GS, Mallet S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC med. 2011; doi:10.1186/1741-9-103.
10. Wen Z, Guo Y, Xu B, et al. Developing risk prediction models for postoperative pancreatic fistula: a systematic review of methodology and reporting quality. Indian J Surg 2016; doi: 10.1007/s12262-015-1439-9
11. Mitchell TM. Machine Learning. New York, NY: McGraw Hill; 1997.
12. Bi Q, Goodman KE, Kaminsky J, et al. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol 2019; doi: 10.1093/aje/kwz189.
13. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019; doi: 10.1186/s12874-019-0681-4.
14. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2.-4.
15. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015 doi:10.7326/M14-069717.
16. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 2015; doi:10.7326/M14-0698
17. Wolff RF, Moons KGM, Riley RD; PROBAST Group. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019; doi:10.7326/M18-1376
18. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med 2019; doi:10.7326/M18-1377
19. Liu Y, Chen PC, Krause J, et al. How to read articles that use Machine Learning: Users' guides to the medical literature. JAMA 2019; doi: 10.1001/jama.2019.16489.
20. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019; doi:

10.1038/s41591-019-0603-3.
21. Collins GS, Moons KGM. Reporting of artificial inteligence prediction models. *Lancet* 2019; doi: 10.1016/S0140-6736(19)30037-6.
22. Heus P, Damen JAAG, Pajouheshnia R, *et al*. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ open* 2019; doi: 10.1136/bmjopen-2018-025611.
23. 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; doi: 10.1186/2046-4053-4-1.
24. Geersing GJ, Bouwmeester W, Zuithoff P, *et al*. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. *PLoS one 2012*; doi: 10.1371/journal.pone.0032844.
25. Howard BE, Phillips J, Miller K, et al. SWIFT-Review: a text-mining workbench for systematic review. *Syst Rev* 2016; doi: 10.1186s/s13643-016-0263-z.
26. Fontaine JF, Barbosa-Silva A, Schaefer M, *et al*. MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res* 2009; doi: 10.1093/nar/gkp353.
27. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan- a web and mobile app for systematic reviews. Syst Rev 2016; doi: 0.1186/s13643-016-0384-4.
28. Harris PA, Taylor R, Thielker R, *et al*. Research electronic data capture (REDCap). A metadata-drien methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; doi: 10.1016/j.jbi.2008.08.010.
29. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Ann Intern Med* 2009; doi: 10.1371/journal.pmed.1000100.
30. Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med* 2014; doi:10.1371/journal.pmed.1001744.
31. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the peak of inflated expectations. *N Engl J Med* 2018; doi: 10.1056/NEJMp1702071.
32. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; doi: 10.1016/j.jclinepi.2019.02.004.
33. Shillan D, Sterne JAC, Champneys A, *et al*. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit Care* 2019; doi: 10.1186/s13054-019-2564-9.
34. Wang W, Kiik M, Peek N, *et al*. A systematic review of Machine Learning models for predicting outcomes of stroke with structured data. *PLoS One* 2020; 15:e0234722.



Chapter 3



Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review

Constanza L Andaur Navarro
Johanna AA Damen
Toshihido Takada
Steven WJ Nijman
Paula Dhiman
Jie Ma
Gary S Collins
Ram Bajpai
Richard D Riley
Karl GM Moons
Lotty Hooft

BMC Medical Research Methodology 2022;22(1):12

ABSTRACT

Background. While many studies have consistently found incomplete reporting of regression-based prediction model studies, evidence is lacking for machine learning-based prediction model studies. We aim to systematically review the adherence of Machine Learning (ML)-based prediction model studies to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement.

Methods. We included articles reporting on development or external validation of a multivariable prediction model (either diagnostic or prognostic) developed using supervised ML for individualized predictions across all medical fields. We searched PubMed from 1 January 2018 to 31 December 2019. Data extraction was performed using the 22-item checklist for reporting of prediction model studies (www.TRIPOD-statement.org). We measured the overall adherence per article and per TRIPOD item.

Results. Our search identified 24 814 articles,of which 152 articles were included: 94 (61.8%) prognostic and 58 (38.2%) diagnostic prediction model studies. Overall, articles adhered to a median of 38.7% (IQR 31-46.4) of TRIPOD items. No articles fully adhered to complete reporting of the abstract and very few reported the flow of participants (3.9%, 95% CI 1.8 to 8.3), appropriate title (4.6%, 95% CI 2.2 to 9.2), blinding of predictors (4.6%, 95% CI 2.2 to 9.2), model specification (5.2%, 95% CI 2.4 to 10.8), and model's predictive performance (5.9%, 95% CI 3.1 to 10.9). There was often complete reporting of source of data (98.0%, 95% CI 94.4 to 99.3) and interpretation of the results (94.7%, 95% CI 90.0 to 97.3).

Conclusion. Similar to prediction model studies developed using conventional regression-based techniques, the completeness of reporting is poor. Essential information to decide to use the model (i.e. model specification and its performance) is rarely reported. However, some items and sub-items of TRIPOD might be less suitable for ML-based prediction model studies and thus, TRIPOD requires extensions. Overall, there is an urgent need to improve the reporting quality and usability of research to avoid research waste.

Systematic review registration: PROSPERO, CRD42019161764

INTRODUCTION

Clinical prediction models are used extensively in healthcare to aid patient diagnosis and prognosis of disease and health status. A diagnostic model combines multiple predictors or test results to predict the presence or absence of a certain disorder, whereas a prognostic model estimates the probability of future occurrence of an outcome.[1–3] Studies developing, validating, and updating prediction models are abundant in most clinical fields and their number will continue to increase as prediction models developed using artificial intelligence (AI) and machine learning (ML) are receiving substantial interest in the healthcare community.[4]

ML, a subset of AI, offers a class of models that can iteratively learn from data, identify complex data patterns, automate model building, and predict outcomes based on what has been learned using computer-based algorithms.[5,6] ML is often described as more efficient and accurate than conventional regression-based techniques. ML-based prediction models, correctly developed, validated, and implemented, can improve patient benefit, and reduce disease and health system burden. There is increasing concern of the methodological and reporting quality of studies developing prediction models, with research till date focusing on models developed with conventional statistical techniques such as logistic and Cox regression.[7–11a] Recent studies have found limited application of ML-based prediction models because of poor study design and reporting.[12,13]

Incomplete (or unclear) reporting makes ML-based prediction models difficult to interpret and impedes validation by independent researchers, thus creating barriers to their use in daily clinical practice. Complete and accurate reporting of ML-based prediction model studies will improve its interpretability, reproducibility, risk of bias assessment, and applicability in daily medical practice and is, therefore, essential for high-quality research.[14] To improve transparency and reporting of prediction model studies, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement, a checklist of 22 items, was designed (www.tripod-statement.org).[15,16] Specific guidance for ML-based prediction model studies is currently lacking and has initiated the extension of TRIPOD for prediction models developed using ML or AI (TRIPOD-AI).[17,18]

We conducted a systematic review to assess the completeness of reporting of ML-based diagnostic and prognostic prediction model studies in recent literature using the TRIPOD Statement.[15,16] Our results will highlight specific reporting areas that can inform reporting guidelines for ML, such as TRIPOD-AI.[17,18]

METHODS

Our systematic review protocol was registered (PROSPERO, CRD42019161764) and published.[19] We reported this systematic review following the PRISMA statement.[20]

Data source and search

We searched PubMed on 19 December 2019 to identify primary articles describing prediction models (diagnostic or prognostic) using any supervised ML technique across all clinical domains published between 1 January 2018 and 31 December 2019. The search strategy is provided in the supplemental material.

Study selection

We included articles that described the development or validation of one or more multivariable prediction models using any supervised ML technique aiming for individualized prediction of risk or outcomes. As there is still no consensus on a definition of ML, we defined a 'study using ML' as a study that describes the use of a non-generalized linear models to develop or validate a prediction model (e.g. tree-based models, ensembles, deep learning). Extensions to traditional statistical techniques such as generalized additive models and multivariable adaptive regression splines were considered as non-machine learning for this study. Hence, studies that claimed to have used ML, but they reported only regression-based statistical techniques were excluded from this systematic review (e.g. logistic regression, lasso regression, ridge regression and elastic net). Specifically, we focused on supervised ML, a subdomain of ML, that is characterized by the development of an algorithm that can predict (the risk of) outcomes for new observations (individuals) after learning from existing individuals and their labelled outcomes. For example, random forests, support vector machine, neural network, naïve bayes, and gradient boosting machines.

Articles reporting on the incremental value or model extension were also included. We included all articles regardless of study design, data source, or patient-related health outcome. Articles that investigated a single predictor, test or biomarker, or its causality with an outcome were excluded. Articles using ML to enhance reading of images or signals, or articles where ML models only used genetic traits or molecular markers as predictors, were also excluded. We also excluded systematic reviews, conference abstracts, tutorials, and articles for which full-text was unavailable via our institution. We restricted the search to human subjects and English-language articles. Further details are stated in our protocol.[19]

Two researchers, from a group of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD), independently screened titles and abstracts to identify potentially eligible studies. Full-text articles were then retrieved, and two independent researchers reviewed

them for eligibility using Rayyan.[21] One researcher (CLAN) screened all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively screened the same articles. Disagreements between reviewers were resolved by a third researcher (JAAD).

Data extraction

The data extraction form was based on the TRIPOD adherence assessment form (www.tripod-statement.org).[22] This form contains several adherence statements (hereafter called sub-items) per TRIPOD item. Some items and sub-items are applicable to all types of studies, while others are only applicable to model development only or external validation only (Table 1). To judge reporting of the requested information, sub-items were formulated to be answered with 'yes', 'no', 'not applicable'. We amended the published adherence form by omitting the 'referenced' option because we checked the information in the references, supplemental material, or appendix. Sub-items 10b and 16 were extracted per model rather than at study-level, as they refer to model performance. We limited our extraction and assessment to the first model reported in the Methods section so we could achieve a consistent evaluation of the items related to the Result section as well (item 13-17).

We performed a double data extraction for included articles. Two reviewers independently extracted data from each article using the standardized form which was available in REDCap, a data capture tool.[23] To accomplish consistent data extraction, the form was piloted by all reviewers on five articles. One researcher (CLAN) extracted data from all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively extracted data from the same articles. Discrepancies in data extraction were discussed and resolved between each pair of reviewers.

Data synthesis and analysis

We categorized prediction model studies as prognosis or diagnosis and classified studies by research aim: development (with or without internal validation), development with external validation (same model), development with external validation (different model), and external validation only. Detailed definition of research aims can be found in the supplemental material. Where articles described the development and/or validation of more than one prediction model, we chose the first ML model reported in the methods section for extraction.

We scored each TRIPOD item as 'reported' and 'not reported' based on answers to corresponding sub-items. If the answer to all sub-items of a TRIPOD item is scored 'yes' or 'not applicable', the corresponding item was considered 'reported'. Two analyses were conducted: adherence per item and overall adherence per article. We calculated the adherence per TRIPOD item by dividing the number of studies that adhered to a specific item by the number of studies in which the item was applicable. The total number of TRIPOD items varies by the type of prediction model study

(Table 1). We calculated the overall adherence to TRIPOD per article by dividing the sum of reported TRIPOD items by the total number of applicable TRIPOD items for each study. If an item was 'not applicable' for a particular study, it was excluded when calculating the overall adherence, both in the numerator and denominator.[22]

Analyses were performed using R version 3.6.2 (R Core Team, Vienna, Austria). Results were summarized as percentages with confidence intervals calculated using the Wilson score interval. In addition, we also used medians, IQR ranges, and using visual plots.

Table 1. TRIPOD adherence reporting items

Reporting Items		Study design	If applicable to studies	Reporting items for TRIPOD adherence	
				Development only	Development and validation
1. Title		D,V		✓	✓
2. Abstract		D,V		✓	✓
Introduction					
3. Background and objectives	**a.** Context and rationale	D,V		✓	✓
	b. Objectives	D,V		✓	✓
Methods					
4. Source of data	**a.** Source of data	D,V		✓	✓
	b. Key dates	D,V		✓	✓
5. Participants	**a.** Study setting	D,V		✓	✓
	b. Eligibility criteria	D,V		✓	✓
	c. Details of treatment	D,V	✓	✓	✓
6. Outcome	**a.** Outcome definition	D,V		✓	✓
	b. Blinding of outcome assessment	D,V		✓	✓
7. Predictors	**a.** Predictors definition	D,V		✓	✓
	b. Blinding of predictor assessment	D,V		✓	✓
8. Sample size	Arrival at study size	D,V		✓	✓
9. Missing Data	Handling of missing data	D,V		✓	✓
10. Statistical analysis	**a.** Handling of predictors in the analysis	D		✓	✓
	b. Specification of the model, all model building procedures, and internal validation methods	D		✓	✓
	c. For validation, description of how predictions were calculated	V		✓	n.a.
	d. Specification of all measures used to assess model performance	D,V		✓	✓
	e. Description of model updating	V	✓	✓	n.a.
11. Risk groups	Details of how risk groups were created	D,V	✓	✓	✓
12. Development vs. validation	For validation, description of differences between development and validation data	V		✓	✓
Results					
13. Participants	**a.** Flow of participants through the study	D,V		✓	✓
	b. Description of characteristics of participants	D,V		✓	✓
	c. For validation, comparison with development data	V		✓	✓
14. Model development	**a.** Number of participants and outcome in each analysis	D		✓	✓
	b. Unadjusted association between each candidate predictor and outcome	D	✓	✓	✓
15. Model specification	**a.** Presentation of full prediction model	D	✓	✓	✓
	b. Explanation of how to use the prediction model	D		✓	✓
16. Model performance	Report of model performance measures	D,V		✓	✓
17. Model updating	Results from any model updating	V	✓	✓	n.a.
Discussion					
18. Limitations	Limitations	D,V		✓	✓
19. Interpretation	**a.** For validation, interpretation of performance measure results	V			✓
	b. Overall interpretation of results	D,V		✓	✓
20. Implications	Potential clinical use of the model and implications for future research	D,V		✓	✓
Other information				✓	✓
21. Supplementary information	Availability of supplementary resources	D,V		✓	✓
22. Funding	Source of funding and role of funders	D,V		✓	✓
	Total number of applicable items for TRIPOD adherence score			**31**	**37**

(n.a) No included studies reported external validation only or model updating (Item 10c, 10e, and 17)

RESULTS

We identified 24 814 unique articles, of which we sampled ten random sets of 249 articles each with sampling replacement for screening. We screened the title and abstracts of 2 482 articles, screened full text of 312 articles and included 152 eligible articles (Figure 1).

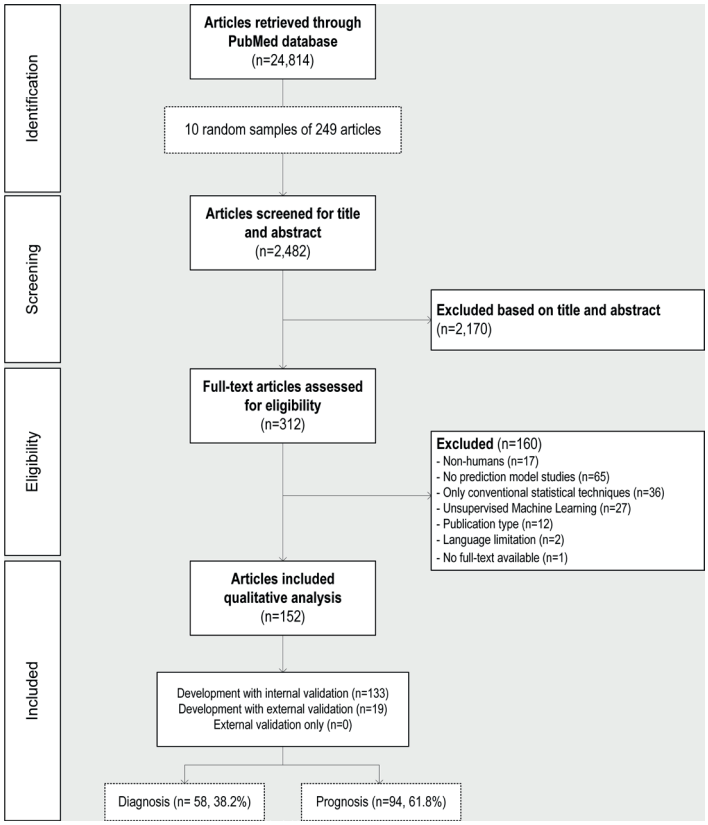


Figure 1. Flowchart of included studies

We included 94 (61.8%) prognostic and 58 (38.2%) diagnostic prediction model studies. 132 (86.8%) articles described development with internal validation and 19 (12.5%) development with external validation (same model). One (0.6%) article was development with external validation (different model) and was included as a development with internal validation study in the present analysis. Prediction models were developed most often in oncology (21/152 [13.8%]). Detailed description of the included studies is provided in supplemental material.

Across the 152 studies, 1429 models were developed and 219 were validated, with a range of 1 to 156 for both types of studies. The most commonly used ML

techniques for the first reported model were Classification and Regression Tree (CART [10.1%]), Support Vector Machine (SVM [9.4%]) and Random Forest (RF [9.4%]). Alongside ML techniques, 19.5% of studies reported the development of a model using conventional statistical techniques, such as logistic regression. Five out of 152 studies (3.3%, 95% CI 1.4% to 7.5%) stated following the recommendations of the TRIPOD Statement.

Overall adherence per TRIPOD item

Five TRIPOD items reached at least 75% adherence (background, objectives, source of data, limitations, and interpretation), whilst 12 TRIPOD items were below 25% adherence (Figure 2). Results for the overall adherence per TRIPOD item stratified by study type, diagnosis and prognosis, and publication year are shown in Table 2.

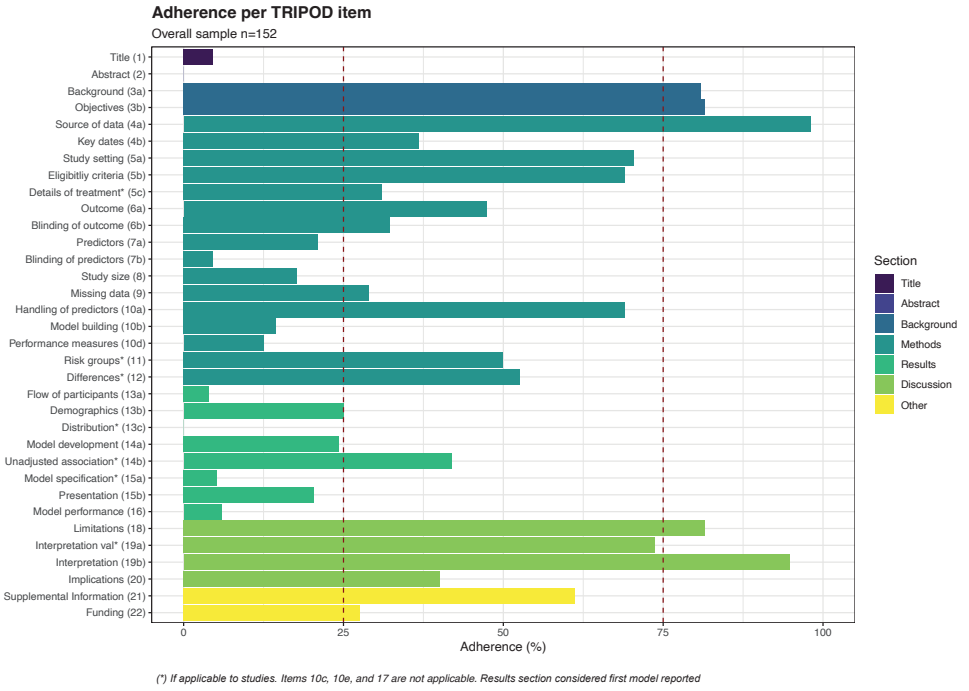


Figure 2. Overall adherence per TRIPOD item

Title and abstract (item 1 and 2)

Seven out of 152 studies (4.6%, 95% CI 2.2 to 9.2) completely adhered to title recommendations. Description of type of prediction model study (sub-item 1.i) was poorly reported (11.2%, CI 7 to 17.2), but outcome to be predicted (sub-item 1.iv) was well reported (91.4%, CI 85.9 to 94.9). No study fully reported item 2, abstract (0%, CI 0%to 2.5).

Introduction (item 3)

Background and objectives were most often reported TRIPOD items. Background was provided in 123 studies (80.9%, 95% CI 73.9 to 86.4), and the objectives were reported in 124 studies (81.6%, CI 74.6 to 86.9).

Methods (item 4-12)

Source of data was the most often reported item in the methods section, and across all TRIPOD items (98%, 95% CI 94.4 to 99.3). Study setting was reported in 107 studies (70.4%, CI 62.7 to 77.1), eligibility criteria in 105 (69.1%, CI 61.3 to 75.9), and handling of predictors in 105 out of 152 studies (69.1%, CI 61.3 to 75.9). Ten studies assessed risk groups and five reported complete information (50%, CI 23.7 to 76.3). Differences between development and validation set were reported in 10 out of 19 applicable studies (52.6%, CI 31.7 to 72.7). For 72 studies, definition of outcome was reported (47.4%, CI 39.6% to 55.3). Key study dates such as start and end date of accrual, and length of follow-up were completely reported in 56 studies (36.8%, CI 29.6 to 44.7). Details of treatment were reported in 36 out of applicable 116 studies (31%, CI 23.3 to 39.9). Blinding of outcome and predictors were reported in 49 (32.2%, CI 25.3 to 40) and 7 studies (4.6%, CI 2.2 to 9.2), respectively.

Forty-four studies reported how missing data were handled (28.9%, 95% CI 22.3 to 36.6). The missing data item consists of four sub-items of which three were rarely addressed in included studies. Within 28 studies that reported handling of missing data: three studies reported the software used (10.7%, CI 3.7 to 27.2), four studies reported the variables included in the procedure (14.3%, CI 5.7 to 31.5) and no study reported the number of imputations (0%, CI 0 to 39). Predictor definitions were given in 32 out of 152 studies (21.1%, CI 15.3 to 28.2), and justification of study size was reported in 27 studies (17.8%, CI 12.5 to 24.6). Model building procedures, such as predictor selection and internal validation, were reported in 22 out of 152 studies (14.5%, CI 9.8 to 20.9). Internal validation, a sub-item of item 10b, was one of the most reported sub-items across studies (91.4%, CI 85.9 to 94.9).

Reporting of measures used to assess and quantify the predictive performance was complete in 19 studies (12.5%, 95% CI 8.2 to 18.7). Though 106 studies (69.7%, CI 62 to 76.5) reported discrimination (sub-item 10d.i), only 19 studies (12.5%, CI 8.2 to 18.7) reported calibration (sub-item 10d.ii). Definitions of discrimination and calibration are stated in supplemental material. Other performance measures (sub-item 10d.iii), for example sensitivity, specificity, or predictive values, were reported in 124 studies (81.6%, CI 74.7 to 86.9).

Results (item 13-17)

Study participant characteristics were reported in 38 out of 152 studies (25.0%, 95% CI 18.8 to 32.4). Basic demographics, at least age and gender (sub-item 13b.i), were

provided in 117 studies (77.0%, CI 69.7 to 83), while summary information of the predictors (sub-item 13b.ii) was reported in 67 studies (44.1%, CI 36.4 to 52). Number of study participants with missing data for predictors (sub-item 13b.iii) was reported in 15 studies (24.2%, CI 15.2 to 36.2). Unadjusted associations were reported in 41 out of the 74 studies that reported regression-based models alongside with ML-models (41.9%, CI 31.3 to 53.3). The number of participants and events were described in 37 studies (24.3%, CI 18.2 to 31.7). In 31 out of 152 studies, an explanation on how to use the developed model to make predictions for new individuals was provided, often in the form of a scoring rule or online calculator (20.4%, CI 14.8 to 27.5). Flow of participants was reported in 6 studies (3.9%, CI 1.8 to 8.3) and model specification was reported in 6 out of 116 applicable studies (5.2%, CI 2.4 to 10.8). Model predictive performance was completely reported in 9 out of 152 studies (5.9%, CI 3.1 to 10.9).

Discussion (items 18-20)

Overall interpretation of results was reported in 124/152 studies (81.6%, 95% CI 74.7 to 86.9). Limitations of the study were reported in 144 studies (94.7%, CI 90 to 97.3). An interpretation of model performance in the validation set in comparison with the development set was given in 14/19 studies (73.7%, CI 51.2 to 88.2). Potential clinical use and implications for future research was reported in 61 studies (40.1%, CI 32.7 to 48.1).

Other information (items 21 and 22)

Availability of supplementary resources was mentioned in 93/152 studies (61.2, 95% CI 53.3 to 68.6). Funding information was reported in 42 studies (27.6%, CI 21.1 to 35.2).

Overall adherence per article

Overall adherence of studies to items of the TRIPOD Statement ranged between 13% and 65%; median adherence was 38.7% (IQR 31 to 46.5). The completeness reporting in prognostic model studies was higher (median adherence=40% (IQR 33.3 to 46.8)) than diagnostic model studies (median adherence=35.7% (IQR 30.2 to 45)) (Figure 3). Moreover, median adherence was 40.6% (CI 28.6 to 46.1) in development (with internal validation) studies, compared to 37.9% (CI 31 to 46.4) in development with external validation studies.

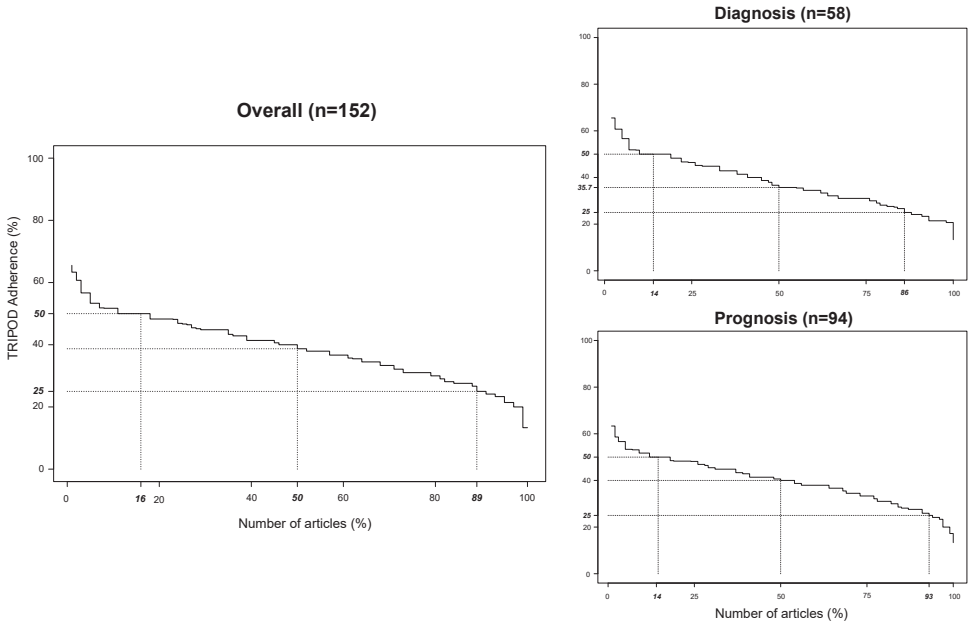


Figure 3. Overall adherence per article

DISCUSSION

We conducted a systematic review of ML-based diagnostic and prognostic prediction model studies and assessed their adherence to the TRIPOD Statement. We found that ML-based prediction model studies adhere poorly to the reporting items of the TRIPOD Statement.

Complete reporting in titles and abstracts is crucial to identify and screen articles. However, titles and abstracts were fully reported in less than 5% of articles. In addition, information about methods was infrequently reported. Complete and accurate reporting of the methods used to develop or validate a prediction model facilitates external validation, as well as replication of study results by independent researchers. For example, to enhance transparency and risk of bias assessment, it is recommended to report the number of participants with missing data and report how missing data were handled in the analysis. Handling of missing data was seldom reported, but this may be partially explained by the fact that some ML techniques can handle missing data by design (e.g. sparsity aware splitting in XGBoost and surrogate splits in decision trees).[24,25] Also most studies divided a single dataset into three: training, validation and test set; the last is used for internal validation. The split sample approach for internal validation was among the most reported sub-items in our sample, but several methodological studies and guidelines have long discouraged this approach.[26]

Overall, most articles adhered to less than half of the applicable items considered essential for complete reporting. Authors may have avoided reporting specific details about methods and results because their objective may be to explore the data and modeling technique accuracy, rather than build models for individualized predictions in "real world" clinical settings. However, high-quality reporting is also essential for reproducibility and replication. Also, most developed models were unavailable for replication, assessment, or clinical application. Only five studies referred to the TRIPOD Statement for reporting their research. Although TRIPOD was published and disseminated in 2015, it is infrequently used for reporting of ML-based prediction model studies.

We stratified studies by type (diagnosis vs prognosis), aim (development vs development with external validation), and year (2018 vs 2019). We included diagnostic model studies developed with deep learning if they used images in combination with demographic and clinical variables. Often, these studies use several numerical variables based on pixels or voxels and build prediction models based on multiple layers of statistical interaction. Both topics are challenging to report due to number of variables used and poor interpretability of interactions. This may explain why diagnostic ML-based model studies were slightly worse reported compared to prognostic studies in our sample. However, we did not observe clear

differences across stratified groups as most confidence intervals overlapped.

Previous systematic reviews have shown poor reporting of regression-based prediction model studies.[7,8,10,11] One study assessed the completeness of reporting in articles published in high impact journals during 2014 within 37 different clinical fields. In 146 prediction model studies, over half of TRIPOD items were not fully reported, obtaining an overall adherence of 44% (IQR 35 to 52%). Although authors excluded models using machine learning, the review found poor reporting of the title, abstract, model building, model specification and model performance, similar to our study.[7] In a sample of prediction model studies published in general medicine journals with the top 7 highest impact factor, the overall reporting adherence was 74% before, and 76% after the implementation of the TRIPOD Statement. Authors included only prediction models developed with regression techniques but also found poor reporting of model building, specification, and performance.[11] A recent study assessed the completeness of reporting of deep learning-based diagnostic model studies. Although they developed their own data extraction for reporting quality, authors found poor reporting of demographics, distribution of disease severity, patient flow, and distribution of alternative diagnosis.[27] These items were also inappropriately reported in our study with a median adherence between 0% and 47.3%. Another systematic review that assessed studies comparing the performance of diagnostic deep learning algorithms for medical imaging versus expert clinicians reported the overall adherence to TRIPOD was poor with a median of 62% (IQR 45 to 69%).[28] In line with our results, a study about the performance of ML models showed that 68% of included articles had unclear reporting.[12]

To our knowledge, this is the first systematic review evaluating the completeness of reporting of supervised ML-based prediction model studies in a broad sample of articles. We ran a validated search strategy and performed paired screening. We also used a contemporary sample of studies in our review (2018-2019). Though some eligible articles may have been missed, it is unlikely they would change the conclusions of this review. We used a systematic scoring-system enhancing the objectivity and consistency for the evaluation of adherence to a reporting guideline.[22] We used the formal TRIPOD adherence form and checklist for data extraction and assessment; however, these were developed for studies developing prediction models with regression techniques. Although we applied the option 'not applicable' for items that were unrelated to ML and items were excluded when calculating overall adherence, our results should be interpreted within this context.

While some items and sub-items may be less relevant for prediction models developed with ML techniques, other items are more relevant for transparent reporting in these studies. For example, source of data (4a), study size (8), missing data (9), transformation of predictors (10a.i), internal validation (10b.iv), and availability

of the model (15b) acquire new relevance within the context of ML-based prediction model studies. As ML techniques are prone to overfitting, we recommend extending item 10b of the TRIPOD adherence form to include a new sub-item specifically related to penalization or shrinkage techniques. New reporting items such as the hardware (i.e. technical aspects) that was used to develop or validate an algorithm in images studies are needed, as well as data clustering. New practices such as explaining models through feature importance plot or tuning of hyper-parameters could be also added to the extension of TRIPOD for ML-based prediction models. Items such as testing of interaction terms (Item 10b-iv), unadjusted associations (14b), and regression coefficients (15a) require updating. Despite these recommendations, most TRIPOD items and sub-items are still applicable for both, regression and ML techniques and should be used to improve reporting quality.

We identified nearly 25 000 articles with prediction and ML-related terms within 2 years, similar to previous systematic reviews about deep learning models.[29,30] The literature has become saturated with ML-based studies; thus, their identification, reporting and assessment becomes even more relevant. If studies are presented without essential details to make predictions in new patients, subsequent researchers will develop a new model, rather than validating or updating an existing model. Reporting guidelines aim to increase the transparent evaluation, replication, and translation of research into clinical practice.[31] Some reporting guidelines for ML clinical prediction models have already been developed.[32,33] However, these guidelines are limited and do not follow the EQUATOR recommendations for developing consensus-based reporting guidelines.[34] The improvement in reporting after the introduction of a guideline has shown to be slow.[31] We acknowledge that the machine learning community developing predictive algorithm for healthcare might be unaware of the TRIPOD Statement. Improving the completeness of reporting of ML-based studies might be even more challenging given the number of techniques and associated details that need to be reported. There are also practical issues, like terminology used, word limits, or journal requirements, that are acting as barriers to complete reporting. To overcome these barriers, the use of online repositories for data, script, and complete pipeline could help researchers share their models with enough details to make predictions in new patients and to allow external validation of the model. Further journal endorsement, training, and tailored guidelines might be required to improve the completeness of reporting. Our results will provide input and support for the development of TRIPOD-AI, an initiative launched in 2019.[17,18] We call for a collaborative effort between algorithm developers, researchers, and journal editors to improve the adoption of good scientific practices related to reporting quality.

CONCLUSION

ML-based prediction model studies currently do not adhere well to the TRIPOD reporting guideline. More than half of the TRIPOD items considered essential for transparent reporting were inadequately reported, especially regarding details of title, abstract, blinding, model building procedures, model specifications and model performance. Whilst ML brings new challenges to the development of tailored reporting guidelines, our study serves as a baseline measure to define future updates or extensions of TRIPOD tailored to ML modelling strategies.

Acknowledgements
The authors would like to thank and acknowledge the support of René Spijker, information specialist.

Authors' contributions
The study concept and design were conceived by CLAN, JAAD, PD, LH, RDR, GSC, and KGMM. CLAN, JAAD, TT, SN, PD, JM, and RB conducted article screening and data extraction. CLAN performed data analysis and JAAD verified the underlying data. CLAN wrote the first draft of this manuscript, which was critically revised for important intellectual content by all authors who have provided the final approval of this version. CLAN, the corresponding author, is the guarantor of the review. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding
This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. GSC is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is supported by the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health.

Competing interest
GSC, RDR and KGMM are members of the TRIPOD Group. All authors have nothing to disclose.

Availability of data and materials
The study protocol is available at doi: 10.1136/bmjopen-2020-038832. The search strategy is available in supplemental material; detailed extracted data are available upon reasonable request to the corresponding author.

Ethical approval and consent to participate
Not required.

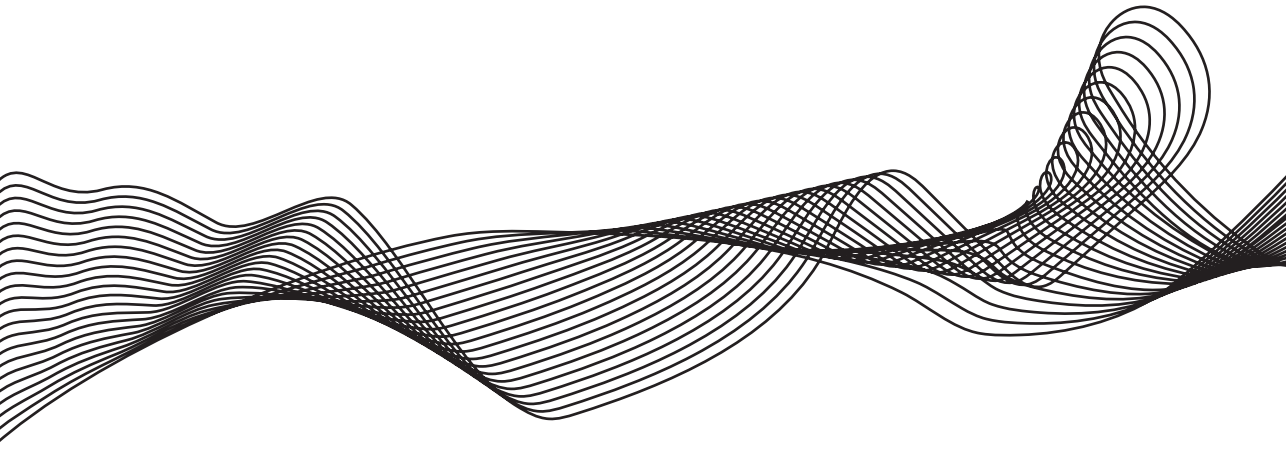
Consent for publication
Not applicable.

Supplementary data
The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01469-6

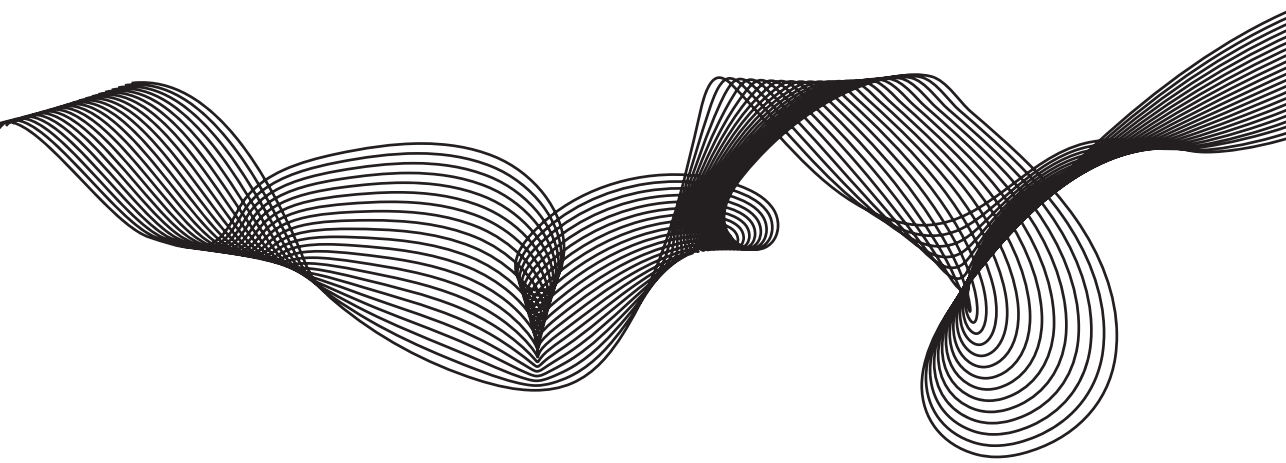
REFERENCES

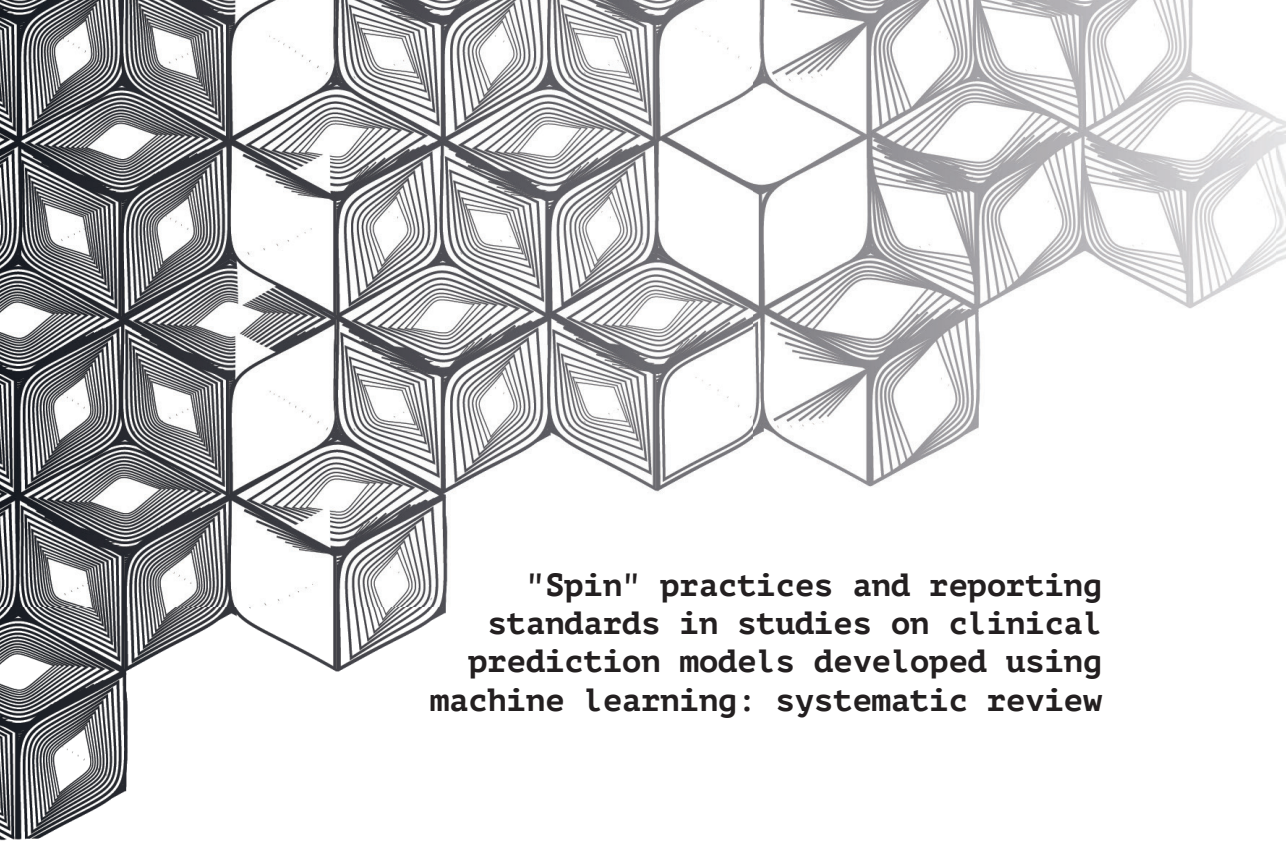
1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ*. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
2. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10(2). doi:10.1371/journal.pmed.1001381
3. Riley, Richard D; van der Windt, Danielle; Croft, Peter; Moons KGM. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. Oxford University Press; 2019. doi:10.1093/med/9780198796619.001.0001
4. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/bmj.i2416
5. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189
6. Mitchell T. *Machine Learning*. McGraw Hill; 1997.
7. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. *BMC Med*. 2018;16(1):1-12. doi:10.1186/s12916-018-1099-2
8. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Med*. 2012;9(5). doi:10.1371/journal.pmed.1001221
9. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC Med*. 2011;9. doi:10.1186/1741-7015-9-103
10. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
11. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
12. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
13. Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol*. 2020;122:95-107. doi:10.1016/j.jclinepi.2020.03.005
14. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X
15. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55. doi:10.7326/M14-0697
17. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Published online 2019. doi:10.1016/S01406736(19)302351
18. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*.

2021;11(7):e048008. doi:10.1136/BMJOPEN-2020-048008
19. Andaur Navarro CL, Damen JAAG, Takada T, 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(11):1-6. doi:10.1136/bmjopen-2020-038832
20. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7). doi:10.1371/journal.pmed.1000097
21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4
22. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9(4). doi:10.1136/bmjopen-2018-025611
23. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
24. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
25. Therneau TM, Atkinson EJ. *An Introduction to Recursive Partitioning Using the RPART Routines*.; 1997.
26. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26(2):796-808. doi:10.1177/0962280214558972
27. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*. 2020;10(3):e034568. doi:10.1136/bmjopen-2019-034568
28. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368. doi:10.1136/bmj.m689
29. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. *Transl Vis Sci Technol*. 2020;9(2):7-7. doi:10.1167/tvst.9.2.7
30. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2
31. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Med*. 2010;8(1):24. doi:10.1186/1741-7015-8-24
32. 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(12). doi:10.2196/jmir.5870
33. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
34. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7(2). doi:10.1371/journal.pmed.1000217



Chapter 4





"Spin" practices and reporting standards in studies on clinical prediction models developed using machine learning: systematic review

Constanza L Andaur Navarro
Johanna AA Damen
Toshihido Takada
Steven WJ Nijman
Paula Dhiman
Jie Ma
Gary S Collins
Ram Bajpai
Richard D Riley
Karl GM Moons
Lotty Hooft

Published in Journal of Clinical Epidemiology

ABSTRACT

Background. The misuse of words when describing scientific results, also known as spin, is a well-stablished and abundant phenomenon in biomedical literature, as demonstrated for randomized therapeutic intervention, diagnostic test accuracy, and prognostic factor studies. Practices such as overinterpretation of study results and linguistic spin might also occur in studies on prediction models, e.g., due to suboptimal reporting and methodological conduct. We evaluated the presence and frequency of spin practices and poor reporting standards in studies that developed and/or validated clinical prediction models using supervised machine learning techniques.

Methods. We systematically searched PubMed from 01-2018 to 12-2019 to identify diagnostic and prognostic prediction model studies using supervised machine learning. No restrictions were placed on data source, outcome, or clinical specialty. To identify spin in prediction model studies, we modified items collected in a previous study about spin practices in prognostic factor studies. We focused on incomplete reporting, the use of linguistic spin to inflate study results, and inappropriate claims of clinical applicability. We estimated the frequency of each item with 95% confidence intervals (CIs), supplemented by a narrative summary.

Results. We included 152 studies: 38% reported diagnostic models and 62% prognostic models. Most studies reported model development only (n=133, 87.5% [95% CI 81.3 - 91.8]). The most frequent item suggesting a risk for spin was the absence of a study protocol (n=150, 98.7% [95% CI 95.3-99.6]). Most studies did not report calibration of the developed or validated prediction model, either in the abstract (n=150, 98.7% [95% CI 95.3-99.6]) nor in the main text (n=134, 88.2%, [95% CI 82.1 - 92.4]). When reported, discrimination was described without precision estimates in 53/71 abstracts (74.6%, [95% CI 63.4 - 83.3]) and 53/81 main texts (65.4%, [95% CI 54.6 - 74.9]). Of the 21 abstracts that recommended the model to be used in daily practice, 20 (95.2% [95% CI 77.3 - 99.8]) lacked any external validation of the developed models. Likewise, 74/133 (55.6% [95% CI 47.2 - 63.8]) studies made recommendations for clinical use in their main text without any external validation. Reporting guidelines for prediction model studies were cited in 13/152 (8.6% [95% CI 5.1 - 14.1]) studies.

Conclusion. Spin practices and poor reporting standards are also present in studies on prediction models using machine learning techniques. Establishing a framework for a rigorous identification of spin practices in prediction model studies will enhance adequate, transparent, and sound reporting of prediction model studies and ultimately promote the implementation of reliable prediction models in medical practice, regardless of the modelling approach.

Systematic review registration: PROSPERO, CRD42019161764

INTRODUCTION

To facilitate transparent and complete reporting of study methodology and findings, reporting guidelines are available to authors of biomedical research. However, there is still room for authors to frame or emphasize a particular interpretation of study findings.[1,2] The misuse of language, intentionally or unintentionally, affects the interpretation of study findings and has been described as 'spin'.[3–8] Spin has also been referred to as the discordance between study results and conclusion, or overextrapolation.[9] Spin is prevalent in biomedical literature and evidence shows that it can have an impact on reader's interpretation and decision-making.[9,10] Inaccurate reporting and misinterpretation of study findings might have consequences on research dissemination and public trust on scientific findings.

Prediction models in healthcare generally use individual data to estimate the probability of the presence of an existing disorder (i.e. a diagnostic model) or of the occurrence of a future outcome (i.e. a prognostic models).[11] To benefit patients and healthcare providers in clinical practice, studies on clinical prediction models should be conducted following the best available methodological evidence and reported in a transparent and complete manner. However, studies on prediction models are often developed using inappropriate methods and are incompletely reported.[12–14] Most published research on prediction models is never used in daily medical practice, contributing largely to research waste.[15]

In recent years, supervised machine learning has gained considerable attention as a flexible suite of data analytic methods for predictions in healthcare.[16] Neural networks, random forest, and support vector machines are some examples.[17] Nonetheless, studies using machine learning techniques are often questioned about their true effectiveness within the clinical workflow.[18,19] The pressure to publish and the intense commercialization agenda may contribute to the exaggeration of the real benefit of machine learning-based prediction models, while underplaying the costs, risks, and limitations. Whether a study applied regression or machine learning techniques, the use of spin to describe model development and validation could provide a false impression of the real performance of the model, thus hampering their further independent validation and transportation to daily healthcare settings.

'Spin' or overinterpreted scientific findings are a well-established phenomenon in randomized therapeutic intervention trials, observational studies, biomarker studies, diagnostic test accuracy studies, prognostic factor studies, and systematic reviews, however, its form and frequency in prediction model studies is unknown.[4,5,20–22] We conducted a systematic review to estimate the frequency of spin practices and reporting standards that might play a role in how the findings of a study are interpreted in studies on prediction models developed using supervised machine learning across clinical domains.

METHODS

For the reporting of this study, we adhered to the PRISMA 2020 statement.[23]

Literature search

We aimed to identify primary studies describing the development or validation of prediction models using supervised machine learning techniques across all medical fields published between 1 January 2018 to 31 December 2019. Hence, we searched in PubMed on 19 December 2019 using a comprehensive search strategy that is provided as supplemental file 1.

Eligibility criteria

We included studies if they met any of the following criteria: 1) Described the development or validation of one or more multivariable prediction models using any supervised machine learning technique aiming for individualized predictions; or 2) Reported on the incremental value or model extension aiming to develop a prediction model. A multivariable prediction model was defined as a model aiming to predict a health outcome by using two or more predictor variables. For this study, we considered a study to be an instance of supervised machine learning when reported any statistical learning technique, except when reporting only models that were strictly regression-based, regardless of whether authors referred to them as machine learning.

We excluded studies if they 1) Investigated a single predictor, test or biomarker, or its causality with an outcome; 2) Used machine learning to enhance the reading of images or signals; 3) Used as predictors only genetic traits or molecular markers; 4) Reported on systematic reviews, conference abstracts, or tutorials. The search was restricted to human subjects, English-language articles, and articles available via our institution. Further details about eligibility criteria have been described in the study protocol.[24]

Literature selection

Two independent reviewers screened all titles and abstracts in parallel. One reviewer (CLAN) screened all studies while the second reviewer came from a group of six (TT, SWJN, PD, JM, RB, JAAD). Full-text reading of selected articles was performed by one reviewer (CLAN) in combination with one of the other six reviewers (TT, SWJN, PD, JM, RB, JAAD). In case of disagreement, a third author (JAAD) was involved.

Data extraction

We defined 'spin practice' as any issue that could make the clinical usefulness of the developed or validated prediction model look more favorable than the study design and results can underpin.[4] A previous article about spin in prognostic factor studies

already identified several practices which we modified for our data extraction.[20] We also added spin practices identified in other study designs, as well as practices to reduce research waste (e.g., presence of/reference to a study protocol, references to previous evidence) which we grouped under 'poor reporting standards'.[5,6,9,15,25] For detailed description of extracted items, see supplemental file 2.

Data extraction was performed in duplicate; one independent reviewer (CLAN) extracted all articles, and the second extraction was carried out by randomly allocating articles to each of the other six reviewers (TT, SWJN, PD, JM, RB, JAAD). We examined spin practices in Abstract and Main text separately, and across sections (Title, Introduction, Results, and Discussion). Discrepancies were discussed between reviewers until agreement was reached.

Our extraction form also included the general characteristics of each study: aim of the study, type of publication (diagnosis vs prognosis), year of publication (2018 vs 2019), journal name, clinical specialty, funding source, disclosure of authors' conflicts of interest (COIs), and mentioning of the TRIPOD Statement.[1,2] The extraction form was pilot-tested in five articles and implemented using Research Data Capture (REDCap).[26]

Synthesis of results

We estimated the frequency of the extracted items with 95% confidence intervals (CIs). We present results separately for Abstract, Main text, and across section of the manuscript. Furthermore, we followed the scheme presented in previous articles wherein spin practices are classified within on these three strategies: misleading reporting (i.e., incomplete, and selective reporting), misleading interpretation (i.e., unreliable statistical analysis, linguistic spin), and misleading extrapolation (i.e., ignoring uncertainty, claiming irrelevant clinical applicability).[20,25] We summarized results using descriptive statistics alongside a narrative summary and visual plot. Analyses were carried out using R version 4.0.3 (R Core Team, 2020).

Ethical approval. This study was performed on published studies, thus ethical approval is not required.

RESULTS

After our search, we retrieved 24 814 articles. Given time and resources constraints, we randomly sampled 2482 (10%) studies for screening. After screening, 312 studies were reviewed in full text. A total of 152 studies were found eligible and included in the final analysis. A flowchart of the screening process is provided in Figure 1.

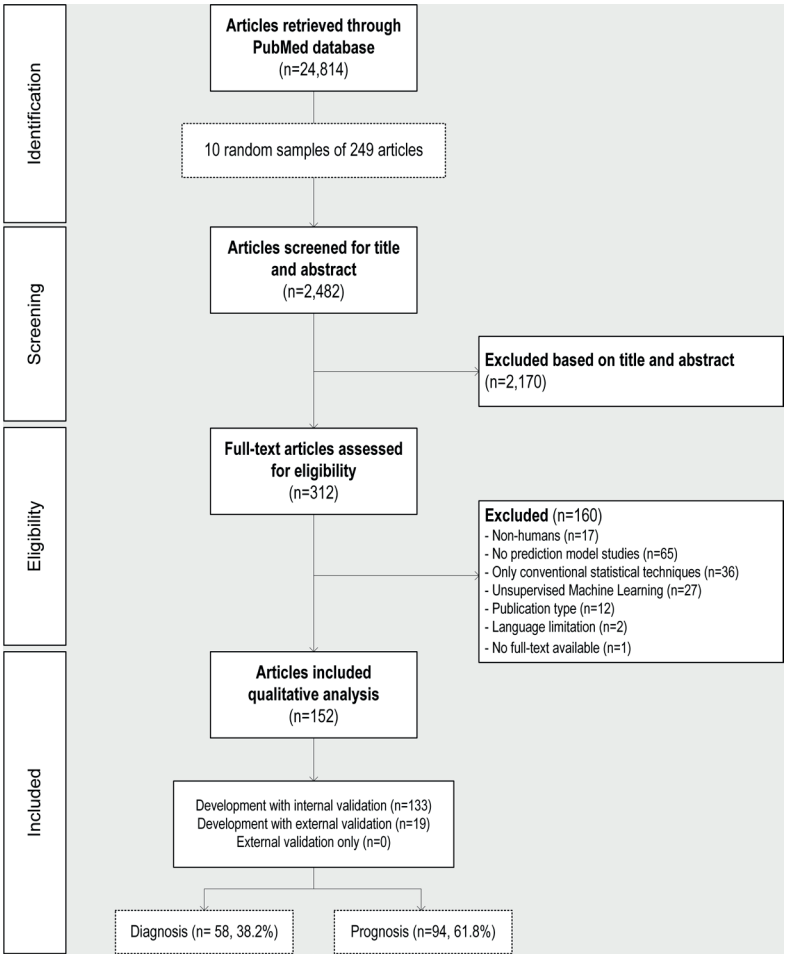


Figure 1. Flowchart of included articles

General characteristics of included studies

Of the 152 articles, 94 (61.8% [95% CI 53.9 - 69.2]) focused on prognostic models and 58 (38.2% [95% CI 30.8 – 46.1]) on diagnostic models. Most studies reported the development of prediction models including internal validation (n=133/152, 87.5% [95%CI 81.3 - 91.8]), and 19 (12.5% [95% CI 8.2 - 18.7]) performed external validation.

The clinical specialties with the most publications were oncology (n=21/152, 14% [95% CI 9.2 - 20.2]), surgery (n=20/152, 14% [95% CI 8.7 - 19.5]), and neurology (n=20/152, 14% [95% CI 8.7 - 19.5]). Table 1 shows the characteristics of the included articles. For details on the included articles, see supplemental file 3.

Table 1. General characteristics of included articles (n=152)

	n (%) [95% CI]
Study type	
Diagnosis	58 (38.2) [30.8-46.1]
Prognosis	94 (61.8) [53.9-69.2]
Study aim	
Development only	133 (87.5) [81.3-91.8]
Development with external validation	19 (12.5) [8.2-18.7]
Clinical specialty	
Oncology	21 (14) [9.2-20.2]
Surgery	20 (14) [8.7-19.5]
Neurology	20 (14) [8.7-19.5]
Origin*	
Europe	37 (24.3) [18.2-31.7]
North America	59 (38.8) [31.4-46.7]
Asia	46 (30.3) [23.5-38]
Other (Oceania, Latin America)	5 (3.3) [1.4-7.5]
Unclear/Not reported	8 (5.3) [2.7-10]
Affiliation with clinical department**	
Yes	85 (55.9) [48-63.6]
No	67 (44.1) [36.4-52]
Conflict of interest	
Yes	20 (13.2) [8.7-19.5]
No	102 (67.1) [59.3-74.1]
Not reported	30 (19.7) [14.2-26.8]
Funding source	
Profit	3 (2) [0.7-5.6]
Non-profit	92 (60.5) [52.6-67.9]
Both	4 (2.6) [1-6.6]
Unclear	8 (5.3) [2.7-10]
Not reported	45 (29.6) [22.9-37.3]
Reference to reporting guidelines	
Yes	13 (8.6) [5.1-14.1]
No	139 (91.4) [85.9-94.9]

Abbreviations: CI, confidence interval
*Three studies originated in more than one continent.
**Reported affiliation of first author.

Most articles originated in North America (n=59/152, 38.8% [95% CI 31.4 - 46.7]) and the first author was often affiliated to a clinical department (n=85/152 (55.9% [95% CI 48 - 63.6]). Source of funding was often reported (n=107/152, 70.4% [95% CI 62.7 - 77.1]), in which 92/107 (86% [95% CI 78.2 - 91.3]) were supported by non-profit organizations. Moreover, 122/152 (80.3% [95% CI 73.2 - 85.8]) studies were published in journals containing a section for conflicts of interest (COI) but only 20/122 (16.4% [95% CI 10.9 - 24]) studies reported at least one COI. Reporting guidelines were cited in 13/152 (8.6% [95% CI 5.1 - 14.1]) studies. Of these 13 studies, 8 (61.5% [95% CI 35.5 - 82.3]) mentioned TRIPOD[1,2], 3 (23.1% [95% CI 8.2 - 50.3])

STROBE[27], 2 (15.4% [95% CI 4.3 - 42.2]) STARD[28], and 2 (15.4% [95% CI 4.3 - 42.2]) the *Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research*.[29]

In total, we evaluated 19 practices in Abstract (including Title) and 26 in Main text (Table 2). The most frequent practice was the absence of a study protocol (n=150/152, 98.7% [95% CI 95.3 - 99.6]). We found a median of 8 (IQR 7 to 9) practices in the abstract, as well as in the main text (IQR 7 to 10) (Figure 2).

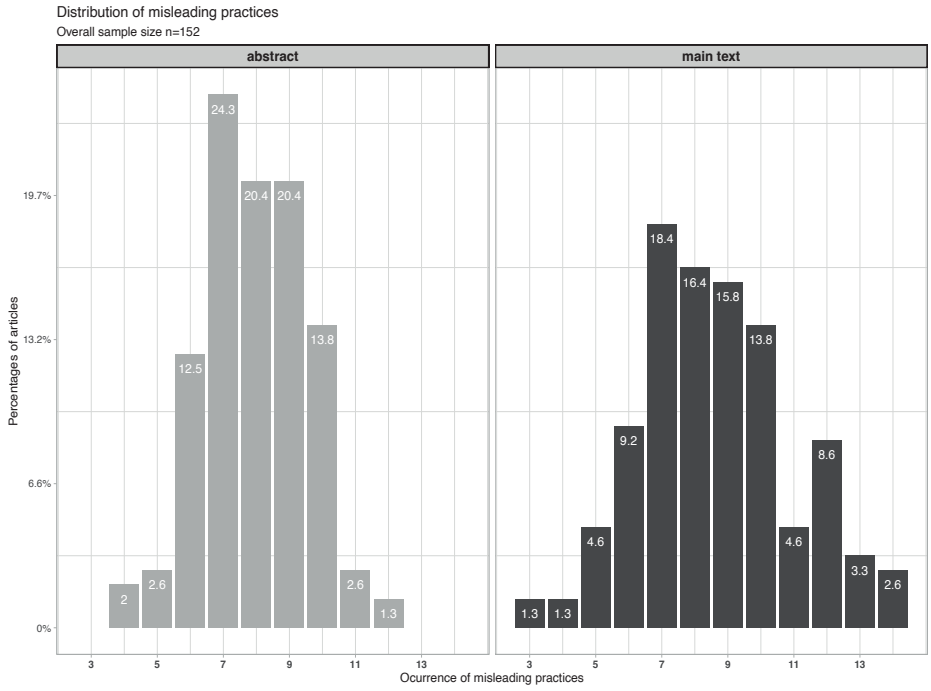


Figure 2. Distribution of misleading practices

Misleading reporting

We classified 13 practices as misleading reporting, of which five were assessed in the Abstract and eight in the Main text (Table 2). Most included studies had four to five misleading reporting practices (n=118/152, 77.6% [95% CI 70.4 - 83.5]) in the Abstract. In 43/152 (28.3% [95% CI 21.7 - 35.9]) abstracts, the term 'machine learning' was used rather than the specific term for the algorithm (i.e., support vector machine, k-nearest neighbor). While 81/152 (53.3% [95% CI 21.7 - 35.9]) abstracts did not report discrimination, 150/152 (98.7% [95% CI 21.7 - 35.9]) abstracts reported

no calibration measure. In 146/152 (96.1% [95% CI 21.7 - 35.9]) abstracts, no study limitation was mentioned. Likewise, 138/152 (90.8% [95% CI 21.7 - 35.9]) abstracts did not mention the availability of previous prediction models.

Similarly, most studies had at least four misleading reporting practices (n=64/152, 42.1% [95% CI 34.5 - 50.1) in the Main text. We identified 7/152 (4.6% [95% CI 2.2 - 9.2]) studies that reported different performance measures in Methods compared to Results. In 68/152 (44.7% [95% CI 37.1 - 52.7]) studies, authors did not provide rationale to support the use of machine learning to address the research question. Similarly, 29/152 (19.1% [95% CI 13.6 - 26.1]) studies ignored models developed previously. Almost all studies did not provide or reference to a study protocol (n=150/152, 98.7% [95% CI 95.3 - 99.6]). Further details can be found in Table 2.

Misleading interpretation

We classified 21 practices as misleading interpretation, of which eight were identified in the Abstract and 13 in the Main text (Table 2). Of the 152 studies, 48/152 (31.6% [95% CI 24.7 - 39.3]) had two misleading interpretation practices in the Abstract. Out of the 71 abstracts that reported discrimination measures, 53 (74.6% [95% CI 63.4 - 83.3]) described them without precision estimates. Strong statements to describe model performance were found in 62/152 (40.8% [95% CI 33.3 - 48.7]) abstracts and 40 (26.3% [95% CI 20 - 33.8]) used at least one leading word. In 38/152 (25% [95% CI 18.8 - 32.4]) abstracts, authors emphasize model relevance while results were not predictive.

Most studies had three to four misleading interpretation practices across sections in the Main text (50/152, 32.9%). When reported, discrimination was presented without precision estimates in 53/101 (52.5% [95% CI 42.8 - 61.9]) studies. Likewise, calibration lacked precision estimates in 7/18 (38.9%) studies. In 59/152 (38.9% [95% CI 20.3 - 61.4]) studies, we identified strong statements to describe model performance. Further details can be found in Table 2.

Misleading extrapolation

We classified 11 practices as misleading extrapolation, of which six were assessed in Abstract and five in Main text (Table 2). Across abstracts, recommendation to use the model in clinical practice was provided in 21 studies, however, 20 (95.2% [95% CI 77.3 - 99.8]) of them lacked any form of external validation despite small sample size. Likewise, recommendation to use the model in different setting of population was given in nine studies, of which all lacked external validation in the same study.

In the main text, 86/152 (56.6% [95% CI 48.6 - 64.2]) studies made recommendations to use the model in clinical practice, however, 74/86 (86% [95% CI 77.2 - 91.8]) lacked external validation in the same article. Out of the 13/152 (8.6%

[95% CI 5.1 - 14.1]) studies that recommended the use of the model in a different setting or population, 11/13 (84.6% [95% CI 57.8 - 95.7]) studies lacked external validation. Finally, qualifiers (such as "very", "may") were used frequently to describe findings in the Main text (n=64/152, 42.1% [95% CI 34.5 - 50.1]). Further details can be found in Table 2.

Extent of spin practices across sections

Most articles contained no spin practice in title (n=132/152, 86.8% [95% CI 80.5 - 91.3]), three spin practices in Results (n=61/152, 40.1% [95% CI 32.7 - 48.1]) and three in Discussion (n=61/152, 40.1% [95% CI 32.7 - 48.1]). Regarding the Main text, articles contained two spin practices in Results (n=48/152, 31.6% [95% CI 24.7 - 39.3]), four in the Discussion (n=36/152, 23.7% [95% CI 17.6 - 31]), and one in another section (n=69/152, 45.4% [95% CI 37.7 - 53.3]). We showed the extent of occurrence of spin per sections in Figure 3.

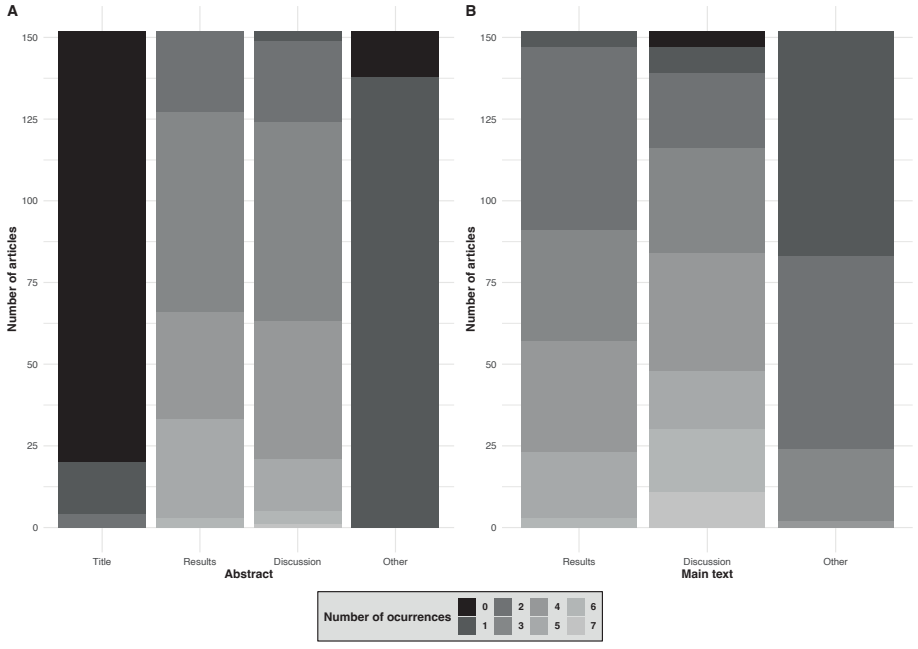


Figure 3. Extent of occurrences per section

Table 2. Frequency of 'spin' practices and poor reporting standards in Title, Abstract, and Main Text

		No. (%) [95% CI of percentage]	
		Abstract (n=152)	**Main text (n=152)**
Misleading reporting			
Results section			
Machine learning techniques used are unreported	Poor reporting standard	43 (28.3) [21.7-35.9]	NE
Differences between performance measures pre-specified in Methods and reported in Results section	Spin	NA	7 (4.6) [2.2-9.2]
Discrimination is not reported	Poor reporting standard	81 (53.3) [45.4-61]	51 (33.6) [26.5-41.4]
Calibration is not reported	Poor reporting standard	150 (98.7) [95.3-99.6]	134 (88.2) [82.1-92.4]
Discussion and Conclusion section			
Limitations are not reported	Poor reporting standards	146 (96.1) [91.7-98.2]	28 (18.4) [13.1-25.3]
Other sections			
Rationale to use machine learning techniques to address the objective in introduction is unavailable	Poor reporting standard	NA	68 (44.7) [37.1-52.7]
No references to existing models	Poor reporting standard	138 (90.8) [85.1-94.4]	29 (19.1) [13.6-26.1]
Main results are reported as supplemental file	Poor reporting standard	NA	14 (9.2) [5.6-14.9]
The study protocol is unavailable	Poor reporting standard	NA	150 (98.7) [95.3-99.6]
Misleading interpretation			
Title			
Title is inconsistent with the study results	Spin	6 (3.9) [1.8-8.3]	NA
Use of leading words	Spin	18 (11.8) [7.6-17.9]	NA
Novel		3 (2) [0.7-5.6]	
Excellent		0	
Accurate		1 (0.7) [0-3.6]	
Optimal		0	
Perfect		0	
Significant		0	
Improved		7 (4.6) [2.2-9.2]	
Other[a]		7 (4.6) [2.2-9.2]	
Results section			
Discrimination is reported without precision estimates	Poor reporting standards	53 (74.6) [63.4-83.3]*	53 (52.5) [42.8-61.9]*
Calibration is reported without precision estimates	Poor reporting standards	2 (100) [34.2-100]*	7 (38.9) [20.3-61.4]*
Use of strong statements to describe the model and/or model performance / accuracy / effectiveness	Spin	62 (40.8) [33.3-48.7]	59 (38.8) [31.4-46.7]
Use of leading words	Spin	40 (26.3) [20-33.8]	59 (38.8) [31.4-46.7]
Novel		4 (2.6) [1-6.6]	0
Excellent		1 (0.7) [0-3.6]	5 (3.3) [1.4-7.5]
Accurate		13 (8.6) [5.1-14.1]	7 (4.6) [2.2-9.2]
Optimal		3 (2) [0.7-5.6]	0
Perfect		0	1 (0.7) [0-3.6]
Significant		10 (6.6) [3.6-11.7]	30 (19.7) [14.2-26.8]
Promising		6 (3.9) [1.8-8.3]	0
Improved		4 (2.6) [1-6.6]	4 (2.6) [1-6.6]
Outperform		4 (2.6) [1-6.6]	12 (7.9) [4.6-13.3]
Other[b]		26 (17.1) [11.9-23.9]	14 (9.2) [5.6-14.9]

Spin in tables or figures		NA	10 (6.6) [3.6-11.7]
Discussion and Conclusion section			
Use of strong statements to describe model and/or model performance / accuracy / effectiveness	Spin	NA	64 (42.1) [34.5-50.1]
Use of leading words	Spin	NA	64 (42.1) [34.5-50.1]
Novel			3 (2) [0.7-5.6]
Excellent			8 (5.3) [2.7-10]
Accurate			10 (6.6) [3.6-11.7]
Optimal			2 (1.3) [0.4-4.7]
Perfect			1 (0.7) [0-3.6]
Significant			15 (9.9) [6.1-15.6]
Superior			7 (4.6) [2.2-9.2]
Outperform			4 (2.6) [1-6.6]
Other[c]			14 (9.2) [5.6-14.9]
Invalid comparison of results to previous development and/or validation studies is given	Spin	NA	99 (65.1) [57.3-72.2]
The comparison in favour of similar prediction models			52 (52.5) [42.8-62.1]*
Some outcomes in favour and not in favour for other			22 (22.2) [15.2-31.4]*
Unclear			11 (11.1) [6.3-18.8]*
Non relevant models are not discussed	Spin	NA	28 (23) [16.4-31.2]*
Authors make use of leading words to reject those non relevant models	Spin		10 (38.5) [22.4-57.5]*
Emphasis on model relevance while results are not predictive	Spin	38 (25) [18.8-32.4]	NE
Discrepancy between full-text and abstract explanation of the study findings	Spin	7 (4.6) [2.2-9.2]	NA
Misleading extrapolation			
Discussion and Conclusion section			
Recommendation to use the model in clinical practice without external validation in same study	Spin	20 (95.2) [77.3-99.8]*	74 (86) [77.2-91.8]*
Recommendation to use the model in different setting or population without external validation in same study	Spin	9 (100) [70.1-100]*	11 (84.6) [57.8-95.7]*
No recommendation for further studies	Poor reporting standard	15 (9.9) [6.1-15.6]	38 (25) [18.8-32.4]
Qualifiers are used	Spin	50 (32.9) [25.9-40.7]	64 (42.1) [34.5-50.1]
Other benefits not prespecified in Methods are addressed	Spin	NA	10 (6.6) [3.6-11.7]
Conclusions are inconsistent with the reported study results	Spin	23 (16.4) [11.2-23.4]*	NE
Conclusion focuses solely on significant results	Spin	85 (55.9) [48-63.6]	NE

Abbreviations: CI, confidence interval; NA, not applicable; NE, not extracted.
*Valid percentage: with respect to the articles which reported the information.
[a] Predictive, well-calibrated, promote, intelligent, outperform, improved.
[b] Best, efficient, superior, satisfactory, greater, substantial, well, effective.
[c] Remarkable, substantial, better, robust, satisfied, superior, huge.

DISCUSSION

We systematically assessed how often spin practices occurred in 152 prediction model studies using supervised machine learning. Our study revealed that spin (or the potential for spin/over or misinterpretation) was widely present in studies on prediction models developed using supervised machine learning.

Principal findings

We found that the occurrence of spin practices was similar between Abstract and Main Text. The most frequent poor reporting standard was the absence of a predefined protocol or registration. Moreover, the use of reporting guidelines was scarce. Although infrequent, we also observed a few discrepancies between Methods and Results in the Main text, as well as discrepancies between Abstract and Main text conclusions. However, a protocol and the use of reporting guidelines could reduce selective and incomplete reporting. TRIPOD, the reporting guideline for studies on prediction models, also includes a version for proper reporting of Abstracts.[1,2,30]

We found that studies often made inappropriate recommendations to use the prediction model in daily clinical practice, ignored limitations, and reported performance measures without precision estimates in the Abstract. Previous research on non-randomized studies has identified that abstracts are the most frequent section with spin.[25] As primary source of dissemination, the content of an abstract must be accurate and useful, not only for evidence users but also for the general audience. Furthermore, research shows that the main factor associated with spin in a press release was the presence of spin in the abstract.[8] Spin in abstracts could partially be explained by the limited word count and the need to attract potential readers, however, any recommendation in concluding statements in an abstract should be consistent with the study design, findings, and limitations to avoid misleading the readers, especially those who can only access to this information.

We further noticed that a considerable number of studies neither report their limitations nor their findings within the context of previously developed models. Given the high number of developed models already available in the biomedical literature, researchers should focus on carry-out systematic reviews and validating most promising models to avoid further research waste. [12,15]

Strengths and limitations

To our knowledge, there has been no systematic review about spin practices and reporting standards in prediction model studies and, particularly not in studies on machine learning based prediction models. We appraised a sample of articles covering a wide range of outcomes and clinical domains. In addition, we evaluated spin in the Title, Abstract, and across several sections of the Abstract and Main text.

However, several limitations are worth highlighting. In our study, we modified the pre-existing tool used in prognostic factor studies as such, but faced certain challenges during data extraction. Although this tool enabled us to capture several practices, it failed to identify aspects particularly related to prediction model studies (i.e., selection of predictors, categorization of continuous predictors, threshold definition). Furthermore, we focused on the use of leading words (i.e., linguistic spin) rather than to allow certain degree of rhetoric and evaluate it within its specific context. Similarly, we could not determine if the use of qualifiers was detrimental because we only counted the occurrence rather than to evaluate its use to show uncertainty. The appraisal of spin practices relied mostly on the subjective judgement of reviewers; thus, it is possible that others will interpret the authors' statements differently as we did, especially the linguistic spin. Although we reduced interpretation bias by resolving any discrepancies through discussion, reviewers were not blinded to authors, funding source, or journal. Likewise, this appraisal depended on what was reported on articles thus, some of our findings might be the consequence of poor reporting quality rather than misleading practices.

As we did not cover the full range of potential spin practices, our findings should be interpreted bearing this in mind. Despite these limitations, we still provided exploratory evidence about the presence of spin in prediction model studies.

Comparison with other studies

A systematic review including 35 publications assessing misleading practices showed that spin evaluation varies per study design.[9] Unfortunately, no study assessing spin practices in prediction model studies was found by this review. Within prognostic research, studies on prognostic factors in oncology frequently overinterpret their findings hampering clinical applicability.[20]

Unanswered questions, recommendations, and future research

There are several obstacles to ensure an accurate interpretation and dissemination of research. The reward system within academia and the increasing amount of published research makes spin in research to some extent necessary and therefore, more frequent. As authors, we naturally want our studies to be published and will consciously and subconsciously use language to increase credibility and readability of our findings. For example, authors reporting exploratory analysis such as studies describing model development only and not any form of evaluation, might allow themselves to overinterpret and extrapolate their results, as it might not be expected to become available in daily clinical practice. But a growing concern is that spin in primary studies is linked to inappropriate reporting of press releases and news media.[8,31] The reach of spin in biomedical research therefore also extends to general audiences, potentially biasing behaviour and jeopardizing public trust. Authors

should make every effort to avoid distortion and hype, and should focus on overall quality, transparency, and further research.

Spin is prevalent in all biomedical literature and therefore, further evaluation of spin practices and reporting standards will benefit those who rely upon biomedical research findings and evidence. However, to some extent, it requires subjective judgement. There is a need to develop an instrument or classification scheme with clear definitions tailored to evaluate spin practices in studies on prediction models. Likewise, further guidance and interventions on how to write study findings could also be helpful.[32] This can guide junior researchers, peer-reviewers, and journal editors to be cautious on how study findings are written, while achieving transparency, accuracy, and conciseness. Similarly, readers should be aware of practices that can mislead their interpretation of findings before deploying models into daily healthcare settings. Spin practices and its association with methodological quality and risks of bias still needs to be systematically assessed to provide evidence of its effect on overall quality of biomedical evidence. A severity scale for spin in prediction models still needs to be developed.

CONCLUSION

Authors have several opportunities to frame the impression their findings will produce in readers. We provide a description of the existence of spin practices and poor reporting standards in studies on prediction models and we indicate the need for strategies to improve how study results are portrayed to increase prediction model validations and uptake in daily clinical practice.

Acknowledgment
The authors would like to acknowledge René Spijker for his assistance in developing the search strategy.

Author Contributions
Constanza L. Andaur Navarro: Conceptualization, Methodology, Investigation, Data Curation, Formal analysis, Writing - original draft, Writing - review & editing; Johanna A.A. Damen: Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision; Toshihiko Takada: Investigation, Writing - review & editing; Steven WJ Nijman: Investigation, Writing - review & editing; Paula Dhiman: Conceptualization, Methodology, Investigation, Writing - review & editing; Jie Ma: Investigation, Writing - review & editing; Gary S Collins: Conceptualization, Methodology, Writing - review & editing; Ram Bajpai: Investigation, Writing - review & editing; Richard D Riley: Conceptualization, Methodology, Writing - review & editing; Karel GM Moons: Conceptualization, Methodology, Writing - review & editing, Supervision; Lotty Hooft: Conceptualization, Methodology, Writing - review & editing, Supervision.

Support
There is no specific funding to disclosure for this study. GSC is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is funded by the NIHR Oxford BRC. RB is affiliated to the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) West Midlands. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Competing interests
Authors declare no competing interests.

Registration and protocol
This review was registered in PROSPERO (CRD42019161764). The study protocol can be accessed in doi:10.1136/bmjopen-2020-038832.

Supplementary data
Supplemental file 1. Search strategy
https://surfdrive.surf.nl/files/index.php/s/MIAt0F0V0SJ4tw2

Supplemental file 2. Table S1. Detailed description of extracted items
https://surfdrive.surf.nl/files/index.php/s/h6iO3HK3V49m9Qv

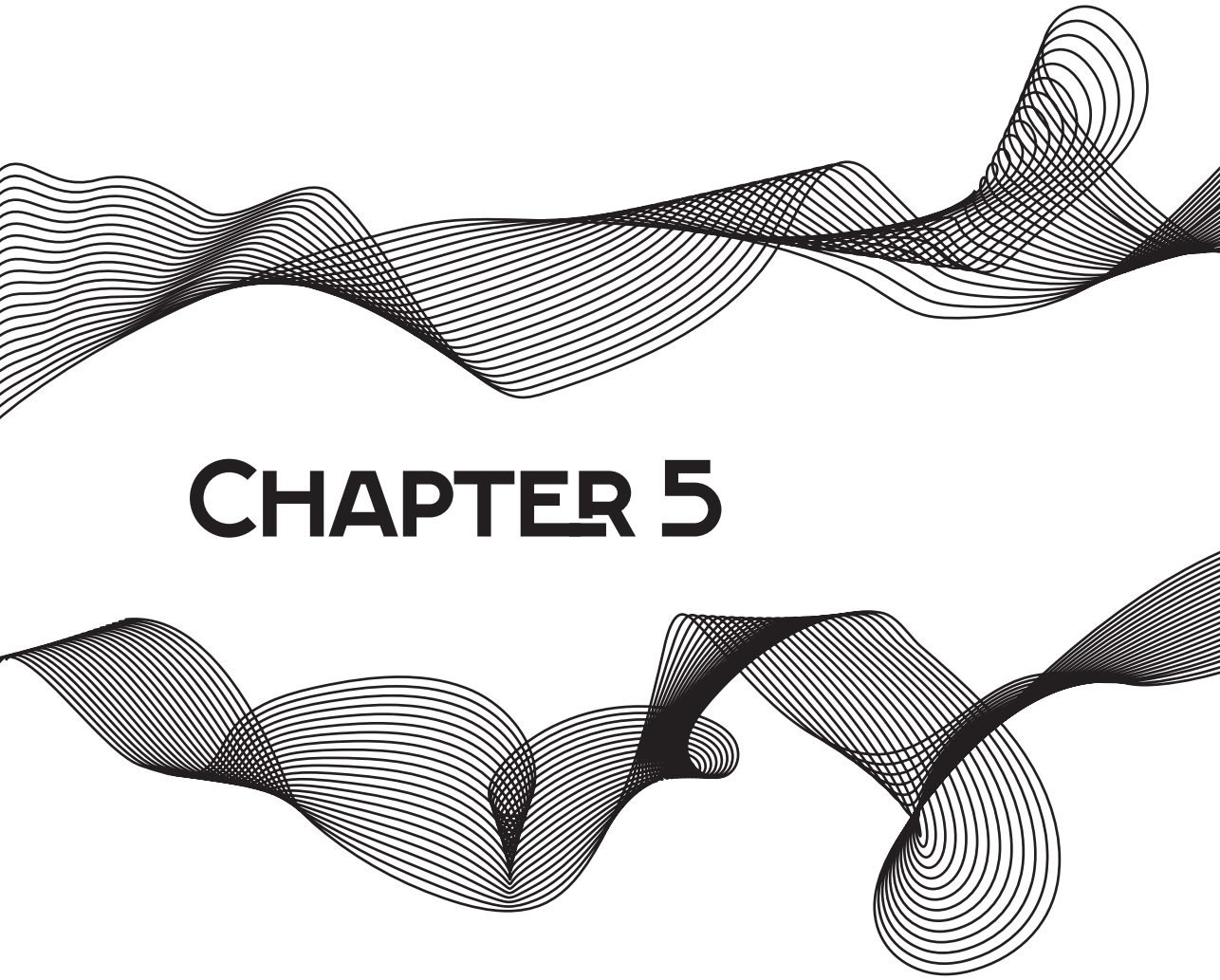
Supplemental file 3. Detailed characteristics and citations of included studies
https://surfdrive.surf.nl/files/index.php/s/zVgoirIFkjrMtco

Availability of data, code, and other materials
Data and analytical code are available upon reasonable request to corresponding author.

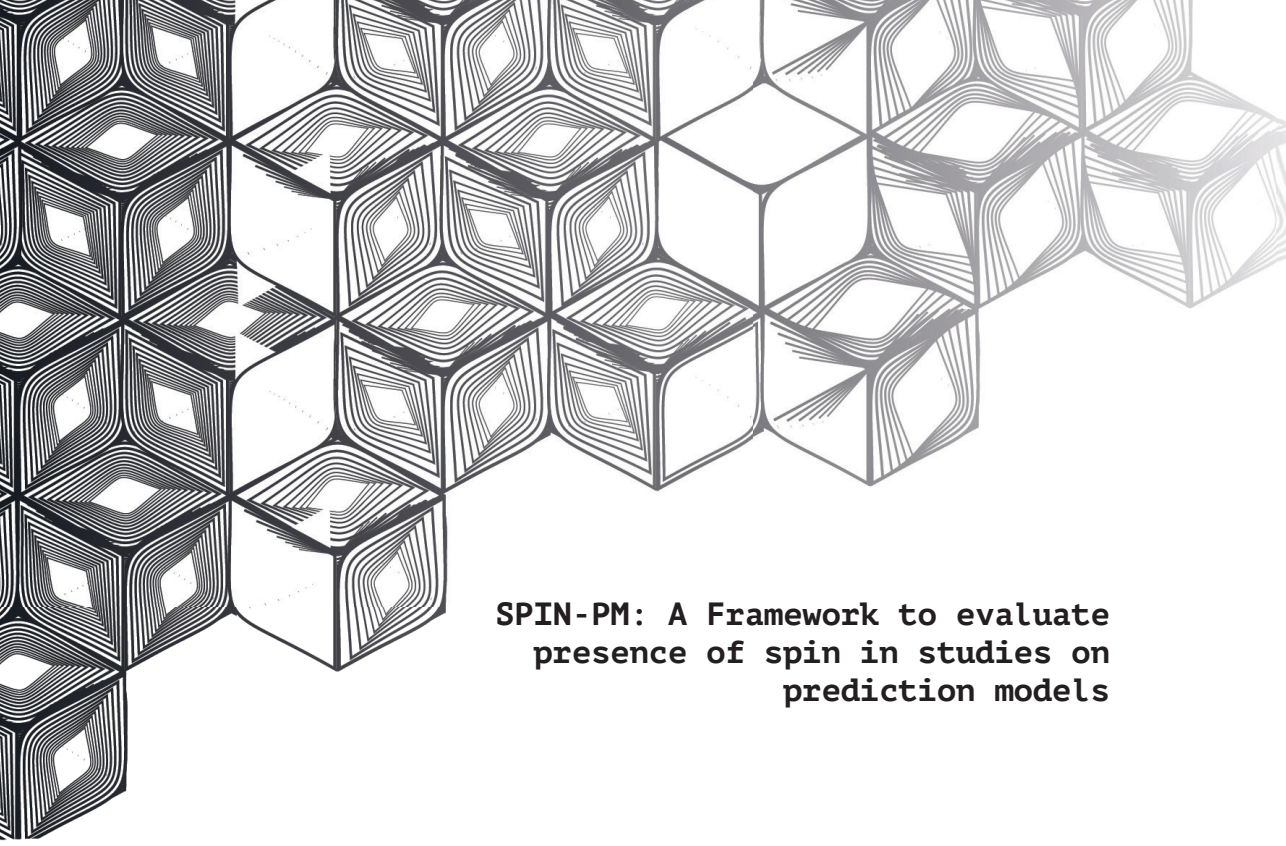
REFERENCES

1. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73. doi:10.7326/M14-0698
2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55. doi:10.7326/M14-0697
3. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-2619. doi:10.1073/PNAS.1710755115
4. Ghannad M, Olsen M, Boutron I, Bossuyt PM. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. J Clin Epidemiol. 2019;116:9-17. doi:10.1016/j.jclinepi.2019.07.011
5. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. J Clin Epidemiol. 2016;77:44-51. doi:10.1016/j.jclinepi.2016.04.012
6. Yavchitz A, Ravaud P, Altman DG, et al. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. J Clin Epidemiol. 2016;75:56-65. doi:10.1016/j.jclinepi.2016.01.020
7. Boutron I, Haneef R, Yavchitz A, et al. Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. BMC Med. 2019;17(1):1-10. doi:10.1186/s12916-019-1330-9
8. Yavchitz A, Boutron I, Bafeta A, et al. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. PLoS Med. 2012;9(9). doi:10.1371/journal.pmed.1001308
9. Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biol. 2017;15(9):1-16. doi:10.1371/journal.pbio.2002173
10. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: The SPIIN randomized controlled trial. J Clin Oncol. 2014;32(36):4120-4126. doi:10.1200/JCO.2014.56.7503
11. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? BMJ. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
12. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. BMJ. 2016;353. doi:10.1136/BMJ.I2416
13. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40. doi:10.1186/1471-2288-14-40
14. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369. doi:10.1136/bmj.m1328
15. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X
16. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189
17. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1). doi:10.1186/S12911-019-1004-8

18. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:1-12. doi:10.1136/bmj.l6927
19. Morley J, Floridi L, Goldacre B. The poor performance of apps assessing skin cancer risk. BMJ. 2020;368. doi:10.1136/bmj.m428
20. Kempf E, de Beyer JA, Cook J, et al. Overinterpretation and misreporting of prognostic factor studies in oncology: a systematic review. Br J Cancer. 2018;119(10):1288-1296. doi:10.1038/s41416-018-0305-5
21. Haneef R, Lazarus C, Ravaud P, Yavchitz A, Boutron I. Interpretation of results of studies evaluating an intervention highlighted in google health news: A cross-sectional study of news. PLoS One. 2015;10(10):1-15. doi:10.1371/journal.pone.0140889
22. McGrath TA, Bowdridge JC, Prager R, et al. Overinterpretation of Research Findings: Evaluation of "Spin" in Systematic Reviews of Diagnostic Accuracy Studies in High-Impact Factor Journals. Clin Chem. 2020;66(7):915-924. doi:10.1093/clinchem/hvaa093
23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71
24. Andaur Navarro CL, Damen JAAG, Takada T, 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(11):1-6. doi:10.1136/bmjopen-2020-038832
25. Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. BMC Med Res Methodol. 2015;15(1):1-8. doi:10.1186/s12874-015-0079-x
26. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
27. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statevon Elm, Erik, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. 2008. "The Strengthening the Reporting of Observational. J Clin Epidemiol. Published online 2008. doi:10.1016/j.jclinepi.2007.11.008
28. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. BMJ Open. 2016;6(11):1-17. doi:10.1136/bmjopen-2016-012799
29. 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(12). doi:10.2196/jmir.5870
30. Heus P, Reitsma JB, Collins GS, et al. Transparent Reporting of Multivariable Prediction Models in Journal and Conference Abstracts: TRIPOD for Abstracts. Ann Intern Med. 2020;173(1):43. doi:10.7326/M20-0193
31. Adams RC, Challenger A, Bratton L, et al. Claims of causality in health news: A randomised trial. BMC Med. 2019;17(1):1-11. doi:10.1186/s12916-019-1324-7
32. Ghannad M, Yang B, Leeflang M, et al. A randomized trial of an editorial intervention to reduce spin in the abstract ' s conclusion of manuscripts showed no significant effect. J Clin Epidemiol. 2021;130:69-77. doi:10.1016/j.jclinepi.2020.10.014



Chapter 5



SPIN-PM: A Framework to evaluate presence of spin in studies on prediction models

Constanza L Andaur Navarro
Johanna AA Damen
Paula Dhiman
Mona Ghannad
Marteen van Smeden
Johannes B Reitsma
Gary S Collins
Richard D Riley
Karl GM Moons
Lotty Hooft

Manuscript in preparation

ABSTRACT

Systematic reviews have consistently shown that reporting and methodological conduct of clinical prediction model studies could be improved. While some initiatives like the TRIPOD checklist aim to increase the completeness of reporting, authors could still, unconsciously or consciously, mislead readers by being more optimistic about the predictive performance and transportability of their investigated model than the study design, data analysis, and study findings, can justify. This is a well-known and harmful phenomenon in the biomedical literature, usually referred to as 'spin'. In this article, Andaur Navarro and colleagues present SPIN-Prediction Models: a framework to identify and evaluate spin practices and its facilitators in studies on clinical prediction model development and validation. We proposed this guidance aiming to facilitate not only the accurate reporting but also an accurate interpretation and extrapolation of clinical prediction models which will likely improve the quality of subsequent research, as well as reduce research waste.

Key messages

- Overoptimistic reporting and misinterpretation of studies on clinical prediction model has received relatively limited attention in biomedical literature. These practices are known as 'spin' and are a wide phenomenon in other study designs.
- SPIN-Prediction Models is a consensus-based framework that identifies seven spin practices and 14 facilitators of spin to avoid when communicating findings in studies on prediction models. This article provides an overview of spin in prediction modelling research with further explanation, guidance, and accompanying examples.
- SPIN-PM is intended to researchers reporting prediction models for peer reviewed journals as well as for peer reviewers and journal editors assessing these studies up for publication. SPIN-PM is not a scoring tool nor a tool to assess quality of studies.
- We hope this guidance will enhance the overall reporting quality and support the critical appraisal of studies on prediction model and their findings to ensure other researchers can independently validate models, before clinicians and others can implement models safely within healthcare.

INTRODUCTION

Prediction models are often used to complement clinical reasoning and (shared) decision-making by estimating the risk of an individual to have a particular outcome (diagnostic model) or to develop an outcome in the future (prognostic model).[1–4] Well-known examples are the Framingham risk score and colonflag.[5–7] In recent years, the number of studies on prediction models has been increasing strongly, and this growth is expected to continue given the growing popularity of artificial intelligence and machine learning methods, and the increasing availability of larger datasets, such as from routine care. Consequently, there are often several prediction models available for a particular health outcome or target population, however very few are used in routine clinical practice.[8–10]

Shortcomings in the design and statistical analyses found in many studies on prediction models make any of their findings vulnerable to overinterpretation and exaggerated claims on transportability and clinical usefulness.[11–13] Moreover, particular hype and expectation has grown around machine learning-based prediction models.[14–17] With a common rhetoric of virtually endless potential and excellent performance, inadvertent pitfalls in study design and analyses may facilitate an inaccurate interpretation and communication of findings which consequently might lead to the implementation of suboptimal prediction models in clinical practice.

To ultimately improve patient outcomes in practice, studies on prediction models should be conducted following the best methodological evidence available and reported in a transparent and complete manner. Identification and evaluation of misleading practices is necessary, as the realization of the benefits and harms of prediction models could remain limited while the investment of resources increases.

What is spin?

Spin is defined as any (reporting) practice, consciously or unconsciously, that leads to mis- or overinterpretation of the findings of a study, usually emphasizing more favourable findings than the study design, analyses, and results warrant.[18,19] While misinterpretation refers to an inconsistent interpretation of the study findings (i.e., incorrect interpretation), overinterpretation refers to when authors take a strong position stemming from their opinion rather than on the study findings. Both practices are closely linked and contribute to the misrepresentation (i.e., distorted presentation) of scientific findings. Examples of spin practice are the use of exaggerated language, highlighting (only) the findings based on selected subgroups, or choosing a particular statistical analysis to shape the impression of their results to readers.[20]

In biomedical research, the concept of spin was introduced by The BMJ in 1995.[21] Since then, there have been several articles addressing the implications of spin across different study designs and settings, showing its high prevalence and detrimental effect.[20,22–26] Incorrect interpretation of studies and their findings can have serious and undesirable consequences in clinical practice (including potential harm to patients), development of clinical guidelines, health policies, funding of subsequent research, and engagement with general audiences.

Why do we need a spin framework for studies on prediction models?

While spin has been extensively studied for trials on therapeutic intervention, and to a lesser extent in studies on diagnostic test accuracy and prognostic factors, neither the extent of spin nor its implications have yet been addressed for studies on prediction models.[20,22–24,27–32]

Evidence suggests that the nature of spin differs depending on the study design.[25] Unlike randomized trials wherein the most common type on spin can be found on the estimate of the treatment effect, in prediction model studies, it is the interpretation of a prediction model's estimated performance (e.g., discrimination and calibration) where the action of spin may be placed. For example, model's performance might be overinterpret in studies on model development, while studies on model validation, model's performance can be overinterpreted or underinterpreted (e.g., to justify the development of a new model. Spin therefore also needs to be considered in the context of the study type (Box 1). Additionally, studies on prediction models are also not typically designed to answer questions on aetiology, association, or causality, and are rarely registered or have a publicly available protocol.

Given the differences between study designs and statistical analyses, spin frameworks previously developed for randomised trials and studies on diagnostic test accuracy are not directly suitable to identify spin in studies on prediction model. In this article, we present SPIN-Prediction Models: a consensus-based framework to assist readers in identifying and evaluating spin practices, as well as guide authors to accurately communicate the findings in studies on development and validation of prediction models with healthcare application.

Box 1. Studies on clinical prediction models

Development studies

A multivariable model is developed to estimate an outcome probability.[33] This type of study involves a process to produce the model equation or algorithm (e.g. including the identification of the most important predictors, assigning relative weights to each of them, etc.) and estimating the model's predictive performance through calibration, discrimination and potentially clinical utility. Discrimination refers to the measure of how well a prediction model can distinguish individuals with the outcome from those without the outcome (e.g., using Area Under the Receiver Operator Curve, c-statistic).[34] Calibration refers to the measure of agreement between predicted and observed probabilities (e.g., calibration plot, calibration curves, observed:expected ratio).[34]

When evaluated in the same data in which it was developed, the performance of a model (called apparent performance) will often be optimistic due to overfitting, notably when development data sets are relatively small (typically with a small number of outcome events). Development studies will therefore include an internal validation to quantify and correct for any 'optimism' in model performance. [34] Examples of internal validation techniques are cross-validation and bootstrapping.

External validation studies

Validation studies consist of assessing the performance of a prediction model in new individuals whose data were not used during model development.[35] There are several types of external validation, most commonly (1) temporal validation: individuals from the same institution as in the development sample, but in a different (usually later) time; (2) geographical validation: individuals from different institutions or countries to the development sample; and (3) domain or setting validation: for example, individuals from secondary care are used to validate a model developed in primary care. Depending on the findings of the external validation, the prediction model may be recalibrated or updated to better fit the population and setting of interest.

DEVELOPMENT OF SPIN-PM

To identify and evaluate spin in studies on prediction model development and validation, we reviewed key publications on established practices of spin in other study designs to build a preliminary list to be included in the framework.[22,24,25,36]

This preliminary list was first discussed with one researcher with expertise in spin (MG). The revised list was further discussed with a panel of eight researchers with demonstrable expertise in spin (JBR, GSC, KGMM, LH) and prediction model studies (CLAN, JAAD, PD, GSC, JBR, MvS, KGMM, LH).

Researchers were invited to provide comments and contextualize the proposed practices during six group meetings. We held five meetings with the panel of researchers, while the sixth was open to researchers working in one of our affiliated institutions with expertise on prediction model methodology. The concept of the framework was also presented at the annual epidemiological conference in the Netherlands carried out in 2021. We openly discussed clarifications, wording, applicability, and useful examples. After each meeting, the list was adapted by the lead researcher (CLAN), accordingly. Through this iterative consensus process, a final list of spin practices and facilitators was derived after all panel's researchers agreed.

Spin is one of several factors contributing to altering the perception of scientific findings on readers.[19] Here we define spin as mismatch between reported information, namely between the actual design and findings and how they have been interpreted or described by the authors within the manuscript, and that might misdirect the interpretation of readers. A mismatch can occur, for example, when 'positive' words (i.e., useful, effective) are over- or misused to describe study findings while the study design, conduct, data analysis and results do not support such optimistic interpretation. Evaluation of spin, therefore, requires two steps: 1, identify potential spin practices and 2, confirm such mismatch.

In case such mismatch cannot directly be verified, the practice may still constitute a *facilitator of spin* — that is, any reporting practices that interferes with the critical appraisal and requires readers to make assumptions. For example, authors reporting performance measures without stating whether these correspond to apparent or internally validated/optimism-corrected measures in studies on model development. And as previous research has suggested, we also therefore propose to classify misleading practices into two categories: *'actual' spin* and *facilitators of spin*.[20]

Previous classification systems have incorporated selective and incomplete reporting as spin practices.[24,36,37] To objectively identify selective reporting, it is necessary to be able to compare the reported information against a protocol or registration. However, unfortunately studies on prediction models still usually lack public availability of study protocols.[38] In this situation, there is no guarantee that, because for example, datasets, subgroups, or thresholds were mentioned in the methods section, these were indeed pre-specified in any form of registration or protocol. Similarly, incomplete reporting refers to when authors leave out essential information which hinder the objective critical appraisal of a study, its findings, interpretation, and conclusions.

The challenge when reporting studies on prediction models in manuscripts with limited word count is that analyses are usually conducted as a funnel —that is several models may be, for example, developed with different modelling strategies

and even different designs or predicted outcomes, until one model is achieved which is solely reported (usually based on the 'best' performance). One cannot assess whether, for example, missed information on other models is crucial for the critical appraisal of the 'best' performing model which is reported as the only model. Moreover, current adherence of studies on prediction models to reporting guidelines is deficient in two ways. Authors may be unaware of reporting guidelines (e.g., the TRIPOD Statement, www.tripod-statements.org), while those who are, may still report insufficient information.[3,34,39,40] Hence, for studies on prediction model development and validation, we categorized practices related to selective and incomplete reporting as facilitators of spin.

In our framework, we identified two approaches through which authors, can consciously or unconsciously, generate spin in studies on prediction model development and validation:

- Misleading interpretation: Claims with overestimation or underestimation of the performance of the developed or validated prediction model. This can be the consequence of either the application of inappropriate methods or statistical analyses, or the use of overly optimistic language to describe the methods or study findings.[37,41] Accordingly, the readers' perception about the quality and quantity of evidence is unsupported by study design, methods, and analyses reported.
- Misleading transportability: Unjustified claims regarding the applicability or generalizability of the reported prediction models to routine healthcare practice, or even to other (than actually studied) populations, settings, or domains. Without a meaningful external validation, inferences on the actual performance or transportability of a prediction model may be misleading and may cause prediction models to be implemented in settings and populations where there is no robust evidence yet to support this.

INTRODUCING SPIN-PM

The conceptualized framework is presented in Table 1 with examples, which can be applied to assess spin in studies of either model development or validation. We also provide further examples of facilitators of spin in Table 2.

Misleading interpretation

We identified four spin practices through the misleading interpretation approach:

1. *Ignoring risks of optimism in model performance* — Concluding paragraphs in main text do not mention methodological limitations (if present) that might have led to, e.g., an overestimation of a models' predictive performance and thus, limiting its generalizability and applicability. Authors evaluating the performance of their 'own' developed or validated model are likely to be overly optimistic in interpreting their findings. To critically appraisal this practice, scrutiny of the methods, analyses, and reporting is necessary to identify any potential limitations.

2. *Unjustified use of strong affirmative statements to support selected study design and methods* — The manuscript contains statements that unjustifiably emphasize or support selected methods where a rationale for their choice is lacking. Actual limitations of the methods applied in a study may be portrayed as advantages. For example, when authors provide references to support the use of a particular method to correct for class imbalance, that is – when one class outcome outnumbers the other class. However, to correct for class imbalance is inappropriate when developing risk prediction models, unless probabilities are later recalibrated afterwards.[42] Used methods are then presented as more robust or adequate than they might be, thus facilitating misinterpretation of inexperience readers.

3. *Unjustified use of strong affirmative statements to describe the model or the model's performance* — The manuscript contains statements with a tone insinuating a strong model or model performance which is not supported by either the study design, analysis, or findings. Examples include '*clearly shows*', '*strongly recommend*', '*definitely suggest*', and '*very important*'.

4. *Unjustified use of optimistic or positive words to describe the model or model's performance* — Using words to positively describe or embellish the interpretation of study findings without the support of either study design or the actual findings.[43–45] Examples include *'outperformed'*, *'improved'*, *'superior'*, *'better'*, *'novel'*, *'unique'*, etc. This differs from the previous spin practices in the senses that here only 'positive' words are used rather than adverbs to emphasize statements.

Misleading transportability

We identified three spin practices in the misleading transportability approach:

1. *Stating a prediction model can be used in routine medical practice without the need for (further) validation* — Studies reporting only the development of a prediction model, especially if the sample size and notably, the number of outcome events is limited, should ideally avoid using statements suggesting the direct applicability of the investigated model to routine clinical practice in concluding paragraphs.[46] After model development, external validation is needed to determine the accuracy and generalizability of a model using data that was not used for the model development (see box 1).

   Furthermore, studies evaluating in some manner (e.g., empirically or using a decision-modelling strategy) the impact on decision-making are necessary to determine whether prediction models indeed improve decision-making and consequently, health outcomes of the targeted individuals.[35,47] Therefore, conclusions should account for the lack of external validation of the model and thus, of the resulting uncertainty in the estimated model's (apparent or only internally validated) performance and its usefulness.

2. *Stating any lack of clinical applicability/effectiveness based solely on the poor performance in the specific validation sample* — In validation studies, authors evaluate the performance (discrimination and calibration) of a model in data not used for development. If the performance is not as good as the performance in the development study, which is very likely and to be expected, the model may still be useful, as the assessment of usefulness of a model still requires clinical judgment and depends on context, especially if there is no other model available.[35,47,48] Furthermore, if the case-mix in the validation sample differs greatly from that of the development sample, the model may exhibit poor performance but could still have some merit by recalibrating or even by including new predictors, thereby avoiding the potentially unnecessary development of a new prediction model.[1,35]

3. *Stating the use of a prediction model for a different outcome, setting or population without any evaluation* — Studies reporting on a prediction model should avoid extrapolating their findings to a different indication of use (e.g., to predict different outcomes, or use in different settings or populations) than the one stated in the aim of the manuscript. To support any of these type of extrapolation statements, evaluation of the performance of that model for and in these other outcomes, setting or population is required. Any extrapolation should be clearly hypothetically framed within a theoretical context in the discussion and ideally avoided in concluding paragraphs.

Table 1. Proposed framework: spin practices in studies on prediction models

Category of spin	Form of spin	Criteria	Examples
Unreliable statistical analysis	Ignoring the risk of optimism in model performance	Conclusion or concluding paragraph in *main text* omits statements addressing methodological concern during model development and/or validation that might lead to an overfitted model performance. This can occur if: a. Limited study size b. Limited number of events relative to number of candidate predictors c. Inappropriate dichotomization of continuous predictors d. Traditional stepwise predictor selection strategies	*"Our results suggest that machine learning can predict myopia onset in children. We used the available dataset in our hospital, and we obtained high accuracies"* [§] — *The study has used a very restrictive sample of participants enrolled (and with a few events relative to the number of candidate predictors) in only one location. Limited sample size is not address as an important limitation that could had led estimates to be too optimistic.*
Misleading interpretation — Linguistic spin	Using strong affirmative statements to support selected study design and methods	*Main text* or *abstract* contains statements about design and selected methods that could be considered *unjustifiable*. An example is 'Given the lack of imputation methods, we carried out complete case-analysis'	*"Classic methods of dealing with missing data such as complete case analysis, …and multiple imputation can potentially bias the estimates of effect of each variable.(ref) …To avoid losing predictive power, the missing data were imputed using the missForest package."*[49] — The cited references to support the statement recommends the use of multiple imputation and provides no evidence on missForest imputation.
	Using strong affirmative statements to describe the model or the model's performance	Conclusion or concluding paragraph in *main text* or *abstract* contains statements using a tone inferring a strong result. Examples: 'clearly shows', 'strongly recommend', 'definitely suggest', 'very important', 'remarkably greater' etc.	*"Although sensitivity was 100% with and without the new biomarker, within the first case, specificity and accuracy were remarkably greater."* [§] — The study reports changes on specificity and accuracy but the increase is low.
	Using overly optimistic or positive words to describe the model or the model's performance	Statements in *title* and conclusion in *main text* and/or in *abstract* contain terms that positively frame model's predictive performance. Examples: outperformed, improved, superior, better, novel, unique, etc.	*"The predictive models performed excellent in predicting EOC recurrence."*[50] — Although reported AUC is high, the study has several methodological limitations, so it is likely that the reported AUC is optimistic.
Misleading transportability	Stating a prediction model can be used in routine medical practices without the need for (further) validation.	Conclusion or concluding paragraph in *main text* or *abstract* contains statements that claim clinical applicability without stating the need to perform proper validation and/or clinical impact studies.	*"Our finding suggests that random forest model would be best option to implement a system for predicting fatty liver disease patients appropriately and effectively"*[51] — Development only study.
	Stating any lack of clinical applicability/effectiveness based solely on the poor performance in the specific validation sample	Conclusion or concluding paragraph in *main text* or *abstract* contains statements that limits the applicability of the validated model and thus, supporting the developed of a new prediction model.	*"Previous developed models showed low accuracies on our external validation. We therefore developed a better performing model"* [§] — The validation sample contain participants with a different case-mix. Instead of recalibrating or adding new predictors, authors have re-developed the model.
	Stating the use of prediction model in a different outcome, setting or population without any evaluation.	Conclusion in *main text* or *abstract* contains statements that generalize models' performance to different outcome, setting or population without stating the need to perform proper evaluation.	*"This can be extended to predict other type of ailments which arise from metabolic syndrome"*[52] – Development only study.

[§] Fictitious example

Facilitators of spin

We identified 14 facilitators of spin which may contribute to misdirect reader's interpretation and potential transportability. The facilitators are described below, in the order they would usually appear in a manuscript:

1. *Study aim is unclear or not reported* — The aim needs to be clearly defined in the abstract and main text, to facilitate the evaluation of the presence of potential spin practices and of other spin facilitators (see below).

2. *Key details of the dataset are partially or un-reported* — Unclear provenance and details of how the participant data were collected might leave room for unsupported conclusions regarding the performance or generalizability of the prediction model. To judge data representativeness, potential biases on data collection, and practical applicability of the prediction model(s) under study, detailed information on how data were obtained is necessary.

   Moreover, a raising issue in studies on prediction model development and validation is whether models' recommendations are *fair.*[53]Data should be examined regarding the representativeness of demographic groups wherein health inequities might exist (e.g., gender, ethnicity) and for which differences in data collection or outcome definition are known. For a thorough assessment is also necessary to evaluate how these sensitive variables have been recorded.

3. *Citation of the original article that describe the development of the prediction model being validated is missed* — In validation only studies, information on how a prediction model was originally developed is important to place the evaluation of the model performance in context. However, often there are multiple models known by a single name (or acronym) or even multiple models for the same or similar outcome developed by the same authors, or models that have been recalibrated or updated. To critically appraise the validation study, information on how the prediction model was developed is essential.

4. *Inappropriate exclusion of participants from the analysis* — Analyses may have been conducted and reported in a selective group of the entire sample of studied participants, e.g., erroneously restricting to those with complete information on all (candidate or final) predictors or excluding participants that were censored during follow-up. This could potentially impact the representativeness of the sample. Given the potential biases in the final analyses, an honest acknowledgment and discussion of these limitations in the discussion section would be needed to inform readers of the potential poorer performance or generalizability of the investigated model.

5. *Additional complexities in the analysis are ignored*. — The development and

validation of prediction models must use statistical analyses that are appropriate for the study design, intended use, and type of outcome analysed.[4] Some of the complexities researchers might have to deal with in studies on prediction models include: sampling of participants, censoring, competing risks, clustering, and recurrent events per participants. For example, prediction models to predict long-term outcomes in which censoring occurs, a time-to-event analysis (e.g., Cox regression) may be used to include censored participants up to the end of their follow-up. Moreover, in studies with long follow-up, participants might face a different and interfering event (e.g., death) that prevents the outcome of interest to happen. The competing event is often ignored in studies on prediction models but it can lead to an overestimation of predicted risk if these are not accounted for in the time-to-event analysis.[54] It is good practice to address in the discussion the potential impact of ignoring these additional complexities might have on the predictive performance of the developed or validated model.

6. *Inappropriate method for internal validation is used* — Researchers might internally validate their prediction models by splitting their total dataset into development (training) and validation (test) set. Splitting a dataset is problematic unless the dataset is extremely large.[55] Splitting reduces the information used to develop a model, increases the risk of the prediction model being overfitted, and leads to test data that is not large enough for precise internal validation. Ideally, the model should rather be developed on all the available and resampling internal validation techniques, such as cross-validation and bootstrapping applied to assess performance.[4,46]

7. *Reported results are not in accordance with study aim and methods* — The reported findings need to match the aim and methods. It is misleading for readers when the results are not based on any specified methods – design or statistical analysis – which do not (evidently) relate to the aims of the study unless it is specifically stated as post-hoc or other form of extra analysis.

8. *Inaccurate reporting of performance measures in development studies* — The potential for model overfitting should be addressed in studies developing a model. The apparent performance (i.e., performance of the model in the same data as used for development, Box 1) will typically be optimistic and an assessment of this optimism (through internal validation) should be carried out and reported, notably in small study samples. Findings described in the *main text* (or abstract) should be clearly reported if the measure of performance is apparent, or optimism corrected to ensure readers do not overinterpret the predictive ability of the model.

9. *Measures of performance are partially reported* — The predictive performance of a model is typically evaluated using measures of calibration and discrimination

(Box 1). While discrimination is widely reported, calibration is usually not evaluated in studies on prediction models. The lack of information will mislead the reader into thinking a model is 'good' based solely in one measure of performance or uninformative measures such as accuracy or confusion matrices. Calibration plots display the direction and magnitude of any model miscalibration across the entire predicted probability range, providing relevant information for decision-makers in healthcare. Furthermore, to make decision based on thresholds, models are required to be well-calibrated. Thus, both discrimination and calibration should be reported alongside when developing risk prediction models.

10. *Performance measures without confidence intervals are reported* — Alongside performance measures, confidence intervals need to be reported so readers can appreciate the (im-)precision of the estimated measures. A model may demonstrate and be interpreted as having reasonable performance based solely on an estimate of the c-statistic (discrimination) of (for example) 0.75, yet when adding a confidence interval associated with this estimate, let's say goes from 0.55 to 0.95, would alter our interpretation on the precision and reliability of the reported estimates.

11. *Inappropriate presentation of plots* — Any plots should be presented using the appropriate scales on their axes. For example, calibration is preferably presented in a plot, with the estimated risk on the horizontal axis and the observed risk on the vertical axis. The figure should be 'square' as both the x and y-axes are on the same scale. Stretching one of these axes can distort how well a model is (or isn't) calibrated. Similarly, for any plots of receiver operating characteristic curves, with 1-specificity (horizontal axis) against sensitivity (vertical axis) should be presented 'square' as they both have the same scale and are subject to distortion if one axis has been stretched. Another concern are calibration or ROC plots that only show a sub-region of the full plot (e.g., risk predictions in the range 0 to 0.1), to potentially hide poor findings in other regions.

12. *Unsubstantiated claims of superiority of one modelling approach over another are stated* — Authors will often claim superiority of one modelling approach over another. For example, given the increased enthusiasm to apply artificial intelligence or machine learning to develop prediction models in healthcare, it is prudent to avoid unilateral belief in superior predictive performance of one modelling approach over another if it is not supported with appropriate study design and analyses. Unsupported claims can mislead readers who do not have the technical expertise to identify potential methodological limitations and biases in the study. However as with any prediction modelling technique, including machine learning, subsequent validation studies carried out by independent

researchers are needed. Without the proper context and support by empirical data, superiority statements should be avoided across the whole manuscript.

13. *Unfair comparison between models* — Comparing the reported prediction model to previous studies based solely on improvements in predictive performance (e.g., changes in the c-index) can be misleading. Higher predictive performance might not always imply a better model. Models might have been developed using dataset with different distribution of characteristic (including outcome occurrence) or different hyperparameters optimisation in the case on machine learning-based prediction models. When making model comparison in main text, authors need to contextualize the reported performance measures so readers can evaluate to what extent comparison are appropriate. Similarly, we recommend authors of validation studies to provide not only the results of the external validation of previously developed model, but also to report the external validation of the new developed model, if done so.

14. *Unsubstantiated claims of clinical usefulness* — The aim of a study on prediction models is to develop and/or validate a model to be applied in routine healthcare practice. However, measures that address the clinical relevance of a prediction model are often overlooked, while the focus remains only on discrimination and calibration.[56] When authors do not report any measure of clinical usefulness related to the developed or validated prediction model (e.g., using a decision curve analysis), the conclusion can contain unsupported claims regarding their true usefulness in daily practice, as well as hide potential harms. [57,5]

Table 2. Examples of facilitators of spin in studies on prediction models

Category of spin	Facilitators	Criteria	Example
Misleading interpretation	Study aim is unclear or not reported.	Study aim is partially described or unspecified within *abstract* or *main text.*	*"This paper proposes an importance-driven approach to identify key markers/features for the detection of early Parkinson's disease."* [59] — The study later reports findings of a classification model based on key predictors.
	Key details of dataset are partially reported or unreported.	Information about origin and/or collection of data, OR enrolment of participants is not provided within the manuscript, supplementary material, and/or referenced. This can occur when using, for example: a. Online open data repositories b. Biobank data c. Population cohorts d. Well-known randomized controlled trials	*"Patients diagnosed with COVID-19 from March 4th to April 5th from eight large University Hospitals were eligible if they had positive reverse transcription polymerase chain reaction (PCR-RT) and signs of COVID-19 pneumonia on unenhanced chest CT."* [60] — Insufficient description of data sources. Further references are not provided.
	Citation of the original article that describe the development of the prediction model being validated is missed.	Information about the development of the model being validated is not properly provided within the manuscript or referenced. It applies to validation studies only.	*"The CHADS2 score ranging from 0 to 6 was calculated for each patient as congestive heart failure or left ventricular ejection fraction _75 years (1 point); and history of stroke, transient ischaemic attack (TIA), or systemic embolism (2 points)."* [61] — Reference of the original study is not provided.
	Inappropriate exclusion of participants from the analysis.	Discussion and/or Conclusion in *main text* lacks addressing potential risk of using an unrepresentative sample of participants during analysis. Examples where this can occur include: a. Inappropriate handling of missing values b. Excluding participants with incomplete follow-up	*"The results were robust to inclusion of participants with known risk factors for cardiovascular disease…"* [62] —Participants with fatal myocardial infarction were excluded in this study. Participants are thus a lower-risk sample of the original population at risk. Limitations regarding the unrepresentativeness of the sample are not further discussed.
	Additional complexities in the analysis are ignored (if applicable).	Discussion and/or Conclusion in *main text* lacks addressing potential limitation due to unaddressed complexities in the data, such as long-term outcomes, competing risks or clustering. An example is death in elderly patients before a second event of interest (competing risk).	*"The main finding of our study is the high specificity for accidental fall prediction reported in older inpatients."* [63] — The aim is to predict fall risk in older patients, however, the outcome is treated dichotomously (i.e., fallers have one or more falls) while data was collected accounting for all fall events per patient. Potential imitations are not further discussed.
	Inappropriate method for internal validation is used.	Discussion and/or Conclusion in *main text* lacks addressing potential limitations due to splitting the original data to obtain a dataset for internal validation (i.e., testing).	*"We randomly split the original data set into 70% of patients for a training subset and 30% for a testing or validation subset."* [64] — The total sample size was relative small. No further concerns are mentioned in discussion.

	Reported results are not in accordance with study aim and methods.	Statements in Results describe findings based solely on analysis that go beyond the aim and methods reported in *main text*. This can occur when reporting on, for example: a. Prognostic factors b. Unplanned subgroups c. Results based on an analysis that was not pre-planned.	*"To develop and validate a prognostic model applicable to high, middle, low income countries".* [65] — Validation was done using data from high income countries alone. However, this is highlighted later as limitation.
	Inaccurate reporting of performance measures	Results in *main text* are described in tables or text without stated if reported performance is apparent or optimism corrected.	*"Additional prediction metrics (eg, recall and precision) are shown in Table 2."*[66] — Unclear in main text whether reported measures are apparent or optimism corrected. No further details are provided in Table 2 either.
	Measures of performance are partially reported.	Results in *main text* are partially reported. Reporting only discrimination might mislead reader into appraising a model as "good" without knowing if the model provides accurate individual probabilities (calibration). Both calibration and discrimination should be reported when developing risk prediction models.	*"In general, NN-based models show better performance when predicting readmission, except for CHF (where GBM outperforms NN)."*[67] — The aim was to identified patients at high risk of readmission, however, calibration measures are not reported. Judgement of "better" performance is based solely in discrimination ability.
	Performance measures without confidence intervals are reported.	Results in *main text* are reported without confidence intervals to indicate the level of precision of reported performance measures.	*"GLMN outperforms the other MLTs among those implemented with CC analysis, with higher values of all the measures used to compare the algorithms."*[68] — Table 2 does not report confidence intervals for any of the performance measures presented.
	Inappropriate presentation of plots	Receiver operating characteristics curves and calibration plots are presented with their axes squashed or truncated.	*"Fig.5 ROC Curve for classification of Gaussian K-Base NB Classifier"*[69] — Left plot presents the y-axis squashed. Also, there is no clear definition of what plot A nor B represent in the main text.
	Unsubstantiated claims of superiority of one modelling approach over another are stated	Statements in *main text* and/or in *abstract* claim superiority of one modelling approach over another.	*"The performance of most of the ML-based models was significantly better than that of conventional methods using a single clinical feature, Knosp grade, which is commonly used to predict TTS response"*[70]
	Unfair comparison between models	Statements in main text and/or in abstract compare models based solely on their predictive performance ignoring methodological differences, changes on patient's characteristics, or clinical context.	*"In previous research, accuracies of up to 94% were reported. Our model achieved 95% and is therefore better."* [§] — *Comparison is made based on 1% increased on model performance. Remains unknown whether this difference was due to design, methods, or conduct applied to during models' development.*
Misleading transportability	Unsubstantiated claims of clinical usefulness are reported.	Model performance measures to determine relevant threshold to support clinical decisions based on prediction models are not reported in *main text*. Examples are net benefit (NB), decision curve analysis (DCA), net reclassification improvement (NRI).	*"The results demonstrate that the proposed technique is suitable with optimal discrimination … producing accurate, specific and decision oriented rules to facilitate physician and make informed choices about their management and improve health condition."*[52] — Measures to assess clinical usefulness are not reported throughout the manuscript.

[§] Fictitious example

DISCUSSION

Readers of scientific literature go through an interpretative process which is often influenced by how authors have framed the findings of a study. To reduce the chance of over- or undervaluing evidence based on a particular framing alone, we developed SPIN-PM, a consensus-based framework to assist readers to identify and evaluate spin practices, as well as to guide authors into reducing reporting practices that may contribute to the mislead interpretation of studies on prediction model development and validation. We defined two approaches, identified seven spin practices, and provided examples of 14 spin facilitators.

Our proposed spin practices and facilitators are consistent with those previously identified in other study designs, as well as the categorization between actual spin and facilitator of spin.[24,71] We focused our spin on concluding paragraphs of a manuscript, because these are especially susceptible to contain mis- or overinterpreted statements as well as unsubstantiated claims of transportability, as readers tend to focus on them to judge the utility of the study. The interpretation of study findings needs to be frame with the inherent limitations, contextualizing the reported prediction model and its performance. Moreover, we incorporated a practice in an opposite direction, that is – the use of words to downgrade findings on external validation studies to support the development of a new prediction model (Misleading transportability, ii). We found this practice equally detrimental, particularly in validation of prediction models performed by groups of independent researchers. Although extrapolation allows to set the research into 'real-world' context by highlighting the potential final application of the prediction model, we suggest authors to avoid such claims in concluding paragraphs, especially if the study has an explorative rather than applied aim and furthermore, when it is not even supported by their analyses and findings.

Authors may overinterpret their findings because of the current academic reward system, methodological illiteracy, and the prioritization of positive and novel studies by journals and funders.[72–74] On the other hand, readers may find overinterpreted articles as consequence of poor peer-review, publication and citation bias, or lack of related expertise.[74] Several more factors and even unconscious ones, are likely to play a role in the complex system of interpreting and communicating scientific findings. Previous studies have addressed factors such as conflict of interest, industry-based research, authorship, affiliation, and journal's impact factor as potential determinants of spin.[24,25,75] Our proposed framework is the first step towards increasing awareness about spin practices in studies on prediction model development and validation, further updates may explore and incorporate such influences.

A growing concern is that spin in primary studies is linked to inappropriate reporting of press releases and news media.[27,76] The reach of spin in biomedical research therefore also extends to general audiences, potentially biasing behaviour and jeopardizing public trust. Authors should make every effort to avoid distortion and hype, and should focus on overall quality, transparency, and further research.

Spin evaluation requires background knowledge about studies on prediction models and to be weighted in the light of the context at hand. Authors and reviewers are still required to judge how detrimental the spin practice is within the context of their particular research question, study type (model development only, development with external validation, external validation only) or publication type (pre-print, peer-reviewed, conference proceedings). Similarly, they need to determine whether the use of qualifiers (i.e., very, clearly) or 'hedging' (i.e., may, could) relativizes the certainty of a statement based on the findings that have been reported. Reviewers might still disagree regarding the likely effect of certain criteria; thus, comprehensive evaluation of spin practices will remain partially subjective.

Strengths and limitations

We conducted an iterative process consisting of several meetings with methodologists and statisticians with related expertise to construct the framework. However, despite this stepwise process, some practices and facilitators are likely to have been overlooked. Furthermore, our framework does not allow to discern if the spin practice is the result of inexperience, deliberate misconduct, or both. Also, the proposed practices do not determine the optimal degree of proper framing for the communication of prediction models. Instead, we provide guidance on how to identify and avoid practices that may have a detrimental effect on readers' interpretations. We incorporated wide criteria looking to increase the applicability of the framework to both development and external validation studies.

Implication for researchers, editorial offices, and future research

Spin practices are embedded in the process of writing, reviewing, and publishing scientific literature, in which different players share a collective responsibility. All authors and editors commonly use language to emphasize or 'spin' the certainty of the results. Behaviour that is often encouraged by today's volume of research publication in which results will hardly speak by themselves.[19] The consequence is a biased representation of science that will almost always suggests robust solutions to healthcare problems. On the other hand, authors of well-conducted studies with scientific novelty and importance may appropriate use spin to frame their research finding and to one extent, it might be necessary to stand out and allow further

research.

Researchers' and reviewers' inexperience regarding studies on prediction models, lack of guidance, and language barriers may explain the presence of spin. Guidance on 'what to write' is available through reporting guidelines, instead guidance on 'how to write' is scarce.[3,34,77] We recommend the use of the TRIPOD checklist and its adherence form to reduce the risk of unintentional mismatch or missed essential information.[3,34,78,79]

Future research may expand the framework by identifying further facilitators of spin or by developing a severity score based on the likelihood to distort reader's interpretation of each practice (low, moderate, high, unclear).[36] Similarly, an overall 'spin-measure' per study could contribute to the critical appraisal when conducting systematic reviews of prediction models. While research suggests that spin could incorrectly drive clinical practice, we theorized that spin practices in prediction model studies might have a larger impact in medical guidelines and research funds.[37] Prevalence of spin, the effect on reader's interpretation, role of peer-reviewers, number of citations, and assignments of research funds still needs to be assessed within studies on prediction models.[27,71,80,81] Moreover, there is an urgent need to implement effective long-term interventions to reduce spin practices across all study designs.[82,83]

When and how should SPIN-PM be used?

SPIN-PM is primarily intended for researchers reporting prediction model's development and validation for peer-reviewed journals as well as for peer-reviewers and journal editors assessing them for publication. We stress that SPIN-PM is not a scoring tool nor a tool to assess overall quality. We anticipate SPIN-PM could be used by decision-makers when assessing potential utility of prediction models for routine clinical care. However, the use of SPIN-PM should be limited to appraise the quality of reporting of a study rather than overall quality or conduct. For this, we recommend to use PROBAST, a tool for methodological quality and risk of bias assessment or an appropriate systematic review.[4] We encourage authors, peer reviewers, and editors to provide further feedback on how we can improve SPIN-PM or those who see opportunities to expand the framework's items to work with authors of this manuscript.

CONCLUDING REMARKS

Spin is a widely recognized phenomenon in the scientific biomedical literature. We call researchers who develop and validate prediction models into publishing robust and consistent findings, that will likely improve the quality of subsequent research. We hope our standardized approach to identify and evaluate spin practices, will contribute to reducing vague and biased reporting of findings and its interpretation in scientific publications and thus, help increase the uptake of prediction models in clinical practice.

Contributors and sources
CLAN had the idea for the article and led the development of SPIN-PM. JAAD has extensive experience in systematic reviews of prediction models studies. MG has expertise in developing intervention strategies to reduce spin. PD has experience in developing and validating prediction models. MvS, JBR, GSC, RDR, KGMM, and LH have extensive experience on methodological aspects of prediction model development and validation. GSC and KGMM co-lead TRIPOD, an international collaboration for the development of consensus reporting guidelines for prediction model studies. JBR and LH have participated in the development of a classification scheme for spin in diagnostic accuracy test studies, while GSC in prognostic factor studies. CLAN wrote this manuscript with substantial contribution and revisions from all the authors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests
All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: all authors have no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Funding
No specific funding was given to this study. GSC is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is funded by the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS nor NIHR. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication.

Provenance and peer review
Not commissioned, externally peer reviewed.

REFERENCES

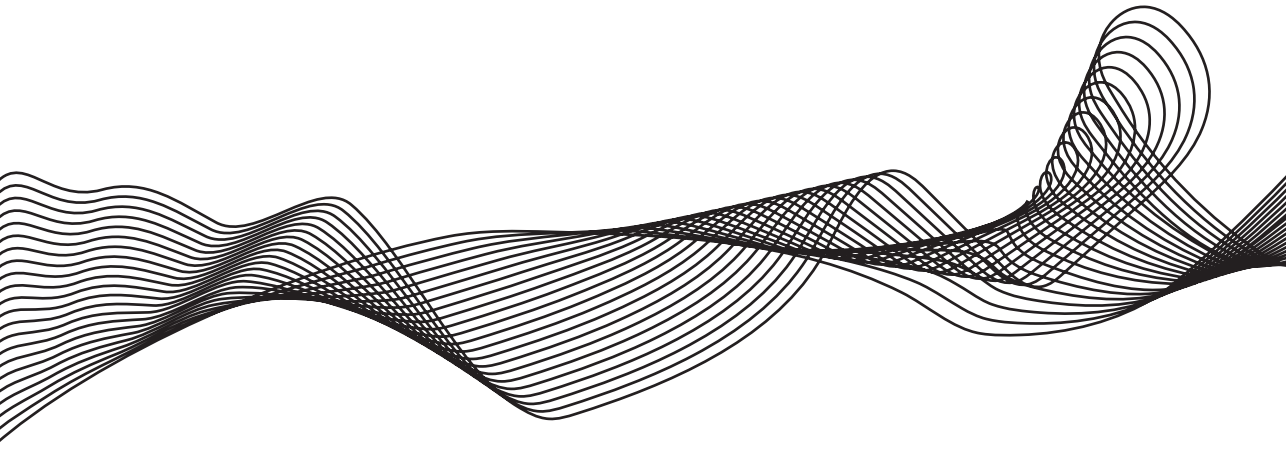
1. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2). doi:10.1371/journal.pmed.1001381
2. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009;338(7707):1373-1377. doi:10.1136/bmj.b604
3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55. doi:10.7326/M14-0697
4. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med. 2019;170(1):W1-W33. doi:10.7326/M18-1377
5. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med. 1992;21(4):384-390. doi:10.1016/S0196-0644(05)82656-3
6. Kinar Y, Kalkstein N, Akiva P, 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(5):879-890. doi:10.1093/jamia/ocv195
7. 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(10):2453-2460. doi:10.1002/cam4.1183
8. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. BMJ. 2016;353. doi:10.1136/BMJ.I2416
9. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. 2006;6:1-10. doi:10.1186/1472-6947-6-38
10. Van Dieren S, Beulens JWJ, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: A systematic review. Heart. 2012;98(5):360-369. doi:10.1136/heartjnl-2011-300734
11. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Med. 2011;9(1):103. doi:10.1186/1741-7015-9-103
12. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369. doi:10.1136/bmj.m1328
13. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021;375:n2281. doi:10.1136/bmj.n2281
14. Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Heal. 2020;2(12):e677-e680. doi:10.1016/S2589-7500(20)30200-4
15. Modine T, Overtchouk P. Machine Learning Is No Magic: A Plea for Critical Appraisal During Periods of Hype. JACC Cardiovasc Interv. 2019;12(14):1339-1341. doi:10.1016/j.jcin.2019.06.004
16. Chen JH, Asch SM. Machine learning and prediction in medicine-beyond the peak of inflated expectations. N Engl J Med. 2017;376(26). doi:10.1056/NEJMp1702071
17. El Hechi M, Ward TM, An GC, et al. Artificial Intelligence, Machine Learning, and Surgical Science: Reality Versus Hype. J Surg Res. 2021;264:A1-A9. doi:10.1016/j.jss.2021.01.046

18. Fletcher RH, Black B. Spin in scientific writing: Scientific mischief and Legal Jeopardy. Med Law. 2007;26(3):511-525.
19. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-2619. doi:10.1073/PNAS.1710755115
20. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Misreporting of Diagnostic Accuracy Studies : Evidence of "Spin." Radiology. 2013;267(2):581-588. doi:10.1148/radiol.12120527/-/DC1
21. Horton R. The rhetoric of research. BMJ. 1995;310(6985):985. doi:10.1136/bmj.310.6985.985
22. Kempf E, de Beyer JA, Cook J, et al. Overinterpretation and misreporting of prognostic factor studies in oncology: a systematic review. Br J Cancer. 2018;119(10):1288-1296. doi:10.1038/s41416-018-0305-5
23. McGrath TA, Bowdridge JC, Prager R, et al. Overinterpretation of Research Findings: Evaluation of "Spin" in Systematic Reviews of Diagnostic Accuracy Studies in High-Impact Factor Journals. Clin Chem. 2020;66(7):915-924. doi:10.1093/clinchem/hvaa093
24. Ghannad M, Olsen M, Boutron I, Bossuyt PM. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. J Clin Epidemiol. 2019;116:9-17. doi:10.1016/j.jclinepi.2019.07.011
25. Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biol. 2017;15(9):1-16. doi:10.1371/journal.pbio.2002173
26. Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. BMC Med Res Methodol. 2015;15(1):1-8. doi:10.1186/s12874-015-0079-x
27. Yavchitz A, Boutron I, Bafeta A, et al. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. PLoS Med. 2012;9(9). doi:10.1371/journal.pmed.1001308
28. Lockyer S, Hodgson R, Dumville JC, Cullum N. "Spin" in wound care research: The reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. Trials. 2013;14(1):1. doi:10.1186/1745-6215-14-371
29. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA - J Am Med Assoc. 2010;303(20):2058-2064. doi:10.1001/jama.2010.651
30. Dwan K, Altman DG, Clarke M, et al. Evidence for the Selective Reporting of Analyses and Discrepancies in Clinical Trials: A Systematic Review of Cohort Studies of Clinical Trials. PLoS Med. 2014;11(6):1-22. doi:10.1371/journal.pmed.1001666
31. Won J, Kim S, Bae I, Lee H. Trial registration as a safeguard against outcome reporting bias and spin ? A case study of randomized controlled trials of acupuncture. Plos. 2019;10:1-19. doi:10.1371/journal.pone.0223305
32. Ioannidis JPA. Spin, Bias, and Clinical Utility in Systematic Reviews of Diagnostic Studies. Clin Chem. 2020;66(7):863-865. doi:10.1093/CLINCHEM/HVAA114
33. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-690. doi:10.1136/heartjnl-2011-301246
34. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73. doi:10.7326/M14-0698
35. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247
36. Yavchitz A, Ravaud P, Altman DG, et al. A new classification of spin in systematic re-

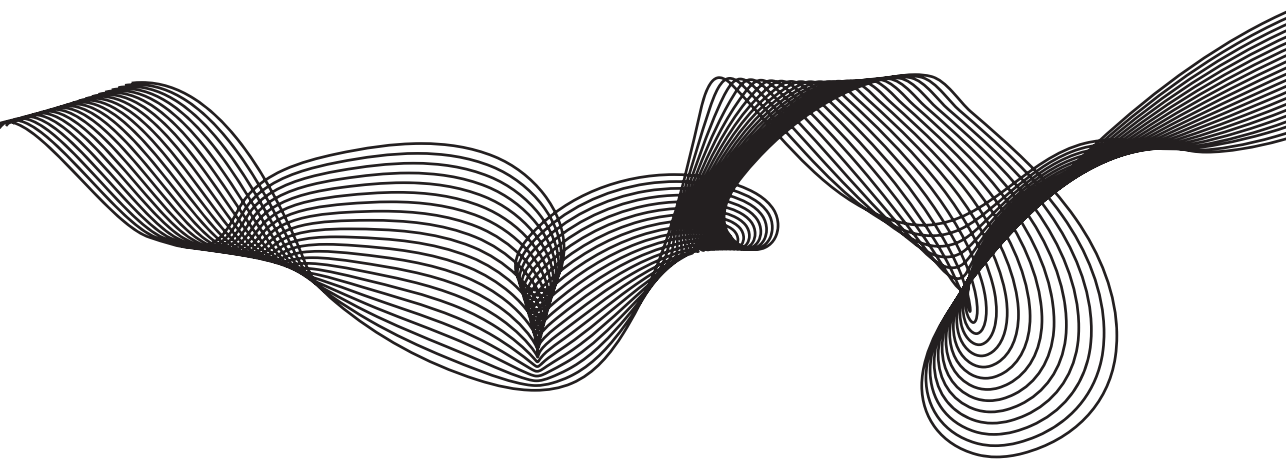
views and meta-analyses was developed and ranked according to the severity. J Clin Epidemiol. 2016;75:56-65. doi:10.1016/j.jclinepi.2016.01.020
37. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: The SPIIN randomized controlled trial. J Clin Oncol. 2014;32(36):4120-4126. doi:10.1200/JCO.2014.56.7503
38. Peat G, Riley R, Croft P, et al. Improving the Transparency of Prognosis Research: The Role of Reporting, Data Sharing, Registration, and Protocols. PLoS Med. 2014;11(7). doi:10.1371/journal.pmed.1001671
39. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22(1):12. doi:10.1186/s12874-021-01469-6
40. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. BMC Med. 2018;16(1):1-12. doi:10.1186/s12916-018-1099-2
41. Krishnamurti T, Woloshin S, Schwartz LM, Fischhoff B. A Randomized Trial Testing US Food and Drug Administration "Breakthrough" Language. JAMA Intern Med. 2015;175(11):1856-1858. doi:10.1001/JAMAINTERNMED.2015.5355
42. Goorbergh R van den, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Informatics Assoc. 2022;00(0):1-10.
43. Vinkers CH, Tijdink JK, Otte WM. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. BMJ. 2015;351(December):1-6. doi:10.1136/bmj.h6467
44. Cepeda MS, Berlin JA, Glasser SC, Battisti WP, Schuemie MJ. Use of Adjectives in Abstracts When Reporting Results of Randomized, Controlled Trials from Industry and Academia. Vol 15.; 2015. doi:10.1007/s40268-015-0085-9
45. Lerchenmueller MJ, Sorenson O, Jena AB. Gender differences in how scientists present the importance of their research: Observational study. BMJ. 2019;367. doi:10.1136/bmj.l6573
46. Steyerberg E, Jr FH. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005
47. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. BMJ. 2009;338(7709):1487-1490. doi:10.1136/bmj.b606
48. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5
49. Chen D, Goyal G, Go RS, Parikh SA, Ngufor CG. Improved Interpretability of Machine Learning Model Using Unsupervised Clustering: Predicting Time to First Treatment in Chronic Lymphocytic Leukemia. JCO Clin Cancer Informatics. 2019;(3):1-11. doi:10.1200/cci.18.00137
50. Zhang F, Zhang Y, Ke C, et al. Predicting ovarian cancer recurrence by plasma metabolic profiles before and after surgery. Metabolomics. 2018;14(5):1-9. doi:10.1007/s11306-018-1354-8
51. Wu CC, Yeh WC, Hsu WD, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed. 2019;170:23-29. doi:10.1016/j.cmpb.2018.12.032
52. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression. Sci Rep. 2018;8(1):1-12. doi:10.1038/s41598-018-20166-x

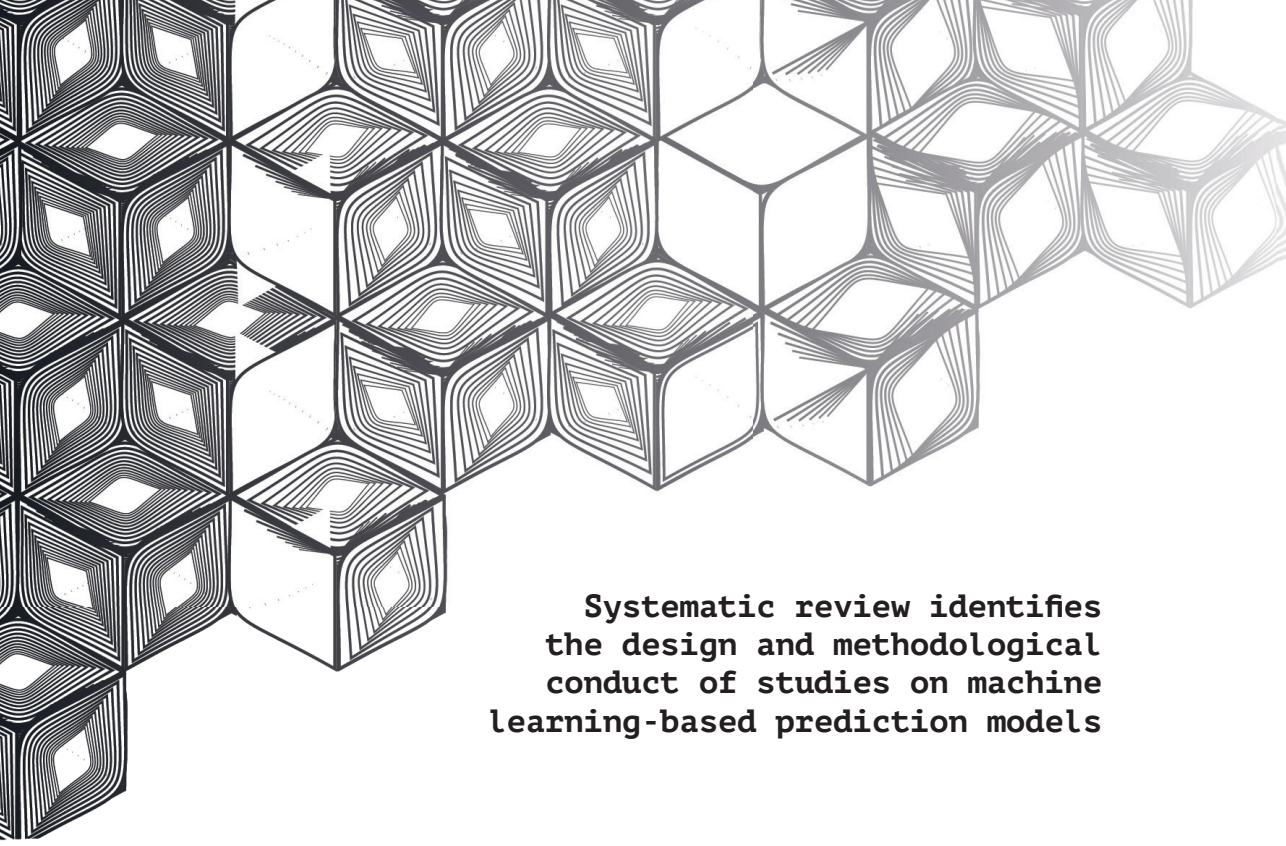
53. Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. Front Artif Intell. 2021;0:116. doi:10.3389/FRAI.2020.561802
54. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. Bmj. Published online 2022:e069249. doi:10.1136/bmj-2021-069249
55. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans RM, Vergouwe Y, Habbema J. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Proc ICED 2007, 16th Int Conf Eng Des. 2007;DS 42:774-781.
56. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138. doi:10.1097/EDE.0B013E3181C30FB2
57. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038
58. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic Progn Res. 2019;3(1):1-8. doi:10.1186/s41512-019-0064-7
59. Xiao C, Liu Y, Feng DD, Wang X. Key Marker Selection for the Detection of Early Parkinson' s Disease using Importance-Driven Models. Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf. 2018;2018:6100-6103. doi:10.1109/EMBC.2018.8513564
60. Chassagnon G, Vakalopoulou M, Battistella E, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Med Image Anal. 2021;67:101860. doi:10.1016/j.media.2020.101860
61. Caro-Codón J, Lip GYH, Rey JR, et al. Prediction of thromboembolic events and mortality by the CHADS2and the CHA2DS2-VASc in COVID-19. Europace. 2021;23(6):937-947. doi:10.1093/europace/euab015
62. Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A. Development of a cardiovascular risk score for use in low- and middle-income countries. J Nutr. 2011;141(7):1375-1380. doi:10.3945/jn.110.133140
63. Beauchet O, Noublanche F, Simon R, et al. Falls risk prediction for older inpatients in acute care medical wards: is there an interest to combine an early nurse assessment and the artificial neural network analysis? J Nutr Heal Aging. 2018;22(1):131-137. doi:10.1007/s12603-017-0950-z.
64. Sanchez Fernandez I, Sansevere AJ, Gainza-Lein M, Kapur K, Loddenkemper T. Machine Learning for Outcome Prediction in Electroencephalograph (EEG)-Monitored Children in the Intensive Care Unit. J Child Neurol. 2018;33(8):546-553. doi:10.1177/0883073818773230
65. Perel P, Prieto-Merino D, Shakur H, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012;345(August):1-12. doi:10.1136/bmj.e5166
66. Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Informatics Assoc. 2019;26(12):1427-1436. doi:10.1093/jamia/ocz171
67. Garcia-Arce A, Rico F, Zayas-Castro JL. Comparison of Machine Learning Algorithms for the Prediction of Preventable Hospital Readmissions. J Healthc Qual. 2018;40(3):129-138. doi:10.1097/JHQ.0000000000000080
68. Lorenzoni G, Sabato SS, Lanera C, et al. Comparison of Machine Learning Techniques for Prediction of Hospitalization in Heart Failure Patients. J Clin Med. 2019;8(9):1298. doi:10.3390/jcm8091298

69. Kaviarasi R, Gandhi Raj R. Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. J Med Syst. 2019;43(7). doi:10.1007/s10916-019-1297-2
70. Fan Y, Li Y, Li Y, et al. Development and assessment of machine learning algorithms for predicting remission after transsphenoidal surgery among patients with acromegaly. Endocrine. 2020;67(2):412-422. doi:10.1007/s12020-019-02121-6
71. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. J Clin Epidemiol. 2016;77:44-51. doi:10.1016/j.jclinepi.2016.04.012
72. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021;138:219-226. doi:10.1016/j.jclinepi.2021.05.018
73. Koletsi D, Karagianni A, Pandis N, Makou M, Polychronopoulou A, Eliades T. Are studies reporting significant results more likely to be published? Am J Orthod Dentofac Orthop. 2009;136(5):632.e1-632.e5. doi:10.1016/j.ajodo.2009.02.024
74. Young NS, Ioannidis JPA, Al-Ubaydli O. Why Current Publication Practices May Distort Science. PLoS Med. 2008;5(10):e201. doi:10.1371/journal.pmed.0050201
75. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34. doi:10.1016/j.jclinepi.2014.09.007
76. Adams RC, Challenger A, Bratton L, et al. Claims of causality in health news: A randomised trial. BMC Med. 2019;17(1):1-11. doi:10.1186/s12916-019-1324-7
77. Boers M. Graphics and statistics for cardiology: Designing effective tables for presentation and publication. Heart. 2018;104(3):192-200. doi:10.1136/heartjnl-2017-311581
78. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open. 2019;9(4). doi:10.1136/bmjopen-2018-025611
79. Heus P, Reitsma JB, Collins GS, et al. Transparent Reporting of Multivariable Prediction Models in Journal and Conference Abstracts: TRIPOD for Abstracts. Ann Intern Med. 2020;173(1):43. doi:10.7326/M20-0193
80. Haneef R, Lazarus C, Ravaud P, Yavchitz A, Boutron I. Interpretation of results of studies evaluating an intervention highlighted in google health news: A cross-sectional study of news. PLoS One. 2015;10(10):1-15. doi:10.1371/journal.pone.0140889
81. Boutron I, Haneef R, Yavchitz A, et al. Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. BMC Med. 2019;17(1):1-10. doi:10.1186/s12916-019-1330-9
82. Ghannad M, Yang B, Leeflang M, et al. A randomized trial of an editorial intervention to reduce spin in the abstract ' s conclusion of manuscripts showed no significant effect. J Clin Epidemiol. 2021;130:69-77. doi:10.1016/j.jclinepi.2020.10.014
83. Blanco D, Schroter S, Aldcroft A, et al. Effect of an editorial intervention to improve the completeness of reporting of randomised trials: A randomised controlled trial. BMJ Open. 2020;10(5). doi:10.1136/bmjopen-2020-036799



Chapter 6





Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models

Constanza L Andaur Navarro
Johanna AA Damen
Maarten van Smeden
Toshihido Takada
Steven WJ Nijman
Paula Dhiman
Jie Ma
Gary S Collins
Ram Bajpai
Richard D Riley
Karl GM Moons
Lotty Hooft

Journal of Clinical Epidemiology 2023;154:8-22

ABSTRACT

Objective. We sought to summarize the study design, modelling strategies, and performance measures reported in studies on clinical prediction models developed using machine learning techniques.

Study Design and Setting. We search PubMed for articles published between 01/01/2018 and 31/12/2019, describing the development or the development with external validation of a multivariable prediction model using any supervised machine learning technique. No restrictions were made based on study design, data source, or predicted patient-related health outcomes.

Results. We included 152 studies, 58 (38.2% [95%CI 30.8-46.1]) were diagnostic and 94 (61.8% [95%CI 53.9-69.2]) prognostic studies. Most studies reported only the development of prediction models (n=133, 87.5% [95%CI 81.3-91.8]), focused on binary outcomes (n=131, 86.2% [95%CI 79.8-90.8), and did not report a sample size calculation (n=125, 82.2% [95%CI 75.4-87.5]). The most common algorithms used were support vector machine (n=86/522, 16.5% [95%CI 13.5-19.9]) and random forest (n=73/522, 14% [95%CI 11.3-17.2]). Values for area under the Receiver Operating Characteristic curve ranged from 0.45 to 1.00. Calibration metrics were often missed (n=494/522, 94.6% [95%CI 92.4-96.3]).

Conclusions. Our review revealed that focus is required on handling of missing values, methods for internal validation, and reporting of calibration to improve the methodological conduct of studies on machine learning-based prediction models.

Systematic review registration: PROSPERO, CRD42019161764.

INTRODUCTION

Clinical prediction models aim to improve healthcare by providing timely information for shared decision-making between clinician and their patients, risk stratification, changes in behaviour, and to counsel patients and their relatives.[1] A prediction model can be defined as the (weighted) combination of several predictors to estimate the likelihood or probability of the presence or absence of a certain disease (diagnostic model), or the occurrence of an outcome over a time period (prognostic model).[2] Traditionally, prediction models were developed using regression techniques, such as logistic or time-to-event regression. However, in the past decade, the attention and use of machine learning approaches to developing clinical prediction models has rapidly grown.

Machine learning can be broadly defined as the use of computer systems that fit mathematical models that assume non-linear associations and complex interactions. Machine learning has a wide range of potential applications in different pathways of healthcare. For example, machine learning is applied in stratified medicine, triage tools, image-driven diagnosis, online consultations, medication management, and to mine electronic medical records.[3] Most of these applications make use of supervised machine learning whereby a model is fitted to learn the conditional distribution of the outcome given a set of predictors with little assumption on data distributions, non-linear associations, and interactions. This model can be later applied in other but related individuals to predict their (yet unknown) outcome. Support vector machines (SVM), random forests (RF), and neural networks (NN) are some examples of these techniques.[4]

The number of studies on prediction models published in the biomedical literature increases every year.[5,6] With more healthcare data being collected and increasing computational power, we expect studies on clinical prediction models based on (supervised) machine learning techniques to become even more popular. Although numerous models are being developed and validated for various outcomes, patients' populations, and healthcare settings, only a minority of these published models are successfully implemented in clinical practice.[7,8]

The use of appropriate study designs and prediction model strategies to develop or validate a prediction model could improve their transportability into clinical settings.[9] However, currently there is a dearth of information about which study designs, what modelling strategies, and which performance measures do studies on clinical prediction models report when choosing machine learning as modelling approach.[10–12] Therefore, our aim was to systematically review and summarise the characteristics on study design, modelling steps, and performance measures reported in studies of prediction models using supervised machine learning.

METHODS

We followed the PRISMA 2020 statement to report this systematic review.[13]

Eligibility criteria

We searched via PubMed (search date 19 December 2019) for articles published between 1 January 2018 and 31 December 2019 (Supplemental file 1). We focused on primary studies that described the development and/or validation of one or more multivariable diagnostic or prognostic prediction model(s) using any supervised machine learning technique. A multivariable prediction model was defined as a model aiming to predict a health outcome by using two or more predictors (features). We considered a study to be an instance of supervised machine learning when reporting a non-regression approach to model development. If a study reported machine learning models alongside regression-based models, this was included. We excluded studies reporting only regression-based approaches such as unpenalized regression (e.g., ordinary least squares or maximum likelihood logistic regression), or penalized regression (e.g., lasso, ridge, elastic net, or Firth's regression), regardless of whether they referred to them as machine learning. Any study design, data source, study population, predictor type or patient-related health outcome was considered.

We excluded studies investigating a single predictor, test, or biomarker. Similarly, studies using machine learning or AI to enhance the reading of images or signals, rather than predicting health outcomes in individuals, or studies that used only genetic traits or molecular ('omics') markers as predictors, were excluded. Furthermore, we also excluded reviews, meta-analyses, conference abstracts, and articles for which no full text was available via our institution. The selection was restricted to humans and English-language studies. Further details about eligibility criteria can be found in our protocol.[14]

Screening and selection process

Titles and abstracts were screened to identify potentially eligible studies by two independent reviewers from a group of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD). After selection of potentially eligible studies, full text articles were retrieved and two independent researchers reviewed them for eligibility; one researcher (CLAN) screened all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively screened the same articles for agreement. In case of any disagreement during screening and selection, a third reviewer was asked to read the article in question and resolve.

Extraction of data items

We selected several items from existing methodological guidelines for reporting and critical appraisal of prediction model studies to build our data extraction form

(CHARMS, TRIPOD, PROBAST).[15–18] Per study, we extracted the following items: characteristics of study design (e.g. cohort, case-control, randomized trial) and data source (e.g. routinely collected data, registries, administrative databases), study population, outcome, setting, prediction horizon, country, patient characteristics, sample size (before and after exclusion of participants), number of events, number of candidate and final predictors, handling of missing data, hyperparameter optimization, dataset splitting (e.g. train-validation-test), method for internal validation (e.g. bootstrapping, cross validation), number of models developed and/or validated, and availability of code, data and model. We defined country as the location of the first author's affiliation. Per model, we extracted information regarding the following items: type of algorithm used, selection of predictors, reporting of variable importance, penalization techniques, reporting of hyperparameters, and metrics of performance (e.g., discrimination and calibration).

Items were recorded by two independent reviewers. One reviewer (CLAN) recorded all items, whilst the other reviewers collectively assessed all articles (CLAN, TT, SWJN, PD, JM, RB, JAAD). Articles were assigned to reviewers in a random manner. To accomplish consistent data extraction, the standardized data extraction form was piloted by all reviewers on five articles. Discrepancies in data extraction were discussed and solved between the pair of reviewers. The full list of extracted items is available in our published protocol.[14]

We extracted information on a maximum number of 10 models per article. We selected the first 10 models reported in the methods section of articles and extracted items accordingly in the results section. For articles describing external validation or updating, we carried out a separate data extraction with similar items. If studies referred to the supplemental file for detailed descriptions, the items were checked in those files. Reviewers could also score an item as not applicable, not reported, or unclear.

Summary measures and synthesis of results

Results were summarized as percentages (with confidence intervals calculated using the Wilson score interval and the Wilson score continuity-corrected interval, when appropriated), medians, and interquartile range (IQR), alongside a narrative synthesis. The reported number of events was combined with the reported number of candidate predictors to calculate the number of events per variable (EPV). Data on a model's predictive performance was summarized for the apparent performance, corrected performance, and externally validated performance. We defined "apparent performance" when studies reported model performance assessed in the same dataset or sample in which the model was developed and in case no re-sampling methods were used; "corrected performance" when studies reported model performance assessed in test dataset and/or using re-sampling methods;

and “externally validated performance” when studies reported model performance assessed in another sample than the one use for model development. As we wanted to identify the methodological conduct of studies on prediction models developed using machine learning, we did not evaluate the nuances of each modelling approach or its performance, instead we kept our evaluations at study level. We did not perform a quantitative synthesis of the model’ performance (i.e., meta-analysis), as this was beyond the scope of our review. Analysis and synthesis of data was presented overall. Analyses were performed using R (version 4.1.0, R Core Team, Vienna, Austria).

RESULTS

Among 24,814 articles retrieved, we drew a random sample of 2482 articles. After title and abstract screening, 312 references potentially met the eligibility criteria. After full-text screening, 152 articles were included in this review: 94 (61.8% [95%CI 53.9-69.2]) prognostic and 58 (38.2% [95%CI 30.8-46.1]) diagnostic prediction model studies (Figure 1). Detailed description of the included articles is provided in Supplemental file 2.

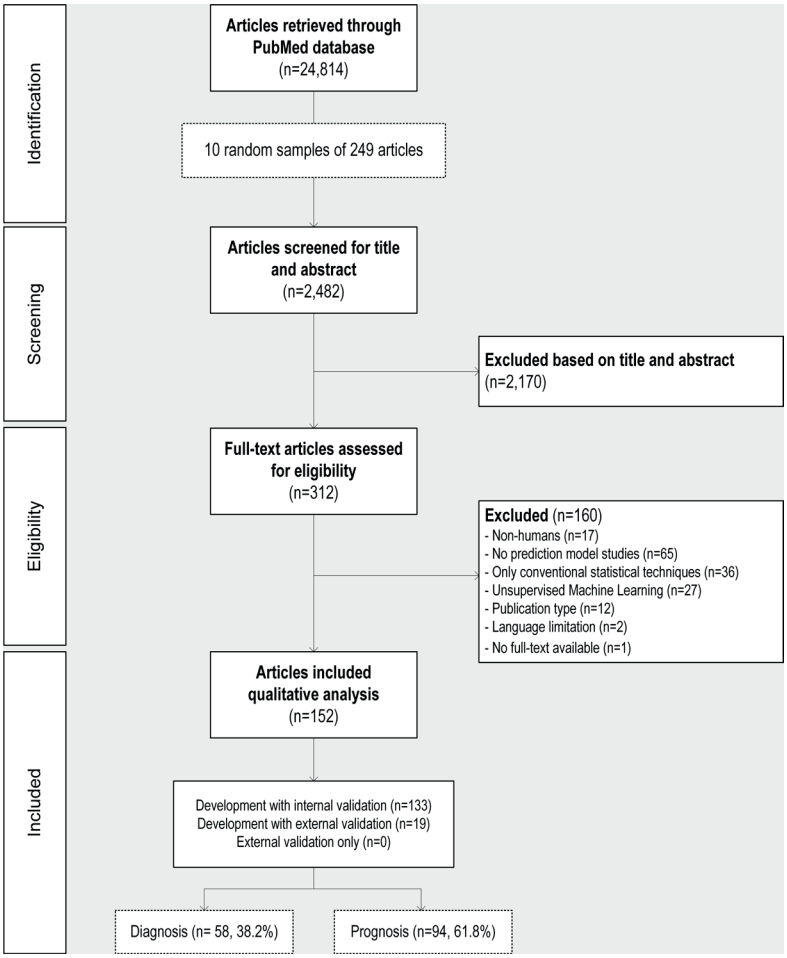


Figure 1. Flowchart of included studies

In 152 articles, 132 (86.8% [95%CI 80.5-91.3]) studies developed prediction models and evaluated their performance using an internal validation technique, 19 (12.5% [95%CI 8.2-18.7]) studies developed and externally validated the same ML based prediction model, and 1 (0.6%) study included model development

with external validation of another comparative model (eventually included as development with internal validation). Eighty-seven studies (57% [95% CI 49.3-64.8]) were published in 2019 and 65/152 studies (42.8% [95% CI 35.2-50.7]) in 2018. The three clinical fields with the most articles were oncology (n=21/152, 13.8% [95%CI 9.2–20.2), surgery (n=20/152, 13.5% [95%CI 8.7-19.5), and neurology (n=20/152, 13.5% [95%CI 8.7-19.5]). Most articles originated from North America (n=59/152, 38.8% [95%CI 31.4-46.7]), followed by Asia (n=46/152, 30.3% [95%CI 23.5-38]) and Europe (n=37/152, 24,3% [95%CI 18.2-31.7]). Half of the studies had a first author with a clinical affiliation (n=85/152, 56% [95%CI 48-63.6]). Other characteristics are shown in Table 1.

Table 1. General characteristics of included studies

	Total (n=152)
	n (%) [95% CI]
Study aim	
Diagnosis	58 (38.2) [30.8-46.1]
Prognosis	94 (61.8) [53.9-69.2]
Study type	
Model development only	133 (87.5) [81.3-91.8]
Model development with external validation	19 (12.5) [8.2-18.7]
Outcome aim	
Classification	120 (78.9) [71.8-84.7]
Risk probabilities	32 (21.0) [80.5-91.3]
Setting†	
General population	17 (11.2) [7.1-17.2]
Primary care	15 (9.9) [6.1-15.6]
Secondary care	32 (21.1) [15.3-28.2]
Tertiary care	78 (51.3) [43.4-59.1]
Unclear	13 (8.6) [5.1-14.1]
Outcome format	
Continuous	7 (4.6) [2.2-9.2]
Binary	131 (86.2) [79.8-90.8]
Multinomial	7 (4.6) [2.2-9.2]
Ordinal	2 (1.3) [0.4-4.7]
Time-to-event	3 (2.0) [0.7-5.6]
Count	2 (1.3) [0.4-4.7]
Type of outcome	
Death	21 (13.8) [9.2-20.2]
Complications	65 (42.8) [35.2-50.7]
Disease detection	30 (19.7) [14.2-26.8]
Disease recurrence	9 (5.9) [3.1-10.9]
Survival	3 (2.0) [0.7-5.6]
Readmission	4 (2.6) [1-6.6]
Other [a]	20 (13.2) [8.7-19.5]
Mentioning of reporting guidelines†	
TRIPOD	8 (5.3) [2.7-10]
STROBE	3 (2.0) [0.7-5.6]
Other [b]	5 (3.3) [1.4-7.5]
None	139 (91.4) [85.9-94.9]
Model availability†	
Repository for data	18 (11.8) [7.6-17.9]
Repository for code	13 (8.6) [5.1-14.1]
Model presentation [c]	31 (20.4) [14.8-27.5]
None	121 (79.6) [72.5-85.2]

† Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option.
[a] This includes length of stay, medication dose, patient's disposition, order type, lesion extension, laboratory results, cancer stage, treatment option, attendance, equipment usage, operative time.
[b] Guidelines for developing and reporting machine learning models in biomedical research (n=2), STARD (n=2), BRISQ (n=1).
[c] This includes simplified scoring rule, chart, nomogram, online calculator, or worked examples.

Overall, 1,429 prediction models were developed (Median: 9.4 models per study, IQR: 2-8, Range: 1-156). As we set a limit on data extraction to 10 models per article, we evaluated 522 models. The most common applied modeling techniques were support vector machine (n=86/522, 16.5% [95%CI 13.5-20]), logistic regression (n=74/522, 14.2% [95%CI 11.4-17.5]), and random forest ([n=73/522, 14% [95%CI 11.2-17.3]). Further modelling algorithms are described in Table 2. In 120/152 (78.9% [95%CI 71.8-84.7]) articles, authors recommended at least one model usually based on model performance (i.e., AUC).

Participants

Participants included in the reviewed studies were mostly recruited from secondary (n=32/152, 21.1% [95%CI 15.3-28.2]) and tertiary care (n=78/152, 51.3% [95%CI 43.4-59.1]) settings (Table 1). Approximately half of the studies involved data from one center (n=73/152, 48% [95%CI 40.2-55.9]) (Table 3).

Data sources

The prediction models were most frequently developed using cohort data, either prospective (n=50/152, 32.9% [95%CI 25.9-40.7]) or retrospective (n=48/152, 31.6% [95%CI 24.7-39.3]). Electronic medical records were used in 30/152 studies (19.7% [95%CI 14.2-26.8]). Data collection was conducted on average for 41.9 months (IQR 3 to 60 months) when used to develop models, while for externally validation this was 44.4 months (IQR 1.75 to 42 months). In 101 out of 152 studies (66.4% [95%CI 58.6-73.5]), the time horizon for the predictions was mostly unspecified. However, when reported (n=51/152, 33.6% [95%CI 26.5-41.4]), the time horizon of prediction ranged from 24 hours to 8 years (Table 3).

Outcome

Most models were developed to predict a binary outcome (n=131/152, 86.2% [95%CI 79.8-90.8]). The most frequent predicted outcome was complications after a certain treatment (n=66/152, 43.4% [95%CI 35.8-51.4]). Mortality was also a common endpoint (n=21/152, 13.8% [95%CI 9.2-20.2]) (Table 1).

Candidate predictors

Candidate predictors frequently involved demographics, such as age and sex (n=120/152, 78.9% [95%CI 71.8-84.7]), clinical history (n=111/152, 73% [95%CI 65.5-79.4]), and blood and urine parameters (n=63/152, 41.4% [95%CI 33.9-49.4]). When applicable, treatment modalities were also considered as predictors (n=36/116, 31.0% [95%CI 17.6-31]). Studies included a median of 24 candidate predictors (IQR 13 – 112). Most studies included continuous variables as candidate predictors (n=131/152, 86.2% [95%CI 79.8-90.8]). Whether continuous predictors were categorized during data preparation was often unclear (n=104/152, 68.4% [95%CI 60.7-75.3]) (Table 4).

Table 2. Modelling algorithms for all extracted models

Modelling algorithm	All extracted models (n=522)
	n (%) [95%CI]
Unpenalized regression models	**101 (19.3) [16.1-23.1]**
Ordinary least squares regression [a]	27 (5.2) [3.5-7.5]
Maximum likelihood logistic regression	74 (14.2) [11.4-17.5]
Penalized regression models	**29 (5.6) [3.8-8]**
Elastic Net	9 (1.7) [0.8-3.4]
LASSO	13 (2.5) [1.4-4.3]
Ridge	7 (1.3) [0.6-2.9]
Tree-based models	**166 (31.8) [28-36]**
Decision trees (e.g., CART) [c]	46 (8.8) [6.6-11.7]
Random forest [d]	73 (14) [11.2-17.3]
Extremely randomized trees	1 (0.2) [0.01-1.2]
Regularized Greedy Forest	1 (0.2) [0.01-1.2]
Gradient boosting machine [e]	34 (6.5) [4.6-9.1]
XGBoost	11 (2.1) [1.1-3.9]
Neural Network (incl. deep learning) [b]	**75 (14.4) [11.5-17.7]**
Support Vector Machine	**86 (16.5) [13.5-20]**
Naïve Bayes	**22 (4.2) [2.7-6.4]**
K-nearest neighbor	**15 (2.9) [1.7-4.8]**
Superlearner ensembles	**14 (2.7) [1.5-4.6]**
Other [f]	**10 (1.9) [1-3.6]**
Unclear	**4 (0.8) [0.2-2.1]**

CART, Classification And Regression Tree; LASSO, Least Absolute Shrinkage and Selection Operator; XGBoost, extreme gradient boosting; CI, confidence interval.
[a] Discriminant analysis, generalized additive models (GAM), partial least squares were extracted as OLS regression.
[b] Multilayer perceptron, denseNet, convolutional, recurrent, and Bayesian neural networks were extracted as neural networks.
[c] This includes conditional inference tree (n=3), optimal tree (n=1).
[d] This includes Random Survival Forest (n=2).
[e] This includes lightGBM (n=1), adaBoost (n=8), catBoost (n=1), logitboost (n=1), RUSBoost (n=1), and stochastic (n=1).
[f] This includes bayesian network (n=3), rule-based classifier (n=1), highly predictive signatures (n=1), Kalman filtering (n=1), fuzzy soft set (n=1), adaptive neuro-fuzzy inference system (n=1), stochastic gradient descent (n=1), fully corrective binning (n=1).

Sample size

Studies had a median sample size of 587 participants (IQR 172 – 6328). The number of events across the studies had a median of 106 (IQR 50 – 364). Based on studies with available information (n=28/152, 18.4% [95%CI 13.1-25.3]), a median of 12.5 events per candidate predictors were used for model development (IQR 5.7 – 27.7) (Table 5). Most studies did not report a sample size calculation or justification for sample size (n=125/152, 82.2% [95%CI 75.4-87.5]). When sample size justification was provided, the most frequent rationale given was based on the size of existing/available data used (n=16/27, 59.3% [95%CI 40.7-75.5]) (Table 3).

Table 3. Study design of included studies, stratified by type of prediction model study

	Total (n=152)	Development only (n=133)	Development with external validation (n=19)
	n (%) [95%CI]	n (%) [95%CI]	n (%) [95%CI]
Data sources [†a]			
Prospective cohort	50 (32.9) [25.9-40.7]	43 (32.3) [25-40.7]	7 (36.8) [19.1-59]
Retrospective cohort	48 (31.6) [24.7-39.3]	45 (33.8) [26.3-42.2]	4 (21.1) [8.5-43.3]
Randomized Controlled Trial	3 (2.0) [0.7-5.6]	2 (1.5) [0.4-5.3]	1 (5.3) [0.3-24.6]
EMR	30 (19.7) [14.2-26.8]	28 (21.1) [15-28.7]	0
Registry	18 (11.8) [7.6-17.9]	15 (11.3) [7-17.8]	4 (21.1) [8.5-43.3]
Administrative claims	4 (2.6) [1-6.6]	4 (3.0) [1.2-7.5]	0
Case-control	18 (11.8) [7.6-17.9]	15 (11.3) [7-17.8]	3 (15.8) [5.5-37.6]
Number of centers	**110 (72.4)**	**98 (73.7)**	**12 (63.2)**
Median [IQR] (range)	1 [1-3], 1 to 51920	1 [1-3], 1 to 712	1 [1-10], 1 to 51920
Follow-up (months) [‡]	**47 (30.9)**	**39 (29.2)**	**8 (42.1)**
Median [IQR] (range)	41.9 [3-60], 0.3 to 307	43.6 [4.5-60], 0.3 to 307	33.5 [1.75-42], 1 to 144
Predictor horizon (months) [‡]	**49 (32.2) [25.3-40]**	**61 (45.9) [37.6-54.3]**	**7 (36.8)**
Median [IQR] (range)	8.5[1-36], 0.03 to 120	6 [1-33.5], 0.03 to 120	36 [6.5-60], 1 to 60
Sample size justification	**27 (17.8) [12.5-24.6]**	**24 (18.0) [12.4-25.4]**	**3 (15.8)**
Power	5 (18.5) [8.2-36.7]	5 (20.8) [9.2-40.5]	0
Justified time interval	5 (18.5) [8.2-36.7]	3 (12.5) [4.3-31]	2 (66.7)
Size of existing/available data	16 (59.3) [40.7-75.5]	15 (62.5) [42.7-78.8]	1 (33.3)
Events per variable	1 (3.7) [0.2-18.3]	1 (4.2) [0.2-20.2]	0
Internal validation [†]			
Split sample with test set	**86 (56.6) [48.6-64.2]**	NA	NA
(Random) split	49 (57) [46.4-66.9]		
(Non-random) split	9 (10.5) [5.6-18.7]		
Split [b]	28 (32.6) [23.6-43]		
Bootstrapping	**5 (3.3) [1.4-7.5]**	NA	NA
With test set	3 (60.0) [23.1-88.2]		
With cross-validation	1 (20) [1-62.4]		
Cross-validation	**70 (46.1) [38.3-54]**	NA	NA
Non-nested (single)	32 (45.7) [34.6]		
Nested	10 (14.3) [7.9-24.3]		
With test set	24 (34.3) [24.2-46]		
External validation [†]			
Chronological	NA	NA	5 (26.3) [11.8-48.8]
Geographical	NA	NA	3 (15.8) [5.5-37.6]
Independent dataset	NA	NA	11 (57.9) [36.3-76.9]
Fully independent dataset	NA	NA	8 (42.1) [23.1-63.7]

[†]Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one measure. We report then the raw percentages. NA, not applicable.
[‡]We collected the longest follow-up and longest prediction horizon.
[a] Data sources also included surveys (n=2), cross-sectional studies (n=2).
[b] Unclear whether split sample was performed random or non-random.

Missing values

Missing values were an explicit exclusion criterion of participants in 56 studies (n=56/152, 36.8% [95%CI 29.6-44.7]). To handle missing values, complete-case analysis was the most common method (n=30/152, 19.7% [95%CI 14.2-26.8]). Other methods were median imputation (n=10/152, 6.6% [95%CI 3.6-11.7), multiple imputation (n=6/152, 3.9% [95%CI 1.9-8.3]) and k-nearest neighbor imputation (n=5/152, 3.3% [95%CI 1.4-7.5]). Further methods to handle missing values are presented in Table 6.

Table 4. Predictors in included studies

	Total (n=152)
	n (%) [95% CI]
Type of candidate predictors†	
Demography	120 (78.9) [71.8-84.7]
Clinical history	111 (73.0) [65.5-79.4]
Physical examination	0
Blood or Urine parameters	63 (41.4) [33.9-49.4]
Imaging	49 (32.2) [52.3-40]
Genetic risk score	7 (4.6) [2.2-9.2]
Pathology	16 (10.5) [6.6-16.4]
Scale score	31 (20.4) [14.8-27.5]
Questionnaires	0
Treatment as candidate predictor	
Yes	36 (23.7) [17.6-31]
No	80 (52.6) [44.7-60.4]
Not applicable	36 (23.7) [17.6-31]
Continuous variables as candidate predictors	
Yes	131 (86.2) [79.8-90.8]
Unclear	17 (11.2) [7.1-17.2]
A-priori selection of candidate predictors‡	
Yes	63 (41.4) [33.9-49.4]
No	47 (30.9) [24.1-38.7]
Unclear	42 (27.6) [21.1-35.2]
Methods to handle continuous predictors†‡	
Linear (no change)	13 (8.6) [5.1-14.1]
Non-linear (planned)	2 (1.3) [0.4-4.7]
Non-linear (unplanned)	4 (2.6) [1-6.6]
Categorised (some)	16 (10.5) [6.6-16.4]
Categorised (all)	18 (11.8) [7.6-17.9]
Unclear	104 (68.4) [60.7-75.3]
Categorization of continuous predictors‡	
Data dependent	4 (2.6) [1-6.6]
No rationale	17 (11.2) [7.1-17.2]
Based on previous literature or standardization	13 (8.6) [5.1-14.1]
Not reported	118 (77.6) [70.4-83.5]

†Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies can report more than one measure.
‡As data preparation

Class imbalance

In our sample, 27/152 (17.8% [95%CI 12.5-24.6]) studies applied at least one method to purportedly address class imbalance, that is – when one class of the outcome outnumbers the other class (Table 7). The most applied technique was Synthetic Minority Over-sampling Technique (SMOTE), a method that combines oversampling the minority class with undersampling the majority class. [19,20]

Modelling algorithms

Tree-based methods were applied in 166/522 (31.8% [95%CI 27.9-36]) models with random forest being the most popular (n=73/522, 14% [95%CI 11.2-17.3]). Alongside machine learning algorithm, unpenalized regression methods (n=101/522, 19.3% [95%CI 16.1-23.1]), and particularly logistic regression (n=74/522, 14.2 [95%CI 11.4-17.5]) were often applied. Few studies reported models built with penalized regression (n=29/522, 5.6% [95%CI 3.8-8]). NNs (n=74/522, 14.2% [95%CI 11.4-17.5]) and Naïve Bayes (n=22/522, 4.2% [95%CI 2.7-6.4]) were also applied in our sample of articles.

Table 5. Sample size of included studies (n=152)

	Total (n=152)	
	n (%)	Median [IQR], range
Initial sample size	**93 (61.2)**	**999 [272-24522], 8 to 1093177**
External validation [a]	13 (68.4)	318 [90-682], 19 to 1113656
Final sample size	**151 (99.3)**	**587 [172-6328], 8 to 594751**
Model development	83 (54.6)	641 [226-10512], 5 to 392536
Internal validation [b]	83 (54.6)	230 [75-2892], 2 to 202215
External validation [a]	18 (94.7)	293 [71-1688], 19 to 59738
Initial number of events	**10 (6.6)**	**66 [15-207], 15 to 4370**
External validation [a]	1 (5.3)	107
Final number of events	**37 (24.3)**	**106 [50-364], 15 to 7543**
Model development	19 (13.2)	156 [47-353], 10 to 5054
Internal validation [b]	19 (13.2)	35 [26-109], 4 to 2489
External validation [a]	4 (21.1)	250 [121-990], 107 to 2834
Number of candidate predictors	119 (78.3)	24 [13-112], 2 to 39212
Number of included predictors	90 (59.2)	12 [7-23], 2 to 570
Events per candidate predictor [c]	28 (18.4)	12.5 [5.7-27.7], 1.2 to 754.3

[a] External validation was performed in 19 studies.
[b] Combines all internal validation methods, e.g., split sample, cross validation, bootstrapping.
[c] For model development.

Table 6. Handling of missing values, stratified by study type

	Total (n=152)	Development only (n=133)	Development with external validation (n=19)
	n (%) [95%CI]	n (%) [95%CI]	n (%) [95%CI]
Missingness as exclusion criteria for participants			
Yes	56 (36.8) [29.6-44.7]	51 (38.3) [30.5-46.8]	2 (10.5) [2.9-31.4]
Unclear	36 (23.7) [17.6-31]	33 (24.8) [18.2-32.8]	6 (31.6) [15.4-54]
Number of patients excluded	**36 (23.7) [17.6-31]**	**34 (25.6) [18.9-33.6]**	**0**
Median [IQR] (range)	191 [19-4209], 1 to 627180	224 [16-4699], 1 to 627180	0
Methods of handling missing data†			
No missing data	4 (2.6) [1-6.6]	3 (2.3) [0.8-6.4]	1 (5.3) [0.3-24.6]
No imputation	4 (2.6) [1-6.6]	4 (3) [1.2-7.5]	0
Complete case-analysis	30 (19.7) [14.2-26.8]	28 (21.1) [15-28.7]	2 (10.5) [2.9-31.4]
Mean imputation	4 (2.6) [1-6.6]	3 (2.3) [0.8-6.4]	1 (5.3) [0.3-24.6]
Median imputation	10 (6.6) [3.6-11.7]	10 (7.5) [4.1-13.3]	0
Multiple imputation	6 (3.9) [1.8-8.3]	6 (4.5) [2.1-9.5]	0
K-nearest neighbor imputation	5 (3.3) [1.4-7.5]	5 (3.8) [1.6-8.5]	0
Replacement with null value	3 (2.0) [0.7-5.6]	1 (0.8) [0-4.1]	2 (10.5) [2.9-31.4]
Last value carried forward	4 (2.6) [1-6.6]	4 (3) [1.2-7.5]	0
Surrogate variable	1 (0.7) [0-3.6]	1 (0.8) [0-4.1]	0
Random forest imputation	4 (2.6) [1-6.6]	3 (2.3) [0.8-6.4]	1 (5.3) [0.3-24.6]
Categorization	3 (2) [0.7-5.6]	2 (1.5) [0.4-5.3]	1 (5.3) [0.3-24.6]
Unclear	6 (3.9) [1.8-8.3]	5 (3.8) [1.6-8.5]	1 (5.3) [0.3-24.6]
Presentation of missing data			
Not summarized	129 (84.9) [78.3-89.7]	114 (85.7) [78.8-90.7]	16 (84.2) [62.4-94.5]
Overall	6 (3.9) [1.8-8.3]	4 (3) [1.2-7.5]	2 (10.5) [2.9-31.4]
By all final model variables	3 (2) [0.7-5.6]	3 (2.3) [0.8-6.4]	0
By all candidate predictors	13 (8.6) [5.1-14.1]	11 (8.3) [4.7-14.2]	1 (5.3) [0.3-24.6]
By number of variables	1 (0.7) [0-3.6]	1 (0.8) [0-4.1]	0

†Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies can report more than one technique.

Selection of predictors

The strategy to build models was unclear in 168 out of 522 models (32.2% [95%CI 28.2-36.4]). Most models reported a data-driven approach for model building (n=192/522, 36.8% [95%CI 32.7-41.1]). One study reported the use of recursive feature elimination for model building (n=3/522, 0.6% [95%CI 0.1-1.8]). Selection of candidate predictors based on univariable predictor–outcome associations was used in 27/522 (5.2% [95%CI 3.5-7.5]) of the models. Further details on modelling strategies are presented in Table 8. Of the three studies that reported time-to-event outcomes none reported how they dealt with censoring.

Variable importance and hyperparameters

Variable importance scores show insight into how much each variable contributed to the prediction model.[21] For 316/522 (60.5% [95%CI 56.2-64.7]) models, authors did not provide these scores, while in 115/522 (22% [95%CI 18.6-25.9]) models these scores were reported without specifying the methods applied to obtain such calculations (Table 8). When reported, the mean decrease in node impurity was the most popular method (n=31/522, 5.9% [95%CI 4.1-8.4]). Hyperparameters (including default settings) were reported in 160/552 (30.7% [95%CI 26.8-34.8]) models. Strategies for hyperparameter optimization were described in 44/152 studies (28.9% [95%CI 22.3-36.3]). The most common method reported was cross-validation (n=15/152) [9.9% [95%CI 6.1-15.6]. Nine studies (n=9/152, 5.9% [95%CI 3.1-10.9]) split their dataset into a validation set for hyperparameter tuning (Table 7).

Performance metrics

Most models used measures of the area under the Receiver Operating Characteristic curve (AUC/ROC or the concordance (c)-statistic) (n=358/522, 68.6% [95%CI 64.4-72.5]) to describe the discriminative ability of the model (Table 9). A variety of methods were used to describe the agreement between predictions and observations (i.e., calibration), the most frequent being a calibration plot (n=23/522, 4.4% [95%CI 2.9-6.6]), calibration slope (n=17/522, 3.3% [95%CI 2-5.3]), and calibration intercept (n=16/522, 3.1% [95%CI 1.8-5]). However, for the large majority no calibration metrics were reported (n=494/522, 94.6% [95%CI 92.2-96.3]). Decision curve analysis was reported for two models (n=2/522, 0.4% [95%CI 0.1-1.5]).[22]

Uncertainty quantification

In 53/152 (34.9% [95% CI 22.8-42.7]) studies, discrimination was reported without precision estimates (i.e., confidence intervals or standard errors). Likewise, 7/152 (4.6% [95%CI 2.2-9.2] studies reported model calibration without precision estimates.

Table 7. Machine learning aspects in the included studies

	Total (n=152)
	n (%) [95% CI]
Data preparation†	**58 (38.2) [30.8-46.1]**
Cleaning	21 (36.2) [25.1-49.1]
Aggregation	6 (10.3) [4.8-20.8]
Transformation	6 (10.3) [4.8-20.8]
Sampling	2 (3.4) [1-11.7]
Standardization/Scaling	11 (19) [10.9-30.9]
Normalization	22 (37.9) [26.6-50.8]
Integration	0
Reduction	12 (20.7) [12.3-32.8]
Other [a]	9 (15.5) [8.4-26.9]
Data splitting	**86 (56.6) [48.6-64.2]**
Train-test set	77 (50.7) [42.8-58.5]
Train-validation-test set	9 (5.9) [3.1-10.9]
Dimensionality reduction techniques	**9 (5.9) [3.1-10.9]**
CART	1 (11.1) [0.6-43.5]
Principal component analysis	3 (33.3) [12.1-64.6]
Factor analysis	1 (11.1) [0.6-43.5]
Image decomposition	1 (11.1) [0.6-43.5]
Class imbalance†	**27 (17.8) [12.5-24.6]**
Random undersampling	4 (14.8) [5.9-32.5]
Random oversampling	5 (18.5) [8.2-36.7]
SMOTE	11 (40.7) [24.5-59.3]
RUSBoost	1 (3.7) [0.2-18.3]
Other [b]	7 (25.9) [13.2-44.7]
Strategy for hyperparameter optimization†	**44 (28.9) [22.3-36.6]**
Grid search (no further details)	5 (3.3) [1.4-7.5]
Cross-validated grid search	14 (9.2) [5.6-14.9]
Randomized grid search	1 (0.7) [0-3.6]
Cross-validation	15 (9.9) [6.1-15.6]
Manual search	1(0.7) [0-3.6]
Pre-defined values/default	3 (2) [0.7-5.6]
Bayesian optimization	2 (1.3) [0.4-4.7]
Tree-structured parzen estimator method	1(0.7) [0-3.6]
Unclear	4 (2.6) [1-6.6]

†Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies can report more than one measure.
CART - Classification And Regression Tree.
[a] This includes matching, augmentation, noise filtering, merging, splitting, binning.
[b] This includes matching, resampling, class weighting, inverse class probability.

Predictive performance

Most models achieved discriminative ability better than chance (i.e., AUC 0.5) with a median apparent AUC of 0.82 (IQR 0.75-0.90; range 0.45 to 1.00), while internally validated AUC was also 0.82 (IQR: 0.74-0.89; range 0.46 to 0.99). For external validation, the median AUC was 0.73 (IQR: 0.70-0.78, range: 0.51-0.88). For calibration and overall performance metrics, see Table 10.

Table 8. Model building of all included studies

	Total (n=522)
	n (%) [95% CI]
Selection of predictors	
Stepwise	8 (1.5) [0.7-3.1]
Forward selection	31 (5.9) [4.1-8.4]
Backward selection	5 (1) [0.4-2.4]
All predictors	72 (13.8) [11-17.1]
All significant in univariable analysis	27 (5.2) [3.5-7.5]
Embedded in learning process	192 (36.8) [32.7-41.1]
Other	19 (3.6) [2.3-5.7]
Unclear	168 (32.2) [28.2-36.4]
Hyperparameter tunning reported	
Yes	160 (30.7) [26.7-34.8]
No	283 (54.2) [49.8-58.5]
Not applicable/Unclear	79 (15.1) [12.2-18.6]
Variable importance reported	
Mean decrease in accuracy	26 (5) [3.3-7.3]
Mean decrease in node impurity	31 (5.9) [4.1-8.4]
Weights/correlation	10 (1.9) [1-3.6]
Gain information	24 (4.6) [3-6.9]
Unclear method	115 (22) [18.6-25.9]
None	316 (60.5) [56.2-64.7]
Penalization methods used	
None	481 (92.1) [89.4-94.2]
Uniform shrinkage	3 (0.6) [0.1-1.8]
Penalised estimation	27 (5.2) [3.5-7.5]
Other	11 (2.1) [1.1-3.9]

Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies can report more than one measure.
[a] Both unpenalized and penalized regression models.
[b] Only for penalized regression techniques.

Internal validation

In total, 86/152 studies (56.6% [95%CI 48.6-64.2]) internally validated their models, most often splitting the dataset into a training and test set. The train-test sets were often split randomly (n=49/86, 57% [95%CI 46.4-66.9]) and in a few studies a temporal (non-random) split was applied (n=9/86, 10.5% [95%CI 5.6-18.7]). The proportion of the data used for test sets ranged from 10% to 50% of the total dataset. Seventy studies also performed cross-validation (46.1% [95%CI 38.3-54]) with ten studies reporting nested cross-validation (6.6% [95%CI 3.6-11.7]). Out of five studies performing bootstrapping (n=5/152, 3.3% [95%CI 1.4-7.5]), one reported 250 iterations, three reported 1000 iterations and one did not report the number of iterations. For further details see Table 3.

External validation

Few studies (n=19/152, 12.5% [95%CI 8.2-18.7]) performed an external validation. Eleven studies (n=11/19, 57.9% [95%CI 36.3-76.9]) used data from independent cohorts and eight (n=8/19, 42.1% [95%CI 23.1-63.7]) used subcohorts within the main cohort to validate their developed models. From the independent cohorts, three studies (n=3/19, 15.8% [95%CI 5.5-37.6]) used data from a different country. Five studies (n=5/19, 26.3% [95%CI 11.8-48.8]) described an external validation based on temporal differences on the inclusion of participants. Seven studies (36.8% [95%CI 19.1-59]) reported differences and similarities in definitions between the development and validation data.

Model availability

Some studies shared their prediction model either as a web-calculator or worked example (n=31/152, 20.4% [95%CI 14.8-27.5]). Furthermore, in a minority of studies datasets and code were accessible through repositories, which were shared as supplemental material (n=18/152, 11.8% [95%CI 7.6-17.9]; n=13/152, 8.6% [95%CI 5.1-14.1]). Details in Table 1.

Table 9. Performance measures reported, stratified by model development and validation

		All extracted models (n=522)	
		n (%) [95% CI]	
		DEV	VAL
Calibration†			
	Calibration plot	23 (4.4) [2.9-6.6]	1 (0.2) [0.01-1.2]
	Calibration slope	17 (3.3) [2-5.3]	1 (0.2) [0.01-1.2]
	Calibration intercept	16 (3.1) [1.8-5]	1 (0.2) [0.01-1.2]
	Calibration in the large	1 (0.2) [0.01-1.2]	0
	Calibration table	1 (0.2) [0.01-1.2]	0
	Kappa	10 (1.9) [1-3.6]	0
	Observed/expected ratio	1 (0.2) [0.01-1.2]	0
	Homer-Lemeshow statistic	4 (0.8) [0.3-2.1]	0
	None	494 (94.6) [92.3-96.3]	
Discrimination			
	AUC/ AUC-ROC	349 (66.9) [62.6-70.9]	46 (8.8) [6.6-11.7]
	C-statistic	9 (1.7) [0.8-3.4]	0
	None	164 (31.4) [27.5-35.6]	
Classification †			
	NRI	9 (1.7) [0.8-3.4]	0
	Sensitivity/Recall	239 (45.8) [41.5-50.2]	30 (5.7) [4-8.2]
	Specificity	193 (37) [32.8-41.3]	22 (4.2) [2.7-6.4]
Decision-analytic †			
	Decision Curve Analysis	2 (0.4) [0.01-1.5]	0
	IDI	1 (0.2) [0.01-1.2]	0
Overall †			
	R2	14 (2.7) [1.5-4.6]	0
	Brier score	19 (3.6) [2.3-5.7]	6 (1.1) [0.5-2.6]
	Predictive values*	160 (30.7) [26.8-34.8]	10 (1.9) [1-3.6]
	AUC difference	2 (0.4) [0.01-1.5]	0
	Accuracy**	234 (44.8) [40.5-49.2]	26 (5) [3.4-7.3]
	F1-score	79 (15.1) [12.2-18.6]	0
	Mean square error	21 (4) [2.6-6.2]	0
	Misclassification rate	9 (1.7) [0.8-3.4]	0
	Mathew's correlation coefficient	5 (1) [0.4-2.4]	0
	AUPR	21 (4) [2.6-6.2]	0

† Counts are absolute numbers, with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies can report more than one performance measure.
*This includes models reporting positive predictive value as precision.
**This includes models reporting balance accuracy.
DEV, developed model; VAL, validation; AUC-ROC, area under the receiver operation characteristic curve; NRI, net reclassification index; IDI, integrated discrimination improvement; AUPR, area under the precision-recall curve; CI, confidence interval.

Table 10. Predictive performance of all extracted models*

	All extracted models (n=522)					
	Reported, n (%)	Apparent performance	Reported, n (%)	Corrected performance**	Reported, n (%)	Externally validated performance
		Median [IQR], range		Median [IQR], range		Median [IQR], range
Calibration						
Slope	11 (1.9)	1.05 [1.02 - 1.07], 0.53 to 1.46	15 (2.9)	1.3 [1-4], 0.52 to 17.6	4 (0.8)	9.9 [7.87-12.8], 5.7 to 17.6
Intercept	10 (1.9)	0.07 [0.05 - 0.12], -0.08 to 2.32	15 (2.9)	-0.01 [-1.85 - 0.15], -8.3 to 2.74	4 (0.8)	-4.5 [-5.7- -3.8], -8.3 to -3
Calibration-in-the-large	1 (0.2)	-0.008	0		0	
Observed:expected ratio	1 (0.2)	0.993	4 (0.8)	0.99 [0.98 - 1.01], 0.98 to 1.04	0	
Homer-Lemeshow	2 (0.2)	Not significant	0		0	
Pearson chi-square	1 (0.2)	Not significant	0		0	
Mean Calibration Error	4 (0.8)	0.81 [0.7 – 0.88], 0.51 to 0.99	0		0	
Discrimination						
AUC	249 (47.7)	0.82 [0.74-0.90], 0.45 to 1.00	154 (29.5)	0.82 [0.74-0.90], 0.46 to 0.99	46 (8.8)	0.82 [0.73-0.98], 0.52 to 0.97
Accuracy						
	128 (24.5)	79.8 [72.6-89.8], 44.2 to 100	117 (22.4)	81.4 [76-89.9], 17.8 to 97.5	9 (1.7)	70 [64-87], 55 to 90
Sensitivity						
	156 (29.9)	74 [58.6-87.8], 0 to 100	103 (19.7)	80 [66.3-89.7], 14.8 to 100	12 (2.3)	77.5 [63.9-83.5], 0.7 to 91
Specificity						
	122 (23.4)	82.2 [73.3-89.7], 17 to 100	80 (15.3)	83.2 [73.6-90.8], 46.6 to 100	10 (1.9)	74.4 [64.8-86.7], 42 to 90.5

*Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because some studies did not report performance measure for all models pre-specified.

**We considered corrected performance only when authors stated results as such. Otherwise, performance measures were considered apparent performance by default.

DISCUSSION

Principal findings

In this study, we evaluated the study design, data sources, modelling steps, and performance measures in studies on clinical prediction models using machine learning across. The methodology varied substantially between studies, including modelling algorithms, sample size, and performance measures reported. Unfortunately, longstanding deficiencies in reporting and methodological conduct previously seen in studies with a regression-based approach, were also extensively found in our sample of studies on machine learning models.[9,23]

The spectrum of supervised machine learning techniques is quite broad.[24,25] In this study, the most popular modelling algorithms were tree-based methods (RF in particular) and SVM. RF is an ensemble of random trees trained on bootstrapped sub-sets of the dataset.[26] On the other hand, SVM first map each data point into a feature space to then identify the hyperplane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors.[27]

Various other well-known methodological issues in prediction model research need to be further discussed. Our reported estimate on EPV is likely to be overestimated given than we were unable to calculate it based on number of parameters, and instead we used only the number of candidate predictors. A simulation study concluded that modern modelling techniques such as SVM and RF might even require 10 times more events.[28] Hence, the sample size in most studies on prediction models using machine learning remains relatively low. Furthermore, splitting datasets persists as a method for internal validation (i.e., testing), reducing even more the actual sample size for model development and increasing the risk of overfitting.[29,30] Whilst AUC was a frequently reported metric to assess predictive performance, prediction calibration or prediction error was often overlooked.[31] Moreover, a quarter of studies in our sample corrected for class imbalance without reporting recalibration, although recent research has shown that correcting for class imbalance may lead to poor calibration and thus, prediction errors.[32] Finally, therapeutic interventions were rarely considered as predictors in the prognostic models, although these can affect the accuracy and transportability of models.[33]

Variable importance scores, tuning of hyperparameters, and data preparation (i.e., data pre-processing) are items closely related to machine learning prediction models. We found that most studies reporting variable importance scores did not specify the calculation method. Data preparation steps (i.e., data quality assessment, cleaning, transformation, reduction) were often not described in enough transparent detail. Complete-case analysis remains a popular method to handle missing values

in machine learning based models. Detailed description and evaluation on how missing values were handled in our included studies has been provided elsewhere.[34] Last, only one third of models reported their hyperparameters settings, which is needed for reproducibility purposes.

Comparison to previous studies

Regression methods were not our focus (as we did not define them to be machine learning methods), but other reviews including both approaches show similar issues with methodological conduct and reporting.[12,35–37] Missing data, sample size, calibration, and model availability remain largely neglected aspects.[7,12,37–40]A review looking at the trends of prediction models using electronic health records (EHR) observed an increase in the use of ensemble models from 6% to 19%.[41] Another detailed review on prediction models for hospital readmission shows that the use of algorithms such as SVM, RF, and NN increased from none to 38% over the last five years.[10]

Strengths and limitations of this study

In this comprehensive review, we summarized the study design, data sources, modelling strategies, and reported predictive performance in a large and diverse sample of studies on clinical prediction model studies. We focused on all types of studies on clinical prediction models rather than on a specific type of outcome, population, clinical specialty, or methodological aspect. We appraised studies published almost three years ago and thus, it is possible that further improvements might have raised. However, improvements in methodology and reporting are usually small and slow even when longer periods are considered.[42] Hence, we believe that the results presented in this comprehensive review still largely apply to the current situation of studies on machine learning-based prediction models. Given the limited sample, our findings can be considered a representative rather than exhaustive description of studies on machine learning models.

Our data extraction was restricted to what was reported in articles. Unfortunately, few articles reported the minimum information required by reporting guidelines, thereby hampering data-extraction.[23] Furthermore, terminology differed between papers. For example, the term 'validation' was often used to describe tuning, as well as testing (i.e., internal validation). An issue already observed by a previous review of studies on deep learning models.[43] This shows the need to harmonize the terminology for critical appraisal of machine learning models.[44] Our data extraction form was based mainly on the items and signaling questions from TRIPOD and PROBAST. Although both tools were primarily developed for studies on regression-based prediction models, most items and signaling questions were largely applicable for studies on machine learning-based models, as well.

Implication for researchers, editorial offices, and future research

In our sample, it is questionable whether studies ultimately aimed to improve clinical care.[45] Aim, clinical workflow, outcome format, prediction horizon, and clinically relevant performance metrics received very little attention. The importance of applying optimal methodology and transparent reporting in studies on prediction models has been intensively and extensively stressed by guidelines and meta-epidemiological studies.[46–48] Researchers can benefit from TRIPOD and PROBAST, as these provide guidance on best practices for prediction model study design, conduct and reporting regardless of their modelling technique.[16,17,46,47] However, special attention is required on extending the recommendations to include areas such as data preparation, tunability, fairness, and data leakage. In this review, we have provided evidence on the use and reporting of methods to correct for class imbalance, data preparation, data splitting, and hyperparameter optimization. PROBAST-AI and TRIPOD-AI, both extensions to artificial intelligence (AI) or machine learning based prediction models are underway.[44,49] As machine learning continues to emerge as a relevant player in healthcare, we recommend researchers and editors to reinforce a minimum standard on methodological conduct and reporting to ensure further transportability.[16,17,46,47]

We identified that studies covering the general population (e.g., for personalized screening), primary care settings, and time-to-event outcomes are underrepresented in current research. Similarly, only a relatively small proportion of the studies evaluated (validated) their prediction model on a different dataset (i.e., external validation).[50] In addition, the poor availability of the developed models hampers further independent validation, an important step before their implementation in clinical practice. Sharing the code and ultimately the clinical prediction model is a fundamental step to create trustworthiness on AI and machine learning for clinical application.[51]

CONCLUSION

Our study provides a comprehensive overview of the applied study designs, data sources, modelling steps, and performance measures used. Special focus is required in areas such as handling of missing values, methods for internal validation, and reporting of calibration to improve the methodological conduct of studies on prediction models developed using machine learning techniques.

Acknowledgements
The authors would like to thank and acknowledge the support of René Spijker, information specialist. The peer-reviewers are thanked for critically reading the manuscript and suggesting substantial improvements.

Authors' contributions
Constanza L. Andaur Navarro: Conceptualization, Methodology, Investigation, Data Curation, Formal analysis, Writing - original draft, Writing - review & editing; Johanna A.A. Damen: Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision; Maarten van Smeden: Conceptualization, Writing - review & editing; Toshihiko Takada: Investigation, Writing - review & editing. Steven WJ Nijman: Investigation, Writing - review & editing; Paula Dhiman: Conceptualization, Methodology, Investigation, Writing - review & editing; Jie Ma: Investigation, Writing - review & editing; Gary S Collins: Conceptualization, Methodology, Writing - review & editing; Ram Bajpai: Investigation, Writing - review & editing; Richard D Riley: Conceptualization, Methodology, Writing - review & editing; Karel GM Moons: Conceptualization, Methodology, Writing - review & editing, Supervision; Lotty Hooft: Conceptualization, Methodology, Writing - review & editing, Supervision

Support
GSC is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is funded by the NIHR Oxford BRC. RB is affiliated to the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) West Midlands. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health and Social Care. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication.

Competing interests
None

Registration and protocol
This review was registered in PROSPERO (CRD42019161764). The study protocol can be accessed in doi:10.1136/bmjopen-2020-038832.

Availability of data, code, and other materials
Articles that support our findings are publicly available. Template data collection forms, detailed data extraction on all included studies, and analytical code are available upon reasonable request.

Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.11.015

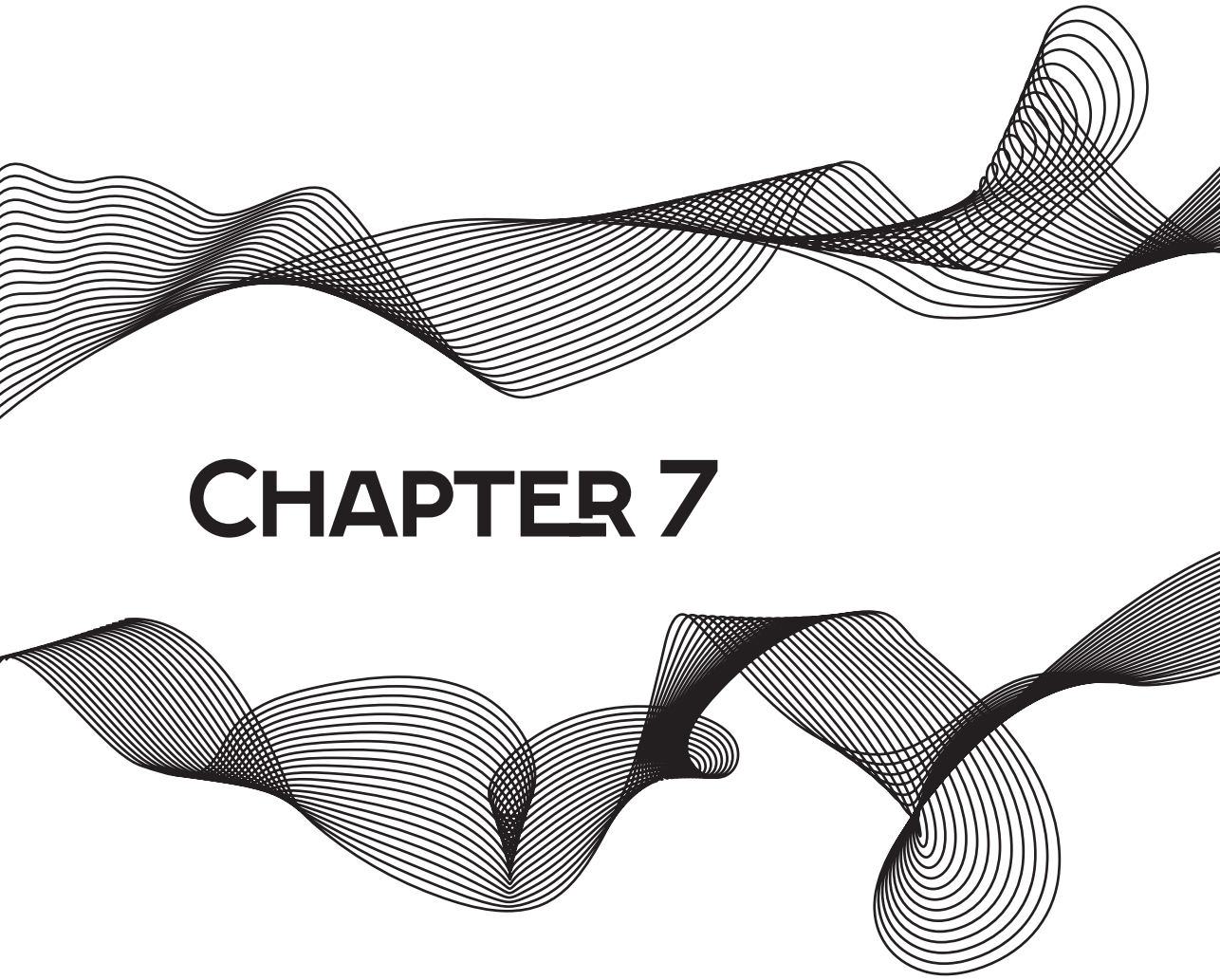
Ethical approval
Not required for this work.

REFERENCES

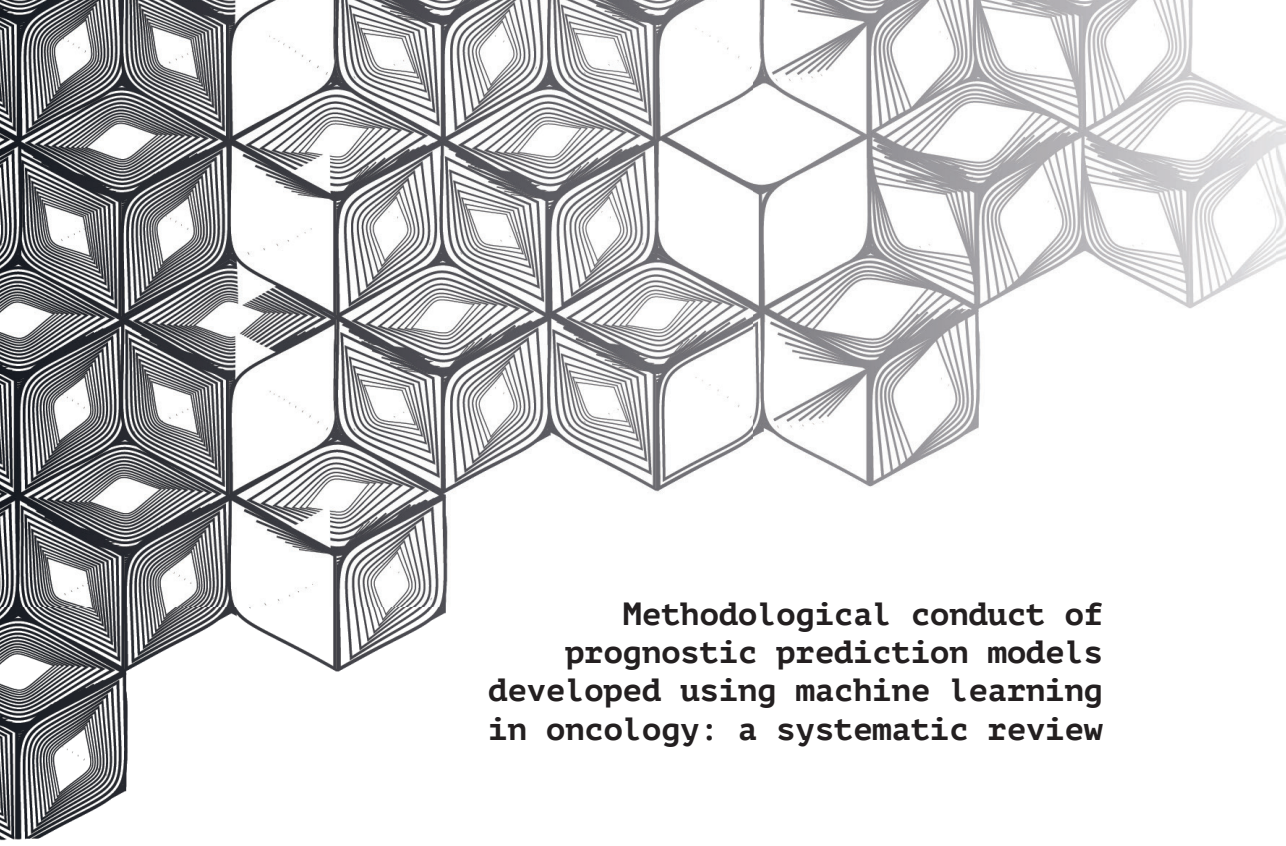
1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ*. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
2. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol*. 2021;132:142-145. doi:10.1016/j.jclinepi.2021.01.009
3. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-00333-z
4. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019;19(1). doi:10.1186/S12911-019-1004-8
5. Macleod MR, Michie S, Roberts I, et al. Biomedical research: Increasing value, reducing waste. *Lancet*. 2014;383(9912):101-104. doi:10.1016/S0140-6736(13)62329-6
6. Jong Y de, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, Diepen M van. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). *Nephrology*. 2021:1-9. doi:10.1111/NEP.13913
7. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/BMJ.I2416
8. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103. doi:10.1186/1741-7015-9-103
9. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281. doi:10.1136/bmj.n2281
10. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: A systematic review of methods. *Comput Methods Programs Biomed*. 2018;164:49-64. doi:10.1016/j.cmpb.2018.06.006
11. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0229-3
12. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol*. 2022;22(1):1-16. doi:10.1186/s12874-021-01469-6
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. 2021;372. doi:10.1136/bmj.n71
14. Andaur Navarro CL, Damen JAAG, Takada T, 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(11):1-6. doi:10.1136/bmjopen-2020-038832
15. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med*. 2014;11(10). doi:10.1371/journal.pmed.1001744
16. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55. doi:10.7326/M14-0697
18. 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(12). doi:10.2196/jmir.5870
19. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. Published online 2002:321-357.
20. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans Syst Man, Cybern Part ASystems Humans*. 2010;40(1):185-197. doi:10.1109/TSMCA.2009.2029559
21. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*. 2019;20:1-81.
22. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol*. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038
23. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22(1):12. doi:10.1186/s12874-021-01469-6
24. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA - J Am Med Assoc*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391
25. Hastie T, Tibshirani R, Friedman J. *The Elements of Statstical Learning: Data Mining, Inference, and Prediction*. 2nd editio. Springer-Verlag; 2009.
26. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
27. Scholkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press; 2001.
28. Ploeg T Van Der, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry : a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137. doi:10.1186/1471-2288-14-137
29. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans RM, Vergouwe Y, Habbema J. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Proc ICED 2007, 16th Int Conf Eng Des*. 2007;DS 42:774-781.
30. Steyerberg E, Jr FH. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005
31. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1-7. doi:10.1186/s12916-019-1466-7
32. Goorbergh R van den, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Informatics Assoc*. 2022;00(0):1-10.
33. Pajouheshnia R, Damen JAAG, Groenwold RHH, Moons KGM, Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic Progn Res*. 2017;1(1):1-10. doi:10.1186/s41512-017-0015-0
34. Nijman SWJ, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol*. 2022;142:218-229. doi:10.1016/j.jclinepi.2021.11.023
35. Dhiman P, Ma J, Andaur Navarro C, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol*. 2021;138:60-72. doi:10.1016/j.jclinepi.2021.06.024
36. Heus P, Reitsma JB, Collins GS, et al. Transparent Reporting of Multivariable Prediction Models in Journal and Conference Abstracts: TRIPOD for Abstracts. *Ann Intern Med*. 2020;173(1):43. doi:10.7326/M20-0193
37. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
38. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prog-

nosis of covid-19: Systematic review and critical appraisal. *BMJ*. 2020;369. doi:10.1136/bmj.m1328
39. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. *BMC Med*. 2018;16(1):1-12. doi:10.1186/s12916-018-1099-2
40. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Med*. 2012;9(5). doi:10.1371/journal.pmed.1001221
41. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *J Am Med Informatics Assoc*. 2022;00(0):1-7. doi:10.1093/jamia/ocac002
42. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
43. Kim DW, Jang HY, Ko Y, et al. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. *PLoS One*. 2020;15(9 September):1-10. doi:10.1371/journal.pone.0238908
44. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(e048008):1-7. doi:10.1136/bmjopen-2020-048008
45. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368:1-12. doi:10.1136/bmj.l6927
46. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
47. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
48. Damen JAAG, Debray TPA, Pajouheshnia R, et al. Empirical evidence of the impact of study characteristics on the performance of prediction models: A meta-epidemiological study. *BMJ Open*. 2019;9(4):1-12. doi:10.1136/bmjopen-2018-026160
49. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393. doi:10.1016/S01406736(19)302351
50. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5
51. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? *J Am Med Informatics Assoc*. 2019;26(12):1651-1654. doi:10.1093/JAMIA/OCZ130



CHAPTER 7



Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review

Paula Dhiman
Jie Ma
Constanza L Andaur Navarro
Benjamin Speich
Garrett Bullock
Johanna AA Damen
Lotty Hooft
Shona Kirtley
Richard D Riley
Ben Van Claster
Karl GM Moons
Gary S Collins

BMC Medical Research Methodology 2022:22(1):101

ABSTRACT

Background. Describe and evaluate the methodological quality of prognostic prediction models developed using machine learning methods in oncology.

Methods. We conducted a systematic review in MEDLINE and Embase between 01/01/2019 and 05/09/2019, for studies developing a prognostic prediction model using machine learning methods in oncology. We used the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement, Prediction model Risk Of Bias ASsessment Tool (PROBAST) and CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) to assess the methodological quality of included publications. Results were summarised by modelling type: regression, non-regression-based and ensemble machine learning models.

Results. Sixty-two publications met inclusion criteria developing 152 models across all publications. Forty-two models were regression-based, 71 were non-regression-based and 39 were ensemble models. A median of 647 individuals (IQR: 203 to 4059) and 195 events (IQR: 38 to 1269) were used for model development, and 553 individuals (IQR: 69 to 3069) and 50 events (IQR: 17.5 to 326.5) for model validation. A higher number of events per predictor was used for developing regression-based models (median: 8, IQR: 7.1 to 23.5), compared to alternative machine learning (median: 3.4, IQR: 1.1 to 19.1) and ensemble models (median: 1.7, IQR: 1.1 to 6). Sample size was rarely justified (n=5/62; 8%). Some or all continuous predictors were categorised before modelling in 24 studies (39%). 46% (n=24/62) of models reporting predictor selection before modelling used univariable analyses, and common method across all modelling types. Ten out of 24 models for time-to-event outcomes accounted for censoring (42%). A split sample approach was the most popular method for internal validation (n=25/62, 40%). Calibration was reported in 11 studies. Less than half of models were reported or made available.

Conclusions. The methodological conduct of machine learning based clinical prediction models is poor. Guidance is urgently needed, with increased awareness and education of minimum prediction modelling standards. Particular focus is needed on sample size estimation, development and validation analysis methods, and ensuring the model is available for independent validation, to improve quality of machine learning based clinical prediction models.

INTRODUCTION

Many medical decisions across all clinical specialties are informed by clinical prediction models [1–7], and they are often used in oncology, for example to assess risk of developing cancer, inform cancer diagnosis, predict cancer outcomes and prognosis, and guide treatment decisions.[8–13] Clinical prediction models use individual-level data, such as demographic information, clinical characteristics, and biomarker measurements, to estimate the individualised risk of existing or future clinical outcomes.

However, compared to the number of prediction models that are developed, very few are used in clinical practice and many models contribute to research waste.[14–17] This problem has been further exacerbated with the rapidly growing use of machine learning to develop clinical prediction models as a class of models perceived to provide automated diagnostic and prognostic risk estimation at scale. This has led to the production of a spiralling number of models to inform diagnosis and prognosis including in the field of oncology. Machine learning methods include neural networks, support vector machines and random forests.

Machine learning is often portrayed to offer more flexible modelling, the ability to analyse 'big', non-linear and high dimensional data, and modelling complex clinical scenarios.[18,19] Despite this, machine learning methods are often applied to small and low dimensional settings. [20,21] However, many perceived advantages of machine learning (over traditional statistical models like regression) to develop prediction models have not materialised into patient benefit. Indeed, many studies have found no additional performance benefit of machine learning over traditional statistical models.[22–27]

A growing reason and concern resulting in their lack of implementation in clinical practice leading to patient benefit is the completeness of reporting, methodological quality and risk of bias in studies using machine learning methods. [22,25,26,28,29] Similarly, many regression-based prediction models have also not been implemented in clinical practice due to incomplete reporting and failure to follow methodological recommendations, often resulting in poor quality studies and models due to using sample sizes that are too small, risk of overfitting and lack of external validation of developed models.[14,30–35]

However, there is a lack of information about the methodological conduct of clinical prediction models developed using machine learning methods within oncology. We therefore aim to describe and evaluate the methodological conduct of clinical prediction models developed using machine learning in the field of oncology.

METHODS

We conducted a systematic search and review of prognostic model studies that use machine learning methods for model development, within the oncology clinical field. We excluded imaging and lab-based studies to focus on low dimensional, low signal and high noise clinical data settings. Machine learning was defined as a subset of artificial intelligence allowing for machines to learn from data with and without explicit programming.

The boundaries between machine learning and statistical, regression-based methods of prediction is often unclear and artificial, often seen as a cultural difference between methods and fields.[36] We therefore included studies that typically identify as machine learning, such as random forests and neural networks, and included any study in which the modelling method was declared as machine learning by authors of the included studies. For example, we included studies using logistic regression if they were explicitly labelled by the authors as machine learning, otherwise it was excluded.

Protocol registration and reporting standards

This study is reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline.[37] We registered this umbrella review with PROSPERO (ID: CRD42019140361)[38] that comprises of four distinct studies to evaluate (1) completeness of reporting, (2) risk of bias, (3) methodological conduct, and (4) spin over-interpretation.

Information sources

We searched the MEDLINE (via OVID) and Embase (via OVID) medical literature databases for published clinical prediction modelling studies that use machine learning methods for model development, within the oncology clinical field. We searched for publications from 1 January 2019 to 5 September 2019, the date the searches were executed.

The search strategy comprised of three specific groups of search terms specific focussing on machine learning models, cancer, and prediction. Relevant Mesh and EMTREE headings were included as were free-text terms, searched in the title, abstract or keyword fields. We used general and specific machine learning model search terms such as “machine learning”, “deep learning”, “neural networks”, “random forest” or “support vector machine”. Cancer search terms included “cancer”, “tumour” or “malignancy”. General prediction and specific model performance search terms included “prediction”, “prognosis”, “discrimination”, “calibration” or “area under the curve”. The three specific groups of terms were combined with ‘AND’ to retrieve the final results set. The search was limited to

retrieve studies published in 2019 only to ensure that a contemporary sample of studies were assessed in the review. The Embase search strategy was also limited to exclude conference abstract publications. No other limits were applied to the search, and we also did not limit our search to specific machine learning methods so we could describe the types of models being used to develop prediction models in low dimensional setting and using clinical characteristics. Search strategies for both databases were developed with an information specialist (SK). The full search strategies for both included databases are provided in Supplementary tables 1 and 2.

Eligibility criteria

We included published studies developing a multivariable prognostic model using machine learning methods within oncology in 2019. A multivariable prognostic model was defined as a model that uses two or more predictors to produce an individualised predicted risk (probability) of a future outcome.[39,40] We included studies predicting for any patient health-related outcome measurement (e.g., binary, ordinal, multinomial, time-to-event, continuous) and using any study design and data source (e.g., experimental studies such as randomised controlled trials, and observational studies such as prospective or retrospective cohort studies, case-control studies or studies using routinely collected data or e-health data).

We excluded studies that did not report the development of a multivariable prognostic model and studies that only validated models. We excluded diagnostic prediction model studies, speech recognition or voice pattern studies, genetic studies, molecular studies, and studies using imaging or speech parameters, or genetic or molecular markers as candidate predictors. Prognostic factor studies primarily focused on the association of (single) factors with the outcome were also excluded. Studies were restricted to the English language and to primary research studies only. Secondary research studies, such as reviews of prediction models, conference abstracts and brief reports, and preprints were excluded.

Study selection, data extraction and management

All retrieved publications were imported into Endnote reference software where they were de-duplicated. Publications were then imported into Rayyan web application (www.rayyan.ai) where they were screened. [41,42]

Two independent researchers (PD, JM) screened the titles and abstracts of the identified publications. Two independent researchers, from a combination of five reviewers (PD, JM, GB, BS, CAN) reviewed the full text for potentially eligible publications and extracted data from eligible publications. One researcher screened and extracted from all publications (PD) and four researchers collectively screened and extracted from the same articles (JM, GB, BS, CAN). Disagreements were discussed and adjudicated by a sixth reviewer (GSC), where necessary.

To reduce subjectivity, the data extraction form to assess the methodological conduct was developed using formal and validated tools: the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guideline, the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) and the Prediction model Risk Of Bias ASsessment Tool (PROBAST).[39,40,43–45] We then added specific machine learning items at the study design and analysis levels.

The form was piloted among all the five reviewers using five eligible publications [46]. Results of the pilot were discussed, and data extraction items were clarified amongst all reviewers to ensure consistent data extraction. All reviewers had expertise in the development, validation, and reviewing of prediction model studies using regression-based and machine learning methods. The data extraction form was implemented using Research Data Capture (REDCap) software.[47]

Data items

Descriptive data was extracted on the overall publication, including items for cancer type, study type, data source/study design, target population, type of prediction outcome, number and type of machine learning models used, setting, intended use and aim of the clinical prediction model. The TRIPOD, CHARMS and PROBAST guidance informed methodological items for extraction, including sample size calculation or justification, sampling procedure, blinding of the outcome and predictors, methods to address missing data, number of candidate predictors, model building strategies, methods to address censoring, internal validation methods and model performance measures (e.g. discrimination, calibration).[39,40,43–45]

Items for the results of each developed model were also extracted, including sample size (and number of events), and model discrimination and calibration performance results. For discrimination, we extracted the area under the receiver operating characteristic curve (AUC), i.e. the c-index (or c-statistic). For calibration, we extracted how this was evaluated (including whether the calibration slope and intercept were assessed), and whether a calibration plot with a calibration curve was presented. Items were extracted for the development and external validation (where available) of the models. We included additional items to capture specific issues associated with machine learning methods, such as methods to address class imbalance, data pre-processing, and hyperparameter tuning.

Summary measures and synthesis of results

Findings were summarised using descriptive statistics and visual plots, alongside a narrative synthesis. Sample size was described using median, interquartile range (IQR) and range. The number of events reported in studies was combined with the reported number of candidate predictors to calculate the events per predictor.

Analysis and synthesis of data was presented overall and by modelling type (regression-based, non-regression based and ensemble machine learning models). Ensemble models were defined models using a combination of different machine learning methods, including models where bagging or boosting was applied to a machine learning model (e.g., random forests, boosted random forests and boosted Cox regression). As we wanted to identify themes and trends in the methodological conduct of machine learning prediction models, we did not evaluate the nuances of each modelling approach and kept our evaluations at the study design and analysis levels.

Results for discrimination (AUC) and calibration (calibration slope and intercept) were summarised for the developed and validated machine learning models. Data was summarised for the apparent performance, internal validation performance, optimism-corrected performance, and the external validation performance.

All analyses were carried out in Stata v15.[48]

RESULTS

2922 unique publications published between 1 January 2019 and 5 September 2019 were retrieved from MEDLINE and Embase. Title and abstract screening excluded 2729 publications and full text screening excluded a further 131 publications that did not meet the eligibility criteria. Sixty-two publications were included in our review, of which 77% (n=48) were development only studies and 23% (n=14) were development and external validation studies (Figure 1). Study characteristics of included studies are presented in Supplementary table 3.

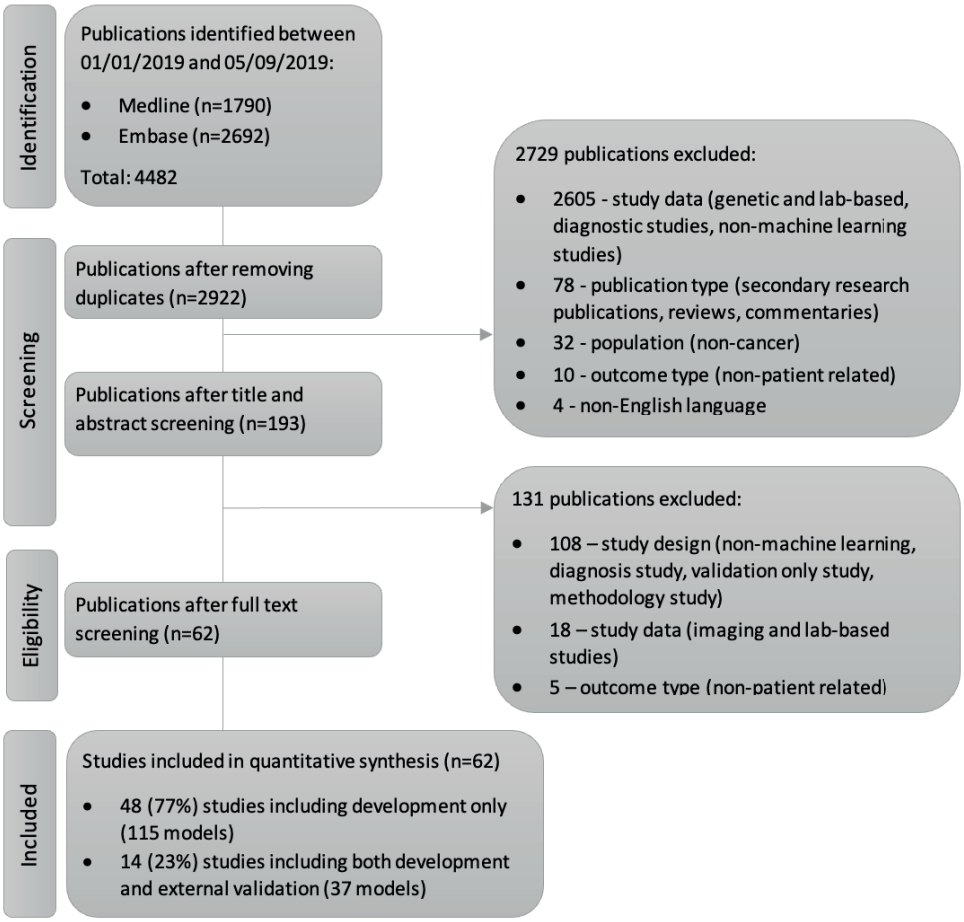


Figure 1. PRISMA flow diagram of studies included in the systematic review.

Model characteristics

A total of 152 prediction models were developed in the 62 publications. 115 (76%) models were from development-only studies and 37 (24%) were from development and validation studies. Overall, a median of two prediction models were developed per publication [range: 1-6] (Table 1). Classification trees (classification and regression trees and decision trees) (n=28, 18%), logistic regression (n=27, 18%), random forest (including random survival forest) (n=23, 15%), neural networks (n=18, 12%) and support vector machines (n=12, 8%) were the most prevalent machine learning methods used. Thirty-nine models were developed using ensemble methods. Rationale for choice of machine learning method was provided for fewer than half of the models (n=66/152, 43%).

Table 1. Model type of the 152 models developed in the 62 included publications.

Model characteristics	All models (n=152)
	n (%)
Regression-based models	**42 (28)**
Logistic regression	26
Cox regression	7
Linear regression	3
LASSO (Logistic regression)	1
LASSO (Cox regression)	1
LASSO (model not specified)	3
Best subset regression with leave-out cross-validation	1
Non-regression-based models	**71 (47)**
Neural network (including deep learning)	18
Classification tree (e.g., CART, decision tree)	28
Support vector machine	12
Naive Bayes	6
K nearest neighbours	3
Other*	4
Ensemble models	**39 (26)**
Random forest (including random survival forest)	23
Gradient boosting machine	8
RUSBoost - boosted random forests	1
Bagging with J48 selected by Auto-WEKA	1
CoxBoost - boosted Cox regression	1
XGBoost: exTreme Gradient Boosting	1
Gradient boosting machine and Nystroem, combined using elastic net	1
Adaboost	1
Bagging, method not specified	1
Partitioning Around Medoid algorithm and complete linkage method	1
Median number of models developed per study [IQR], range	2 [1-4], 1-6

CART - Classification And Regression Tree; LASSO - Least Absolute Shrinkage and Selection Operator

*Other includes voted perceptron; fuzzy logic, soft set theory and soft set computing; hierarchical | clustering model based on the unsupervised learning for survival data using the distance matrix of survival curves; Bayes point machine

Study design features

Data source, sampling, treatment details and blinding

Models were mainly developed using registry data (n=21/62, 34%) and validated using retrospective cohorts (n=4/14, 29%). Consecutive sampling was specified in only eight studies (13%) [49–56] random sampling was used in one study [57] and one study sampled individuals by screening their entire database for eligible individuals. [58] For most studies, however, sampling methods were not reported (n=52/62, 85%). Details of treatments received by patients at baseline were described during development in 53% of studies (n=33/62), compared to 36% during validation (n=5/14).

Blinding of predictor assessment to the outcome is needed to ensure predictors are not influenced by assessors and is especially important for predictors with subjective interpretation (e.g., patient reported outcome measures). However, only seven studies reported blinding predictor assessment to the outcome during model development (n=7/62; 11%) [59–65] and two reported for model external validation (n=2/62; 3%).[61,63] No studies reported blinding predictors assessment from other predictors during development and validation.

Candidate predictors and sample size

Nine studies provided rationale for their choice of candidate predictors (e.g., based on previous research) [60,61,63,66–71] and one study forced a-priori predictors during model development [72] (Table 2). Fifty-six studies (90%) clearly reported their candidate predictors and a median of 16 candidate predictors were considered per study (IQR: 12 to 26, range: 4-33788). Continuous candidate predictors were included in all studies, except one study for which it was unclear.

Categorisation of continuous predictors results in a loss of information and is discouraged for prediction modelling research. [73] However, all continuous predictors were categorised before modelling for nearly a third of models from 24 studies (n=44/152 models, 29%; n=24/62 studies, 39%). For 35 models from 25 studies continuous predictors were implicitly categorised based on the modelling method used (e.g., random forests, CART) (n=35/152 models, 23%; n=25/62 studies, 40%).

A more acceptable approach to handle continuous predictors (for approaches that are not inherently based on categorisation as part of the method) is to assess the linearity assumption with the outcome and to model them non-linearly. Investigation into nonlinearity of predictors was explicitly reported in the methods for only two models (one study), a logistic regression model which included 'interactions between variables and non-linearities' and a support vector machine that included 'different kernels (linear, polynomial and radial) and hyperparameters' in its grid search to 'fine tune the model'. [74] For 33 models from 23 studies, nonlinearity of continuous predictors was considered implicit to modelling method used (e.g., neural networks,

support vector machines and ensemble models), unless categorisation before modelling was specified (n=33/152 models, 22%; n=23/62 studies, 37%). A further eight models (three studies) also implicitly handled nonlinearity of continuous predictors in addition to some continuous predictors being categorised before modelling. For 28 models from 19 studies, continuous predictors were assumed to have a linear relationship with the outcome (n=28/152 models, 18%; n=19/62 studies, 31%). A further two models (one study) also categorised some predictors before modelling.

Methods to categorise predictors were also often unclear (n=65/85, 80%). Methods for categorisation included clinically informed cut points (n=3 studies) [6,75,76], percentiles (n=4 studies) [6,63,70,77], arbitrary dichotomisation (n=3 studies) [63,78,79] and other data driven methods that included classification and regression trees, Monte Carlo simulation (authors report that 'Monte Carlo simulation [was used] to evaluate multiple parameters by accounting for all possible dichotomous cut-offs and interactions between the inputted variables') and fuzzification (n=3 studies).[67,80,81]

Table 2. Methods for predictor selection before and after modelling and hyperparameter tuning for 152 developed clinical prediction models, by modelling type.

	All (n=152)	Regression-based models (n=42)	Non-regression-based models (n=71)	Ensemble models (n=39)
	n (%)	n (%)	n (%)	n (%)
Predictor selection (before modelling) reported	**52 (34)**	**20 (48)**	**23 (32)**	**9 (23)**
A-priori	5	3	1	1
No selection before modelling	3	1	2	-
Univariable	24	12	8	4
Clinically relevant and available data	1	-	1	-
Dropout technique at input layer	1	-	1	-
Random forest with RPA	9	1	6	2
Other modelling approach*	9	3	4	2
Predictor selection (during modelling) reported	**63 (41)**	**25 (59)**	**27 (38)**	**11 (28)**
Stepwise	6	4	2	-
Forward selection	6	5	-	1
Backward elimination	5	3	2	-
Full model approach (no selection)	11	4	5	2
Feed forward/backpropagation	5	-	5	-
Recursive partitioning analysis	7	-	7	-
LASSO	5	5	-	-
Gini index (minimised)	7	1	4	2
Cross validation	4	2	-	2
Other**	7	1	2	4
Hyperparameter tuning methods reported	**31 (21)**	**4 (10)**	**15 (23)**	**12 (31)**
Cross validation	19	4	7	8
Grid search (no further details provided)	6	-	4	2
Max tree depth	2	-	1	1
Adadelta method	2	-	2	-
Default software values	2	-	1	1

RPA=recursive partitioning analysis, LASSO= Least Absolute Shrinkage and Selection Operator

*Modelling approaches included support vector machine, logistic regression, Cox regression, best subset linear regression, decision tree, meta-transformer (base algorithm of extra trees)

** Other includes change in unspecified performance measure, stochastic gradient descent, function, aggregation of bootstrapped decision trees and Waikato Environment for Knowledge Analysis for development-only studies, and hyperbolic tangent function, greedy algorithm for all models and using final chosen predictors from comparator model

Five studies calculated or provided rationale for their sample size for model development and were all based on flawed methodology.[82] This included, one study used 10 events per variable when developing a logistic regression model and a neural network[49], and another study used estimation of a relative hazard ratio between prognostic groups to calculate their sample size.[83] Two studies considered their sample size restricted by the size and availability of the existing data they were using (one randomised controlled trial[84] and one cohort study[66]) and one study justified sample size based on a time interval (e.g., consecutive adult patients over a 2-year period to allow a sufficient sample size for randomization to the training and validation data sets)[54] One study reported traditional statistical sample size calculations are not applicable as 'CART analysis generates nonparametric, predictive models'.[70] Two studies calculated or provided rationale for their sample size for model validation. One study considered their sample size restricted by the size and availability of existing data they were using (randomised controlled trial [84]), and one study based their sample size on a power calculation (but details were not provided).[49]

Overall, a median of 647 individuals (IQR: 203 to 4059, range: 20 to 582398) and 195 events (IQR: 38 to 1269, range: 7 to 45979) was used for model development, and 553 individuals (IQR: 69 to 3069, range: 11 to 836659) and 50 events (IQR: 17.5 to 326.5, range: 7 to 1323) for model validation. The study size informing model development was lower in development-only studies (median: 155 events, IQR: 38 to 392, range: 7 to 10185), compared to development with validation studies (median: 872 events, IQR: 41.5 to 18201, range: 22 to 45797). A higher proportion of individuals with the outcome event were found in the development of regression-based models (median: 236 patients, IQR: 34 to 1326, range: 7 to 35019), compared to non-regression-based machine learning (median: 62, IQR: 22 to 1075, range: 7 to 45797) and ensemble models (median: 37, IQR: 22 to 241, range: 8 to 35019) (Table 3).

Combining the number of candidate predictors with number of events used for model development, a median 7.4 events were available per predictor (IQR: 1.7 to 15.2, range: 0.2 to 153.6) for development only studies and 49.2 events per predictor for development with validation studies (IQR: 2.9 to 2939.1, range 1.0 to 5836.5). A higher number of events per predictor was used for developing regression-based models (median: 8, IQR: 7.1 to 23.5, range: 0.2 to 5836.5), compared to alternative machine learning (median: 3.4, IQR: 1.1 to 19.1, range: 0.2 to 5836.5) and ensemble models (median: 1.7, IQR: 1.1 to 6, range: 0.7 to 5836.5). The distribution of the events per predictor, by modelling type, is provided in Supplementary figures 1 and 2.

Table 3. Sample size and number of candidate predictors informing analyses for 152 developed models, by modelling type.

	Regression-based models (n=42)		Non-regression-based models (n=71)		Ensemble models (n=39)	
	Reported, n (%)	Median [IQR], range	Reported, n (%)	Median [IQR], range	Reported, n (%)	Median [IQR], range
Total sample size						
Model development	42 (100)	561 [203 to 2822], 20 to 582398	70 (99)	447 [156 to 11901], 20 to 582398	39 (100)	768 [203 to 1599], 20 to 582398
Internal validation*	20 (48)	122 [82 to 228], 47 to 291200	35 (49)	145 [90 to 492], 47 to 291200	24 (62)	162 [97 to 1510], 67 to 291200
External validation	12 (29)	511 [67 to 2300], 11 to 836659	14 (20)	793 [59 to 1675], 11 to 836659	11 (28)	313 [229 to 836659], 11 to 836659
Number of events						
Model development	20 (48)	236 [34 to 1326], 7 to 35019	37 (52)	62 [22 to 1075], 7 to 45797	10 (26)	37 [22 to 241], 8 to 35019
Internal validation*	2 (5)	41 [21 to 61], 21 to 61	3 (4)	61 [21 to 62], 21 to 62	1 (3)	61
External validation	8 (19)	81 [18 to 327], 7 to 513	11 (15)	19 [7 to 513], 7 to 1323	5 (13)	81 [81 to 81], 7 to 513
No. candidate predictors	38 (90)	21 [15 to 34], 6 to 33788	64 (90)	16 [12 to 25], 5 to 33788	36 (92)	25 [14 to 37], 4 to 33788
Events per predictor**	20 (48)	8.0 [7.1 to 23.5], 0.2 to 5836.5	35 (49)	3.4 [1.1 to 19.1], 0.2 to 5836.5	10 (26)	1.7 [1.1 to 6.0], 0.7 to 5836.5

* Combines all internal validation methods, e.g., split sample, cross validation, bootstrapping.

**Events per predictor for model development

Validation procedures

When internally validating a prediction model, using the random split sample is not efficient use of the available data as it reduces the sample size available for developing the prediction model more robustly.[39,44] However, a split sample approach was the most popular method to internally validate the developed models (n=25/62, 40%).

Resampling methods, such cross-validation and bootstrapping are preferred approaches as they use all the data for model development and internal validation [39,44]. Bootstrapping was used in seven studies (11%)[61,63,77,79,85–87], and cross-validation in 15 studies (24%). [49–51,53,57,71,74,76,88–94] Four studies used a combination of approaches; one study used split sample and bootstrapping [95], two studies used split sample and cross-validation [64,96], and one study used cross-validation and bootstrapping [97]. For 11 studies, internal validation methods were unclear (18%). [65,70,75,80,83,84,98–102]

Of the 14 development with validation (external) studies, two used geographical validation[49,90], three used temporal validation [63,71,103] and 9 used independent data that was geographically and temporally different from the development data to validate their models.[58,61,69,75,80,84,86,93,95] Seven studies (50%) reported differences and similarities in definitions between the development and validation data. [58,61,69,71,75,84,90]

Analysis methods

Missing data and censoring

Handling of missing data was poor. The assumed mechanism for missingness was not reported in any study. Using a complete case analysis to handle missing data, not only reduces the amount of data available to develop the prediction model but may also lead to biased results with an unrepresentative sample of the target population [104–106] However, nearly half of studies performed a complete case analysis (n=30/62, 48%), of which 87% of studies (n=26/30, 87%) excluded missing data (outcome or predictor) as part of study eligibility criteria. For 12 of the studies reporting the amount of missing data excluded as part of the study eligibility criteria (n=12/62, 19%), a median of 11.1% (IQR: [4.0-27.9], range: 0.5-57.8) of individuals were excluded from the data prior to analysis.[65,71,76–78,81,83,89,99,102,107,108]

For six studies (n=6/62, 10%), mean, median, or mode imputation was used (for three studies this was in addition to exclusion of missing data as part of the study eligibility criteria).[51,56,58,76,102,108] For five studies (n=5/30, 17%) multiple imputation was used (of which one was used in addition to exclusion of missing data as part of the study eligibility criteria)[50,60,66,96,107], including one study using missForest imputation.[96] Procedure methods for multiple imputation was not appropriately described. An imputation threshold was specified in two studies, which only imputed data if missing data was less than 25% and 30%, respectively. [60,96] One study specified the number of repetitions for the multiple imputation.[50] Two studies used

subsequent follow up data and another study used a k-nearest neighbour algorithm.[65,95]

Missing data in the development data was presented by all or some candidate predictors in 13 studies (n=13/62, 20%). Two studies (out of 14) presented missing data for all predictors during validation.

Information regarding loss to follow up and censoring was rarely reported. Only 14 studies explicitly mentioned methods to handle loss to follow-up (n=14/62 studies, 17%), of which six studies excluded patients that were lost to follow up [63,77,86,89,98,107], and one study reported that the 'definition of treatment failure does not capture patients lost to follow-up due to future treatments at other institutions or due to the cessation of treatment for other reasons'.[56] For the remaining seven studies, patients who were lost of follow up were included in the study and outcome definition[53,65,67,83,90,100,109]. For example, Hammer et al reported that 'if no event of interest had occurred, patients were censored at the time of last documented contact with the hospital'.[83]

Eleven studies developed 24 models for a time to event outcome(n=11/62 studies, 18%; n=24/152 models, 16%); these were seven Cox regression models, one logistic regression model, one linear regression model, two neural networks, three random forests (including two random survival forests), four gradient boosting machines, one decision tree, two naïve bayes algorithms, one hierarchical clustering model based on the unsupervised learning for survival data using the distance matrix of survival curves, and two ensemble models (CoxBoost and Partitioning Around Medoid algorithm). Of these, only 10 models explicitly accounted for censored observations (n=10/24, 42%).

Data pre-processing, class imbalance

Only two studies assess collinearity between predictors (3%).[65,69] Nine studies used data pre-processing techniques. One study reduced data variables using automated feature selection [56] and seven studies transformed and/or standardised their predictors (including normalisation).[49,57,58,84,92,95,110] One used one-hot coding to transform categorical data and create dummy predictors in addition to predictor standardisation. [58] One study inappropriately used propensity score to obtain comparable matched groups between events and non-events.[111]

Class imbalance was examined in 19 models (from six development-only studies). One study used Synthetic Minority Oversampling TEchnique (SMOTE) to generate synthetic samples on the minority (positive) class using K-nearest neighbourhood graph[88], another study also used oversampling on the minority (dead) class to balance the number of 'alive and 'dead' cases.[107] Undersampling was used in two studies.[92,102] For two studies, methods to address class imbalance was

unclear and only described 'addressing class imbalance during hyperparameter tuning' [72] and using '5-fold cross validation'.[51] The four studies using oversampling and undersampling methods to address class imbalance failed to then examine calibration or recalibrate their models which would be miscalibrated given the artificial event rate created using these approaches.

Predictor selection, model building and hyperparameter tuning

Univariable and multivariable predictor selection before model building can lead to biased results, incorrect predictor selection for modelling and increased uncertainty in model structure. [112–115] However, methods for predictor selection before modelling were not reported for 66% of models (n=100/152), and of the 52 models that did report predictor selection before modelling, 24 used univariable screening selection to select predictors (46%), and for 18 models, predictors were selected before modelling by using other modelling approaches (35%), for example a multivariable logistic regression was developed, and predictors retained in this model were then entered into a random forest.

Methods for predictor selection during modelling were reported for 41% of developed models (n=63/152). Forward selection, backward elimination and stepwise methods were most commonly used (n=17/63, 27%) and were predominantly for regression-based machine learning models, with only five non-regression machine learning and ensemble model using them. Seven non-regression machine learning models used recursive partitioning and seven models (overall) were based on minimising the Gini index (13%). Only seven models (three regression based, three non-regression based machine learning models and one ensemble model) explicitly planned assessment of interactions.[65,74,80,93]

Thirty-two models reported hyperparameter tuning methods. Most of these models (n=19/32, 59%) used cross-validation (14 used k-fold, two used repeated k-fold and for three models it cross-validation type was unclear), including four regression-based machine learning models. Six non-regression machine learning and ensemble models used grid search for hyperparameter tuning but did not provide any further details (e.g., one study stated that 'an extensive grid search was applied to find the parameters that could best predict complications in the training sample' [78]).

Model performance

Overall fit of the developed model was reported for three studies (two used the Brier Score and one used R-squared). Model discrimination was reported in 76% (n=47/62) of all studies. Discrimination (i.e., c-statistic, c-index) was reported in all studies predicting a binary or time-to-event (survival) outcome. Three studies predicting a time-to-event outcome (n=11 models) incorrectly calculated discrimination and

used an approach which does not account for censored observations. The root mean square log error was reported for the one study predicting a continuous (length of stay) outcome.

Model calibration was only reported in 18% (n=11/62) of studies. Of these, 10 studies presented a calibration plot, including four studies that also reported estimates of the calibration slope and intercept. One study reported the Hosmer Lemeshow test, which is widely discouraged as a measure of calibration as it provides no assessment of the direction or magnitude of any miscalibration. [39]

Of the 11 studies reporting calibration, three studies modelled for a time to event outcome. One study presented 3- and 5-year survival calibration plots [86], one study presented a linear regression and plot of the predicted and actual survival time [52], and one study presented a 1-year calibration plot.[77]

Other performance measures were reported in 69% of studies (n=43/62), which predominantly included classification measures such as sensitivity, specificity, accuracy, precision and F1 score (n=35/43, 81%). For these classification measures, seven reported the associated cut-off values.

Three studies reported results of a net benefit and decision curve analysis, and one study reported the net reclassification index and integrated discrimination improvement. Measures of error were reported in four studies and included mean per class error; absolute relative error, percentage difference between observed and predicted outcomes; root node error; applied root mean square error.

Model performance results

Apparent discrimination (AUC) was reported for 89 models (n=89/152, 59%), optimism corrected AUC was reported for 26 models (n=26/152, 17%) and external validation AUC results were reported for 26 models (n=26/37, 70%). The median apparent AUC was 0.75 (IQR: 0.69-0.85, range: 0.54-0.99), optimism corrected AUC was 0.79 (IQR: 0.74-0.85, range: 0.56-0.93), and validation AUC was 0.73 (IQR: 0.70-0.78, range: 0.51-0.88).

Both apparent and optimism corrected AUC was reported for eight models, in which we found a median 0.05 reduction in AUC (IQR: -0.09 to -0.03, range: -0.14 to 0.005). Both apparent and validation AUC was reported for 11 models, in which we found a median 0.02 reduction in AUC (IQR: -0.04 to -0.002, range: -0.08 to 0.01).

Risk groups and model presentation

Risk groups were explicitly created in four studies, of which three provided cut-off boundaries for the risk groups. Two studies created 3 groups, one created 4 groups and one created 5 groups. To create the risk groups, three studies used data driven methods including one study that used a classification and regression tree, and for

one study it was unclear.

Two development with validation studies created risk groups, both provided cut-off boundaries and created 3 groups. To create the risk groups, one study used data driven methods and for the other it was unclear.

Presentation or explanation of how to use the prediction model (e.g., formula, decision tree, calculator, code) was reported in less than half of studies (n=28/62, 45%) of studies. Presentation of the full (final) regression-based machine learning model was provided in two studies (n=2/28, 7%).[61,87] Decision trees (including CART) were provided in 14 studies (n=14/28, 50%). Code or a link or reference to a web calculator was provided in six studies (n=6/28, 21%), and a point scoring system or nomogram was provided in four studies (n=4/28, 14%). Two studies provided a combination of a point scoring system or code, with a decision tree (n=2/28, 7%). Thirty-six studies (n=36/62, 58%) developed more than 2 prediction models, and a the 'best' model was identified in 30 studies (n=30/36, 83%). Twenty-eight studies identified the 'best' model based on model performance measures (i.e., AUC, net benefit, and classification measures), one study model based it on model parsimony, and for one study it was unclear.

DISCUSSION

Summary of findings

In this review we assessed the methodological conduct of studies developing author defined machine learning based clinical prediction models in the field of oncology. Over a quarter of statistical regression models were considered machine learning. We not only found poor methodological conduct for nearly all developed and validated machine learning based clinical prediction models, but also a large amount of heterogeneity in the choice of model development and validation methodology, including the choice of modelling method, sample size, model performance measures and reporting.

A key factor contributing to the poor quality of these models was unjustified, small sample sizes used to develop the models. Despite using existing data from electronic health records and registries, most models were informed by small datasets with too few events. Non-regression-based machine learning and ensemble models were developed using smaller datasets (lower events per predictor), compared to regression-based machine learning models. Use of smaller datasets for non-regression and ensemble machine learning models is problematic and increases their risk of overfitting further due to increased flexibility and categorisation of prediction inherent to many machine learning methods. [116,117]

The risk of overfitting in the included studies and models was further exacerbated by split sample internal validation approaches, exclusion of missing data, univariable predictor selection before model building and stepwise predictor selection during model building. Few models also appropriately handled complexities in the data, for example, methods for censoring were not reported in many studies and was rarely accounted for in models developed for a time for event outcome.

Model performance measures were often discrimination and classification performance measures and were not corrected for optimism, yet these measures were often used to identify the 'best' model in studies developing and comparing more than one model. Under and over sampling methods were used to overcome class imbalance, however this results in distortion of the outcome event rate resulting in poorly calibrated models.; however, calibration was rarely reported in studies.

Over half the developed models would not be able to be independently validated, an important step for implementation of prediction models in clinical practice, as they were not reported or available (via code or web calculator) in their respective studies.

Literature

Our review supports evidence of poor methodological quality of machine learning clinical prediction models which has been highlighted by cancer and non-cancer reviews 22,26,118,119. Methodological shortcomings have also been found in prediction modelling reviews focussed on only regression-based cancer prediction models. Our findings are comparable to these reviews which highlight inappropriate use of methods and lack of sufficient sample size for development and external validation of prediction models.[120–123]

Li et al reviewed machine learning prediction models for 5-year breast cancer survival and compared machine learning to statistical regression models. [118] They found negligible improvement in the performance of machine learning models and highlighted low sample sizes, lack of pre-processing steps and validation methods and problematic areas for these models. Christodoulou et al conducted a systematic review of studies comparing machine learning models to logistic regression and also found inconclusive evidence of superiority of machine learning over logistic regression, a low quality or indeed high risk of bias associated to model and a need to further reporting and methodological guidance. [22]

Insufficient sample size when developing and validating machine learning based clinical prediction models is a common methodological flaw in studies.[22,23,26] However, it may be a bigger problem for machine learning models with lower events per variable observed, compared to regression-based models and studies have shown that much larger sample sizes are needed when using machine learning methods and so the impact and risk of bias introduced from these insufficient sample sizes may be much larger. [117,124]

Strengths and limitations

This review highlights the common methodological flaws found in studies developing machine learning based clinical prediction models in oncology. Many existing systematic reviews have focussed on the quality of models in certain clinical sub-specialties and cancer types, and we provide a broader view and assessment that focusses on the conduct of clinical prediction model studies using machine learning methods in oncology.

We calculate the event per predictor, instead of the events per predictor parameter as the number of predictor parameters was not possible to ascertain due to the 'black box' nature of machine learning models. This means that the sample size may be more inadequate than is highlighted in our review.

Though we searched MEDLINE and Embase, two major information databases for studies that developed (and validated) a machine learning based

clinical prediction model, we may have missed eligible publications. Our studies are also restricted to models that were published during 01 Jan 2019 and 05 Sept 2019 resulting in missing models published since our search date. However, our aim for this review was to describe a contemporary sample of publications to reflect current practice. Further, as our findings agree with the existing evidence, it is unlikely that additional studies would change the conclusion of this review.

We included a study by Alcantud et al [81] which used fuzzy and soft set theory, traditionally an artificial intelligence method that resembles human knowledge and reasoning, as opposed to a machine learning method that learns from data. This was a result of using a broader search string to describe the types of models being used to develop prediction models in low dimensional setting and using clinical characteristics. Removing this study from our review does not change our findings and conclusions.

Future research

Methodological guidance, better education, and increased awareness on the minimum scientific standards for prediction modelling research is urgently needed to improve the quality and conduct of machine learning models. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) collaboration has initiated the development of a TRIPOD statement and PROBAST quality assessment tool specific to machine learning (TRIPOD-AI and PROBAST-AI) to improve reporting conduct and evaluation of these models. [39,125] Both this review and a sister review of diagnostic and prognostic models have been conducted to inform these guidelines (PROSPERO ID: CRD42019161764).

These guidelines need to be complemented with methodological guidance to support researchers developing clinical prediction models using machine learning to ensure use better and efficient modelling methods. There is a primary need for sample size guidance that will ensure informed and justified use of data and machine learning methods to develop these models.

Development of machine learning based clinical prediction models in general and in oncology is rapid. Periodic reviews and re-reviews are needed so evidence reflects current practice. These reviews should both focus on individual clinical domains and be cancer specific but should also focus on machine learning based clinical prediction models.

CONCLUSION

The methodological conduct of machine learning based clinical prediction models is poor. Reporting and methodological guidance is urgently needed, with increased awareness and education of minimum prediction modelling scientific standards. A particular focus is needed on sample size estimation, development and validation analysis methods, and ensuring the developed model is available for independent validation, to improve quality of machine learning based clinical prediction models.

Acknowledgements
None

Authors' contributions
PD and GSC conceived the study. PD, CAN, BVC, KGMM and GSC designed the study. PD and SK developed the search strategy. PD and JM carried out the screening. PD, JM, CAN, BS, and GB carried out the data extraction of all items from all articles. PD performed the analysis and drafted the first draft. PD, JM, CAN, BS, GB, JAAD, SK, LH, RDR, BVC, KGMM, and GSC critically reviewed and edited the article. All authors read and approved the final manuscript.

Funding information
Gary Collins, Shona Kirtley and Jie Ma are supported by Cancer Research UK (programme grant: C49297/A27294). Benjamin Speich is supported by an Advanced Postdoc. Mobility grant (P300PB_177933) and a return grant (P4P4PM_194496) from the Swiss National Science Foundation. Gary Collins and Paula Dhiman are supported by the NIHR Biomedical Research Centre, Oxford. Ben Van Calster is supported by Internal Funds KU Leuven (grant C24M/20/064), University Hospitals Leuven (grant COPREDICT), and Kom Op Tegen Kanker (grant KOTK TRANS-IOTA). This publication presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the Cancer Research UK, the NHS, the NIHR or the Department of Health and Social Care.

Competing interests
The authors declare no conflict of interest.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Data availability
The datasets generated and/or analysed during the current study are available in the Open Science Framework repository (https://osf.io/3aezj/).

Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1186/s12874-022-01577-x

List of abbreviations
AUC: area under the curve; TRIPOD: The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; PROBAST: Prediction model Risk Of Bias ASsessment Tool; CART: Classification and regression tree; RPA: recursive partitioning analysis; LASSO: Least Absolute Shrinkage and Selection Operator; CHARMS: CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses guideline.

REFERENCES

1. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099.
2. Pulitanò C, Arru M, Bellio L, Rossini S, Ferla G, Aldrighetti L. A risk score for predicting perioperative blood transfusion in liver surgery. Br J Surg. 2007;94(7):860-865.
3. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987-1003.
4. Nashef SAM, Roques F, Sharples LD, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012;41(4):734-745.
5. Thamer M, Kaufman JS, Zhang Y, Zhang Q, Cotter DJ, Bang H. Predicting Early Death Among Elderly Dialysis Patients: Development and Validation of a Risk Score to Assist Shared Decision Making for Dialysis Initiation. Am J Kidney Dis. 2015;66(6):1024-1032.
6. Velazquez N, Press B, Renson A, et al. Development of a Novel Prognostic Risk Score for Predicting Complications of Penectomy in the Surgical Management of Penile Cancer. Clin Genitourin Cancer. 2019;17(1):e123-e129.
7. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100(6):1619-1636.
8. Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower-Thomas K. The Nottingham Prognostic Index: five- and ten-year data for all-cause Survival within a Screened Population. Ann R Coll Surg Engl. 2015;97(2):137-139.
9. Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst. 1998;90(10):766-771.
10. Corbelli J, Borrero S, Bonnema R, et al. Use of the Gail Model and Breast Cancer Preventive Therapy Among Three Primary Care Specialties. J Women's Health. 2014;23(9):746-752.
11. Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A Validated Clinical Risk Prediction Model for Lung Cancer in Smokers of All Ages and Exposure Types: A HUNT Study. EBioMedicine. 2018;31:36-46.
12. Lebrett MB, Balata H, Evison M, et al. Analysis of lung cancer risk model (PLCOM2012 and LLPv2) performance in a community-based lung cancer screening programme. Thorax. 2020;75(8):661-668.
13. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open. 2015;5(3):e007825.
14. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLoS Med. 2012;9(5):e1001221.
15. Bradley A, Meer RVD, McKay CJ. A systematic review of methodological quality of model development studies predicting prognostic outcome for resectable pancreatic cancer. BMJ Open. 2019;9(8):e027192.
16. Fahey M, Crayton E, Wolfe C, Douiri A. Clinical prediction models for mortality and functional outcome following ischemic stroke: A systematic review and meta-analysis. PLoS One. 2018;13(1):e0185402.
17. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
18. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-260.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
20. Banerjee A, Chen S, Fatemifar G, et al. Machine learning for subtype definition and

risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. BMC Med. 2021;19(1):85.
21. Navarro CLA, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021;375:n2281.
22. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
23. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. Dig Dis Sci. 2019;64(8):2078-2087.
24. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. N Engl J Med. 2017;376(26):2507-2509.
25. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019;23(1):284.
26. Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One. 2020;15(6):e0234722.
27. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int J Med Inform. 2021;151:104484.
28. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368.
29. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol. 2021;138:60-70.
30. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9(1):103.
31. Bridge J, Blakey JD, Bonnett LJ. A systematic review of methodology used in the development of prediction models for future asthma exacerbation. BMC Med Res Methodol. 2020;20(1):22.
32. Mushkudiani NA, Hukkelhoven CWPM, Hernández AV, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol. 2008;61(4):331-343.
33. Sahle BW, Owen AJ, Chin KL, Reid CM. Risk Prediction Models for Incident Heart Failure: A Systematic Review of Methodology and Model Performance. J Card Fail. 2017;23(9):680-687.
34. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013;66(3):268-277.
35. Collins SD, Peek N, Riley RD, Martin GP. Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient. J Clin Epidemiol. 2021;133:53-60.
36. Breiman L. Statistical Modeling: The Two Cultures. Statist Sci. 2001;16(3):199-231.
37. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6(7):e1000097.
38. A systematic review protocol of clinical prediction models using machine learning methods in oncology. PROSPERO. Accessed December 19, 2020. Available at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=140361
39. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable

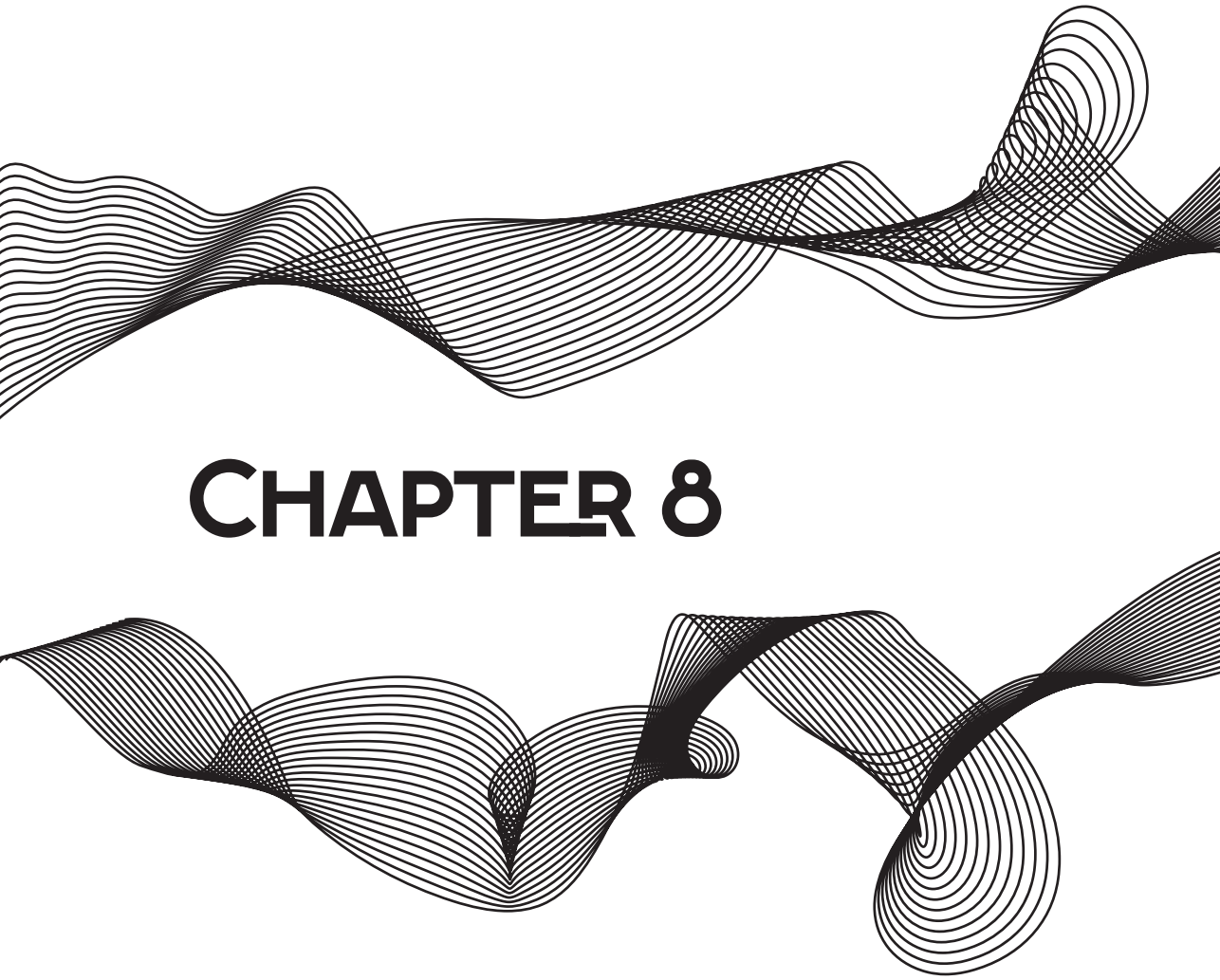
prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
40. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
41. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan — a web and mobile app for systematic reviews. Syst Rev. 2016;5:210.
42. The Endnote Team. Endnote. Published online 2013.
43. Moons KGM, Groot JAH de, Bouwmeester W, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10):e1001744.
44. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019;170(1):W1-W33.
45. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51-58.
46. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open. 2019;9(4):e025611.
47. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde J. Research electronic data capture (REDCap)-metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377e81.
48. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017. Published online 2017.
49. Zhou HF, Lu J, Zhu HD, et al. Early Warning Models to Estimate the 30-Day Mortality Risk After Stent Placement for Patients with Malignant Biliary Obstruction. Cardiovasc Intervent Radiol. 2019;42(12):1751-1759.
50. Dihge L, Ohlsson M, Edén P, Bendahl PO, Rydén L. Artificial neural network models to predict nodal status in clinically node-negative breast cancer. BMC Cancer. 2019;19(1):610.
51. Luna JM, Chao HH, Diffenderfer ES, et al. Predicting radiation pneumonitis in locally advanced stage II-III non-small cell lung cancer using machine learning. Radiother Oncol. 2019;133:106-112.
52. Yang XG, Wang F, Feng JT, et al. Recursive Partitioning Analysis (RPA) of Prognostic Factors for Overall Survival in Patients with Spinal Metastasis: A New System for Stratified Treatment. World Neurosurg. 2019;127:e124-e131.
53. Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. Am J Obstet Gynecol. 2019;220(4):381.e1-381.e14.
54. Khalaf MH, Sundaram V, AbdelRazek Mohammed MA, et al. A Predictive Model for Postembolization Syndrome after Transarterial Hepatic Chemoembolization of Hepatocellular Carcinoma. Radiology. 2019;290(1):254-261.
55. Wong NC, Lam C, Patterson L, Shayegan B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. BJU Int. 2019;123(1):51-57.
56. Lindsay WD, Ahern CA, Tobias JS, et al. Automated data extraction and ensemble methods for predictive modeling of breast cancer outcomes after radiation therapy. Med Phys. 2019;46(2):1054-1063.
57. Wang YH, Nguyen PA, Islam MM, Li YC, Yang HC. Development of Deep Learning Algorithm for Detection of Colorectal Cancer in EHR Data. Stud Health Technol Inform. 2019;264:438-441.
58. Muhlestein WE, Akagi DS, Davies JM, Chambless LB. Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Im-

prove Predictive Performance. Neurosurgery. 2019;85(3):384-393.
59. Iraji ms. Deep stacked sparse auto-encoders for prediction of post-operative survival expectancy in thoracic lung cancer surgery. Journal of Applied Biomedicine. 2019;17:75-75.
60. Karhade AV, Thio QCBS, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. Neurosurgery. 2019;85(1):E83-E91.
61. Chi S, Li X, Tian Y, et al. Semi-supervised learning to improve generalizability of risk prediction models. J Biomed Inform. 2019;92:103117.
62. Xu Y, Kong S, Cheung WY, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. BMC Cancer. 2019;19(1):210.
63. Zhao B, Gabriel RA, Vaida F, Lopez NE, Eisenstein S, Clary BM. Predicting Overall Survival in Patients with Metastatic Rectal Cancer: a Machine Learning Approach. J Gastrointest Surg. 2020;24(5):1165-1172.
64. Günakan E, Atan S, Haberal AN, Küçükyıldız İA, Gökçe E, Ayhan A. A novel prediction method for lymph node involvement in endometrial cancer: machine learning. Int J Gynecol Cancer. 2019;29(2):320-324.
65. Vagnildhaug OM, Brunelli C, Hjermstad MJ, et al. A prospective study examining cachexia predictors in patients with incurable cancer. BMC Palliat Care. 2019;18(1):46.
66. Thapa S, Fischbach L, Delongchanp R, Faramawi M, Orloff M. Using Machine Learning to Predict Progression in the Gastric Precancerous Process in a Population from a Developing Country Who Underwent a Gastroscopy for Dyspeptic Symptoms. Gastroenterol Res Pract. 2019;2019: 8321942.
67. Xu Y, Kong S, Cheung WY, Quan ML, Nakoneshny SC, Dort JC. Developing case-finding algorithms for second events of oropharyngeal cancer using administrative data: A population-based validation study. Head & Neck. 2019;41(7):2291-2298.
68. Auffenberg GB, Ghani KR, Ramani S, et al. askMUSIC: Leveraging a Clinical Registry to Develop a New Machine Learning Model to Inform Patients of Prostate Cancer Treatments Chosen by Similar Men. Eur Urol. 2019;75(6):901-907.
69. Alabi RO, Elmusrati M, Sawazaki-Calone I, et al. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. Virchows Arch. 2019;475(4):489-497.
70. Greene MZ, Hughes TL, Hanlon A, Huang L, Sommers MS, Meghani SH. Predicting cervical cancer screening among sexual minority women using Classification and Regression Tree analysis. Prev Med Rep. 2019;13:153-159.
71. Nartowt BJ, Hart GR, Roffman DA, et al. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. PLoS One. 2019;14(8):e0221421.
72. Taninaga J, Nishiyama Y, Fujibayashi K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. Scientific Reports. 2019;9(1):12384.
73. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35(23):4124-4135.
74. Oyaga-Iriarte E, Insausti A, Sayar O, Aldaz A. Prediction of irinotecan toxicity in metastatic colorectal cancer patients based on machine learning models with pharmacokinetic parameters. J Pharmacol Sci. 2019;140(1):20-25.
75. Yan P, Huang R, Hu P, et al. Nomograms for predicting the overall and cause-specific survival in patients with malignant peripheral nerve sheath tumor: a population-based study. J Neurooncol. 2019;143(3):495-503.
76. Ryu SM, Lee SH, Kim ES, Eoh W. Predicting Survival of Patients with Spinal Ependy-

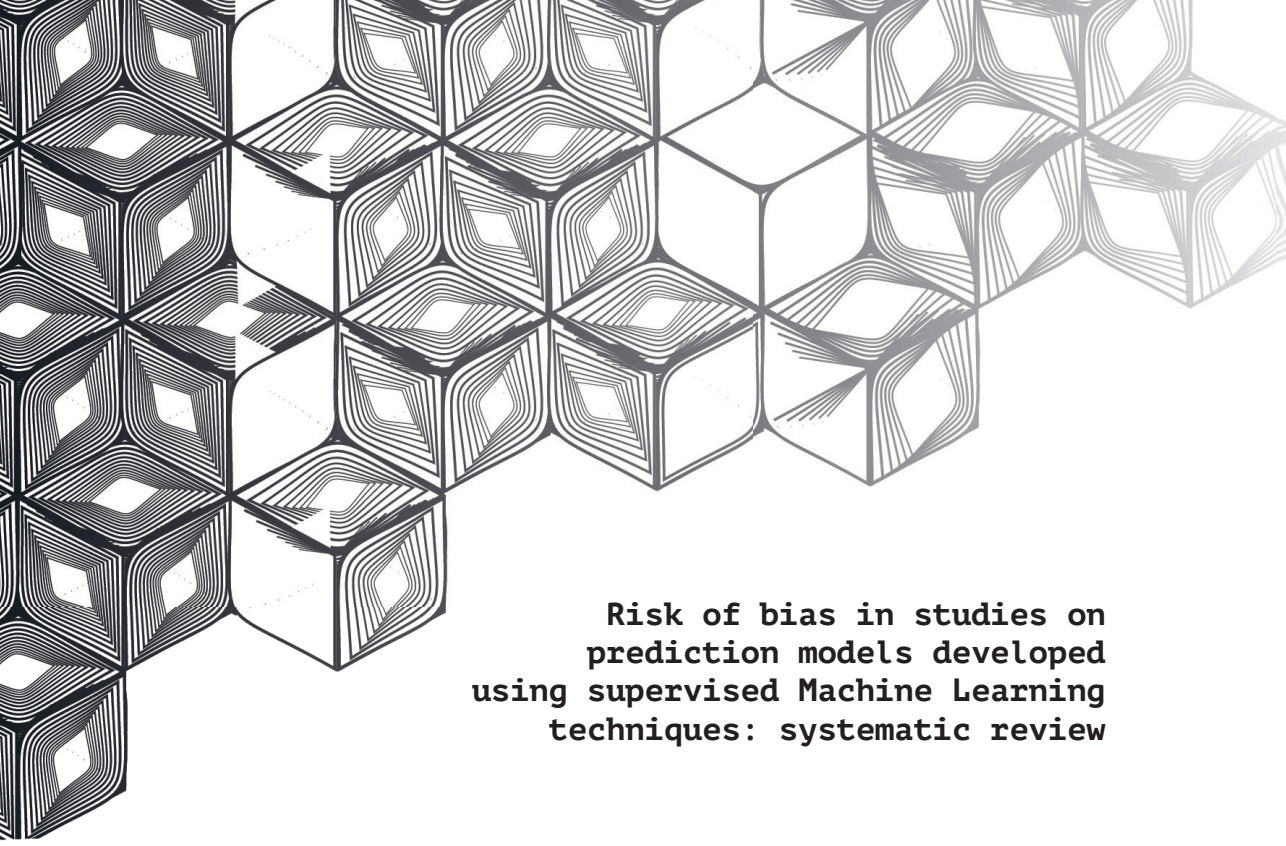
moma Using Machine Learning Algorithms with the SEER Database. World Neurosurg. 2019;124:e331-e339.
77. Feng SS, Li H bao, Fan F, et al. Clinical characteristics and disease-specific prognostic nomogram for primary gliosarcoma: a SEER population-based analysis. Sci Rep. 2019;9(1):10744.
78. van Niftrik CHB, van der Wouden F, Staartjes VE, et al. Machine Learning Algorithm Identifies Patients at High Risk for Early Complications After Intracranial Tumor Surgery: Registry-Based Cohort Study. Neurosurgery. 2019;85(4):E756-E764.
79. Merath K, Hyer JM, Mehta R, et al. Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery. J Gastrointest Surg. 2020;24(8):1843-1851.
80. Egger ME, Stevenson M, Bhutiani N, et al. Age and Lymphovascular Invasion Accurately Predict Sentinel Lymph Node Metastasis in T2 Melanoma Patients. Ann Surg Oncol. 2019;26(12):3955-3961.
81. Alcantud JCR, Varela G, Santos-Buitrago B, Santos-García G, Jiménez MF. Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making. PLoS One. 2019;14(6):e0218283.
82. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163.
83. Hammer J, Geinitz H, Nieder C, et al. Risk Factors for Local Relapse and Inferior Disease-free Survival After Breast-conserving Management of Breast Cancer: Recursive Partitioning Analysis of 2161 Patients. Clin Breast Cancer. 2019;19(1):58-62.
84. Mahmoudian M, Seyednasrollah F, Koivu L, Hirvonen O, Jyrkkiö S, Elo LL. A predictive model of overall survival in patients with metastatic castration-resistant prostate cancer. F1000Res. 2016;5:2674.
85. Zheng B, Lin J, Li Y, et al. Predictors of the therapeutic effect of corticosteroids on radiation-induced optic neuropathy following nasopharyngeal carcinoma. Support Care Cancer. 2019;27(11):4213-4219.
86. Li M, Zhan C, Sui X, et al. A Proposal to Reflect Survival Difference and Modify the Staging System for Lung Adenocarcinoma and Squamous Cell Carcinoma: Based on the Machine Learning. Front Oncol. 2019;9:771.
87. Beachler DC, de Luise C, Yin R, Gangemi K, Cochetti PT, Lanes S. Predictive model algorithms identifying early and advanced stage ER+/HER2- breast cancer in claims data. Pharmacoepidemiol Drug Saf. 2019;28(2):171-178.
88. Tian Z, Yen A, Zhou Z, Shen C, Albuquerque K, Hrycushko B. A machine-learning–based prediction model of fistula formation after interstitial brachytherapy for locally advanced gynecological malignancies. Brachytherapy. 2019;18(4):530-538.
89. Obrzut B, Kusy M, Semczuk A, Obrzut M, Kluska J. Prediction of 10-year Overall Survival in Patients with Operable Cervical Cancer using a Probabilistic Neural Network. J Cancer. 2019;10(18):4189-4195.
90. Fuse K, Uemura S, Tamura S, et al. Patient-based prediction algorithm of relapse after allo-HSCT for acute Leukemia and its usefulness in the decision-making process using a machine learning approach. Cancer Med. 2019;8(11):5058-5067.
91. Tighe D, Lewis-Morris T, Freitas A. Machine learning methods applied to audit of surgical outcomes after treatment for cancer of the head and neck. Br J Oral Maxillofac Surg. 2019;57(8):771-777.
92. Tseng YJ, Huang CE, Wen CN, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform. 2019;128:79-86.
93. Sala Elarre P, Oyaga-Iriarte E, Yu KH, et al. Use of Machine-Learning Algorithms in Intensified Preoperative Therapy of Pancreatic Cancer to Predict Individual Risk of

Relapse. Cancers (Basel). 2019;11(5).
94. Wang HH, Wang YH, Liang CW, Li YC. Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer. JAMA Dermatol. 2019;155(11):1277-1283.
95. Paik ES, Lee JW, Park JY, et al. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. J Gynecol Oncol. 2019;30(4):e65.
96. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-Day and 1-Year Mortality in Spinal Metastatic Disease: Development and Internal Validation. Neurosurgery. 2019;85(4):E671-E681.
97. Facciorusso A, Del Prete V, Antonino M, Buccino VR, Muscatiello N. Response to repeat echoendoscopic celiac plexus neurolysis in pancreatic cancer patients: A machine learning approach. Pancreatology. 2019;19(6):866-872.
98. Lemée JM, Corniola MV, Da Broi M, et al. Extent of Resection in Meningioma: Predictive Factors and Clinical Implications. Sci Rep. 2019;9(1):5944.
99. Yang CQ, Gardiner L, Wang H, Hueman MT, Chen D. Creating Prognostic Systems for Well-Differentiated Thyroid Cancer Using Machine Learning. Front Endocrinol (Lausanne). 2019;10:288.
100. Corniola MV, Lemée JM, Da Broi M, et al. Posterior fossa meningiomas: perioperative predictors of extent of resection, overall survival and progression-free survival. Acta Neurochir (Wien). 2019;161(5):1003-1011.
101. R K, R GR. Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. J Med Syst. 2019;43(7):201.
102. Sasani K, Catanese HN, Ghods A, et al. Gait speed and survival of older surgical patient with cancer: Prediction after machine learning. J Geriatr Oncol. 2019;10(1):120-125.
103. Wang X, Zhang Y, Hao S, et al. Prediction of the 1-Year Risk of Incident Lung Cancer: Prospective Study Using Electronic Health Records from the State of Maine. J Med Internet Res. 2019;21(5):e13260.
104. Knol MJ, Janssen KJM, Donders ART, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol. 2010;63(7):728-736.
105. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184(11):1265-1269.
106. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
107. Sim JA, Yun YH. Predicting Disease-Free Lung Cancer Survival Using Patient Reported Outcome (PRO) Measurements with Comparisons of Five Machine Learning Techniques (MLT). Stud Health Technol Inform. 2019;264:1588-1589.
108. Karadaghy OA, Shew M, New J, Bur AM. Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. JAMA Otolaryngol Head Neck Surg. 2019;145(12):1115-1120.
109. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019;9(1):6994.
110. Al-Bahrani R, Agrawal A, Choudhary A. Survivability prediction of colon cancer patients using neural networks. Health Informatics J. 2019;25(3):878-891.
111. Maubert A, Birtwisle L, Bernard JL, Benizri E, Bereder JM. Can machine learning predict resecability of a peritoneal carcinomatosis? Surg Oncol. 2019;29:120-125.
112. Harrell Jr FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer; 2015.
113. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907-916.

114. Sauerbrei W, Boulesteix AL, Binder H. Stability investigations of multivariable regression models derived from low- and high-dimensional data. J Biopharm Stat. 2011;21(6):1206-1231.
115. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer; 2019.
116. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
117. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14(1):137.
118. Li J, Zhou Z, Dong J, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. PLoS One. 2021;16(4):e0250370.
119. Abreu PH, Santos MS, Abreu MH, Andrade B, Silva DC. Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Comput Surv. 2016;49(3):52:1-52:40.
120. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res (Phila). 2016;9(1):13-26.
121. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20.
122. Grigore B, Lewis R, Peters J, Robinson S, Hyde CJ. Development, validation and effectiveness of diagnostic prediction tools for colorectal cancer in primary care: a systematic review. BMC Cancer. 2020;20(1):1084.
123. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC Cancer. 2019;19(1):230.
124. Balki I, Amirabadi A, Levman J, et al. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can Assoc Radiol J. 2019;70(4):344-353.
125. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet. 2019;393(10181):1577-1579.



CHAPTER 8



Risk of bias in studies on prediction models developed using supervised Machine Learning techniques: systematic review

Constanza L Andaur Navarro
Johanna AA Damen
Toshihido Takada
Steven WJ Nijman
Paula Dhiman
Jie Ma
Gary S Collins
Ram Bajpai
Richard D Riley
Karl GM Moons
Lotty Hooft

BMJ 2021;375:n2281

ABSTRACT

Objective. To assess the methodological quality of machine learning (ML)-based prediction model studies across all medical fields.

Design. Systematic review.

Data sources. PubMed from 1 January 2018 to 31 December 2019.

Eligibility criteria. We included articles reporting on the development or development with external validation of a multivariable prediction model (either diagnostic or prognostic) developed using supervised ML for individualized predictions. No restrictions were made based on study design, data source, or predicted patient-related health outcomes.

Review methods. To determine the methodological quality of the ML-based prediction model studies, we evaluated the risk of bias (RoB) using the Prediction Risk Of Bias ASsessment Tool (PROBAST). We measured RoB per domain (participants, predictors, outcome, and analysis) and per study (overall).

Results. We included 152 studies, 58 (38.2%) diagnostic and 94 (61.8%) prognostic studies. We applied PROBAST to 152 developed models and 19 external validations. Out of these 171 analyses, 148 (86.5%, 95% confidence interval 80.6% to 90.9%) were rated at high RoB. The Analysis domain was the most frequently rated at high RoB. We observed 85/152 (55.9%, 48.0% to 63.6%) models developed with an inadequate number of events per candidate predictor, 62/152 with poor handling of missing data (40.8%, 33.3% to 48.7%) and 59/152 with unproper assessment of overfitting (38.8%, 31.4% to 46.7%). Most models used appropriate data sources to develop (73.0%, 65.5% to 79.4%) and externally validate their ML-based prediction models (73.7%, 51.2% to 88.2%). However, information about blinding of outcome and blinding of predictors was absent in 60/152 (39.5%, 32.1% to 47.4%) and 79/152 (52.0%, 44.1% to 59.8%) developed models, respectively.

Conclusion. Most ML-based prediction model studies show poor methodological quality and are at high risk of bias. Factors contributing to the risk of bias include small study size, poor handling of missing data, and failure to address overfitting. Efforts to improve the design, conduct, reporting, and validation of ML-based prediction model studies are necessary to boost its application in clinical practice.

Systematic review registration PROSPERO, CRD42019161764

ARTICLE SUMMARY

What is already known on this topic?

1. Several publications have highlighted the poor methodological quality of regression-based prediction models studies.
2. The number of clinical prediction models developed using supervised machine learning is rapidly increasing, however, evidence about their methodological quality and risk of bias is scarce.

What this study adds?

1. Prediction model studies developed using supervised machine learning have poor methodological quality. Limited sample size, poor handling of missing data, and inappropriate evaluation of overfitting contributed largely to the overall high risk of bias.
2. Machine learning prediction models often claim superior accuracy compared to regression-based approaches. However, reported performance may be at high risk of bias based on the study design and modelling strategies used. Caution is needed when interpreting these findings.
3. Future research should improve transparency when reporting and the study designs used to develop, validate, and compare prediction models to reduce methodological biases.

INTRODUCTION

A multivariable prediction model is defined as any combination of two or more predictors (i.e. variables, features) for estimating the probability or risk of an individual to have (diagnosis) or will develop (prognosis) a particular outcome.[1–4] Properly conducted and well reported prediction model studies are essential for a proper implementation in clinical practice. Even though prediction model studies are abundant in biomedical literature, a limited amount of them are used in clinical practice. As a result, many published studies contribute to research waste.[5] We anticipate that the rise of modern data-driven modelling techniques will boost the existing popularity of prediction model studies in the biomedical literature.[6,7]

Machine learning (ML), a subset of artificial intelligence (AI), has gained considerable popularity in recent years. Broadly, machine learning refers to computationally intensive methods that use data-driven approaches to develop models that require fewer modelling decisions by the modeler compared to traditional modelling techniques.[8–11] Within machine learning, there are two approaches: supervised and unsupervised learning. While supervised learning is defined as an algorithm that learn to predict using previously labelled outcomes, unsupervised learning learns to find unexpected patterns using unlabelled outcomes.[12] Traditional prediction models in healthcare usually resemble supervised learning: datasets used for development are labelled and the objective is to predict an outcome in new data. Supervised learning includes tree-based methods, such as random forests, naïve bayes, and gradient boosting machines, support vector machines, neural networks. Supervised ML-based prediction model studies have shown promising and even superior predictive performance compared to conventional statistical techniques, however, recent systematic reviews have shown otherwise.[13–16] Although several publications have raised concern about the methodological quality of prediction models developed with conventional statistical techniques[6,17,18], a formal methodological and risk of bias (RoB) assessment of supervised ML-based prediction model studies across all medical disciplines has not yet been carried out.

Shortcomings in study design, methods, conduct, and analysis may set the study at high RoB, which could lead to deviated estimates of models' predictive performance.[19,20] The Prediction model Risk Of Bias Assessment Tool (PROBAST) was developed to facilitate RoB assessment, and thus provides a methodological quality assessment of primary studies that report on development, validation, or update of prediction models, regardless of the clinical domain, predictors, outcomes, or modelling technique used.[19,20] Using a prediction model considered at high RoB, might lead to unnecessary or insufficient interventions, and thus affect patients' health and health systems. Rigorous RoB evaluation of prediction model studies is, therefore, essential to ensure reliability, fast, and valuable application of prediction

models. Therefore, we conducted a systematic review to assess the methodological quality and RoB of supervised ML-based prediction model studies across all medical fields in a contemporary sample of recent literature.

METHODS

Our systematic review was reported following the PRISMA statement.[21] The review protocol was registered (PROSPERO, CRD42019161764) and published.[22]

Identification of prediction model studies

We searched for eligible studies published in PubMed between 1 January 2018 and 31 December 2019. We restricted the search to obtain a contemporary sample of articles that would reflect the current practices in prediction modelling using machine learning to date. The search was performed on 19 December 2019 with a strategy that is provided in Supplemental File 1.

Eligible publications needed to describe the development or validation of at least one multivariable prediction model using any supervised ML technique aiming for individualized prediction of risk or patient-related health outcomes. Details about inclusion and exclusion criteria are stated in our protocol.[22] A publication was also eligible if it aimed to develop a prediction model based on model extension or incremental value of new predictors. No restrictions were made based on study design, data source, or types of patient-related health outcomes. We defined a publication to be an instance of ML when a non-regression statistical technique was used to develop or validate a prediction model. Hence, studies using only linear regression, logistic regression, lasso regression, ridge regression, or elastic net were excluded. Publications that report about the association of a single predictor, test, or biomarker, or its causality with an outcome were excluded. Publications that aimed to use ML to enhance the reading of images or signals or those where ML models only used genetic traits or molecular markers as predictors, were also excluded. We also excluded systematic reviews, methodological articles, conference abstracts, and publications for which full text was unavailable through our institution. The search was restricted to human subjects and English-written articles.

Screening process

Titles and abstracts were screened by two independent reviewers, from a group of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD). A third reviewer was involved when required to resolve any disagreements (JAAD). After selection of potentially eligible studies, full-text articles were retrieved and two independent researchers reviewed them for eligibility; one researcher (CLAN) screened all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively screened the same articles for agreement. In case of any disagreement, a third reviewer was asked to read the article in question and resolve (JAAD).

Data extraction

We developed a data extraction form based on the four domains: participants, predictors, outcome, and analysis (box 1) as well as 20 signalling questions as described in PROBAST.[19 20]

Box 1 Description of domains used in data extraction form

Participants domain Covers potential biases related to the selection of participants and data sources used
Predictors domain Evaluates potential sources of bias from the definition and measurement of the candidate predictors
Outcome domain Assesses how and when the outcome was defined and determined
Analysis domain Examines the statistical methods that authors have used to develop and validate the model, including study size, handling of continuous predictors and missing data, selection of predictors, and model performance measures

Our extraction form contained 3 sections per domain: two to nine specific signalling questions, judgement of RoB, and rationale for the judgment. Signalling questions were formulated to be answered 'yes/probably yes', 'no/probably no', and 'no information'. All signalling questions were phrased so that 'yes/probably yes' indicated absence of bias. Likewise, judgement of RoB was defined as 'high RoB', 'low RoB', and 'unclear RoB'. Also, we requested reviewers to provide a rationale for judgment as free-text comments.

If a study included external validation, we applied the extraction form to both, the development and external validation of the model. Signalling question 4.5 –was selection of predictors based on univariable analysis avoided? –, 4.8 –Were model overfitting and optimism in model performance accounted for? –, and 4.9 – Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? – did not apply to external validation. If a study reported more than one model, we applied PROBAST to the recommended model defined by the authors in the article. If the authors did not recommend a single model, the model with highest accuracy (in terms of discrimination) was selected as the recommended model. The PROBAST tool, its considerations, and related publications are available on the PROBAST website (www.probast.org). A summary table with the criteria to judge risk of bias is provided in Supplemental File 2.

Two reviewers independently extracted data from each article using the constructed form. To accomplish consistent data extraction, the form was piloted

on five articles by all reviewers. During pilot, reviewers clarified differences in interpretation and standardise data extraction. After the pilot, articles used were randomly assigned and screened again in the main data extraction. One researcher (CLAN) extracted data from all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively extracted data from the same articles. Any disagreements in data extraction were settled by consensus among each pair of reviewers.

Data analysis

Prediction model studies were categorized as prognosis or diagnosis and into four types of prediction models studies: development (with internal validation), development with external validation (same model), development with external validation (different model), and external validation only. *Model development studies* aim to develop a prediction model to be used for individualized predictions where its predictive performance is directly evaluated using the same data, either by resampling participant data or random/non-random split sample (internal validation). *Model development studies with external validation (same model)* have the same aim as the previous type, but the development of the model is followed by quantifying the predictive performance of the model in a different dataset. *Model development studies with external validation (different model)* aim to update or adjust an existing model that performs poorly by recalibrating or extending the model. *External validation only* studies aim to assess only the predictive performance of existing prediction models using data external to the development sample. [19,20]

Two independent reviewers each assessed signalling question by the degree of compliance with the PROBAST recommendations. If there was any disagreement, it was discussed until consensus was reached. The RoB judgement per domain was based on the answers to the signalling questions. If the answer to all signalling questions was 'yes/probably yes', the RoB domain was judged as 'low RoB'. If reported information was insufficient to answer the signalling questions, these were judged as 'no information', and the RoB domain scored as 'unclear RoB'. If any signalling question was answered as 'no/probably no', reviewers applied their judgment to rate the domain as 'low RoB', 'high RoB', or 'unclear RoB'.

After judging all the domains, we performed an overall assessment per application of PROBAST. PROBAST recommends rating the study as 'low RoB' if all domains had 'low RoB'. If at least one domain had 'high RoB', overall judgment should be rated as 'high RoB'. 'Unclear RoB' was assigned if 'unclear RoB' was noted in at least one domain and all other domains had 'low RoB'. Judgement rationale was recorded to facilitate discussion among reviewers when solving discrepancies. We removed signalling question 4.9 –Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? – because it is tailored for regression-based studies. Results were summarized

as percentages with 95% confidence intervals and visual plots. Analyses were performed using R version 3.6.2 (R Core Team, 2020).

Patient and public involvement

We conducted a methodological appraisal; thus, no patients were involved in setting the research question, nor were they involved in the design or implementation of the study, or the interpretation or writing up of results.

RESULTS

The search identified 24,814 publications, of which we sampled ten random sets of 249 publications each. Of the 2,482 screened publications, 152 were eligible: 94 (61.8%) prognostic and 58 (38.2%) diagnostic ML-based prediction model studies (Figure 1). Detailed description of the included studies is provided in Supplemental File 3. We classified publications according to their research aims: 132 (86.8%) articles were classified as development with internal validation, 19 (12.5%) as development with external validation of the same model, and 1 (0.6%) as development with external validation of another model (eventually included as development with internal validation). Across the 152 studies, a total of 1429 ML-based prediction models were developed and 219 validated. For our analyses, we selected only the recommended model by the authors for our RoB assessment. Hence, we applied PROBAST 171 times: in 152 developed models and 19 external validations. The most common ML techniques for the first model reported were Classification and Regression Tree (CART [10.1%]), Support Vector Machine (SVM [9.4%]), and Random Forest (RF [9.4%]). Detailed list of techniques assessed is provided in Supplemental File 3. The clinical fields with the most publications were oncology (21/152 [13.8%]), surgery (20/152 [13.5%]), and neurology (20/152 [13.5%]).

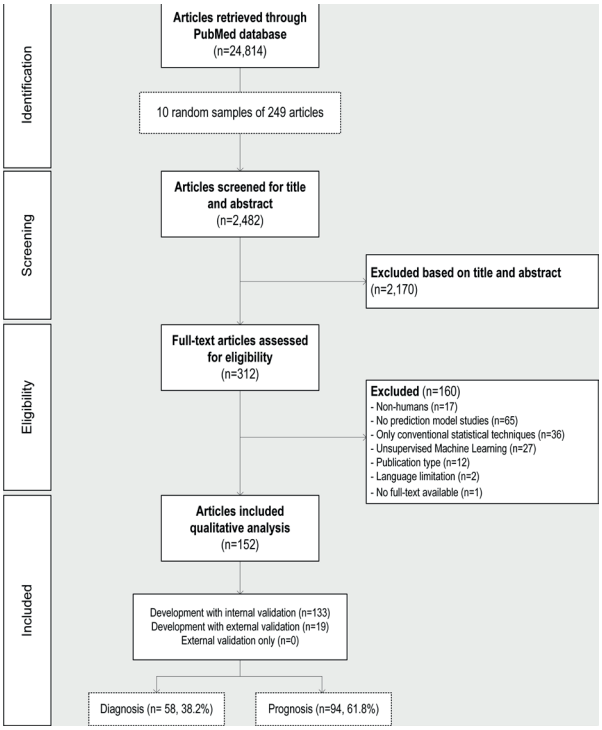


Figure 1. Flowchart of participants

Domain 1: Participants

In total, 36/152 (23.7%) developed models and 3/19 (15.8%) external validations were scored as high RoB for the Participants domain (Figure 2). Prospective and longitudinal data sources (SQ1.1) were properly used for model development in 111/152 (73.0%) and to externally validate in 14/19 (73.7%). We were unable to evaluate whether the inclusion and exclusion of participants (SQ1.2) was representative of the target population in 47/152 (30.9%) developed models and in 12/19 (63.1%) external validations (Table 1).

Domain 2: Predictors

We rated 14/152 (9.2%) developed models and 2/19 (10.5%) external validations to be at high RoB for the Predictors domain (Figure 2). Candidate predictors were defined and assessed in a similar way for all included participants (SQ2.1) in 109/152 (71.7%) developed models and in 8/19 (42.1%) external validations. Information on blinding of predictor assessment to outcome data (SQ2.2) was missing in 60/152 (39.5%) developed models and in 7/19 (36.8%) external validations. All considered predictors should be available at the time the model is intended to be used (SQ2.3), which we found appropriate in 116/152 (76.9%) developed models and in 12/19 (63.1%) external validations (Table 1).

Domain 3: Outcome

The domain Outcome was scored as unclear RoB in 65/152 (42.8%) and 12/19 (63.2%) of developed models and external validations, respectively (Figure 2). We missed information about the outcome being determined without knowledge of predictors' information (SQ3.5) in 79/152 (52.0%) developed models and in 14/19 (73.7%) external validations. Predictors were excluded from the outcome definition (SQ3.3) in 90/152 (59.2%) developed models and in 10/19 (52.6%) external validations. We considered the time interval between predictor measurement and outcome determination appropriate (SQ3.6) in 110/152 (72.4%) developed models and in 11/19 (57.9%) external validations. We observed in 114/152 (75%) developed models and in 12/19 (63.1%) external validations that the outcome was determined using appropriate methods, thus reducing risk of misclassification (SQ3.1). Similarly, 118/152 (77.6%) developed models and 13/19 (68.4%) external validations used prespecified, standard or consensus-based definitions to determine the outcome (SQ3.2). The outcome was defined and measured with the same categories or thresholds for all included participants (SQ3.4) in 118/152 (77.6%) developed models and 10/19 (52.6%) external validations (Table 1).

Domain 4: Analysis

We classified 128/152 (84.2%) developed models and 14/19 (73.7%) external validations as high RoB in the Analysis domain. We considered that the number of participants with the outcome (SQ4.1) was insufficient (i.e. event per predictor parameter <10) in 85/152 (55.9%) developed models and 8/19 (42.1%) external validations (i.e. number of events <100). Information about methods to handle continuous and categorical predictors (SQ4.2) was missed in 81/152 (53.3%) developed models and 18/19 (94.7%) external validations. We found that 84/152 (55.3%) developed models and 10/19 (52.6%) external validation included in their statistical analyses all enrolled participants (SQ4.3).

Handling of missing data (SQ4.4) was inappropriate (i.e. participants with missing data were omitted from the analysis or imputation method was flawed) in 62/152 (40.8%) developed models and in 7/19 (36.8) external validation. We observed that 28/152 (18.4%) developed models used univariable analyses to select predictors (SQ4.5). We were unable to assess if censoring, competing risks or sampling of control participants (SQ4.6) were considered in 54/152 (35.5%) developed models and in 7/19 (36.8%) external validations. Similarly, the reporting of relevant model performance measures (e.g., both discrimination and calibration) (SQ4.7) was missing in 91/152 (59.9%) developed models, while 13/19 (68.4%) external validations lacked this information too. 76/152 (50.0%) developed models accounted for model overfitting and optimism (SQ4.8).

Overall Risk of Bias

Finally, the overall RoB assessed using PROBAST let to 133/152 (87.5%) developed models, and 15/19 (78.9%) external validations being classified as high RoB (Figure 2). Further information about each signalling question answered as 'Yes/probably yes', 'No/probably no', and 'No information' is provided in Table 1.

Diagnostic versus prognostic models

Regarding diagnostic versus prognostic prediction models, the Analysis domain is the major contributor to an overall high RoB in both. We evaluated 56/58 (96.6%) developed models and 7/7 (100%) external validation as high RoB in diagnostic studies, and 77/94 (81.9%) developed models and 8/13 (66.7%) external validation in prognostic studies (Figure 2). External validations of both diagnostic and prognostic models suffer from unclear information to judge RoB. While in diagnostic models, signalling questions in domain Outcome were frequently answered with 'no information' (Table S2), in prognostic models this was the case for both Outcome and Analysis domains (Table S3). Further information about each signalling question is provided in Supplemental file 3.

Table 1. PROBAST Signaling questions for model development and validation analyses in 152 included studies

		Developed models (n=152)			External validations (n=19)		
		Yes, probably yes *n* (%; 95% CI)	No, probably no *n* (%; 95% CI)	No information *n* (%; 95% CI)	Yes, probably yes *n* (%; 95% CI)	No, probably no *n* (%; 95% CI)	No information *n* (%; 95% CI)
Participants							
1.1	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	111 (73.0, 65.5 to 79.4)	32 (21, 15 to 28)	9 (6, 3 to 11)	14 (74, 51 to 88)	5 (26, 12 to 49)	0
1.2	Were all inclusions and exclusions of participants appropriate?	89 (59, 51 to 66)	16 (11, 7 to 16)	47 (31, 24 to 39)	7 (37, 19 to 59)	0	12 (63, 41 to 81)
Predictors							
2.1	Were predictors defined and assessed in a similar way for all participants?	109 (71.7, 64.1 to 78.3)	19 (13, 8 to 19)	24 (16, 11 to 22)	8 (42, 23 to 64)	1 (5, 0 to 25)	10 (53, 32 to 73)
2.2	Were predictor assessments made without knowledge of outcome data?	88 (58, 50 to 66)	4 (3, 1 to 7)	60 (40, 32 to 47)	10 (53, 32 to 73)	2 (11, 3 to 31)	7 (37, 19 to 59)
2.3	Are all predictors available at the time the model is intended to be used?	117 (77.0, 69.7 to 83.0)	4 (3, 1 to 7)	31 (20, 15 to 28)	12 (63, 41 to 81)	1 (5, 0 to 25)	6 (32, 15 to 54)
Outcome							
3.1	Was the outcome determined appropriately?	114 (75.0, 67.6 to 81.2)	6 (4, 2 to 8)	32 (21, 15 to 28)	12 (63, 41 to 81)	0	7 (37, 19 to 6)
3.2	Was a prespecified or standard outcome definition used?	118 (77.6, 70.4 to 83.5)	6 (4, 2 to 8)	28 (18, 13 to 25)	13 (68, 46 to 85)	0	6 (32, 15 to 54)
3.3	Were predictors excluded from the outcome definition?	90 (59, 51 to 67)	8 (5, 3 to 1)	54 (36, 28 to 43)	10 (53, 32 to 73)	0	9 (47, 27 to 69)
3.4	Was the outcome defined and determined in a similar way for all participants?	118 (77.6, 70.4 to 83.5)	11 (7, 4 to 13)	23 (15, 10 to 22)	10 (53, 32 to 73)	1 (5, 0 to 25)	8 (42, 23 to 64)
3.5	Was the outcome determined without knowledge of predictor information?	63 (41, 34 to 49)	10 (7, 4 to 12)	79 (52, 44 to 60)	4 (21, 9 to 43)	1 (5, 0 to 25)	14 (74, 51 to 88)
3.6	Was the time interval between predictor assessment and outcome determination?	110 (72.4, 64.8 to 78.9)	2 (1, 0 to 5)	40 (26, 20 to 34)	11 (60, 36 to 77)	1 (5, 0 to 25)	7 (37, 19 to 59)

Figure 2. Risk of bias of included studies (n=152) and stratified by study type

DISCUSSION

Principal findings

We have conducted a detailed assessment of the methodological quality of supervised ML-based prediction model studies across all clinical fields. Overall, 133/152 (87.5%) developed models and 15/19 (78.9%) external validations showed high RoB. The Analysis domain was most commonly rated as high RoB in developed models and external validations, mainly due to a low number of participants with the outcome (relative to the number of candidate predictors), risk of overfitting, and inappropriate handling of participants with missing data. Although there are still no conclusive studies about sample size calculations for developing prediction models using ML techniques, these usually require (many) more participants and events than conventional statistical approaches.[23,24] One hundred studies failed to either provide the number of events or reported an event per candidate predictor (EPV) lower than 10, which historically is a marker of potentially low sample size. Furthermore, ML studies with a low number of participants with the outcome are likely to suffer from overfitting, that is the model is too much tailored to the development dataset. [23–26] Only half of the included studies examined potential overfitting of models either by using split data, bootstrapping or cross-validation. Random-split was often relied on to internally validate models (i.e. validation based on the same participants' data), whereas bootstrapping and cross-validation are generally considered more appropriate.[27]

Most studies carried out complete-case analyses or mean/median imputation. Multiple imputation is generally preferred as it prevents biased model performance due to deletion or single imputation of participants' missing data. Unfortunately, multiple imputation is still unpopular within models developed with ML techniques.[28,29] Some ML techniques have the power to incorporate this missingness by including a separate category of a predictor variable that has missing values.[30] Therefore, we urge algorithm developers to improve imputation methods and incorporate informative missingness in their models when possible.

Several signalling questions were scored as 'No information' making it impossible for us to judge potential biases. It was often unclear whether all enrolled participants were included in the analyses, how many participants had missing values, and how missing data were handled. ML are powerful and automated techniques that will learn from data, however, if there was selection bias in the dataset, predictions made using the trained ML algorithm will also be biased. Similarly, several signaling questions in PROBAST are tailored to identify lack of blinding (SQ 2.2, SQ 3.3, SQ 3.5); however, almost half of included articles failed to report any information for us to assess blinding. Furthermore, model calibration tables or plots were often not presented, whereas classification measures (i.e. confusion

matrix) were commonly reported with an overreliance on accuracy.[31] Reporting and assessment of discrimination (i.e., ability to discriminate between cases and non-cases) and calibration (i.e., agreement between predictions and observed outcomes) is essential to assess a models' predictive accuracy.[31]

Comparison with other studies

A systematic review of 23 studies about ML for diagnostic and prognostic predictions in emergency departments shows that analysis was the most poorly rated domain with 20 studies at high RoB.[32] This study found deficiencies in how continuous variables and missing data were handled, and found that model calibration was rarely reported. Another publication about ML risk prediction models for triage of patients entering the emergency room also considered 22/25 studies considered at high RoB.[33] A study assessing the performance of diagnostic deep learning algorithms for medical imaging reported 58 of 81 studies being classified as overall high RoB.[7] Similar to our results, major deficiencies were found in the analysis domain including the number of events per variable, inclusion of enrolled participants in the analysis, reporting of relevant model performance measures, and overfitting. Recently, a living systematic review about COVID-19 prediction models indicated that all 57 studies that used ML were at high RoB due to insufficient sample size, unreported calibration, and internal validation based on training-test split.[34]

Strength and limitations of the study

We evaluated the risk of bias of supervised ML-based prediction model studies in a broad sample of articles which included prognostic and diagnostic development only and development with external validation studies. After using a validated search strategy, we retrieved nearly 25,000 publications which is similar to a previous study.[35] We finally screened the tenth part of the whole sample; therefore, our results are presented using confidence intervals to extrapolate them to the whole sample. The present analyses considered results from studies that were published over one year ago; nevertheless, we expect these findings to be still applicable and relevant for the clinical prediction field. We adopted PROBAST as the benchmark to evaluate RoB enhancing the objectivity and consistency, however, this is not without certain limitations. While two signalling question in PROBAST might become less relevant within the ML context (i.e. selection of predictors based on univariable analysis and reporting of weighted estimates in the final model correspond to the results from the reported multivariable analysis), further signalling questions related to data generation, feature selection, and overfitting might be necessary.

Implication for researchers, editorial offices, and future research

The number of ML-based studies is increasing every year; thus, their identification, reporting and assessment become even more relevant. It will remain a challenge to

determine the risk of bias if detailed information about data and modelling approach (including justifications to any decision made that may biases estimates) is not clearly reported in articles. To better judge studies, we recommend researchers to adhere to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.[36,37] Although TRIPOD was not exlicilty developed for machine learning prediction models, all items are applicable. Similarly, while there is yet no RoB assessment tool available specifically for supervised ML models, we suggest researchers to follow PROBAST recommendations to reduce potential biases when planning and modelling primary prediction model studies using either regression or non-regression models. For example, the adoption of multiple imputation to handle missing value and cross-validation or bootstrapping to internally validate the developed models.

Currently, extensions of TRIPOD and PROBAST for prediction models developed using machine learning are under development (TRIPOD-AI, PROBAST-AI).[36,37]As sample size contributed largely to the overall high RoB, future methodological research could focus on determine appropriate sample sizes for each supervised learning technique. Giving the rapid and constant evolution of machine learning, periodic systematic reviews of prediction model studies need to be conducted. Although high quality ML-based prediction model studies are scarce, those who stand out need to be validated, re-calibrated, and promptly implemented in clinical practice.[34] To avoid research waste, we suggest peer-reviewers and journal's editors to promote the adherence to reporting guidelines.[5] Facilitating the documentation of studies (i.e. supplemental material, data, and code) and setting unlimited word count may improve methodological quality assessment, as well as independent validation (i.e. replication). Likewise, requesting external validation of prediction models upon submission might help setting minimum standards to ensure generalizability of supervised ML-based prediction models studies.

CONCLUSION

Most supervised ML-based prediction model studies show poor methodological quality and are at high risk of bias. Factors contributing to the risk of bias include the exclusion of participants, small sample size, poor handling of missing data, and failure to address overfitting. Efforts to improve the design, conduct, reporting, and validation of supervised ML-based prediction model studies are necessary to boost its application in clinical practice and avoid research waste.

Author statements
The authors would like to thank and acknowledge the support of René Spijker, information specialist.

Authors' contributions
The study concept and design were conceived by CLAN, JAAD, PD, LH, RDR, GSC, and KGMM. CLAN, JAAD, TT, SN, PD, JM, and RB conducted article screening and data extraction. CLAN performed data analysis and wrote the first draft of this manuscript, which was revised by all authors who have provided the final approval of this version. CLAN, the corresponding author, is the guarantor of the review. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding
There is no specific funding for this study. GSC is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is supported by the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication.

Competing interests
GSC, RDR and KGMM are members of the PROBAST Steering Group. All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval
*W*e analysed only published data; therefore, ethics approval was not required.

Data sharing
The study protocol is available at ***doi:*** 10.1136/bmjopen-2020-038832 . Detailed extracted data on all included studies are available upon reasonable request to the corresponding author.

Transparency
The guarantor of this review affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination plans
We plan to disseminate the findings and conclusions from this study through social media (such as Twitter), a plain-language summary on www.probast.org, and scientific conferences. In addition, the findings will provide insights to the development of PROBAST-AI.

Provenance and peer review
Not commissioned, externally peer reviewed

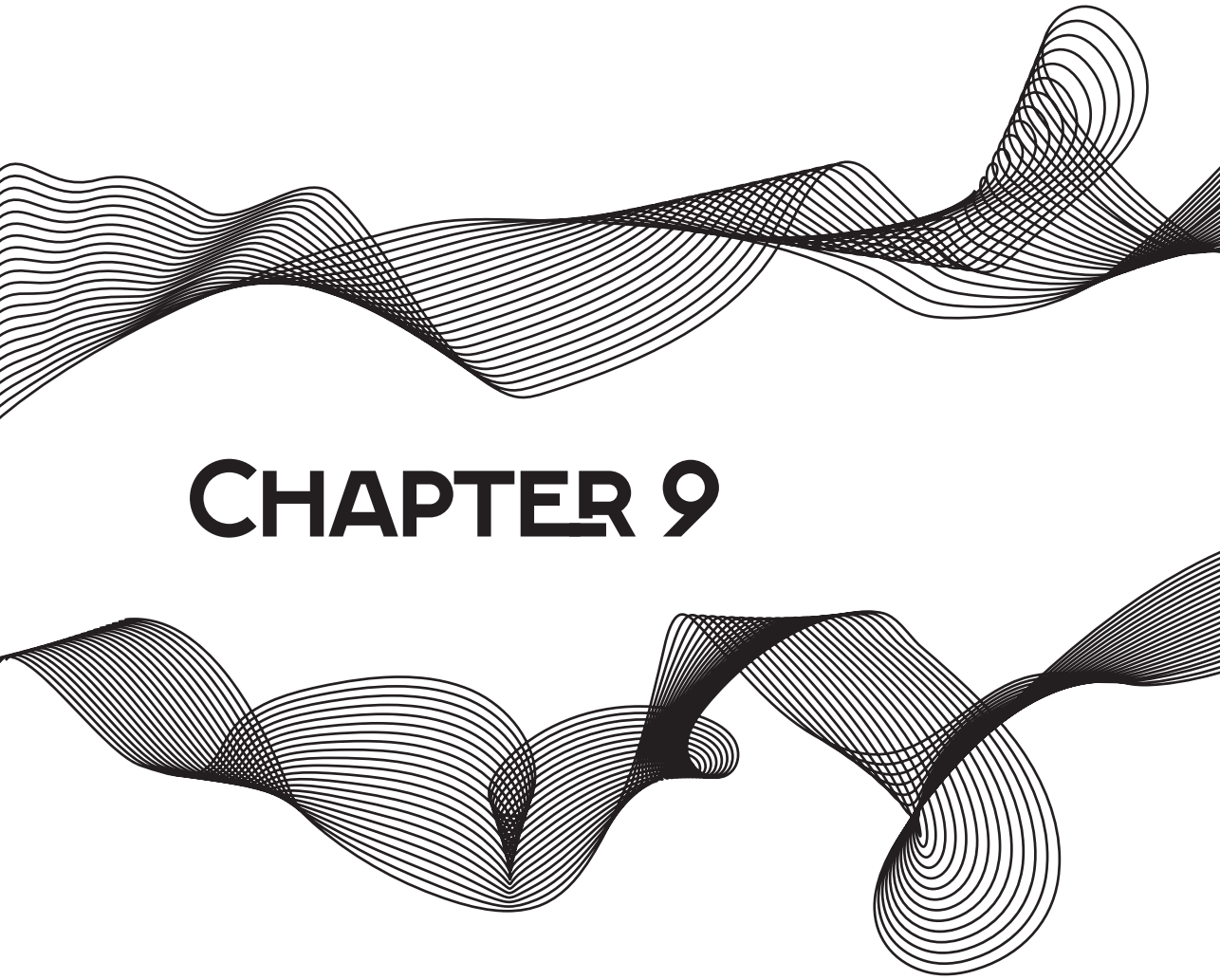
Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1136/bmj.n2281

REFERENCES

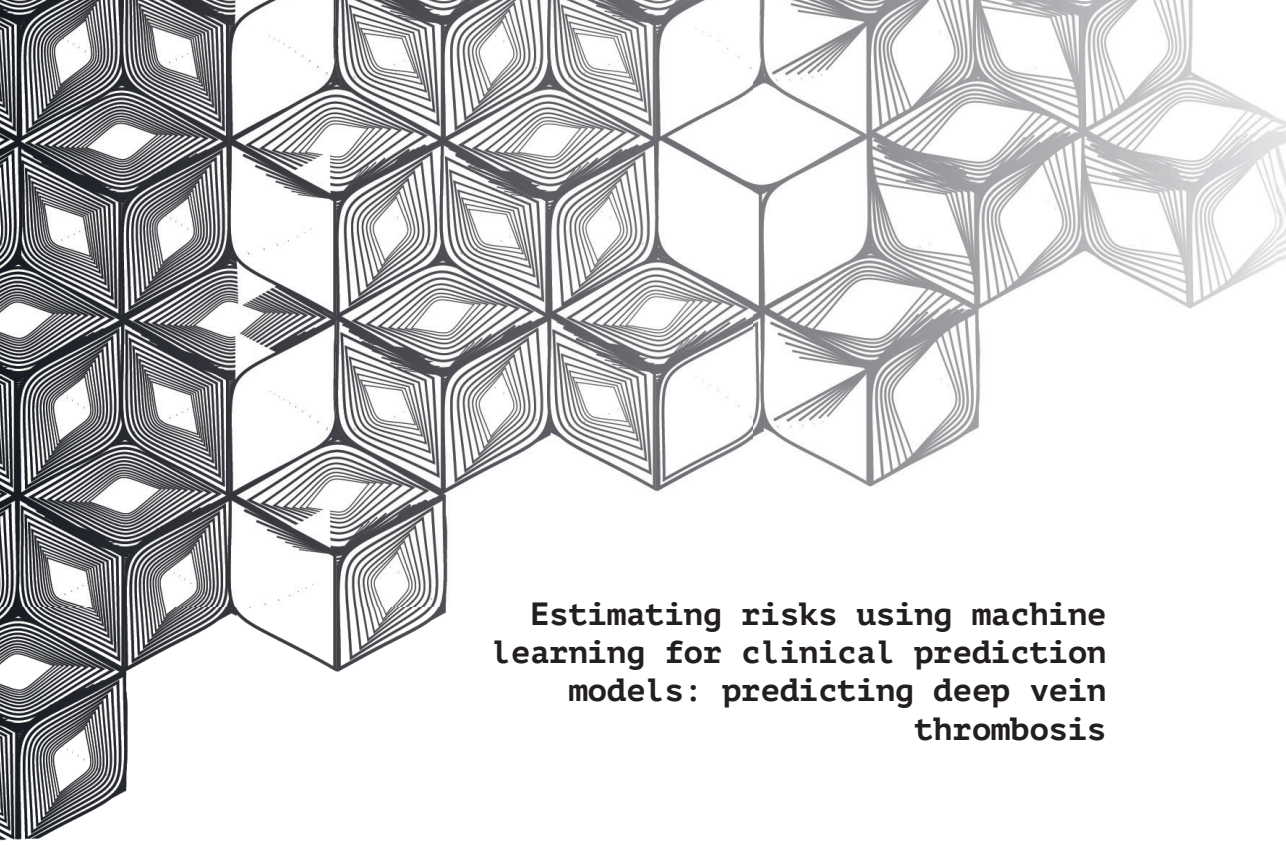
1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ*. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
2. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10(2). doi:10.1371/journal.pmed.1001381
3. Riley, Richard D; van der Windt, Danielle; Croft, Peter; Moons KGM. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. Oxford University Press; 2019. doi:10.1093/med/9780198796619.001.0001
4. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Second. Springer; 2019. doi:10.1007/978-3-030-16399-0
5. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X
6. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/BMJ.I2416
7. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368:1-12. doi:10.1136/bmj.m689
8. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189
9. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):1-18. doi:10.1186/s12874-019-0681-4
10. Mitchell T. *Machine Learning*. McGraw Hill; 1997.
11. Obermeyer, Ziad MD, Emanuel, Ezekiel J., M.D. PD. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1212-1216. doi:10.1056/NEJMp1606181.Predicting
12. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. *J Glob Health*. 2018;8(2). doi:10.7189/jogh.08.020303
13. 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*. 2018;1(1). doi:10.1038/s41746-018-0040-6
14. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Hear Fail*. 2021;8(1):106-115. doi:10.1002/ehf2.13073
15. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
16. Cho SM, Austin PC, Ross HJ, et al. Machine learning compared to conventional statistical models for predicting myocardial infarction readmission and mortality: a systematic review. *Can J Cardiol*. Published online March 5, 2021. doi:10.1016/j.cjca.2021.02.020
17. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
18. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Med*. 2012;9(5). doi:10.1371/journal.

pmed.1001221
19. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
20. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
21. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097
22. Andaur Navarro CL, Damen JAAG, Takada T, 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(11):1-6. doi:10.1136/bmjopen-2020-038832
23. Ploeg T Van Der, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry : a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137. doi:10.1186/1471-2288-14-137
24. Riley RD, Ensor J, Snell KIE. Calculating the sample size required for developing a clinical prediction model. 2020;441(March):1-12. doi:10.1136/bmj.m441
25. Courvoisier DS, Combescure C, Agoritsas T. Performance of logistic regression modeling : beyond the number of events per variable , the role of data structure. 2021;64(2011):993-1000. doi:10.1016/j.jclinepi.2010.11.012
26. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol*. 2016;76:175-182. doi:10.1016/j.jclinepi.2016.02.031
27. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26(2):796-808. doi:10.1177/0962280214558972
28. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*. 2009;339(7713):157-160. doi:10.1136/bmj.b2393
29. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010;63(2):205-214. doi:10.1016/j.jclinepi.2009.03.017
30. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic Progn Res*. 2020;4(1):4-9. doi:10.1186/s41512-020-00077-0
31. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1-7. doi:10.1186/s12916-019-1466-7
32. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. *Acad Emerg Med*. Published online 2020:1-13. doi:10.1111/acem.14190
33. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagnostic Progn Res*. 2020;4(1):1-12. doi:10.1186/s41512-020-00084-1
34. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*. 2020;369. doi:10.1136/bmj.m1328
35. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

36. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393. doi:10.1016/S01406736(19)302351
37. GS C, P D, CL AN, 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(7):e048008. doi:10.1136/BMJOPEN-2020-048008
38. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
39. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55. doi:10.7326/M14-0697



CHAPTER 9



Estimating risks using machine learning for clinical prediction models: predicting deep vein thrombosis

Constanza L Andaur Navarro
Johanna AA Damen
Toshihido Takada
Geert-Jan Geersing
Lotty Hooft
Karl GM Moons
Maarten van Smeden

Manuscript in preparation

ABSTRACT

Objectives. Machine learning (ML) algorithms are increasingly used for risk prediction in healthcare. Previous studies have shown that prediction models developed with ML algorithms often yield similar discriminative performance as those developed with traditional techniques, which may lead to the conclusion that both analytical approaches are exchangeable. The objective of this study was to illustrate differences in individual risk prediction between models developed using logistic regression (LR), random forests (RF), and support vector machine (SVM) in the situation where discrimination of the prediction models is similar.

Methods. A case study on individualized prediction modelling for the diagnosis of deep venous thrombosis (DVT) in suspected patients using a large individual participant dataset is presented. We developed five diagnostic models using three modelling algorithms: LR, RF, and SVM, but using five different implementation methods in the R package 'caret'. Age, sex, d-dimer, previous history of DVT, alternative diagnosis, and cancer were pre-specified as candidate predictors. Model performance was evaluated in terms of discrimination, calibration, and consistency of individual risk prediction for the same patients among models with comparable discrimination.

Results. We included 10,002 individuals of which 1864 were diagnosed with DVT. All prediction models had similar discrimination with AUCs ranging from 0.80 to 0.82. However, the probabilities for individual risk of DVT varied widely between and within the three different modelling algorithms and their implementation in the statistical software R. For the same individual, models provided probabilities ranging from 0 to 0.40. Compared to LR, the RF and SVM models over- and underpredicted individual risk probabilities. The RF model build using the implementation method 'ranger' had the highest correlation with LR model.

Conclusions. We showed that prediction models developed with different ML algorithms even with similar discriminative performance may have different individual risk estimation, thus yielding different calibration performance. Hence, our findings indicate the importance of assessing besides discrimination also calibration and the distribution of individual risk predictions, alongside the transparent reporting of implementation methods when intended for clinical decision making.

INTRODUCTION

Healthcare professionals regularly combine multiple pieces of information to estimate an individual's probability of having an outcome (diagnosis) or developing an outcome in the future (prognosis).[1–3] Some well-known examples are: QRISK3 and the Framingham risk score, models developed to predict the probability of cardiovascular disease in the general population[4] Traditionally, regression models such as logistic regression and Cox regression have been used to develop prediction models that estimate individual risk probabilities. However, recently machine learning have become increasingly popular in studies on prediction models.[8]

Several studies have suggested that prediction models based on machine learning can achieve better, if not, superior performance than regression-based models.[9–12] Evaluation of prediction models has primarily focused on discrimination which is the ability to separate participants with and without a particular outcome.[13,14] Whether regression or machine learning is used, systematic reviews have showed that studies on prediction models often report discrimination, while calibration is often lacking.[13–15] Although discrimination is an important aspect of prediction models' evaluation, it does not assess the accuracy of the individual risk predictions or probabilities, which is most crucial when using prediction models to inform clinical decisions about a particular individual. The accuracy of such probability estimation is rather assessed in terms of calibration, which indicates how well the estimated probability (or risk) of an outcome for a particular individual matches the observed frequency of that outcome among similar individuals.

The objective of this study is to illustrate the consistency among individual risk probabilities estimated using prediction models with comparable discrimination which were developed using logistic regression, random forest, and support vector machine. We used data from a large scale international individual participant data meta-analysis on deep venous thrombosis (DVT) as an example.

METHODS

Data sources

We used individual participant data from prospective diagnostic studies of patients suspected of deep vein thrombosis (DVT).[16] Authors of 13 studies from USA, Canada, the Netherlands, and Sweden provided their original datasets anonymized, and were merged into one dataset. Studies were eligible if they enrolled consecutive patients with suspected DVT. Further details on the construction of the dataset can be found elsewhere.[16]

Predictors and outcome

We selected *a-priori* six predictors: age (continuous), sex (categorical, yes/no), d-dimer (categorical, yes/no), previous history of DVT (categorical, yes/no), alternative diagnosis (categorical, yes/no), and cancer (categorical, yes/no) based on previously developed prediction models for DVT.[16,17] The outcome was dichotomous with values no DVT = 0 (absence), and DVT = 1 (presence).

Classification algorithms

We used two machine learning (ML) algorithms to develop a diagnostic DVT prediction model: random forest (RF) and support vector machine (SVM). RF and SVM algorithms are considered non-probabilistic classifier because they do not 'naturally' produce class probabilities, but rather have been adapted to provide probabilities. RF is an ensemble technique consisting of multiple decision trees trained on bootstrapped sub-sets of the full dataset and different initial variable. The outcomes for each of these decision trees are aggregated and the most popular outcome value is 'voted'.[18,19] SVM aims to find the best separating hyperplane (decision boundary) between classes, which can be either linear or non-linear. SVM then maximizes the marginal distance between classes (support vector) and returns the corresponding hyperplane to carry out predictions on new data.[20] Last, we have applied logistic regression (LR) to develop a probabilistic diagnostic DVT prediction model as a benchmark.

Data analysis

We used the R meta-package "caret" version 6.0.89 to obtain a series of five models (Table 1) based on the common set of predictors (above). The "caret" package (Classification And Regression Training) provides a uniform interface to run several machine learning algorithms (called *implementation methods*) for which packages are available independently but having different syntax for similar tasks. Hence, "caret" provides a syntax convention for tasks such as data preparation, data splitting, parameter tuning, and variable importance in one uniform interface.[21] Modelling options available in "caret" are listed on https://topepo.github.io/caret/available-models.html.

Model development

We used five different implementation methods available in the "caret" meta-package to develop the models (Table 1). Imputed datasets were not pooled for analyses, instead one imputed dataset was used to trained models and another one was used to tested them. We fitted two RF models using the implementation method "ranger" and "rf" which are based on the packages ranger[22] and randomForest[23], respectively. We fitted two SVM models using the implementation method "svmLinear" and "svmLinear2" which are based on the packages kernlab[24] and e1071[25], respectively. Finally, as benchmark, we fitted one LR model using the implementation method "glm" based on the package glmnet.[26] Overall, we fitted a total of five models in which, except for LR, non-linear associations and interactions were done automatically. The inherent clustering of the IPD dataset was ignored for simplicity. We expect discrimination of all fitted models to be in close range to allow for the comparison of calibration and individual risk probabilities. We performed 3 repeats of 10-fold cross-validation to find the best values for tuning parameters based on highest AUC.

Table 1. Details on model development

Algorithm	Independent R packages	Implementation method in "caret" package	Tuning parameters
Logistic regression	glmnet	glmnet	alpha, lambda
Random Forest 1	e1071, ranger, dplyr	ranger	Mtry, splitrule, min.node.size
Random Forest 2	randomForest	rf	mtry, splitrule, min.node.size
Support vector machine 1 with linear kernel	kernlab	svmLinear	C
Support vector machine 2 with linear kernel	e1071	svmLinear2	cost

For RF, we explored two different ways to calculate individual risk probabilities. The package ranger calculates a probability forest whereby each tree returns a probability class estimate and these estimates are later averaged to provide an averaged probability.[27] Conversely, the package randomForest makes use of a process called 'hard' majority voting. For 'hard' majority voting the individual probability equals the number of trees that indicate class 1, i.e., if we have 500 trees in a RF model, and 400 of them indicate class 1, the probability for class 1 would be 80%. Thus, predicted risk probabilities are either (1) the mean terminal leaf probability across all trees or (2) the proportion of trees voting either class. The package ranger allows to estimate both scenarios.

For SVM, both kernlab24 and e107125 packages are based on the software libsvm.28 While e1071 offers a rigid interface to libsvm with visualization and parameter tuning, kernlab provides a SVM based on the optimizers used in libsvm and bsvm29, alongside a variety of kernel-based methods. In both implementation methods, probabilities are obtained after a second regression model has been trained on the SVM outputs.

Sample size

We used the R package pmsampsize30 to calculate the sample size required to develop a new prediction model based on logistic regression. One is required to input the overall fraction of participants expected to develop DVT (18%), the number of candidate predictor parameters (n=6) and the anticipated c-statistic (0.89). The minimum sample size required is of at least 227 participants with 41 events, and 6.81 events per candidate predictor parameter (EPP) while considering a shrinkage factor of 0.906 and R2 Cox-Snell (R2cs) of 0.2957.

Statistical analysis

We applied each model to a test dataset which was one of the imputed datasets. We calculated discrimination graphically by plotting receiver operating characteristics (ROC) curves and quantitatively by calculating Area Under the Curve (AUC). A perfect model would have an AUC equal to 1. We also assessed calibration graphically by plotting reliability curves. Calibration plots (i.e., reliability diagrams) were used to graphically compare the posterior class probabilities among models. If the model is well calibrated, the points will fall near the diagonal line. We assessed classification with sensitivity, specificity, positive predictive value, and negative predictive value. We considered the threshold of 0.5 to decide an instance as positive.

To compare individual risk probabilities for the same individuals across implementation methods, we provided a scatter plot that shows the correlation between the outputs of the algorithms. In the diagonal line, the distribution of probabilities across the entire range of probabilities (density plot).

All statistical analyses were performed in R version 4.0.3 (www.R-project.org) with R base, crossTable, rms, pROC, ggplot, and caret packages. We performed no additional coding or changes to the basic algorithms underlying these libraries, as these already provided the prob option allowing the estimation of posterior individual probabilities.

RESULTS

The final dataset included 10,002 patients, 1864 (18.6%) were confirmed cases of DVT. The mean age was 59.2 (SD±17.3) years for all participants, of which 6155 (61.5%) were female. Further details on the study participants can be found in the original publication16 and in Table 2.

Table 2. Characteristics of study population, stratify by outcome class

	Total	No DVT	DVT
	Values	Values	Values
Age	59.2 (± 17.3)	58.8 (± 17.4)	61.2 (± 17)
Sex			
Female	6155 (61.5)	5180 (63.7)	975 (52.3)
Male	3847 (38.5)	2958 (36.3)	889 (47.7)
D-dimer			
No	4752 (47.5)	4601 (56.5)	151 (8.1)
Yes	5250 (52.5)	3537 (43.5)	1713 (92.9)
Cancer			
No	9168 (91.7)	7621 (93.6)	1547 (83)
Yes	834 (8.3)	517(6.4)	317 (17)
Previous history of DVT			
No	9020 (90.2)	7383 (90.7)	1637 (87.8)
Yes	982 (9.8)	755 (9.3)	227 (12.2)
Alternative diagnosis			
No	5240 (52.4)	3795 (46.6)	1445 (77.5)
Yes	4762 (47.6)	4343 (53.4)	419 (22.5)
Total	10002 (100)	8138 (81.6)	1864 (18.6)

Values are counts and column percentages for categorical variables; means ± SDs for continuous variables with a normal distribution. DVT = deep vein thrombosis.

Tuning parameters

The best values for hyperparameters are presented in Table S1 (Supplemental material).

Discrimination

The five models achieved AUCs ranging from 0.81 to 0.82 on the test dataset. Figure 1 displays the receiver operating characteristic curves for all the models in this study. While LR (glm) and RF model 1 (ranger) models showed the highest discrimination (0.82, 95% CI [0.81-0.83]), RF model 2 (randomForest) and SVM model 1 (kernlab) both showed AUCs of 0.80 (95% CI 0.79-0.81).

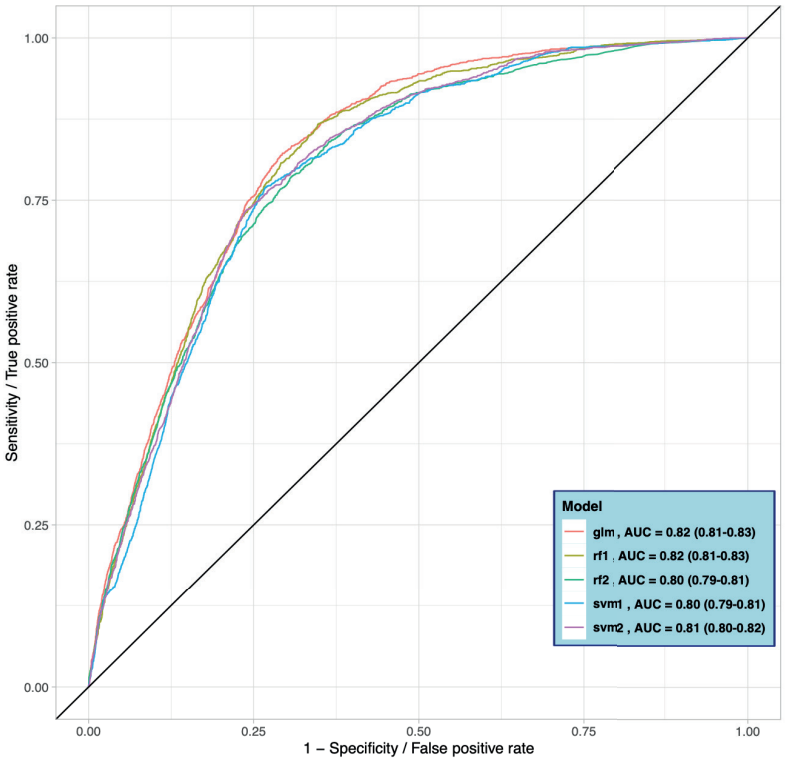


Figure 1. Receiver operating characteristic curves

Calibration

The calibration curves are shown in Figures 3 and 4 and suggest that both RF and SVM are miscalibrated when compared to the benchmark logistic regression model. The calibration plot for the LR model (glm) is shown in Figure 2. The LR model appears to accurately estimate risks across the range of probabilities. The calibration plots for RF models are shown in Figure 2. RF model 1 (ranger) shows miscalibration in the lower as well as the upper end of risks spectrum (left). RF Model 1 slightly underestimated probabilities between the ranges 0.0 to 0.2, while between 0.3 and 0.7, it overestimated risks. The RF model 2 (randomForest) shows miscalibration in both ends, with an extreme underestimation of low risks until 0.7, after which risks are slightly overestimated (right). The calibration plots for SVM models are shown in Figure 3. SVM model 1 (kernlab) shows between 0.2 to 0.4 an underestimation of risk with an overestimation on the low and high extreme. SVM model 2 (e1071) shows probabilities that are not extreme enough, with overestimation in the low range until 0.2, after which there was a consistent underestimation of high risks (right).

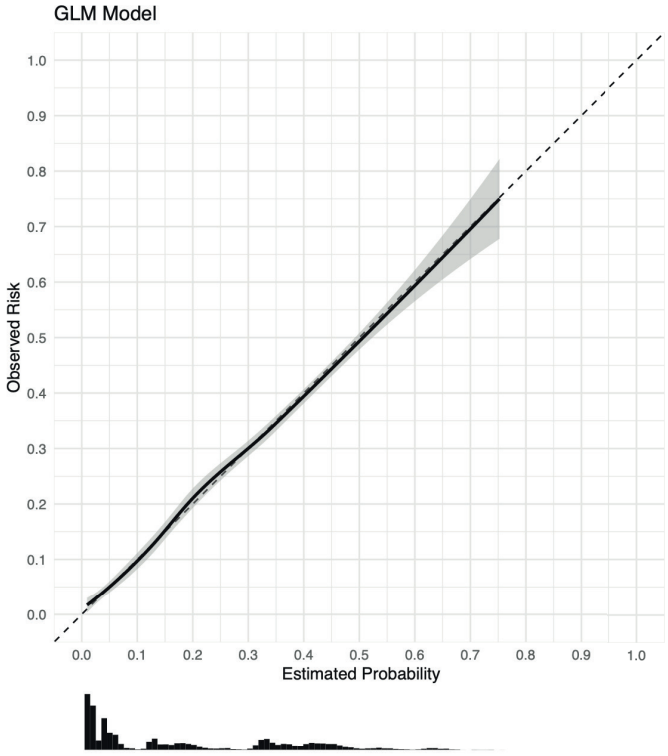


Figure 2. Calibration plot for model based on logistic regression.

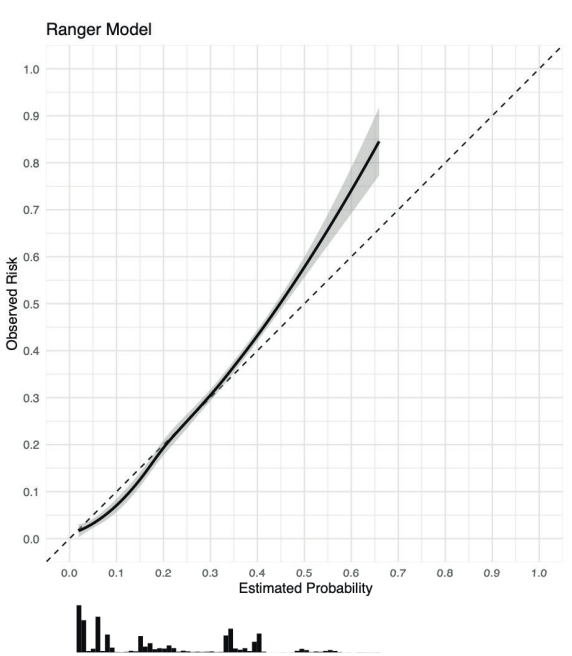
Classification

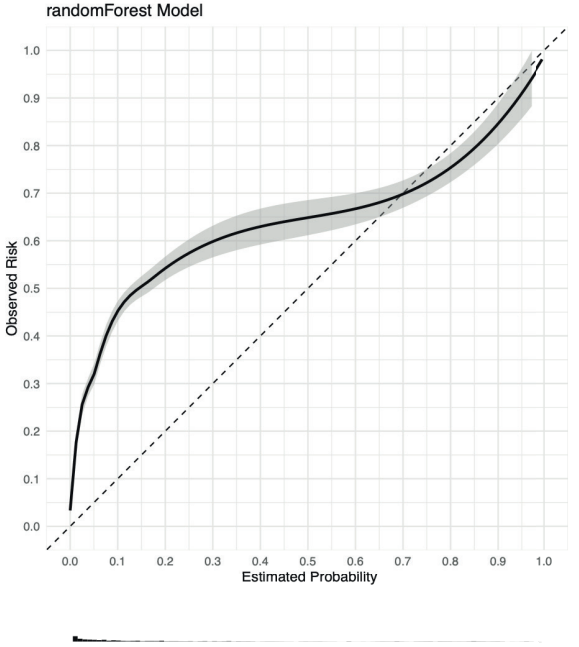
Table 3 summarizes the classification performance of the five fitted models when applied to the test dataset. The sensitivity of both RF and SVM were low (range 0-0.16), but most models had high specificity (range 0.97-1.0).

Table 3. Performance measures in test dataset

Machine	Performance measures				
	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV / Recall (95% CI)	NPV (95% CI)
Logistic regression	0.82 (0.81-0.83)	0.16 (0.15-0.18)	0.97 (0.97-0.98)	0.59 (0.55-0.64)	0.83 (0.83-0.84)
Random Forest 1 (ranger)	0.82 (0.81-0.83)	0.11 (0.10-0.13)	0.98 (0.98-0.99)	0.64 (0.59-0.69)	0.83 (0.82-0.84)
Random Forest 2 (randomForest)	0.80 (0.79-0.81)	0.12 (0.11-0.14)	0.98 (0.98-0.99)	0.62 (0.57-0.67)	0.83 (0.82-0.84)
Support Vector 1 (kernlab)	0.80 (0.79-0.81)	0.11 (0.10-0.12)	0.97 (0.97-0.97)	0.47 (0.42-0.51)	0.83 (0.82-0.83)
Support Vector 2 (e1071)	0.81 (0.80-0.82)	0	1	na	0.81 (0.81-0.82)

CI= confidence interval. Confidence interval for AUC were calculated using DeLong method and for classification measures using Wilson score interval.





Most prediction are equal to 0 in the ranger model, so their distribution is not visible in this plot. To see the distribution, please see figure 5.

Figure 3. Calibration plot of different implementation methods for Random Forest

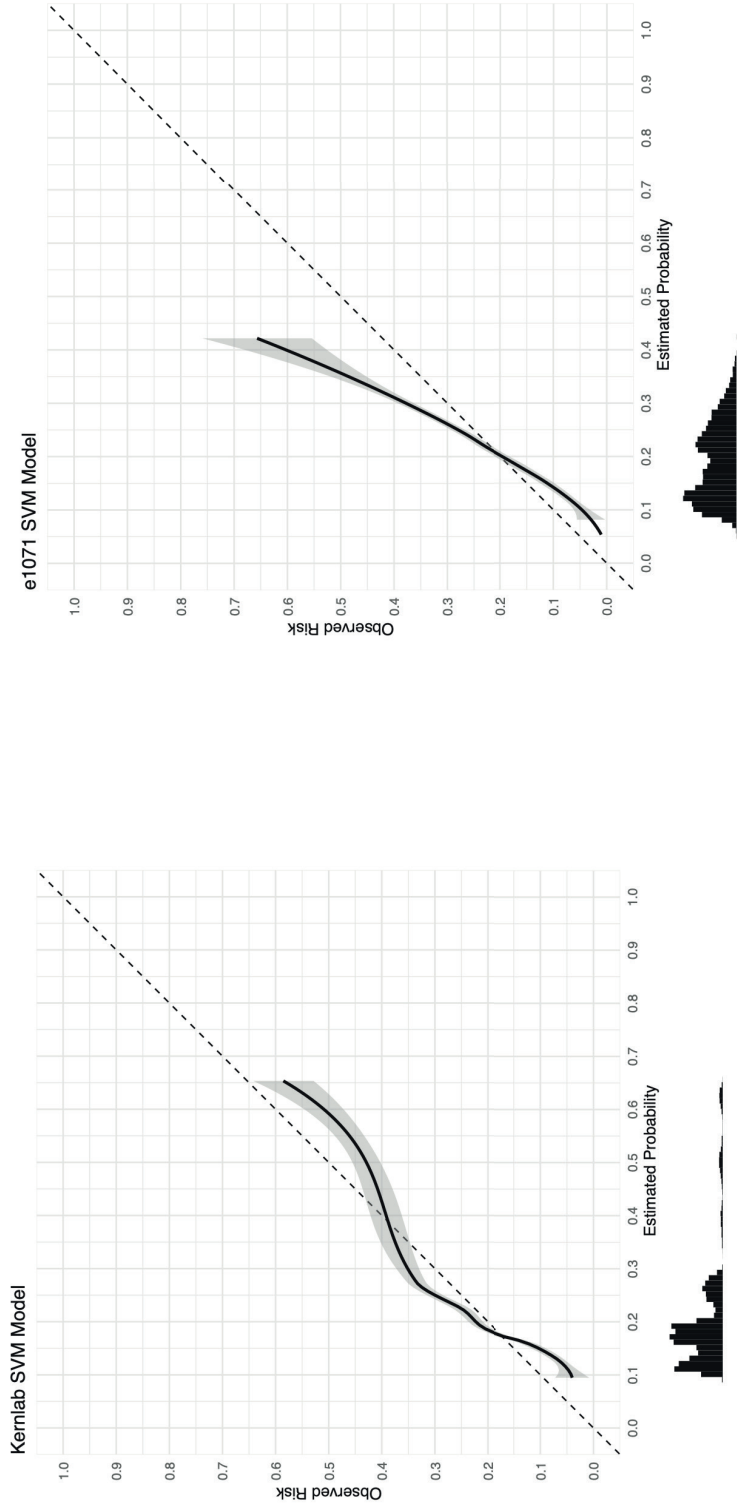


Figure 4. Calibration plot of different implementation methods for Support Vector Machine

Individual risk probabilities

The range of probabilities also differed between and within machine learning techniques. While RF model 2 (ranger) provides a broader range of probabilities (from 0 to 1), SVM model 2 (e1071) had the narrowest range (from 0.05 to 0.4). Likewise, the estimated probabilities are not equally distributed across the range of probabilities, with most individual probabilities found in the lower end of probabilities. For example, most individual probabilities provided by RF model 2 (randomForest) were close to 0 (diagonal line, Figure 5).

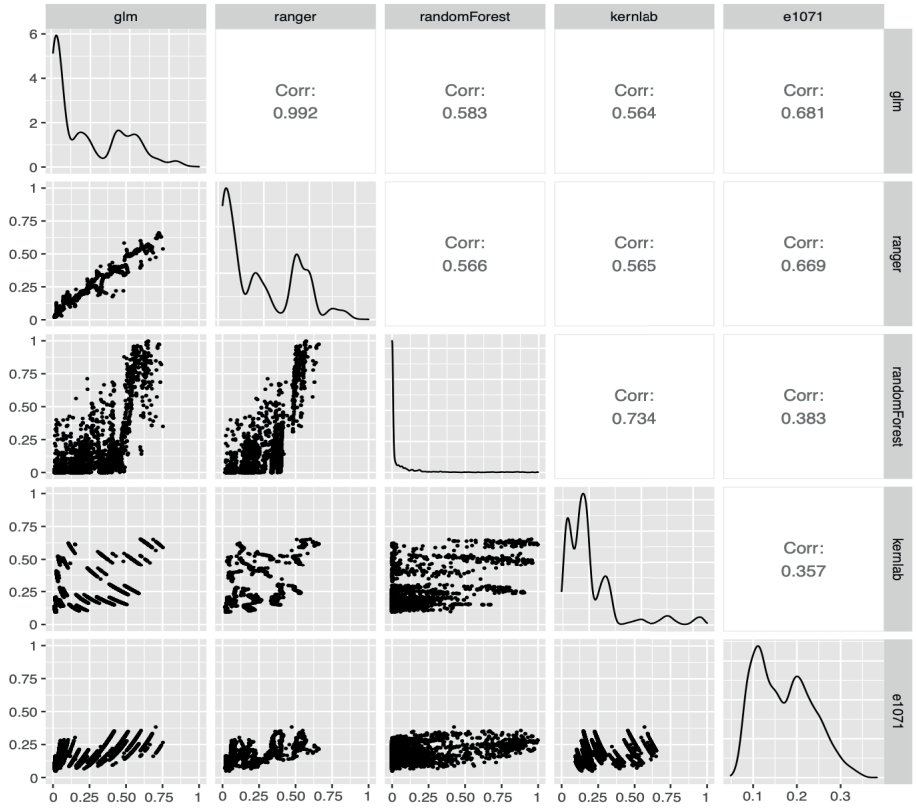


Figure 5. Scatter plot matrix

We also compared the different estimated risk probabilities for ten random individuals (Figure 6). For example, the individual ID 4855 had a 0.04 probability of DVT being present based on the glm model, while the models based on machine learning provided risk probabilities from 0 to 0.21 for the same individual (RF model 2 and SVM model, respectively). Further details are provided on Table S2.

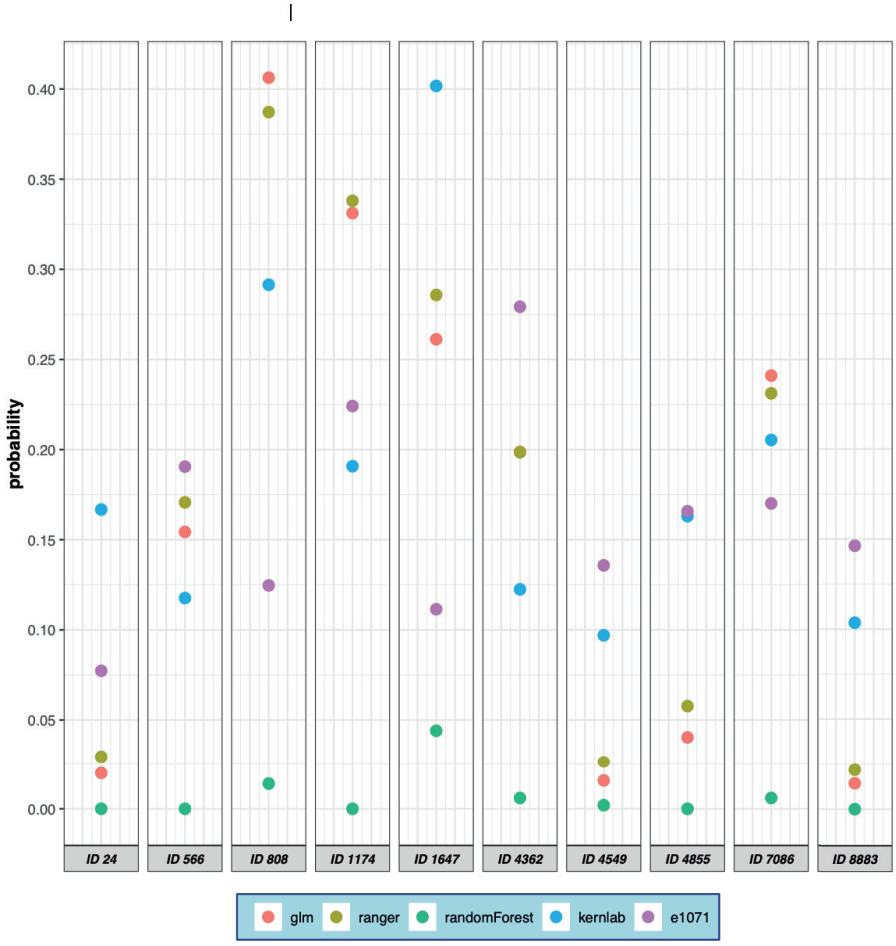


Figure 6. Comparison of risk probabilities for ten random individuals

DISCUSSION

We showed that individual risk probabilities given by prediction models developed using different implementation methods may vary considerably even when models have similar discrimination. In this study, RF and SVM models frequently under- or over-predict risks as compared to conventional logistic regression.

Accurately predicting the risk of DVT in suspected patients can help decide on their further diagnosis management and treatment. Our study compared a conventional modelling approach with machine learning to build several diagnostic models to estimate the risk of DVT based on six well-known predictors. All models (including logistic regression) showed good discriminative ability. Furthermore, models correctly excluded DVT risk on most individuals, however, the true cases of DVT were still difficult to identify.

Miscalibration reduces the clinical utility of a prediction model because errors in individual estimated probabilities can lead to inappropriate management or treatment decisions, or poor allocation of resources which could harm individuals. Issues related and non-related to the modelling technique may distort the calibration of individual risk predictions. For example, when a prediction model is developed in a setting with a high disease incidence, it may systematically overestimate risks when applied in a lower-incidence setting. To support clinical decisions, strong discrimination alone is insufficient, thus one of the challenges remains to develop well-calibrated prediction models.[31] If a model shows poor calibration upon validation, there are two popular approaches to recalibrate probabilities arising from ML classification models: (a) Isotonic regression and (b) Platt Scaling.[32] Isotonic Regression can support outputs with different shapes and recalibrate predictions from naïve bayes, SVM, and decision tree models. [33,34] The Platt scaling method trains a probability model on top of the SVM's outputs under a cross-entropy loss function to rescale them into probabilities (rescaled based on the maximum and minimum seen distances from the hyperplane).[32] To prevent this model from overfitting, it uses an internal five-fold cross validation. Platt Scaling is simpler and suitable for outputs with the S-shape.[32,34]

Comparison to previous research

A previous study showed that predictions for individual risks of cardiovascular disease (CVD) varied between and within different types of machine learning and regression models. In this study, a patient with a risk of 9.5-10.5% predicted by QRISK3, (a well-stablished CVD risk score) had a risk of 2.9-9.2% in a random forest and 2.4-7.2% in a neural network.[35] A study evaluating the performance of three different existing risk prediction models based on regression techniques for cardiovascular disease concluded that their application may result in considerable

misclassification for individuals with highest risk.[36]

Strengths and limitations of this study

This analysis included IPD from more than 10,000 participants with suspected DVT, of which 1864 were diagnosed with DVT. Given that our dataset was based on 13 different studies, data used to build models were inherently clustered. As we have ignored clustering during model development, this could have possibly affected the accuracy (or the associations) of the predictors. Nonetheless, a study using the same IPD dataset for predicting DVT in patients with cancer taking into account clustering of data showed similar discriminative ability to our study.[17] We developed SVM models with linear kernel because only this setting was available in both implementation methods and it allows us to compare the outputs with logistic regression.[37] We used a threshold of 0.5 to determine classes. However, diagnosis of DVT requires high sensitivity because it can potentially lead to dead. Further analyses using clinical significance thresholds need to be carried out.

Implications for researchers and future research

Systematic reviews have found that calibration is assessed far less often compared to discrimination, irrespective of whether models were built using machine learning or regression techniques.[13,14] However in many situations, healthcare staff are better informed by having an estimated individual probability rather than a binary class to provide information about an individual's likelihood on a particular diagnosis or prognosis. The lack of calibration assessment limits the use of clinical prediction models and consequently, several guidelines have stressed the need to report calibration alongside discrimination.[38] Likewise, as shown in this study, the reporting of packages and details on the implementation's methods is necessary for critical appraisal. To provide guidance on reporting, the TRIPOD reporting guideline (www.tripoid-statement.org) and the PROBAST risk of bias assessment tool (www.probast.org) tailored to studies on AI-based prediction models are soon to be launched.[39,40]

The R package "caret" provides an easy interface for the execution of many algorithms with only small changes to the code, thus reducing the requirements on the researcher's expertise and avoiding the researcher's own bias towards a particular algorithm.[21] The implementation method ranger seems to be a more flexible approach to build machine learning models based on RF given that returns a matrix (sample x tree) for classification and regression, a 3d array for probability estimation (sample x class x tree) and survival (sample x time x tree). Hence, it allows you to return not only the individual prediction for each tree but also the aggregated prediction for all trees. To be noted, SVM probabilities originate from a secondary regression model fitted to the predicted classes and thus, the accuracy of the prediction may decrease. To avoid this loss of accuracy, researchers building probabilistic models should focus on testing probabilistic algorithms rather classifiers. Further research could extent

this study to multi-class probability estimation or to illustrate potential difference using other software such as Python with the packages "Sklearn" or "h2o".[35]

CONCLUSION

We found that using RF, SVM, and logistic regression to estimate the diagnostic probability of having deep venous thrombosis in suspected patients, could yield inconsistent individual risk predictions. Researchers should be aware that choices related to selecting the modelling approach and their implementation method affect the predicted risk and can therefore influence clinical decision-making. Our findings indicate the importance of assessing the distributions and calibration of individual risk predictions across modelling techniques and the thoughtful reporting of modelling steps.

Authors' contributions
The study concept and design were conceived by CLAN, JAAD, GJG, MvS and KGMM. TT and GJG provided insight into the dataset and clinical problem. CLAN performed data analysis and wrote the first draft of this manuscript, which was revised by all authors who have provided their final approval for this version. CLAN, the corresponding author, is the guarantor of the review. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding
GJG is supported by a Veni (016.166.030) and Vidi (91719304) grant from the Dutch Research Council (NWO/ZonMw).

Ethical approval
Not required for this work.

Data sharing
Requests for data sharing can be sent to GJG and will then be discussed with all other coauthors. Analytical code is available via repository www.github.com/ConstanzaAndaur/Absolute_risk.

Transparency
The lead author (CLAN) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Supplementary data
Supplementary data to this article can be found online at https://surfdrive.surf.nl/files/index.php/s/6JbSgYWXLj7jVxf

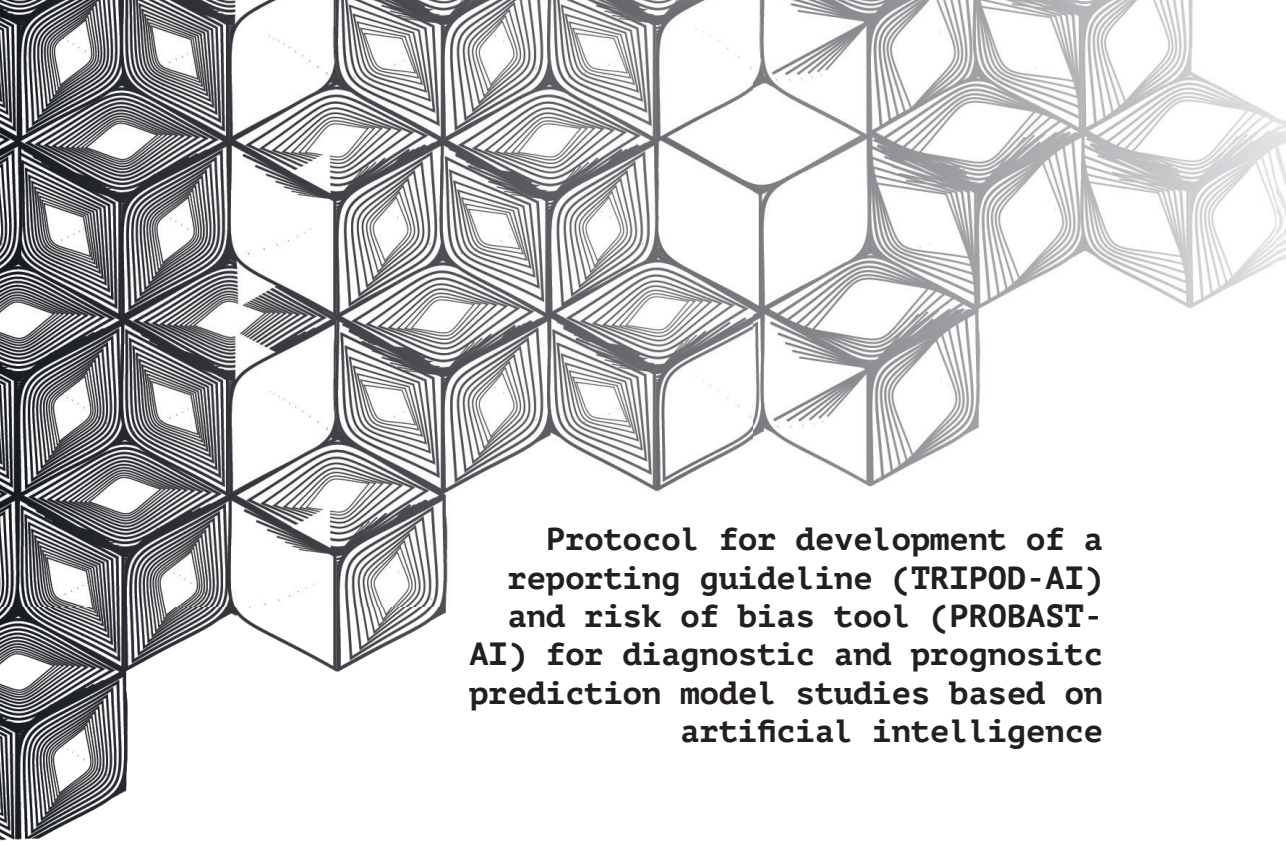
REFERENCES

1. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2). doi:10.1371/journal.pmed.1001381
2. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-690. doi:10.1136/heartjnl-2011-301246
3. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? BMJ. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
4. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. BMJ. 2017;357. doi:10.1136/bmj.j2099
5. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective. Lancet. 2014;383(9921):999-1008. doi:10.1016/S0140-6736(13)61752-3
6. Gail MH, Brinton LA, Byar DP, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI J Natl Cancer Inst. 1989;81(24):1879-1886. doi:10.1093/jnci/81.24.1879
7. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100(6):1619–1636. doi:https://doi.org/10.1378/chest.100.6.1619
8. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Informatics Assoc. 2022;00(0):1-7. doi:10.1093/jamia/ocac002
9. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One. 2018;13(8):e0202344. doi:10.1371/journal.pone.0202344
10. Hale AT, Stonko DP, Brown A, et al. Machine-learning analysis outperforms conventional statistical models and CT classification systems in predicting 6-month outcomes in pediatric patients sustaining traumatic brain injury. Neurosurg Focus. 2018;45(5):1-7. doi:10.3171/2018.8.FOCUS17773
11. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. Acad Emerg Med. Published online 2020:1-13. doi:10.1111/acem.14190
12. 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(4):e0174944. doi:10.1371/journal.pone.0174944
13. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22(1):12. doi:10.1186/s12874-021-01469-6
14. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. BMC Med. 2018;16(1):1-12. doi:10.1186/s12916-018-1099-2
15. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. BMJ Open. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
16. Geersing GJ, Zuithoff NPA, Kearon C, et al. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: Individual patient data meta-analysis. BMJ. 2014;348(March):1-13. doi:10.1136/bmj.g1340

17. Takada T, van Doorn S, Parpia S, et al. Diagnosing deep vein thrombosis in cancer patients with suspected symptoms: An individual participant data meta-analysis. J Thromb Haemost. 2020;18(9):2245-2252. doi:10.1111/jth.14900
18. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324
19. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016;46(12):2455-2465. doi:10.1017/S0033291716001367
20. Cortes C, Vapnik V. Support-Vector Networks. Machine. 1995;20(5):273-297. doi:10.1109/64.163674
21. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1-26. doi:10.18637/jss.v028.i05
22. Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77(1). doi:10.18637/jss.v077.i01
23. Liaw A, Wiener M. Package "randomForest." Published online 2022. doi:10.1023/A
24. Alexandros A, Smola A, Hornik K. Package ' kernlab ': Kernel-Based Machine Learning Lab. Published online 2022:1-108.
25. Meyer D, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C. Package 'e1071.' Published online 2022.
26. Hastie T, Qian J. An Introduction to glmnet. 2022;5(April):1-19.
27. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability Machines: Consistent probability estimation using nonparametric learning machines. Methods Inf Med. 2012;51(1):74-81. doi:10.3414/ME00-01-0052
28. Chang CC, Lin CJ. LIBSVM: A Library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3). doi:10.1145/1961189.1961199
29. Chih-Wei Hsu, Chih-Jen Lin. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Networks. 2002;13(2):415-425. doi:10.1109/72.991427
30. Riley RD, Ensor J, Snell KIEE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;(March):1-12. doi:10.1136/bmj.m441
31. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):1-7. doi:10.1186/s12916-019-1466-7
32. Platt J, others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv large margin Classif. 1999;10(3):61-74.
33. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Icml. Published online 2001:1-8.
34. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. 85th AMS Annu Meet Am Meteorol Soc - Comb Prepr. Published online 2005:4943-4947. doi:https://doi.org/10.1145/1102351.1102430
35. 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. doi:10.1136/bmj.m3919
36. Van Staa TP, Gulliford M, Ng ESW, Goldacre B, Smeeth L. Prediction of cardiovascular risk using framingham, ASSIGN and QRISK2: How well do they predict individual rather than population risk? PLoS One. 2014;9(10). doi:10.1371/journal.pone.0106455
37. Bischl B, Binder M, Lang M, et al. Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges. Published online 2021.
38. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73. doi:10.7326/M14-0698
39. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. Lancet. 2019;393. doi:10.1016/S01406736(19)302351
40. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a re-

porting 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):1-7. doi:10.1136/bmjopen-2020-048008

CHAPTER 10



Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognositc prediction model studies based on artificial intelligence

Gary S Collins
Paula Dhiman
Constanza L Andaur Navarro
Jie Ma
Lotty Hooft
Johannes B Reitsma
Patricia Logullo
Andrew L Beam
Lily Peng
Ben van Calster
Maarten van Smeden
Richard D Riley
Karl GM Moons

BMJ Open 2021;11(7):e048008

ABSTRACT

Introduction. The Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) statement and the Prediction model Risk Of Bias ASsessment Tool (PROBAST) were both published to improve the reporting and critical appraisal of prediction model studies for diagnosis and prognosis. This paper describes the processes and methods that will be used to develop an extension to the TRIPOD statement (TRIPOD- artificial intelligence, AI) and the PROBAST (PROBAST- AI) tool for prediction model studies that applied machine learning techniques.

Methods and analysis. TRIPOD-AI and PROBAST-AI will be developed following published guidance from the EQUATOR Network, and will comprise five stages. Stage 1 will comprise two systematic reviews (across all medical fields and specifically in oncology) to examine the quality of reporting in published machine- learning- based prediction model studies. In stage 2, we will consult a diverse group of key stakeholders using a Delphi process to identify items to be considered for inclusion in TRIPOD- AI and PROBAST- AI. Stage 3 will be virtual consensus meetings to consolidate and prioritise key items to be included in TRIPOD- AI and PROBAST- AI. Stage 4 will involve developing the TRIPOD- AI checklist and the PROBAST- AI tool, and writing the accompanying explanation and elaboration papers. In the final stage, stage 5, we will disseminate TRIPOD- AI and PROBAST- AI via journals, conferences, blogs, websites (including TRIPOD, PROBAST and EQUATOR Network) and social media. TRIPOD- AI will provide researchers working on prediction model studies based on machine learning with a reporting guideline that can help them report key details that readers need to evaluate the study quality and interpret its findings, potentially reducing research waste. We anticipate PROBAST- AI will help researchers, clinicians, systematic reviewers, and policymakers critically appraise the design, conduct and analysis of machine learning based prediction model studies, with a robust standardised tool for bias evaluation.

Ethics and dissemination. Ethical approval has been granted by the Central University Research Ethics Committee, University of Oxford on 10- December-2020 (R73034/RE001). Findings from this study will be disseminated through peer- review publications.

PROSPERO registration number CRD42019140361 and CRD42019161764.

INTRODUCTION

Models that predict clinical outcomes are abundant in the medical literature and are broadly categorised as those that estimate the probability of the presence of a particular outcome (diagnostic) or whether a particular outcome (eg, event) will occur in the future (prognostic).[1] Traditionally, these models (herein referred to as prediction models) have been developed using regression- based methods, typically logistic regression for short- term outcomes and Cox regression for longer- term outcomes.[2] Numerous reviews have observed that studies describing the development and validation (including updating) of a prediction model often fail to report key information to help readers judge the methods and have a complete, transparent and clear picture of the model's predictive accuracy and other relevant details such as the target population and the content of the model itself.[3–6] The absence of full and comprehensive reporting limits the usability of the findings of these studies, for example, in subsequent validation studies, evidence synthesis studies or in daily practice, and therefore, contribute to research waste.[7] In response to this, in 2015, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement was published.[1, 8] The TRIPOD Statement is a checklist of 22 items that authors should report with sufficient detail and clarity to inform how the study was carried out.

Since the publication of the TRIPOD Statement, artificial intelligence (AI) and in particular machine learning, approaches to clinical prediction have evolved and grown in popularity with the number of AI and machine learning publications rapidly rising.[9–14] This is evident within a recent review of COVID-19 related prediction models, where 57 (out of 107 included studies) used machine learning methods to develop their model.[15]

Machine learning, a branch of AI, can be broadly described as data analytical methods that learn from data without being explicitly programmed, with patterns identified based on the data itself. They are often described as having flexibility to capture complex associations particularly in large and unstructured data and complexity in modelling. While the vast majority of the items in the TRIPOD Statement are relevant to machine learning based prediction model studies, there are some unique challenges with machine learning that are not captured. Due to their complexity, these prediction models are typically considered to be 'black box', unlike say regression- based models where the full model can be transparently presented (eg, as an equation containing all the regression coefficients). Also, while many machine learning methods have origins in the statistical literature, two (overlapping) prediction model cultures have emerged as those from a statistical/epidemiological background and those from the computer science/data sciences.[16] Although there is clear overlap, different approaches to model development, validation and updating

have appeared, and different and sometimes conflicting terminology have arisen.

Due to the relative novelty of applying machine- learning methods to clinical prediction modelling, there is little information on the quality of reporting of these studies. However, the few reviews that have examined the completeness of reporting of have concluded that reporting is poor.[17, 18] In response to these concerns, guidance is required to help authors fully describe their prediction model study when machine learning methods were used. Therefore the TRIPOD group initiated a large international project to develop a consensus based extension of TRIPOD with specific focus on reporting of studies that undertake the development, validation or updating of a diagnostic or prognostic prediction model, using machine learning techniques—herein referred to as TRIPOD-AI.[19] The TRIPOD- AI extension, comprising a checklist and an accompanying elaboration and explanation document will provide researchers, authors, reviewers, editors, users and other stakeholders of machine- learning- based prediction model studies, with guidance on the minimal set of items to report, with detailed examples of good reporting for each item.

Complete reporting allows studies to be understood, replicated and used. However, critical appraisal and of the quality of study method is a crucial component of evidence- based medicine as well. Critical appraisal and assessing the quality of studies is a crucial component of evidence- based medicine. In 2019, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) was published[20, 21] to help a variety of stakeholders including, for example, systematic reviewers, researchers, journal editors, manuscript reviewers and policy- makers involved in clinical guideline development, critically appraise the study design, conduct and analysis of prediction model studies. PROBAST comprises four domains (participants, predictors, outcome and analysis) and contains 20 signalling questions to facilitate risk of bias assessment. Clearly risk of bias assessment and reporting are intrinsically linked, in that judging risk of bias is predicated on what has been reported in the primary study. While in principle PROBAST is relevant for prediction model studies using machine learning, different approaches to model development and validation, and terminology have appeared, and the ability to critically appraise these studies is crucial before they are implemented.[22, 23] Therefore, in parallel with the development of TRIPOD- AI, we will also develop PROBAST- AI, a tool to assess risk of bias in machine learning based multivariable prediction model studies.

FOCUS OF TRIPOD-AI AND PROBAST-AI

The focus of both TRIPOD- AI and PROBAST- AI is on reports of research or endeavours in which a multivariable prediction model is being developed (or updated), or validated (tested) using any (supervised) machine learning technique. Conforming to the original TRIPOD and PROBAST publications, a multivariable prediction model is defined as any combination or equation of two or more predictors that is to be used for individualised predictions to estimate an individual's probability of having (diagnosis) or developing (prognosis) a particular health outcome or state. Predictors may have any form and emerge from patient history, physical examination, diagnostic, prognostic or monitoring tests and from undergone treatments. Outcomes may also have any form (dichotomous, categorical, continuous) and of any kind, such as, a particular condition or disorder being present or absent (diagnostic outcome or classification), short- term prognosis outcomes (eg, hospital mortality or postoperative complications), and long- term prognostic outcomes such as 1-year occurrence of treatment complications, 5- year occurrence of metastases or lifelong survival).

As per the original publications, TRIPOD- AI and PROBAST- AI will also address prediction model studies from all medical care settings (public health, primary, secondary, tertiary, and nursing home care) and all corresponding target populations (healthy individuals, suspected and diseased individuals). TRIPOD- AI and PROBAST- AI are not meant to address:

- Comparative studies that quantify the impact of using a prediction model as compared with not using the model.[24]
- So- called predictor finding studies (also known as risk or prognostic factor studies) where multivariable machine learning techniques are used to identify (usually from a wider set of potential predictors) those predictors that are associated with an outcome, but not to develop a model that can be used for individualised predictions in new individuals.
- Single medical test studies that use machine learning or AI techniques aimed to read, for example, CT or MRI, images to find which image parameters are best associated with an outcome (such studies fall under the remit of STARD-AI).[25] If these image parameters are included as predictors in a multivariable model combined with other predictors, TRIPOD- AI and PROBAST- AI may be useful.

METHODS/DESIGN

Both TRIPOD- AI and PROBAST- AI will be developed following published guidance from the EQUATOR Network.[26] We will develop the guideline in five stages: (1) systematic reviews to establish the quality of current reporting, (2) Delphi exercise, (3) consensus meeting, (4) development of the guidance statement and (5) guideline dissemination. We have registered our intent to develop the TRIPOD extension for AI on the EQUATOR Network website (www. equator- network. org), the TRIPOD website (www. tripod- statement. org) and recently announced it in the Lancet,19 while the PROBAST- AI development has been announced on the PROBAST website (www. probast. org).

TRIPOD-AI/PROBAST-AI working group

The TRIPOD/PROBAST working group will include: (1) an executive committee (2) an advisory and working group and (3) a large international Delphi panel. The TRIPOD- AI/PROBAST-AI executive committee will be responsible for the leadership and coordination of all the processes involved in the development and dissemination of the TRIPOD-AI guideline. The executive committee consists of the two lead authors of the TRIPOD reporting guideline and the PROBAST tool, and also prediction model experts and researchers from the machine learning community. Key stakeholders for stage 2 (Delphi survey) will be identified and approached to participate and a subset of these key stakeholders (the advisory group) will participate in stage 3 (consensus meeting).

Here, the term key stakeholder refers to a cross- sector participant (both industry and public sector) who falls into at least one of the following categories:

1. Researchers who have used machine learning in the context of clinical prediction, have clear knowledge and expertise in using machine learning or developed machine learning methods. These include applied (bio)medical investigators, statisticians, epidemiologists, and data scientists).

2. Assessors and approvers of AI or machine learning model, such as regulatory assessors and ethics commit- tee members.

3. Beneficiaries or users of the resultant TRIPOD-AI guidance and PROBAST-AI tool such as journal editors and journal reviewers.

4. Commissioners of research grants, such as funders.

5. Consumers of research results such as healthcare providers and patients and citizens.

Stage 1: systematic review of current reporting

Two parallel systematic reviews are ongoing to evaluate the quality of current reporting in published studies developing, validating or updating machine learning based prediction models in the medical domain. Both systematic reviews will assess adherence of the reporting against the original TRIPOD Statement,[1, 8] using the TRIPOD adherence checklist.[27] The reviews will also examine the methodological conduct of the primary studies, including a risk of bias assessment using the recently issued risk of bias tool (quality appraisal) for diagnostic and prognostic prediction model studies (PROBAST),[20,21] and will draw out specific issues, currently not covered by TRIPOD and PROBAST relating to machine learning. The protocols for the two systematic reviews have been registered with the International Prospective Register of Systematic Reviews (PROSPERO IDs CRD42019140361 and CRD42019161764). One review (CRD42019161764) will examine the quality of reporting of machine-learning- based prediction model studies across all medical fields (between January 2018 to December 2019), while the other review (CRD42019140361) will focus on the quality of reporting of machine learning based prediction model studies published in oncology (between January 2019 and September 2019).

Undertaking these reviews serves two purposes: (1) to understand the completeness of current reporting of machine- learning- based prediction model studies in the medical literature and (2) to identify unique reporting items for consideration for TRIPOD extension, and unique risk of bias or quality items for PROBAST extension. The data collection for this phase is underway. The reviews will evaluate the current completeness of reporting and the quality of the research and identify additional reporting and quality items to be considered for TRIPOD- AI and PROBAST-AI.

These two reviews will evaluate the current completeness of reporting and the quality of the research. Together with other evidence[3, 4, 17, 18, 28] from existing methodological guidance papers, they will provide important information on the transparency and quality of reporting. Using the original TRIPOD and PROBAST checklists as starting points, the executive committee will identify in the literature the preliminary items to consider in stage 2 (the Delphi study) and therefore inclusion in the eventual TRIPOD- AI checklist and PROBAST- AI tool.

Stage 2: Delphi exercise

We will perform an extensive Delphi survey among a large international network of relevant stakeholders, with a maximum of three rounds, to help decide on items that could be modified, added to, or removed from the TRIPOD 2015 checklist to form the TRIPOD- AI checklist, and subsequently the PROBAST- AI checklist.

Design

The Delphi process will comprise of a series of rounds where panellists will independently and anonymously evaluate and achieve consensus on the inclusion or exclusion of the proposed reporting and quality items—in addition to suggesting additional items. The process will be repeated for a maximum of three rounds. Following each round, participants will be provided with structured feedback of the previous round to help reconcile individual opinions and achieve group consensus. Items achieving a high level of agreement (≥70%) will be taken forward to the consensus meeting (stage 3).

Selection of potential items The list of items for TRIPOD-AI (and PROBAST-AI) will be collated by the executive committee, including the results of the two systematic reviews, any other available studies on methodology or reporting of machine learning based prediction models, and expert recommendations from the Delphi panellists. Relevant methodological guidance or methodological papers will be retrieved to identify additional candidate reporting and quality items for machine learning-based prediction model studies. Preselection involves dividing items into those to further consider, those that can be provided as optional guidance (to be outlined in an Explanation and Elaboration accompanying document), or those not to consider for potential inclusion. Delphi participants will have the opportunity to view and provide feedback in each round, and also to suggest new items.

Recruitment process and participants

Delphi participants will be identified through professional networks of the executive committee, participation in the Delphi exercise of the original TRIPOD guideline (and TRIPOD for Abstracts and TRIPOD Cluster Delphi surveys), original PROBAST Delphi exercise, via self- response to the Lancet 2019 paper where TRIPOD-AI was announced,[19] and responses to social media announcements of TRIPOD-AI (eg, Twitter).

We will invite international participants with diverse roles (eg, researchers, healthcare professionals, journal editors, funders, policy makers, healthcare regulators, end users of prediction models) from a range of settings (eg, universities, hospitals, primary care, biomedical journals, non-profit organisations and for-profit organisations). Participants will be invited via personalised email that will describe the TRIPOD-AI extension and PROBAST-AI tool development, and explain the objective, process, and timelines of the Delphi exercise. We plan to invite at least 200 participants to the Delphi survey. In all rounds, the survey will remain open for 3 weeks, with a reminder email sent 1 week after the initial invitation. In round two of the Delphi exercise, additional participants may be sought to ensure fair representation of all key stakeholders.[29]

Informed consent from participants will be obtained using an online consent form and participants can with- draw at any time. Individuals who indicate that they wish to opt out of the survey will be removed from subsequent invitations. Participants will not know the identities of other individuals in the Delphi panel, nor will they know the specific answers that any individual provides.

Procedure for selection of items We plan to ask participants to consider the following guiding principles when reviewing existing, new or modified items for inclusion: (1) reporting of the item should facilitate reproducibility of the study (ie, users should be able to recreate the findings based on the information reported); (2) reporting of the item facilitates assessment of the quality and risk of bias in and applicability of the machine learning study findings, to enhance their uptake and use in subsequent studies, systematic reviews and daily practice; (3) item is likely relevant to nearly all prediction model studies; (4) the set of items represent the minimum that should be reported in all machine learning studies developing, validating or updating a diagnostic or prognostic prediction model.

Round 1

Participants will be asked to rate on a 5-point Likert scale, the extent to which they agree with the inclusion of each checklist item in the TRIPOD-AI extension and PROBAST-AI tool (1=strongly disagree, 2=somewhat disagree, 3=I don't know, 4=somewhat agree, 5=strongly agree). A free- text box will be provided for general comments on each item (to justify their decision or suggest wording changes), and a free- text box will be provided at the end of the survey to suggest additional checklist items or provide general comments on the checklist. The survey will be pilot- tested for usability and clarity to a small number of individuals familiar with prediction models or machine learning but not involved in the TRIPOD-AI guideline extension or PROBAST-AI tool and revised accordingly based on their feedback.

Round 2

The same participants involved in round 1 will be invited to participate in round 2. Participants will be provided with their first- round responses on each item, an anonymised summary of the group ratings and anonymised comments to justify ratings. Using the same format as round 1, participants will be presented with each item, including any new items suggested during round 1, and again express the extent to which they agree with the inclusion of the item in the TRIPOD-AI checklist or PROBAST-AI tool, considering the structured feedback to inform their responses. Participants who were invited to participate in round 1, but who did not respond will be invited to participate in round 2, and will be presented with an anonymised summary of the group ratings. Items that reached a high- level of agreement (scoring 4 or 5) in round 1 (≥70%) will be presented for information purposes only, with no

voting on these items, though a free- text box will be provided for any comments. A third Delphi round will be used if deemed necessary by the Executive Committee.

Results from the Delphi survey

Item scores will be summarised for the entire panel as a whole, as appropriate (eg, frequency and proportions across the rating categories) accompanied by a narrative summary of findings, comments, and suggestions. Results from both rounds of the survey will be discussed by the executive committee. For items where there was no consensus following the second Delphi found will be discussed by the executive committee, and will be considered for discussion at the subsequent consensus meeting.

Stage 3: consensus meeting

Two virtual consensus meetings (separately for TRIPOD-AI and PROBAST-AI), both spread over 2 days, will be held with the objective of discussing the results from the Delphi exercise and finalising items to be included in the reporting guideline and risk of bias tool. The composition of the consensus group will reflect the diversity of the key stakeholders addressed above. Key experts participating in the Delphi exercise will be considered to participate in the consensus meeting. We will also consider inviting experts who did not contribute to the Delphi to participate in the consensus. A total of around 25–30 international participants are expected to contribute to the virtual consensus meeting.

Procedure

The agenda and any material (eg, results from the systematic reviews and Delphi) for the consensus meeting will be prepared by the executive committee and will be shared with attendees in advance. Members of the executive committee will facilitate a structured discussion on the rationale behind each item identified in the Delphi exercise. Consensus meeting participants will then be given the opportunity to discuss each item (reporting item for TRIPOD-AI and signalling question for PROBAST-AI) and vote on each item. The decision to retain an item in the TRIPOD- AI and PROBAST- AI will be based on achieving at least 70% support from the consensus meeting participants. The group will agree on the draft list of reporting items for the final TRIPOD-AI extension and PROBAST-AI tool. Specific item wording will not be discussed during the meeting, though participants can suggest and the group to agree on general intent and meaning of the item. Plans for dissemination will be discussed at the end of the consensus meeting.

Pilot testing

We will invite authors of machine learning prediction model studies in the medical domain, doctoral students undertaking prediction model, machine learning courses

or workshops, and peer- reviewers and editors of journals who frequently publish such prediction model studies, to pilot the use of a draft version of the TRIPOD-AI checklist and PROBAST-AI tool. We will ask those who pilot the checklist and tool whether the wording of items is ambiguous or difficult to interpret.

Stage 4: development of the draft TRIPOD-AI statement, PROBAST-AI and explanation and elaboration documents

The executive committee will lead the development of the TRIPOD-AI reporting guidance and PROBAST-AI signalling questions based on the agreed list of items from the consensus meeting (stage 3). The executive committee will invite a subset of members from the consensus meeting (to form a writing group) to help draft the explanation and elaboration paper. The executive committee will reserve the right to update (ie, remove or add) additional items to the TRIPOD- AI checklist during the development of the TRIPOD-AI statement, if and as necessary (as a result of the pilot testing).

For each of the TRIPOD-AI extension and the PROBAST-AI risk of bias tool, two manuscripts will be developed: (1) the statement paper, presenting the check- list/tool and describing the process of how it was developed and (2) an explanation and elaboration paper. The explanation and elaboration papers will outline the rationale of the reporting items (TRIPOD- AI) and signalling questions (PROBAST-AI), examples of good reporting (TRIPOD-AI) and examples of how to use PROBAST-AI. Drafts of the papers will be circulated to all participants of the consensus meeting for their comments.

Stage 5: guideline dissemination

The dissemination strategy will be informed by discussions at the consensus meeting. We will aim to seek simultaneous publication in key journals to target different readerships. To increase visibility and aid uptake, the TRIPOD-AI checklist and PROBAST-AI tool will be published open access, and made available on the TRIPOD website along with other TRIPOD extensions (www. tripod- statement. org), and on the PROBAST website (www. probast. org) respectively, as well on the PROGRESS website (www. prognosisresarch. com)

PUBLICATION PLAN

It is envisaged that the following publications will arise from the TRIPOD-AI and PROBAST- AI initiative:

- Publication 1: study protocol.
- Publication 2: systematic review protocol (with registration on PROSPERO).
- Publication 3 and 4: Systematic reviews.
- Publication 5 & 6: TRIPOD-AI statement and the Explanation and Elaboration paper.

- Publication 7 & 8: PROBAST-AI tool and the Explanation and Elaboration paper.

CONCLUSION

The number of prediction model studies using machine learning methods is rapidly increasing, including developed, validated or updated prediction models. Ensuring that key details are reported is important so that readers can evaluate the study quality, and interpret its findings including the developed, validated or updated prediction model to enhance their uptake in subsequent research (eg, validation studies), evidence synthesis projects (eg, systematic reviews of prediction models) and in daily practice by healthcare professionals, patients or citizens. We anticipate that TRIPOD-AI will help authors transparently report their study and help reviewers, editors, policy- makers and end- users understand the methods and findings, and thereby reduce research waste. Similarly, we anticipate PROBAST-AI will help researchers, clinicians, systematic reviewers and policy- makers critically appraise the design, conduct and analysis of machine learning-based prediction model studies.

Contributors
GSC, PD, CLAN, JM, LH, JBR, PL, ALB, LP, BVC, MvS, RDR and KGMM were involved in the planning and design of the study. GC drafted the manuscript with all authors contributing to the writing.

Funding
This research was supported by Health Data Research UK, an initiative funded by UKResearch and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities, Cancer Research UK programme grant (C49297/A27294), the NIHR Biomedical Research Centre, Oxford, and the Netherlands Organisation for Scientific Research.

Competing interests
None declared.

Patient and public involvement
Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

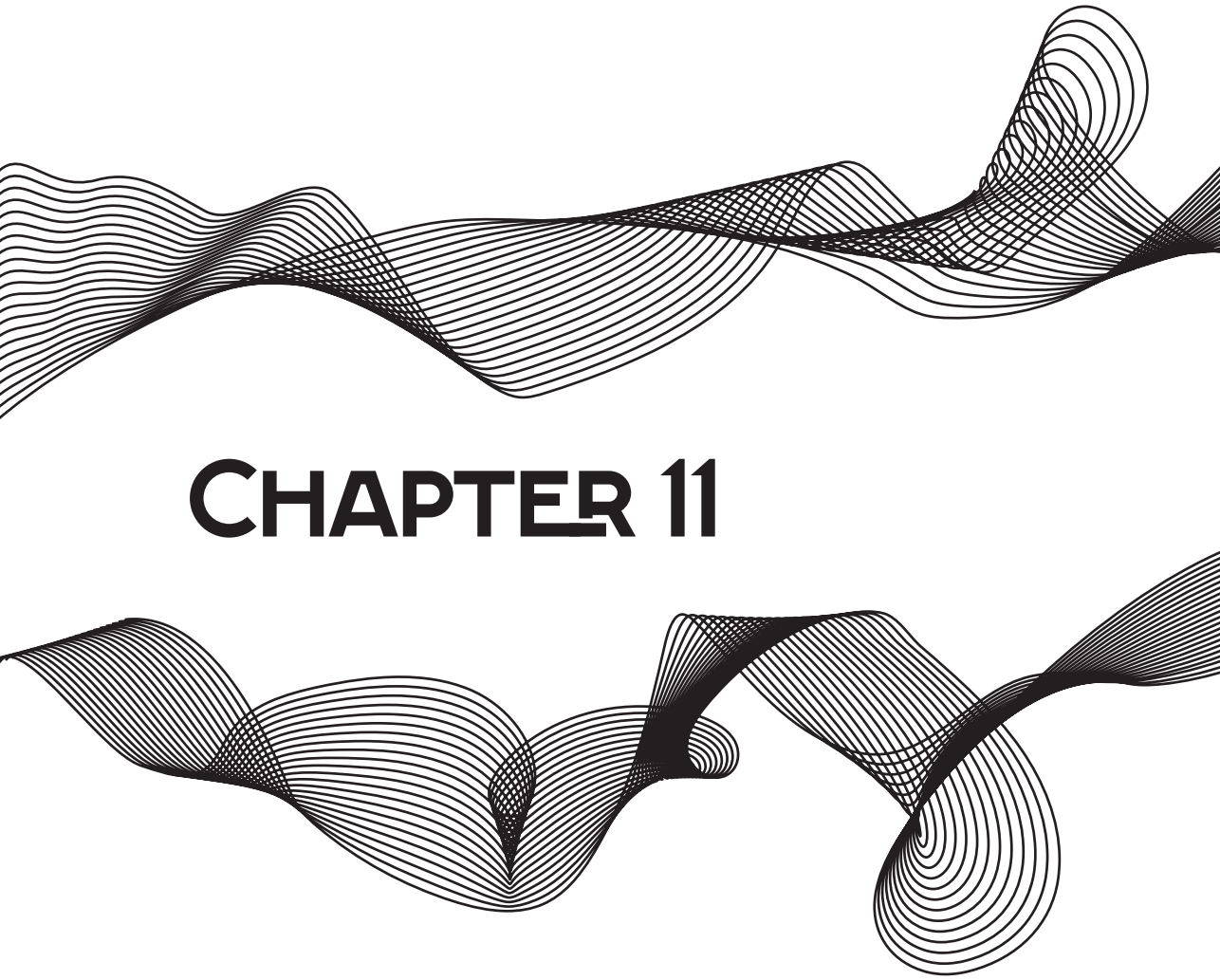
Patient consent for publication
Not required.

Provenance and peer review
Not commissioned; externally peer reviewed.

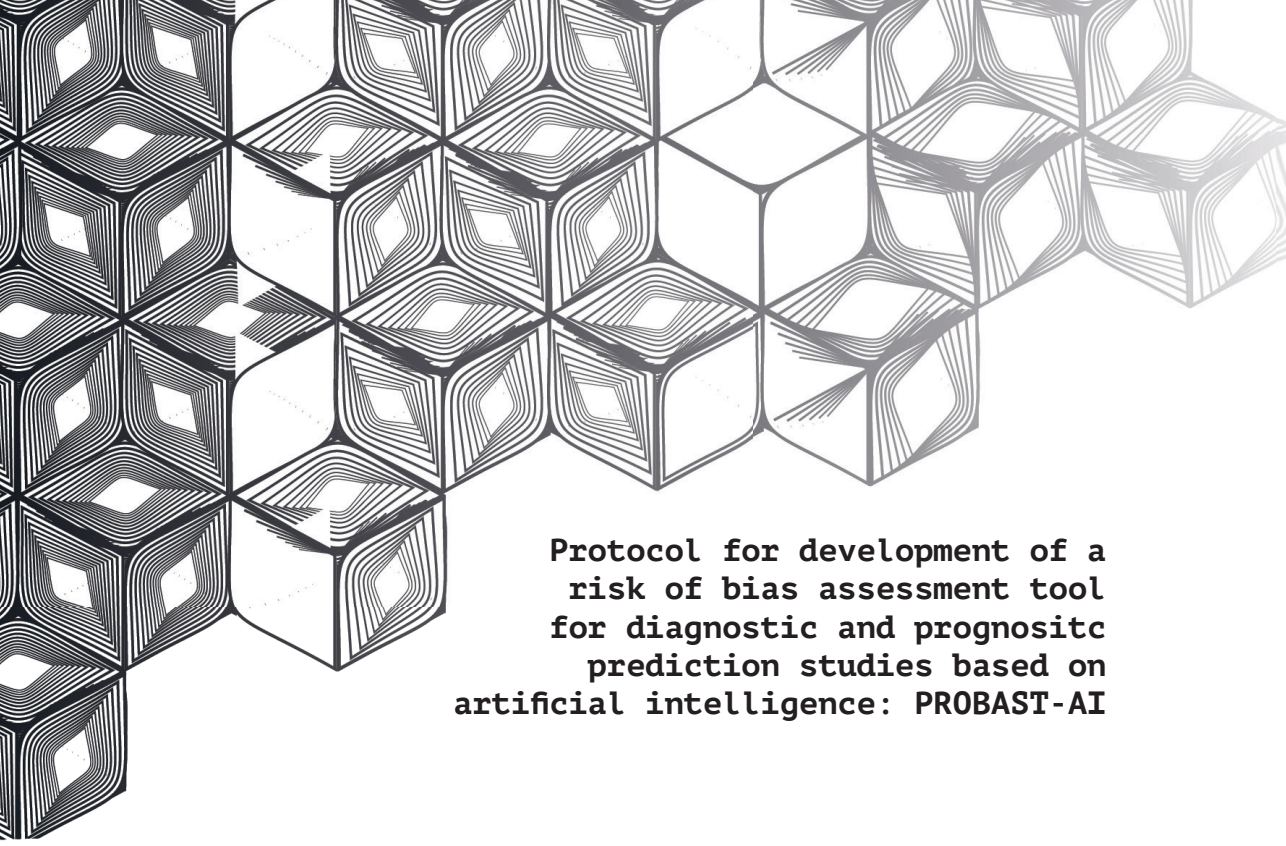
REFERENCES

1. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
2. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90.
3. 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;9:103.
4. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.
5. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9:e1001221.
6. Mallett S, Royston P, Dutton S, et al. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 2010;8:20.
7. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86–9.
8. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.
9. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719–31.
10. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317.
11. Ghassemi M, Naumann T, Schulam P, et al. Practical guidance on artificial intelligence for health- care data. Lancet Digit Health 2019;1:e157–9.
12. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323:305.
13. Sendak M, D'Arcy J, Kashyap S. A path for translation of machine learning products into healthcare delivery. EMJ Innov 2020.
14. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019.
15. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.
16. Breiman L. Statistical modeling: the two cultures. Stat Sci 2001;16:199–231.
17. Shillan D, Sterne JAC, Champneys A, et al. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care 2019;23:284.
18. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.
19. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet 2019;393:1577–9.
20. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8.
21. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–33.
22. Liu Y, Chen P- HC, Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322:1806.
23. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl Vis Sci Technol 2020;9:7.
24. Moons KGM, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: ap-

plication and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
25. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD- AI steering group. Nat Med 2020;26:807–8.
26. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7:e1000217.
27. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 2018;16:120.
28. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health- care professionals in detecting diseases from medical imaging: a systematic review and meta- analysis. Lancet Digit Health 2019;1:e271–97.
29. Boel A, Navarro- Compán V, Landewé R, et al. Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. J Clin Epidemiol 2021;129:31–9.



CHAPTER 11



Protocol for development of a risk of bias assessment tool for diagnostic and prognositc prediction studies based on artificial intelligence: PROBAST-AI

Constanza L Andaur Navarro
Johanna AA Damen
Lotty Hooft
Johannes B Reitsma
Maarten van Smeden
Richard D Riley
Gary G Collins
Karel GM Moons

Available in Open Science Framework

ABSTRACT

Introduction. Clinical prediction models aim to guide clinical decision-making by accurately predicting the probability of an outcome in patients, for either diagnostic (i.e., to predict disease presence) or prognostic purposes (i.e., to predict future outcomes). In recent years, a considerable amount of artificial intelligence (AI)/machine learning (ML)-driven models to predict individual patient outcomes have emerged, creating the sense of endless potential for these new technologies. Hence, evaluation of these techniques is upmost necessary, as their potential benefits for patients could remain limited. PROBAST (Prediction Risk Of Bias ASsesment Tool) was developed to identify potential risk of biases in multivariable clinical prediction model studies. Nevertheless, it lacks detailed recommendation for methodological appraisal of prediction models developed using AI/ML techniques. The aim of this project is to develop an extension to PROBAST to specifically identify potential biases across the spectrum of AI/ML techniques, namely PROBAST-AI. This initiative has been announced on the PROBAST website (www.probast.org).

Methods and analysis. We will carry out a Delphi procedure among a representative group with expertise in developing, regulating, and implementing prediction models developed using AI/ML. Based on this, we will decide whether the proposed signaling questions should be modified, extended, or removed, or if new signaling questions should be added to PROBAST-AI.

Conclusion. A tool to facilitate the evaluation and interpretation of clinical prediction models developed using AI/ML technologies is necessary to enhance subsequent research (e.g., validation, update, added value), systematic reviews of prediction modes, and clinical implementation (e.g., impact studies). We anticipate that PROBAST-AI will help researchers, healthcare providers, clinicians, AI industry, and regulatory authorities to critically appraise the design, conduct, and analysis of AI/ML-based prediction model studies. Furthermore, we expect PROBAST-AI to improve the transparency and quality of prediction models, and ultimately patient care.

Ethics and dissemination. Ethical approval is not required. Findings will be disseminated in parallel through peer-reviewed publications, social media, and conferences.

INTRODUCTION

Clinical prediction models, algorithms, tools or rules (hereafter defined as prediction models) aim to guide clinical decision-making by accurately predicting the probability of an outcome in patients, for either diagnostic (i.e., to predict disease presence) or prognostic (i.e., to predict future outcomes) purposes. PROBAST (www.probast.org) was developed to aid the quality assessment of primary studies reporting on the development, validation, or updating of prediction models.[1,2]and competing prediction models frequently exist for the same outcome or target population. Health care providers, guideline developers, and policymakers are often unsure which model to use or recommend, and in which persons or settings. Hence, systematic reviews of these studies are increasingly demanded, required, and performed.A key part of a systematic review of prediction models is examination of risk of bias and applicability to the intended population and setting. To help reviewers with this process, the authors developed PROBAST (Prediction model Risk Of Bias ASsessment Tool PROBAST consists of 21 signaling questions within 4 domains (Participants, Predictors, Outcome, Analysis). PROBAST is a widely accepted tool to asses both the quality and applicability of multivariable prediction model studies. Furthermore, it has been adopted by the Cochrane Collaboration to support systematic reviews of diagnostic and prognostic prediction models.[3]

In recent years, we have seen an accelerated grown of artificial intelligence (AI)/machine learning (ML)-driven models to predict individual patient outcomes. PROBAST was design to assess any type of prediction model regardless of the predictors used; outcomes being predicted; or modelling methods used to develop, validate, or update a model.[1,2]and competing prediction models frequently exist for the same outcome or target population. Health care providers, guideline developers, and policymakers are often unsure which model to use or recommend, and in which persons or settings. Hence, systematic reviews of these studies are increasingly demanded, required, and performed.A key part of a systematic review of prediction models is examination of risk of bias and applicability to the intended population and setting. To help reviewers with this process, the authors developed PROBAST (Prediction model Risk Of Bias ASsessment Tool Currently, PROBAST provides guidance on how to adapt the signaling questions to address AI/ML-based prediction model studies. However, this guidance might be limited to cover the all the new emerged challenges such as the variety of techniques and performance measures, dissimilar terminology, higher complexities of datasets, interpretation of 'validation', and comparison to human performance. We consider imperative the development of a more detailed and extensive guidance on methodological conduct with a standardized nomenclature for AI/ML-based prediction models studies, as their potential benefits for patients may remain limited.

The PROBAST-AI group is preparing an extension to specifically identify potential biases across the spectrum of AI/ML modelling techniques. Through an established Delphi-based process, we aim to develop a consensus-based tool useful to the biomedical research as well as the AI/ML community involved in setting prediction models with clinical application.[4] This will be achieved by gathering a representative group of experts including methodologists, statisticians, healthcare providers, data scientists, policy makers, and AI industry representatives. We expect PROBAST-AI to aid healthcare, academia, industry, and regulatory authorities to evaluate AI/ML-driven technologies.

The main objective is to develop a tool to determine the quality and applicability of studies developing, validating, and updating prediction models studies using AI/ML techniques.

SCOPE

PROBAST-AI will be developed to identify good (i.e., low risks for any form of bias) and poor (i.e., high risks of bias) quality primary studies reporting on the development, validation, or updating of prediction models developed using AI/ML techniques. We define poor quality as shortcomings in the study design, conduct, or analysis that might lead to distorted estimates of the model's predictive performance, such as its calibration, discrimination or classification. Hence, we expect PROBAST-AI to facilitate the evaluation and interpretation of AI/ML-based prediction model studies and enhance subsequent research (e.g., validation, update, added value, impact studies) regardless of the aim (i.e., solving a healthcare problem or applicability of AI/ML technique) or form of publication (i.e. journal, pre-print). Likewise, PROBAST-AI will support systematic reviews of prediction model studies, healthcare providers, and regulatory authorities by providing guidance on critical appraisal. In addition, we expect to highlight the importance of developing 'fair' AI/ML prediction models and help determine if decisions made based on these models are equitable and free from human biases.[5]

Studies evaluating the implementation of AI-driven tools into clinical settings (i.e. DECIDE-AI), as well as studies evaluating the quality of diagnostic test accuracy using artificial intelligence methods (i.e. QUADAS-AI) will be out of the scope of PROBAST-AI.

METHODS

We will develop PROBAST-AI in five stages: (1) Conduct various systematic reviews to establish the current methodological conduct and potential biases in AI/ML-based prediction model studies, (2) Delphi process, (3) Consensus meeting, (4) Publication of PROBAST-AI and Explanation & Elaboration documents, and (5) Tool dissemination. PROBAST-AI will be developed following the EQUATOR (Enhancing Quality and Transparency of Health Research) Network guidelines.[6]

PROBAST-AI working group

The PROBAST Working Group will include: (1) an Executive Committee and (2) a large international Delphi Panel. The PROBAST-AI Executive Committee will be responsible for the leadership and coordination in the development and dissemination of PROBAST-AI. The Executive Committee consists of various healthcare researchers with vast experience in prediction research, and of whom some also participated in the original PROBAST tool (CLAN,JAAD, LH, HR, MvS, RDR, GSC, KGMM). The Executive Committee will recruit a representative group of experts for Stage 2 (Delphi process). In this study, we refer to 'experts' as someone who falls into at least one of the following categories:

- Researchers who have knowledge and expertise in using AI/ML or developed prediction models using AI/ML. These might include applied (bio)medical investigators, statisticians, epidemiologists, data scientists, technicians, and investigators from other AI related fields.
- Potential users of PROBAST-AI such as systematic reviewers, journal editors, journal reviewers, AI industry stakeholders.
- Healthcare providers who have expertise in implementing artificial intelligence/machine learning-driven technologies in healthcare settings.
- Regulatory authorities, ethical committees, and funders who deal with projects using AI/ML techniques.

Stage 1 – Overview of current methodological conduct and potential biases

Systematic reviews

Recently, four parallel systematic reviews examining the reporting, and methodological conduct and quality of primary studies describing AI/ML-based diagnostic and prognostic prediction model studies have been submitted for publication.[7,8] The protocols have been registered with the International Prospective Register of Systematic Reviews (PROSPERO IDs CRD42019140361 and CRD42019161764). These reviews described the results of a methodological quality

assessment of AI/ML-based prediction models using PROBAST in oncology and across all clinical fields. The findings will provide important information on the methodological quality and potential signaling questions to consider in Stage 2 (Delphi process) and therefore, for inclusion in PROBAST-AI.

Survey

Last year, we invited a large group of PROBAST users (n=49) to rate the extent to which the original signaling question were applicable to AI/ML-based prediction model studies. The survey was launched in August 2020 and we received 44 (89.7%) responses. The findings provided important information on current limitations of PROBAST and potential signaling questions to consider in Stage 2 (Delphi process). The findings of this survey were discussed within the research team. The report can be requested to the lead author (CLAN).

Stage 2 – Modified Delphi process

We will carry out an extensive Delphi process among an international representative group of experts.[9] The Executive Committee will elaborate and propose a preliminary list of potential signaling questions for round 1 of the Delphi survey. We will seek help to decide whether the proposed signaling questions should be modified, extended, or removed, or if new signaling questions should be added to PROBAST-AI.

Ethical approval

This project is considered non-WMO (no medical-related scientific research). According to UMC Utrecht policies, it requires approval by the Quality Coordinator from our division (Geertje de Lange, KwaliteitJuliusCentrum@umcutrecht.nl) . Consent to participate is required and data collected needs to be stored securely. Details are provided in the data management plan (DMP) in Appendix 1.

Design

The Delphi process will consist of a survey with three rounds where participants will independently and anonymously evaluate and achieve consensus on the inclusion or exclusion of the proposed signaling questions. New signaling questions can also be suggested. Participants will have the opportunity to provide feedback in each round. Consensus will be considered as reached if 2/3 (67%) of survey participants rate a signaling question as either high (4,Agree - 5,Strongly Agree) or low (1,Strongly disagree - 2,Disagree). Signaling questions achieving high level of agreement (>67%) will be taken forward to the online consensus meeting (Stage 3). In addition, participants will be provided with structured feedback of the previous round to help reconcile individual opinions and achieve group consensus. In all rounds, the survey will remain open for three weeks, with two reminders send one week after the initial invitation.

Selection of potential signaling questions

The Executive Committee will propose a list of signaling questions for PROBAST-AI consideration. This list will be based on the original PROBAST signaling questions, new signaling questions generated from the aforementioned systematic reviews, further items from any relevant study on AI/ML methodology guidance (e.g. STARD-AI, SPIRIT-AI, CONSORT-AI, DECIDE-AI), and the preliminary survey carried out in August 2020.[2,10–14]

We will provide a structured feed-back after each round to Delphi participants draft by the lead author (CLAN). This will consist of dividing signaling questions into those to further consider in round 2, those that can be provided as optional guidance (to be outlined in an Explanation and Elaboration accompanying document), or those not to consider for potential inclusion.

Recruitment of participants

Delphi participants will be identified through the Delphi processes of the TRIPOD statement for reporting of prediction model studies (www.tripod-statement.org), TRIPOD for Abstracts, TRIPOD Cluster, and PROBAST. Additionally, participants were recruited via self-response to the Lancet 2019 paper where TRIPOD-AI was announced, responses to social media announcements (e.g., Twitter, LinkedIn), and from the professional networks of the Executive Committee.[2,14–18]

Participants will be invited via personalized emails that will describe the overarching aim of the project (i.e., PROBAST-AI development) and explain the objective, process, and timelines of the Delphi process. The e-mail will contain a personalized web link to the first round of the survey. We plan to invite at least 200 participants to the first round of the Delphi survey. Participants will not know the identities of other individuals in the Delphi panel, nor will they know the specific answers that any individual provides. Consent from participants will be obtained using an online consent form and participants will be able to withdraw at any time. Individuals who indicate that they wish to withdraw of the Delphi process will be removed from subsequent invitations.

Data management

We plan a web-based survey using REDCap, an online data capture tool.[19] The lead author (CLAN) will be responsible for data management during project development. There is no long-term value on the data collected, thus it is unlikely the data will be shared and/or preserved. In case of reuse or sharing, data will be treated anonymously and be available for 1 year after publication of main documents. UMC Utrecht will remain the owner of all collected data for this study. Further details are provided in the data management plan.

Procedure

The lead author (CLAN) will draft each of the survey rounds and test it with one co-author (JAAD). The survey will be revised accordingly based on the feedback from the pilot before sending the invitation to all participants in the panel. We will ask participants to consider whether:

- The set of signaling questions represents the minimum required to assess the quality and risk of bias in AI/ML-based prediction model studies.
- The set of signaling questions enhances subsequent studies (i.e., validation, update, added values, impact studies), systematic reviews and lastly, improve daily clinical practice.
- Each signaling question is relevant to nearly all AI/ML-based prediction model studies.
- Each signaling question is relevant to nearly all well-known AI/ML modelling techniques. If not, this can be addressed briefly in the Explanation & Elaboration document.

Round 1

The proposed signaling questions will be structured under each of the current 4 domains (Participants, Predictors, Outcome, Analysis) of PROBAST. After each domain, participants will have the option to provide comments and suggestions using a free-text box. Participants will be asked to rate on a 5-point Likert scale, the extent to which they agree with the inclusion of each proposed signaling questions in the PROBAST-AI tool (1=strongly disagree, 2= disagree, 3=neither agree nor disagree, 4=agree, 5=strongly agree). Participants will have the option to justify their decision, suggest additional questions, or provide general comments in the free-text box. In addition, another free-text box will be provided at the end of the survey for any general comments.

Round 2

Responders from round 1 will be invited to participate in round 2. Participants will be provided with their first-round responses on each item, an anonymized summary of the group ratings and anonymized comments to justify ratings. Using the same format as round 1, participants will be presented with each item, including any new items suggested during round 1, and again express the extent to which they agree with the inclusion of signaling questions to PROBAST-AI tool, considering the structured feedback to inform their responses.

Participants who were invited to participate in round 1, but who did not respond will be invited to participate in round 2 and will be presented with an anonymized summary of the group ratings. Signaling questions that reached a high-

level of agreement in round 1 (≥ 67%) will be presented for information purposes only, with no voting on these items. Again, participants will be able to provide comments in a free-text box to justify their decision or suggest wording changes. If necessary, a third Delphi round will be carried out by the Executive Committee.

Analysis

Rates will be summarized for the entire group of participants, as appropriate (e.g., frequency and proportions across the rating categories) accompanied by a narrative summary of findings, comments, and suggestions.

Outcome of Delphi process

Results from the three rounds will be discussed by the Executive Committee. Signaling questions where there was no consensus following the third Delphi round will be further discussed by the Executive Committee and then, will be considered for discussion at the subsequent consensus meeting.

Stage 3 – Consensus meeting

An online consensus meeting will be held to discuss the results from the Delphi survey and determine the final list of signaling questions to be included in PROBAST-AI. Key experts participating in the Delphi survey will be considered to participate in the consensus meeting. We will also consider inviting experts who did not contribute to the Delphi to participate in the consensus. A total of around 25-30 international participants are expected to contribute to the consensus meeting.

Procedure

- The agenda and any material (e.g., results from the systematic reviews and Delphi) for the consensus meeting will be prepared by the Executive Committee and will be shared with attendees in advance via e-mail.
- Members of the Executive Committee will facilitate a structured discussion on the rationale behind each item identified in the Delphi exercise.
- Participants of the consensus meeting will then be given the opportunity to discuss each signaling question for PROBAST-AI, and vote on each item. Participants will be asked to rate each signaling questions 'yes', 'no', 'no opinion'. The decision to retain a signaling question will be based on achieving at least 67% or more in each signaling question.
- After the consensus meeting a final list of signaling questions for PROBAST-AI will be drafted.
- Discussion of specific item wording will not be discussed during the meeting, though participants can suggest on general intent and meaning of the item.
- Plans for dissemination will be discussed at the end of the consensus meeting.

Stage 4 – Development of PROBAST-AI and Explanation & Elaboration documents

Pilot

The Executive Committee will pilot a draft version of the PROBAST-AI tool based on the consensus meeting (Stage 3). We will ask researchers who frequently publish AI/ML-based prediction model studies to pilot a draft version of the PROBAST-AI tool. We will ask those who pilot the tool whether the wording of the tool is ambiguous or difficult to interpret. The Executive Committee will reserve the right to modify, remove, or add additional signaling questions to PROBAST-AI after this piloting phase, if necessary.

Manuscripts

The Executive Committee will invite a subset of members from the consensus meeting to form a writing group. Two manuscripts will be written: (1) the Statement paper, presenting the tool and describing the process of how it was developed and (2) an Explanation & Elaboration document. The Explanation & Elaboration document will outline the rationale of each signaling question and it will provide examples of how to use PROBAST-AI properly. Each manuscript draft will be sent to all participants of the consensus meeting for their comments, and approval prior to publication.

Stage 5 – Guideline dissemination

We aim to simultaneous publish in several journals to target different readerships. To increase visibility and uptake, the PROBAST-AI tool will be published open access and made available on the PROBAST website (www.probast.org). Social media will be used to help disseminate the tool (e.g., Twitter, LinkedIn). The Executive Committee will (and consensus participants will be encouraged to) publicize the PROBAST-AI tool at key conferences and courses.

Publication plan

It is envisaged that the following publications will arise from PROBAST-AI initiative:

- Publication 1: Protocol
- Publication 2: Systematic review protocol (with registration on PROSPERO)[20]
- Publication 3: Systematic reviews[7,8]
- Publication 4 : PROBAST-AI tool
- Publication 5: Explanation and Elaboration document

CONCLUSION

An exponential number of AI/ML-based prediction model studies have been published in the medical literature in recent years. A tool to facilitate the evaluation and critical appraisal of these studies is, therefore, necessary to enhance subsequent research (e.g., validation, update, added value), systematic reviews of prediction model studies, and clinical implementation (e.g., impact studies). We anticipate that PROBAST-AI will help researchers, clinicians, healthcare providers, the AI- industry, and regulatory authorities to critically appraise the design, conduct, and analysis of AI/ML-based prediction model studies. Furthermore, we expect that PROBAST-AI will improve the transparency and quality of prediction models, and ultimately healthcare in general, as AI/ML-based prediction models will undoubtedly become an integral part of the future healthcare systems worldwide.

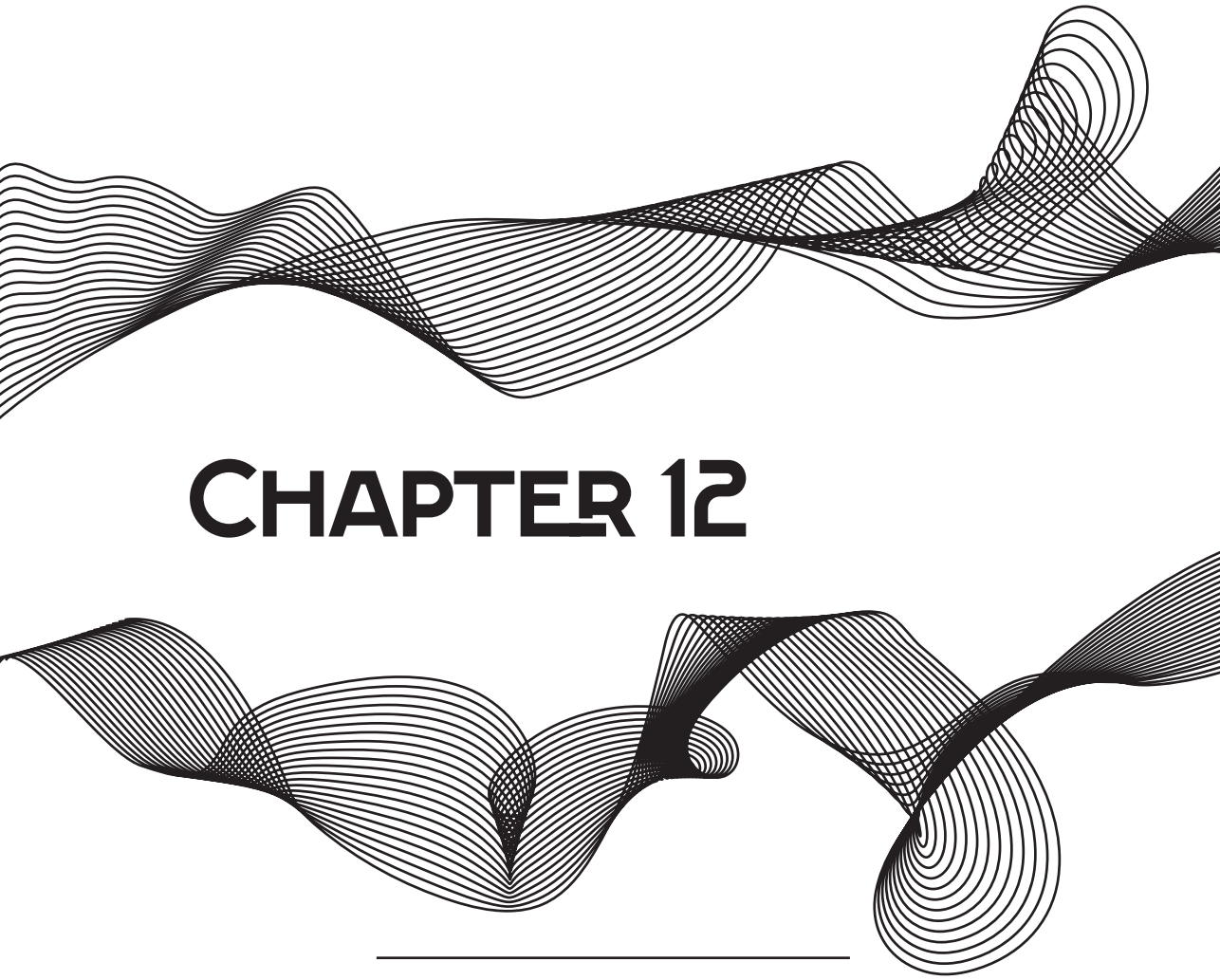
Funding
There are no funders involved in this project.

Abbreviations: AI, artificial intelligence; ML, machine learning; PROBAST, Prediction Risk Of Bias ASsesment Tool; PROBAST-AI, Prediction Risk Of Bias ASsesment Tool-Artificial Intelligence; QUADAS-AI, Quality Assessment of Artificial Intelligence centered diagnostic accuracy studies; STARD-AI, Standard for Reporting of Diagnostic Accuracy Studies; CONSORT-AI, Consolidated Standards of Reporting Trials-Artificial Intelligence; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence; DECIDE-AI, Developmental and Exploratory Clinical Investigation of Decision-support systems driven by Artificial Intelligence.

REFERENCES

1. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
2. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
3. Moons KG, Hooft L, Williams K, Hayden JA, Damen JA, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane database Syst Rev*. 2018;10:ED000129. doi:10.1002/14651858.ED000129
4. Collins GS, Dhiman P, Navarro CLA, et al. Protocol for development of a reporting AI) and risk of bias guideline (TRIPOD- tool (PROBAST- AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. Published online 2021:1-7. doi:10.1136/bmjopen-2020-048008
5. Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. 2018;(Ec). http://arxiv.org/abs/1808.00023
6. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7(2). doi:10.1371/journal.pmed.1000217
7. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning : A systematic review. Published online 2021.
8. Dhiman P, Ma J, Andaur Navarro C, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol*. Published online 2021. doi:10.1016/j.jclinepi.2021.06.024
9. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs*. 2000;32(4):1008-1015. doi:10.1046/j.1365-2648.2000.t01-1-01567.x
10. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med*. 2020;26(6):807-808. doi:10.1038/s41591-020-0941-1
11. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7
12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
13. Vasey B, Clifton DA, Collins GS, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186-187. doi:10.1038/s41591-021-01229-5
14. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
15. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. Published online 2019. doi:10.1016/S01406736(19)302351
16. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55. doi:10.7326/M14-0697
18. Heus P, Reitsma JB, Collins GS, et al. Transparent Reporting of Multivariable Prediction Models in Journal and Conference Abstracts: TRIPOD for Abstracts. *Ann Intern Med*. 2020;173(1):43. doi:10.7326/M20-0193

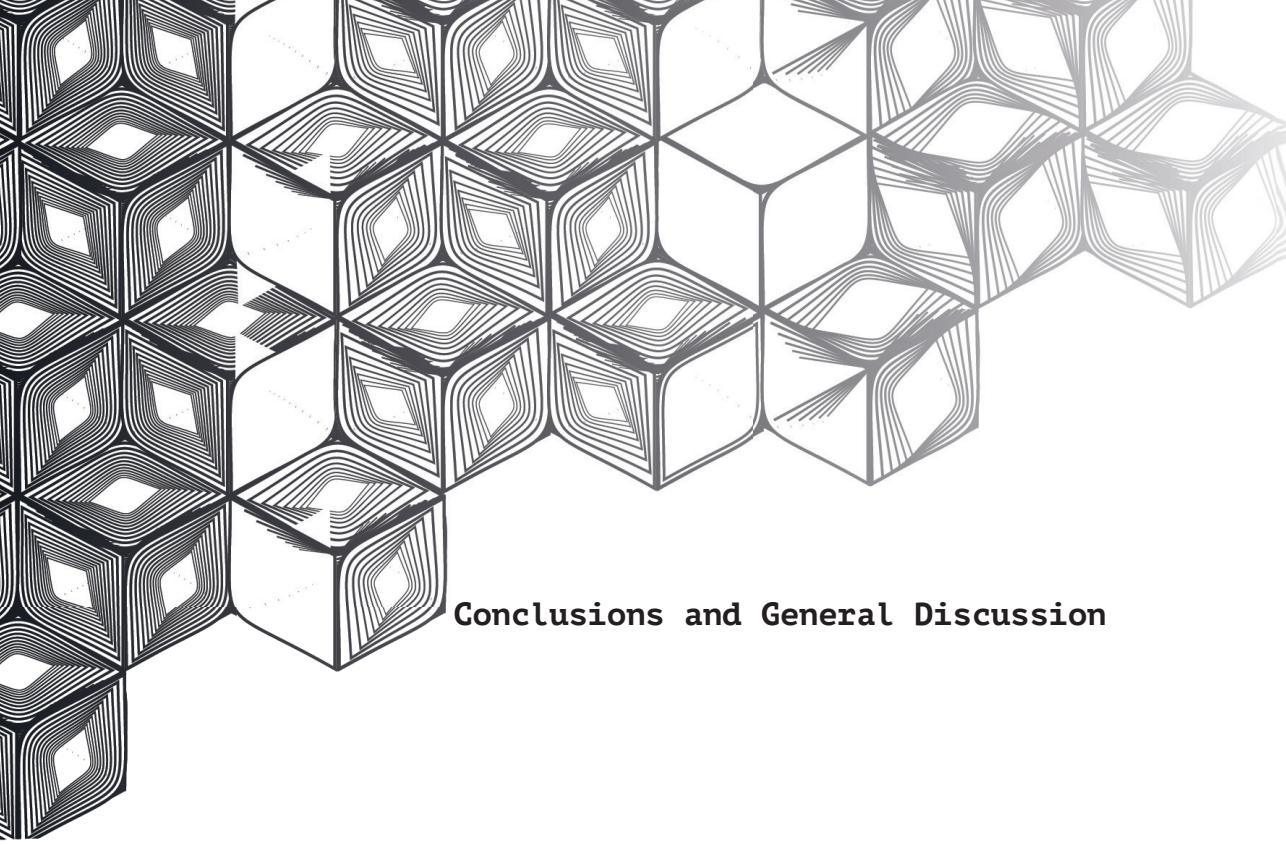
19. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
20. Andaur Navarro CL, Damen JAAG, Takada T, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open*. 2020;10(11):1-6. doi:10.1136/bmjopen-2020-038832



CHAPTER 12

"Results without quality is boring; quality without results is meaningless."

— JOHAN CRUYFF, 1947-2016



Conclusions and General Discussion

In this thesis, we primarily addressed the quality of reporting and methodological conduct of studies on clinical prediction models developed using supervised machine learning. In this final chapter, I first present the lessons learned. Subsequently, I discuss some preliminary results of the ongoing PROBAST-AI survey and the remaining challenges for reporting of studies on artificial intelligence (AI)-based prediction models. To finalize, I elaborate on the need to establish local AI governance committees — that is, a multidisciplinary panel with expertise on deployment of AI systems, in response to one of the upcoming challenges for AI-based healthcare systems.

Lessons learnt

1. We encountered significant gaps in the reporting and several hurdles on the methodological conduct of studies on AI-prediction models.
2. Most TRIPOD items are still applicable to studies on AI-based prediction models. Nonetheless, items such as source of data, study size, missing data, transformation of predictors (input features), internal validation, and availability of the AI model or algorithm bring new challenges that TRIPOD-AI needs to cover.
3. The community developing prediction models for healthcare using AI techniques such as machine learning are unaware of the TRIPOD Statement. Further journal endorsement, training, and a tailored guideline will likely increase awareness and facilitate adherence.
4. 'Spin' and poor reporting standards are frequent in studies on prediction models, AI or non-AI based. Nonetheless, its assessment remains largely subjective. Therefore, we introduced SPIN-PM, a thorough guidance on how to avoid misleading practices when reporting studies on prediction models, regardless of the statistical modelling approach.
5. The design, methodological conduct, and reporting of studies on AI-based prediction models is heterogenous. Most studies reported only the development of prediction models, focused on binary outcomes, and lacked external validation of the models.
6. Tree-based methods such as random forest, and support vector machine are the two most frequent machine learning algorithms used for the development of clinical prediction models. There are several methods available in the R package "caret" to implement these two algorithms which might show inconsistent results. Detailed reporting of the modelling steps and code availability are necessary.
7. Currently, most studies on AI-based prediction models are at high risk of bias given low number of participants with the outcome, poor handling of missing data, and questionable internal validation methods. AI-based prediction models in the field of oncology suffer similar deficiencies.
8. Most likely the reported predictive performance of AI-based prediction models in the healthcare domain is overly optimistic given the studies' high risk of bias; thus, caution is needed when interpreting the reported findings and implementing AI-based prediction models in daily healthcare.
9. It will remain a barrier to determine methodological quality of models if detailed information about data, modelling approach, and evaluation is not clearly reported in articles.
10. The development of an AI-extension for both TRIPOD and PROBAST should involve the view of different stakeholders involved in developing, validating, implementing, and publishing studies on AI-based clinical prediction models.

Tailored tools for quality assessment of AI prediction models are essential: the PROBAST-AI extension

While most applied literature is devoted to improving the performance of AI-based prediction models, there is yet relatively little guidance available regarding how an AI-algorithm should be evaluated and reported prior to its implementation in real-world healthcare, including public health and prevention settings.[1,2] Moreover, identifying valuable models through the large amount of published studies has become a challenging endeavor.[3] The AI-extension for PROBAST has two roles: facilitate the quality assessment of AI-based prediction models in order to identify prediction models with the highest potential for clinical implementation and provide methodological guidance for the design, conduct, and analysis of studies on AI-based prediction models.

In July 2021, we invited 201 international experts to participate in the first round of the online Delphi survey for PROBAST-AI, of which 105 responded to the invitation (Chapter 11). Subsequent Delphi rounds are underway. Participants were asked to vote on each item using a 5-point Likert scale (strongly agree; agree; neither agree nor disagree; agree; and strongly agree) as well as to provide comments in a free-text space. In addition to the signaling questions already in the current PROBAST tool (www.probast.org), we proposed five new, more AI-focused signaling questions in relation to data provenance and preparation, tunability, leakage, and algorithm fairness (Table 1). To better address the PROBAST-AI survey results, we briefly introduce these topics bellow:

Data provenance. A careful review of the origin, creation purposes, and changes to datasets is crucial to understanding what researchers are measuring and how valid.[4] Today's healthcare data generation involves multiple sources and types of data and as such, datasets are increasingly complex and diverse. Tracing of data thus enables readers to evaluate the credibility of a given dataset.

Data preparation. Any changes or transformations to the raw data into a format that machine learning techniques can understand and work with.[5] Examples are data transformation (e.g., normalization) and data reduction (i.e., dimensionality reduction).

Data leakage. This term is used in machine learning-based healthcare research to describe when data used to (internally) validate a model is mixed with the data that is also used for model development. Data leakage can occur during data collection, data sampling, data pre-processing, prediction modelling, or model evaluation.[6] Leakage usually leads to inflated estimates of model performance and hampers

reproducibility. The reporting and assessment of leakage-mitigating strategies are essential.

Tunability. Many machine learning algorithms have hyperparameters —that is, predefined values to control the speed and quality of the learning process. Examples include the kernel function used in SVM, the number of trees in a random forest, and the (number of) layers and their architecture in neural networks. Options for setting hyperparameters are typically default values from software packages, but also manual configuration or tuning. Careful tuning of hyperparameters can significantly improve the performance of a prediction model, and their reporting is essential for reproducibility.[7]

Fairness. Prediction models are often portrayed as objective tools to estimate the presence (diagnosis) or future occurrence (prognosis) of an outcome, however, recent evidence shows that AI-based prediction models may be inherently biased since modelling algorithms might be very good at learning and preserving historical biases based on gender, racial, or socioeconomic disparities which are embedded in our society.[8] Nonetheless, resisting the tendency to view AI-based prediction models as objective is essential to remaining patient-centered and prevent unintended harms.

Table 1. Some proposed updated signaling questions for PROBAST AI-extension*

Domain	PROBAST-AI extension	ML aspect	Agreement
Participants	Were all inclusions and exclusions in data at enrollment appropriate?	Data provenance	78%
Analysis	Were (any) pre-processing steps appropriate (e.g., cleaning, harmonization, sampling, linkage, de-duplication, de-identification, quality checks)?	Data preparation	94%
	Was hyperparameter tuning appropriate?	Tunability	83%
	Was data leakage appropriately avoided throughout modelling stages?	Leakage	83%
	If approaches to address class imbalance were used, was recalibration performed afterwards?	Data preparation	89%

*To see the full list of proposed signaling questions, see figure 1

Although the previous topics are not all restricted to AI-based prediction modeling but in fact to prediction modeling in general, most participants of the first Delphi round of PROBAST-AI indicated that these topics should become more explicitly addressed in a future update of PROBAST (Figure 1). It was widely acknowledged by the experts that PROBAST helps identifying potential biases in data collection and poor methodological quality, regardless of the modelling approach.[12,13]a tool for assessing the risk of bias (ROB However, it was noted that the major focus of PROBAST seems to be on the "development" of predictive algorithms, while the evaluation of the algorithms' performance should receive more explicit attention. Furthermore, although the experts agree on its relevance, they did wonder how to approach fairness or algorithmic bias within PROBAST (Figure 2).

Given the constant development of statistical methods for prediction modelling and the demands for more guidance on model evaluation, an update of PROBAST was considered desired. Whilst model development has been typically associated with the study's methodological quality, model evaluation relates also to the risk of bias in the estimation of a model's predictive performance. Although the assessment of both studies on model development and validation requires largely the same domains, different signaling questions are necessary for model evaluation, notably for the Analysis domain. Currently in PROBAST, the evaluation of the predictive performance of a model can be assessed using only a pair of signaling questions within the analysis domain. However, the assessment of a model's predictive performance needs to account for the following cases:

(1) poorly developed models that might perform well in daily clinical practice (i.e. they show accurate performance on evaluation)

(2) properly developed models that might perform poorly in daily practice (i.e. performance of models remains unknown).

The goal for the updated version of PROBAST will be to achieve a better distinction between quality assessment for model development and the potential risk of bias in performance through assessment of model evaluation.

Patterns of health inequalities and discrimination are perpetuate through unrepresentative datasets, biased and exclusionary model design, and discriminatory use of AI technologies.[14] Besides identifying a systematic tendency in a prediction model to favor one demographic group over another based on their inherent or acquired characteristics, fairness evaluation of a prediction model involves examining the impact of the discrimination based on a set of legal, ethical, and cultural requirements that vary per country. Whether a prediction model is 'fair' may be hard to determine. To date, PROBAST does not provide explicit guidance on how to assess the risk of algorithmic bias in prediction models. However, it does

provide guidance on how to evaluate data representativeness through the assessment of sampling bias by using the signaling questions of the domains Participants, Predictors, and Outcome. In practice, unfairness may be difficult to uncover when studies are improperly reported and specially, when data is not available for scrutiny. TRIPOD-AI and the update of PROBAST with its AI extension will undoubtedly contribute further to address fairness in the development and validation of clinical prediction models.

To conclude, the update of PROBAST with its AI extension is expected to assist multiple stakeholders (i.e. researchers, editors, peer-reviewers, guideline developers, decision makers, and patients and their relatives), to understand, interpret, and critically appraise the quality of studies in which prediction models were developed and validated, regardless of whether models were built using AI or traditional statistical techniques. Overall, it will help promote better methodological conduct of studies on prediction models and thus, potentially speed up the introduction of valuable models into healthcare settings.

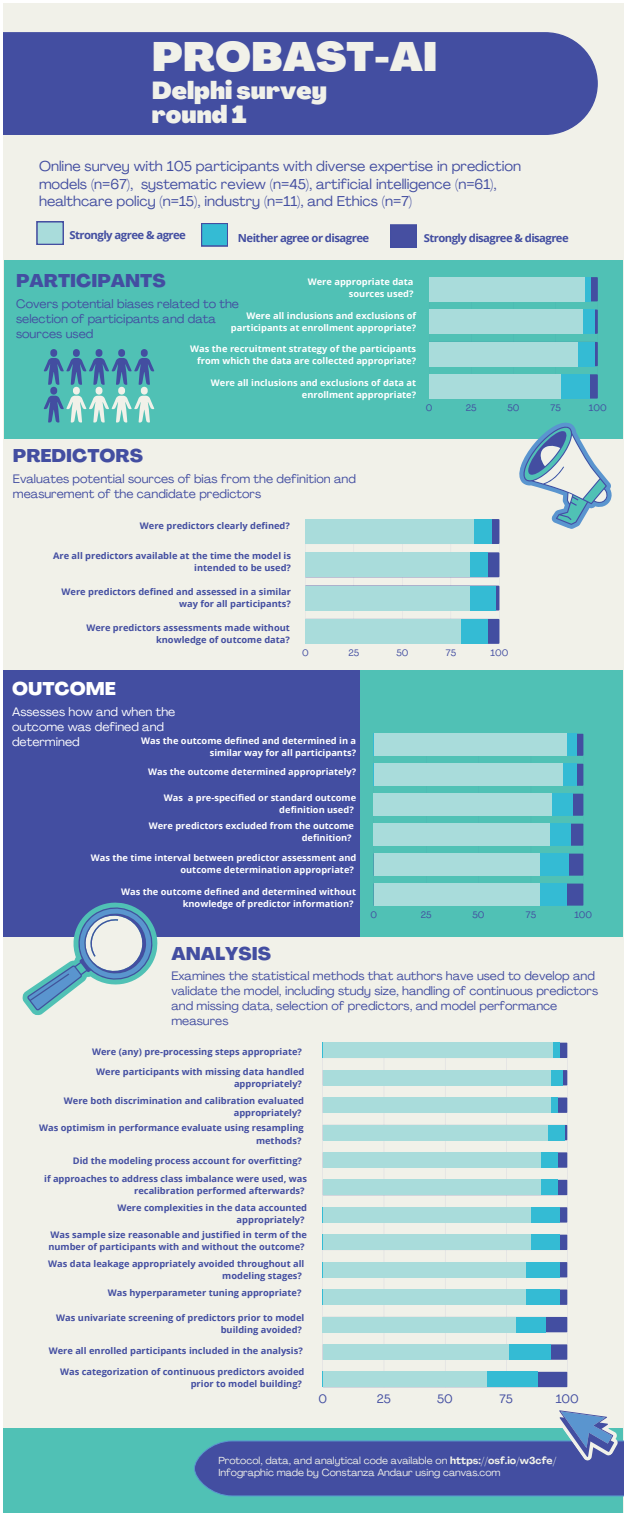


Figure 1. Results from round 1 PROBAST-AI Delphi survey

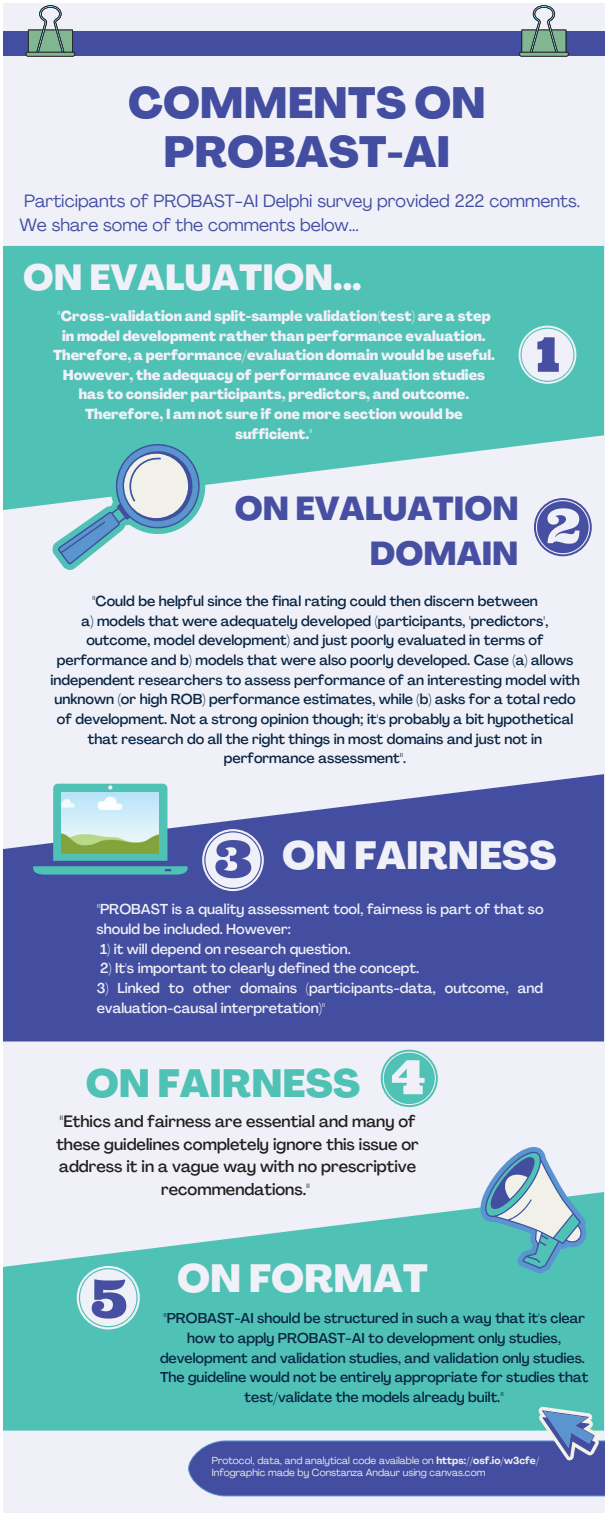


Figure 2. Comments from round 1 PROBAST-AI Delphi survey

Reproducibility is essential for scientific progress

Currently, most prediction models that make use of machine learning are unavailable for extensive critical appraisal, external validation, or even for its application in healthcare practice (Chapter 3). Poor reporting of modelling steps, and poor code and data availability means that most studies on prediction models are non-reproducible and even more important, developed models cannot be applied to new individuals in daily clinical practice. Consequently, several checklists have been published in response to making AI applications in healthcare reproducible.[15–21] TRIPOD is one of the most cited reporting guidelines in the prediction modeling field, most likely due to its extensive explanation and elaboration document[22], adherence form[23], and the endorsement by several journals (www.tripod-statement.org). We can therefore anticipate a positive effect of TRIPOD-AI on the reporting quality of studies on AI-based prediction models.[24,25] The TRIPOD-AI extension is currently in advanced stages of development.[24]

Prediction models based on machine learning present unique challenges to reproducibility, which must be carefully considered to ensure that guidelines are adopted properly.[26] During our umbrella review (Chapter 2), we were challenged by the endless taxonomy of machine learning techniques and the lack of harmonized terminology between statistics and machine learning.[27] Moreover, developers of machine learning-based models usually emphasize the limitation to report healthcare data and complex algorithms using the traditional biomedical reports. Besides creating awareness and promoting tools such as TRIPOD-AI within the machine learning community, it has become indispensable to change the way published research is conceptualized and operationalized in studies on prediction models to reduce research waste.[28]

A new publication format that acknowledged the importance of code and data with a unique interface for the reporting of studies on AI-based prediction models could allow researchers and algorithm developers to automatically register, report, share code and data, harmonize terminology, interact with figures and tables, and collaborate and discuss model development and validation. Executable research article (ERA) is a new research publication format that could also be used for studies on prediction models.[29] ERAs make use of technologies like Jupyter notebook, R Markdown, and repositories such as Zenodo and GitHub to make reporting of research articles more transparent and reproducible. The demo can be found on https://elifesciences.org/articles/30274/executable.

ERAs are computationally reproducible papers that combine text, raw data, and code used for the analysis. For example, static tables and figures are replaced with the code chunks that reproduce them, so the reader can interact with them. ERAs could be upgraded to serve also as online registry for studies on prediction

models. Then each step on prediction model development and validation can become auditable through data stamps, avoiding selective reporting. Systematic reviews and Individual Participant Data meta-analyses could also benefit from the registration of primary studies. Integrating technology such as Penelope-ai (https://www.penelope.ai/), could enable automated checks to the manuscript and thus provide immediate feedback to authors on how to improve the completeness of reporting based on TRIPOD and TRIPOD-AI (www.tripod-statement.org).[12,22,30] Furthermore, this openness could facilitate and improve the peer-review process by allowing readers' input on the code. Funders could monitor research progress and impact easier.

Only the rigorous adherence to reporting standards, namely TRIPOD-AI, and extensive changes to the publication format, will facilitate a smoother translation of studies on prediction models into deployable AI tools and finally improve health care and prevention.

AI Governance committees: a healthcare based on AI technologies

AI- or machine learning-based prediction models will very likely become the core of automated tools and software to assist healthcare professionals, patients, and citizens in making healthcare and lifestyle decisions. However, one needs to be aware that AI-based prediction models are sensitive to historical biases and changes in population and care pathways (i.e., data shifts).[31] Behind the relative 'newness' of AI-based prediction models, there is an urgent need to regulate the scope and monitor the performance of AI-based prediction models in the clinical workflow. AI Governance committees might rise as a key component for safety surveillance and updating of AI-based prediction models in routine clinical practice.[32,33]

Evidence shows that the performance of prediction models – whether developed with AI or not – worsens over time as a consequence of natural and expected data shifts.[34] Therefore, the impact of prediction models on patients' health outcomes and on the clinical workflow should be evaluated constantly. However, given the large investment required to carry out randomized trials, impact or effectiveness studies of prediction models are infrequent.[35] The widespread and near-real-time availability of real-world data might offer opportunities to quickly generate evidence to address these needs.

The adoption of AI-based prediction algorithms by healthcare professionals may be faster than the development of rigorous evidence to support an appropriate, safe, and equal use. A plethora of evidence shows that prediction models run considerable risk of insufficient validation (i.e. predictive performance evaluation) and poor generalizability (chapter 8).[35,36] The major role of AI-based governance committees would be to prevent unintended harm to patients by acknowledging that prediction models are dynamic.[37,38]

AI governance committees might have different responsibilities: (1) monitor performance of implemented prediction models; (2) the development of protocols for an efficient and effective updating and repurposing of AI-based prediction models; (3) update models to the local population and/or to the local practice; (4) investigate potential inequalities to guarantee that models perform consistently across patient cohorts, especially those who may not have been adequately represented in the training cohort; (5) surveillance of adverse events; (6) recalibrate models with up-to-date information; and (7) auditing machine learning models.[1,39]

Most healthcare organizations today lack the data infrastructure required to collect the data that is needed for updating prediction algorithms, representing a barrier for the work of AI governance committees. Likewise, a healthcare workforce train to take careful consideration of the strengths, limitations, and potential biases in the interpretation of the output of AI systems is necessary. There is an urgent

need to create local AI governance committees and train a healthcare workforce on AI technologies (advantages and pitfalls) to obtain sustainable changes after the introduction of AI systems into clinical care.[1,39]

Concluding remarks

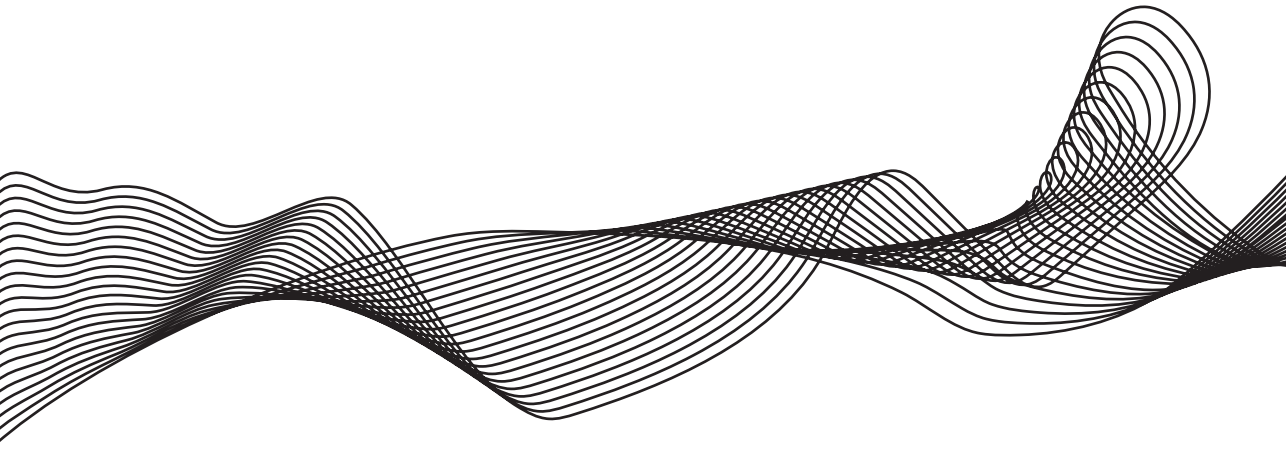
Artificial intelligence, including machine learning, represents a paradigm shift for many sectors and healthcare is no exception. It has become indispensable to develop AI-based clinical prediction models that are not only feasible or accurate but also fair, safe, and cost-effective. The assessment of studies on AI-based clinical prediction models requires deep methodological understanding of the technical challenges and potential biases. In response, the development of TRIPOD-AI and the update of PROBAST with its AI extension will provide guidance on the highest standards for reporting and methodological conduct of studies on early stages of clinical prediction model development and validation with the goal of ensuring transparency and safety.

REFERENCES

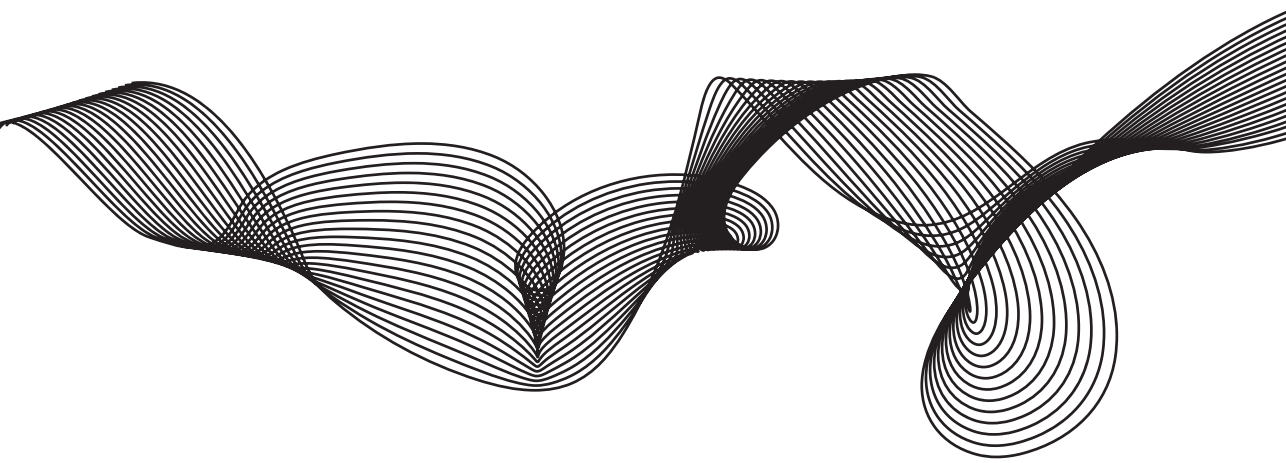
1. van Smeden M, Moons C, Hooft L, Kant I, Van Os H, Chavannes N. Guideline for high-quality diagnostic and prognostic applications of AI in healthcare. Published online 2021.
2. Lones MA. How to avoid machine learning pitfalls: a guide for academic researchers. Published online 2021:1-17. http://arxiv.org/abs/2108.02497
3. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *J Am Med Informatics Assoc*. 2022;00(0):1-7. doi:10.1093/jamia/ocac002
4. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*. 2021;2(11). doi:10.1016/j.patter.2021.100336
5. Kotsiantis SB, Kanellopoulos D. Data preprocessing for supervised leaning. *Int J Comput Sci*. 2006;1(2):1-7. doi:10.1080/02331931003692557
6. Kapoor S, Narayanan A. *Leakage and the Reproducibility Crisis in ML-Based Science*.; 2022. doi:10.48550/arXiv.2207.07048
7. Probst P, Boulesteix AL, Bischl B. Tunability: Importance of hyperparameters of machine learning algorithms. *J Mach Learn Res*. 2019;20:1-32.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-)*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
9. Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front Artif Intell*. 2021;0:116. doi:10.3389/FRAI.2020.561802
10. Pessach D, Shmueli E. A Review on Fairness in Machine Learning. *ACM Comput Surv*. 2023;55(3):1-44. doi:10.1145/3494672
11. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. *npj Digit Med*. 2019;2(1):1-7. doi:10.1038/s41746-019-0103-3
12. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
13. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
14. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ*. 2021;372:1-5. doi:10.1136/bmj.n304
15. 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(12). doi:10.2196/jmir.5870
16. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
17. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0253-3
18. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Heal Care Informatics*. 2021;28(1):1-10. doi:10.1136/bmjhci-2021-100385
19. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform*. 2021;153(June). doi:10.1016/j.ijmedinf.2021.104510

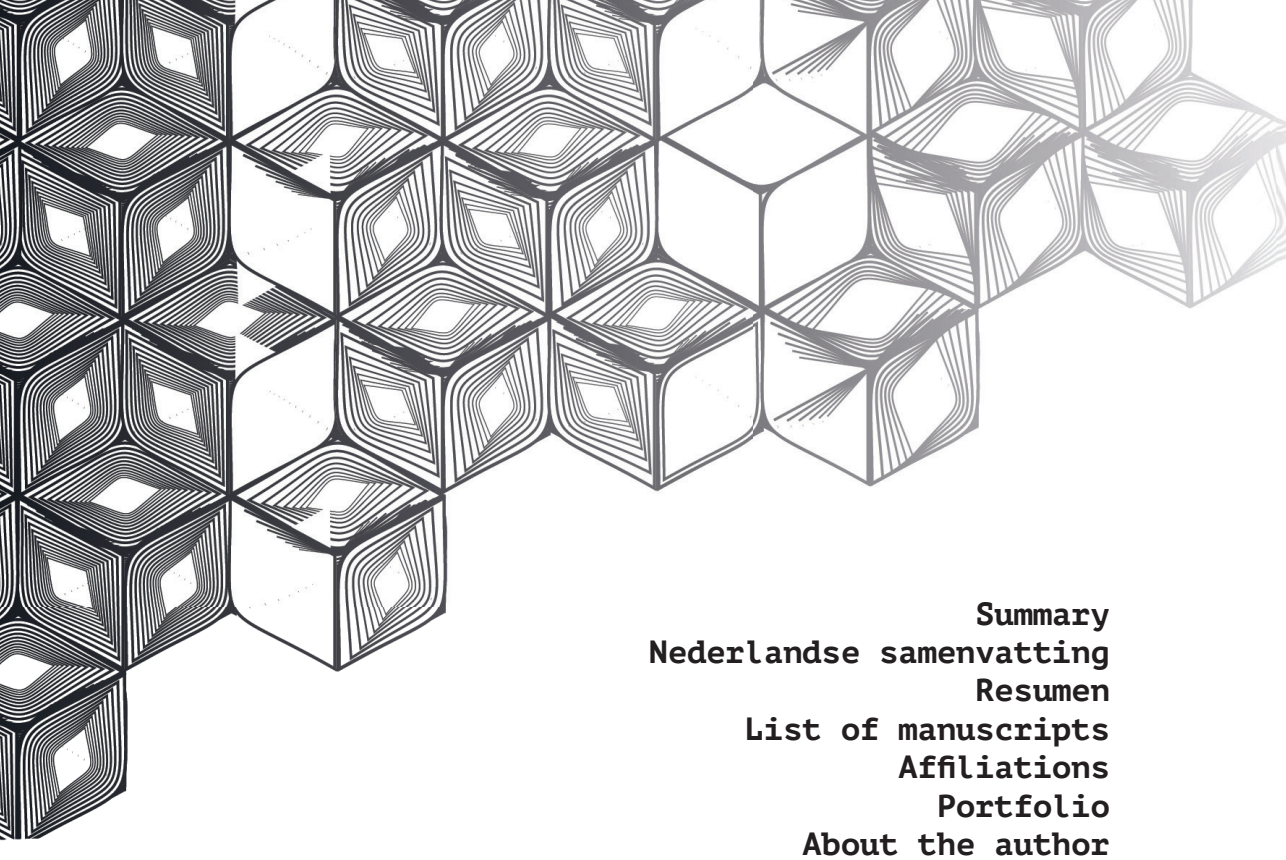
20. Reismann J, Romualdi A, Kiss N, et al. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. *PLoS One*. 2019;14(9):1-11. doi:10.1371/journal.pone.0222030
21. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): Developing reporting standards for artificial intelligence in health care. *J Am Med Informatics Assoc*. 2020;27(12):2011-2015. doi:10.1093/jamia/ocaa088
22. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
23. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9(4). doi:10.1136/bmjopen-2018-025611
24. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(e048008):1-7. doi:10.1136/bmjopen-2020-048008
25. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
26. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA - J Am Med Assoc*. 2020;323(4):305-306. doi:10.1001/jama.2019.20866
27. Faes L, Sim DA, Van Smeden M, Held U, Bossuyt PM, Bachmann LM. Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? *Front Digit Heal*. 2022;1:833912. doi:10.3389/fdgth.2022.833912
28. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X
29. Lasser J. Creating an executable paper is a journey through Open Science. *Commun Phys*. 2020;3(1):1-5. doi:10.1038/s42005-020-00403-4
30. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393. doi:10.1016/S01406736(19)302351
31. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021;385(3).
32. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36. doi:10.1038/s41591-018-0307-0
33. Huisman M, Ranschaert E, Parker W, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. *Eur Radiol*. 2021;31(9):7058-7066. doi:10.1007/s00330-021-07781-5
34. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Informatics Assoc*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030
35. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68(1):25-34. doi:10.1016/j.jclinepi.2014.09.007
36. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
37. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn Res*.

2021;5(1):1-7. doi:10.1186/s41512-020-00090-3
38. Andrinopoulou E-R, Harhay MO, Ratcliffe SJ, Rizopoulos D. Reflection on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data. *Int J Epidemiol*. 2021;50(5):1731-1743. doi:10.1093/ije/dyab047
39. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit Med*. 2022;5(1). doi:10.1038/s41746-021-00549-7



APPENDICES





SUMMARY

The role of prediction models for clinical decision-making is becoming increasingly important. For the implementation of valuable prediction models in clinical practice, properly conducted and well reported studies on early stages of model development and validation are essential.

In **Chapter 2**, we present the protocol for the umbrella review that constitutes the core of this doctoral thesis. Our objective was to evaluate (1) completeness of reporting, (2) quality and risk of bias, (3) methodological conduct, and (4) spin or over-interpretation in studies on artificial intelligence (AI)-based diagnostic and prognostic prediction models. We provided the protocol registration [PROSPERO, ID: CRD42019161764], review questions, and described the methods for literature search, data extraction, and the critical appraisal of included studies.

Previous studies have consistently found poor completeness of reporting of studies on regression-based prediction models. In **Chapter 3**, we systematically reviewed the adherence of 152 studies on machine learning-based prediction models to the 22-item checklist with the minimum standards for high quality reporting, TRIPOD. Overall, articles adhered to a median of 38.7% of applicable TRIPOD items. Reporting of background, objectives, source of data, limitations, and interpretation of findings reached at least 75% of adherence, whilst for 12 items mostly related to methods and results, adherence was below 25%. No articles fully adhered to complete reporting of the abstract and very few reported the flow of participants, appropriate title, blinding of predictors, model specification, and model's predictive performance. Furthermore, we identified that TRIPOD requires new items to cover AI-related aspects such as model tunability.

Published research may also be biased when results are described in a more favourable way than they deserve or when harms and limitations are downplayed. These so-called 'spin' or misleading practices in the reporting of studies on prediction models have been demonstrated empirically in this thesis. In **Chapter 4**, we systematically reviewed the included studies for 15 spin practices and 11 poor reporting standards. A considerable number of studies lack a pre-specified protocol, make claims of clinical applicability (without further validation), and limitations are neither reported nor discussed in the context of previously developed models. Given that a first approach to spin evaluation using a classification scheme for studies on prognostic factors proved to be inefficient, we present SPIN-PM in **Chapter 5**, a new framework for spin identification tailored to studies on prediction models, regardless of the modelling approach.

Chapter 6 provides a detailed overview of the study design, modelling strategies, and performance measures reported in studies on machine learning-

based prediction models. Most studies reported only the development of prediction models and focused on binary outcomes. Within the 152 studies, we evaluated 522 models (on average 9.4 models per study), in which the most common modelling algorithms used were support vector machine and random forest. Special attention is required to areas such as handling of missing values, methods for internal validation, and reporting of calibration. Several aspects remained poorly reported, constituting the major barrier for detailed critical appraisal. Moreover, we deepen on the methodological conduct of prognostic prediction models in oncology (**Chapter 7**). We reviewed 62 publications reporting 152 models. Authors of studies on clinical prediction models in oncology need to pay attention to sample size estimation, censoring, and ensure that models are available for independent validation to improve the methodological conduct of machine learning based clinical prediction models in oncology.

We comprehensively reviewed the methodological quality and risk of bias of studies on prediction models developed using machine learning techniques across medical specialties (**Chapter 8**). We applied PROBAST to 152 studies on model development and 19 external validations. Of these 171 analyses, 148 were rated at high risk of bias due to deficiencies in the domain analysis mainly because of small study size, poor handling of missing data, and failure to deal with overfitting. Efforts to improve the design, conduct, reporting, and validation of such studies are necessary to boost the introduction of machine learning based prediction models into clinical practice.

In **Chapter 9**, we compared the absolute risk probabilities of three different modelling techniques: logistic regression, random forest, and support vector machine. For the last two techniques, we applied two different implementation methods within the package 'caret' in the statistical software R. Using logistic regression as benchmark, we showed that risk probabilities for deep venous thrombosis vary substantially between modelling techniques and implementation methods.

TRIPOD and PROBAST were published to facilitate the critical appraisal of studies on diagnostic and prognostic prediction models. In **Chapter 10**, we described the five stages for the development of both extensions to machine learning-based models. The systematic reviews presented in this thesis compromised stage one. A survey using the Delphi methodology constitute stage 2 and the results of round one are briefly discussed in **Chapter 12**. We shared in **Chapter 11** the detailed protocol for PROBAST-AI.

In conclusion, we have thoroughly evaluated the methodological conduct and reporting of studies on machine learning-based prediction models. The findings described above will contribute to the development of both PROBAST-AI and TRIPOD-AI. Furthermore, we have proposed a framework for spin identification

in studies on prediction models, SPIN-PM. Overall, we have upgraded the current guidance for quality evaluation and interpretation of findings in studies on prediction models, potentially helping reduce vague and biased research outputs. We can anticipate that future studies on prediction models will benefit from these guidelines.

SAMENVATTING

De rol van voorspellingsmodellen voor klinische besluitvorming wordt steeds belangrijker. Voor de implementatie van waardevolle voorspellingsmodellen in de klinische praktijk zijn goed uitgevoerde en goed gerapporteerde studies naar de vroege stadia van modelontwikkeling en -validatie essentieel.

In **hoofdstuk 2** presenteren we het protocol voor de overkoepelende review die de kern vormt van dit proefschrift. Ons doel was om (1) volledigheid van rapportage, (2) kwaliteit en risico op bias, (3) methodologische karakteristieken, en (4) spin of over-interpretatie te evalueren in studies over AI-gebaseerde diagnostische en prognostische voorspellingsmodellen. We beschrijven de protocolregistratie [PROSPERO, ID: CRD42019161764], de reviewvragen, de literatuursearch, de data-extractie, en de kritische beoordeling van de geïncludeerde studies.

Eerdere studies hebben consistent beschreven dat de rapportage van studies over regressie-gebaseerde voorspellingsmodellen onvolledig is. In **hoofdstuk 3** hebben wij systematisch onderzocht in hoeverre 152 studies over voorspellingsmodellen op basis van machine learning voldeden aan TRIPOD checklist. TRIPOD bestaat uit 22 items voor de minimum standaarden voor rapportage van hoge kwaliteit. In het algemeen voldeden de artikelen aan een mediaan van 38,7% van de TRIPOD-items. De rapportage van achtergrond, doelstellingen, gegevensbron, beperkingen en interpretatie van de bevindingen was voor ten minste 75% in overeenstemming met de TRIPOD-items, terwijl 12 items in minder dan 25% van de studies voldoende beschreven waren. Geen enkel artikel voldeed aan de volledige rapportage van de samenvatting en zeer weinig artikelen rapporteerden de selectie van deelnemers, de volledige titel, blindering van voorspellers, modelspecificatie, en de voorspellende waarde van het model. Bovendien stelden we vast dat de TRIPOD uitgebreid dient te worden met nieuwe items vereist om bijvoorbeeld de bruikbaarheid van het model te beschrijven.

Wanneer de resultaten van onderzoek gunstiger worden beschreven dan ze verdienen of wanneer beperkingen van het onderzoek worden gebagatelliseerd kan dit resulteren in een vertekende publicatie van het onderzoek. Deze zogenaamde 'spin' of misleidende praktijken in de rapportage van studies over voorspellingsmodellen zijn in dit proefschrift empirisch beschreven. In **hoofdstuk 4** hebben we de geïncludeerde studies systematisch beoordeeld op 15 'spin'-praktijken en 11 slechte rapportagestandaarden. In een aanzienlijk aantal studies ontbreekt een vooraf gespecificeerd protocol, worden claims gemaakt van klinische toepasbaarheid (zonder verdere validatie), en worden beperkingen niet gerapporteerd of besproken in de context van eerder ontwikkelde modellen. Aangezien deze eerste evaluatie van 'spin' op basis van een schema ontwikkeld voor studies naar prognostische factoren niet passend bleek, presenteren wij in **hoofdstuk 5** SPIN-PM, een nieuw raamwerk

voor de identificatie van 'spin' in studies naar voorspellingsmodellen.

Hoofdstuk 6 geeft een gedetailleerd overzicht van de studieopzet, de modelleringsstrategieën en de rapportage van de voorspellende kwaliteiten in studies over voorspellingsmodellen die ontwikkeld zijn met machine learning methodes. De meeste studies rapporteerden alleen de ontwikkeling van voorspellingsmodellen en richtten zich op binaire uitkomsten. Binnen de 152 studies evalueerden wij 522 modellen (9,4 modellen per studie), waarbij support vector machine en random forest de vaakst gebruikte modelleringsalgoritmen waren. Speciale aandacht is vereist voor gebieden zoals de behandeling van missende waarden, methoden voor interne validatie, en de rapportage van kalibratie. Verscheidene aspecten werden gebrekkig gerapporteerd, wat de belangrijkste belemmering vormt voor een gedetailleerde kritische beoordeling van de kwaliteit van een studie. Daarnaast hebben we gekeken naar de methodologische kwaliteit van prognostische voorspellingsmodellen in de oncologie (**hoofdstuk 7**). We hebben 62 publicaties beoordeeld waarin 152 modellen werden gerapporteerd. Auteurs van studies naar klinische voorspellingsmodellen in de oncologie moeten aandacht besteden aan steekproefgrootte, censoring, en ervoor zorgen dat modellen beschikbaar zijn voor onafhankelijke validatie.

We hebben de methodologische kwaliteit en het risico op vertekening van studies over voorspellingsmodellen ontwikkeld met behulp van machine learning algoritmen in alle medische specialismen uitvoerig onderzocht (**hoofdstuk 8**). Wij pasten PROBAST toe op 152 studies over modelontwikkeling en 19 externe validaties. Van deze 171 analyses kregen er 148 een hoog risico op vertekening door tekortkomingen in het domein analyse, onder andere door kleine studieomvang, slechte behandeling van missende gegevens, en het negeren van overfitting. Acties om het ontwerp, de uitvoering, de rapportering en de validatie van dergelijke studies te verbeteren zijn noodzakelijk om de invoering van voorspellingsmodellen op basis van machine learning in de klinische praktijk te stimuleren.

In **hoofdstuk 9** vergeleken we de absolute kansen berekend volgens drie verschillende modelleringstechnieken: logistische regressie, random forest, en support vector machine. Voor de laatste twee technieken pasten we twee verschillende implementatiemethoden toe binnen het pakket 'caret' in de statistische software R. Met logistische regressie als benchmark toonden we aan dat de absolute kansen op diep veneuze trombose substantieel variëren tussen modelleertechnieken en binnen implementatiemethoden.

TRIPOD en PROBAST werden gepubliceerd om de kritische beoordeling van studies over diagnostische en prognostische voorspellingsmodellen te vergemakkelijken. In **hoofdstuk 10** hebben we de vijf stadia beschreven voor de ontwikkeling van uitbreidingen op deze checklists voor voorspellingsmodellen gebaseerd op machine learning. De systematische reviews die in dit proefschrift

worden gepresenteerd, zijn onderdeel van het eerste stadium in de ontwikkeling van deze uitbreidingen. Een enquête met behulp van de Delphi methodologie vormde stadium 2 en de resultaten voor ronde één worden kort besproken in **hoofdstuk 12**. In **hoofdstuk 11** hebben we het gedetailleerde protocol voor de ontwikkeling van PROBAST-AI beschreven.

Concluderend hebben we de methodologische kenmerken en de rapportage van studies over voorspellingsmodellen gebaseerd op machine learning grondig geëvalueerd. De hierboven beschreven bevindingen dragen bij aan de ontwikkeling van zowel PROBAST-AI als TRIPOD-AI. Voorts hebben wij een kader voorgesteld voor de identificatie van 'spin' in studies over voorspellingsmodellen, SPIN-PM. In het algemeen hebben wij de huidige richtsnoeren voor de kwaliteitsevaluatie en de interpretatie van bevindingen in studies over voorspellingsmodellen verbeterd, wat kan helpen om de presentatie van vertekende onderzoeksresultaten te verminderen. Wij verwachten dat toekomstige studies over voorspellingsmodellen baat zullen hebben bij deze richtlijnen.

RESUMEN

El papel de los modelos de predicción en la toma de decisiones clínicas es cada vez más importante. Para la aplicación de modelos de predicción valiosos para la práctica clínica, es esencial conducir y reportar estudios sobre las primeras fases de desarrollo y validación de los modelos de manera apropiada.

En el **capítulo 2**, presentamos el protocolo de la revisión sistemática que constituye la base de esta tesis doctoral. Nuestro objetivo era evaluar (1) la exhaustividad del reporte, (2) la calidad y el riesgo de sesgo, (3) la conducta metodológica, y (4) el spin o la sobreinterpretación en los estudios sobre modelos de predicción diagnóstica y pronóstica basados en inteligencia artificial (IA). Proporcionamos el registro del protocolo [PROSPERO, ID: CRD42019161764], las preguntas de revisión, y se describen los métodos para la búsqueda bibliográfica, la extracción de datos y la evaluación crítica de los estudios incluidos.

En estudios anteriores se ha constatado sistemáticamente la escasa exhaustividad del reporte de los estudios sobre modelos de predicción basados en regresión. En el **capítulo 3**, revisamos sistemáticamente la adherencia de 152 estudios sobre modelos de predicción basados en machine learning a TRIPOD, una lista de 22 ítems con los estándares mínimos para el reporte de alta calidad. En general, los artículos cumplieron una mediana del 38,7% de los ítems aplicables de TRIPOD. Antecedentes, objetivos, la fuente de los datos, limitaciones e interpretación de los hallazgos alcanzan una adherencia de al menos 75%, mientras que, en 12 ítems, la mayoría relacionados con los métodos y los hallazgos, el cumplimiento fue inferior al 25%. Ningún artículo cumplió plenamente con la información mínima y muy pocos estudios informaron del flujo de participantes, un título apropiado, cegamiento de los predictores, especificación del modelo y rendimiento predictivo del modelo. Además, identificamos que TRIPOD requiere nuevos ítems para cubrir aspectos relacionados con la IA, como tunear el modelo.

La investigación publicada también puede estar sesgada cuando los resultados se describen de forma más favorable de lo que merecen o cuando se resta importancia a los daños y las limitaciones. Estas prácticas, llamadas también spin, en el reporte de estudios sobre modelos de predicción se han demostrado empíricamente en esta tesis. En el **capítulo 4**, revisamos sistemáticamente los estudios incluidos en busca de 15 prácticas de spin y 11 formas de reporte deficientes. Un número considerable de estudios carece de un protocolo, hacen afirmaciones de aplicabilidad clínica (sin validación externa), y no se informa de las limitaciones ni se discute el modelo en el contexto de modelos desarrollados previamente. Dado que una primera aproximación a la evaluación de spin utilizando un esquema de clasificación para estudios sobre factores pronósticos resultó ineficaz, en el capítulo 5 presentamos SPIN-PM, un nuevo marco para la identificación de spin adaptado a estudios sobre

modelos de predicción, independientemente si los modelos fueron desarrollados con regresión o machine learning.

En el **capítulo 6** se ofrece un panorama detallado del diseño de los estudios, las estrategias de modelización y las métricas de rendimiento utilizadas en los estudios sobre modelos de predicción basados en machine learning. La mayoría de los estudios sólo informan sobre el desarrollo de modelos de predicción y se centran en la clasificación. En los 152 estudios, se evaluaron 522 modelos (una media de 9,4 modelos por estudio), en los que los algoritmos más utilizados fueron support vector machine y random forest. Es necesario prestar especial atención a áreas como el tratamiento de los valores faltantes, los métodos de validación interna y la calibración. Varios aspectos permanecen mal reportados, lo que constituye el principal obstáculo para una evaluación crítica detallada. Adicionalmente, profundizamos en la conducta metodológica de los modelos de predicción pronóstica en oncología (**capítulo 7**). Revisamos 62 publicaciones que informaban sobre 152 modelos. Los autores de estudios sobre modelos de predicción clínica para oncología deben prestar atención a la estimación del tamaño de la muestra, censoring y garantizar que los modelos estén disponibles para su validación independiente a fin de mejorar la conducta metodológica de los modelos de predicción clínica basados en el machine learning en oncología.

Se revisó exhaustivamente la calidad metodológica y el riesgo de sesgo de los estudios sobre modelos de predicción desarrollados basados en machine learning en varias especialidades médicas (**capítulo 8**). Aplicamos PROBAST a 152 estudios sobre desarrollo de modelos y 19 estudios de validaciones externas. De estos 171 análisis, 148 se calificaron con alto riesgo de sesgo debido a deficiencias en el dominio análisis, principalmente por un tamaño muestral pequeño, manejo deficiente de los datos faltantes y falta de tratamiento del overfitting. Esfuerzos para mejorar el diseño, la conducción, el reporte y la validación de tales estudios son necesarios para potenciar la introducción de la predicción basada en machine learning en la práctica clínica.

En el **capítulo 9**, comparamos las probabilidades absolutas de riesgo de tres algoritmos diferentes: regresión logística, random forest y support vector machine. Para los dos últimos algoritmos, aplicamos dos métodos de implementación diferentes en el paquete "caret" del programa estadístico R. Utilizando la regresión logística como referencia, demostramos que las probabilidades absolutas de riesgo de trombosis venosa profunda varían sustancialmente entre los algoritmos y entre los métodos de implementación.

TRIPOD y PROBAST fueron publicados para facilitar la apreciación critica de los estudios sobre modelos predictivos diagnósticos y pronósticos. En el capítulo 10, describimos las 5 etapas para el desarrollo de ambas extensiones para modelos

basados en machine learning. Las revisiones sistemáticas presentadas en esta tesis constituyen la etapa 1. Una encuesta utilizando la metodología Delphi constituyó la etapa 2 y los resultados de la ronda uno es discutidos brevemente en el **capítulo 12**. Discutimos en el **capítulo 11** el protocolo detallado para PROBAST-AI.

En conclusión, hemos evaluado detalladamente la conducta metodológica y el reporte de los estudios sobre modelos predictivos basados en machine learning. Los resultados descritos anteriormente contribuirán al desarrollo de ambos PROBAST-AI y TRIPOD-AI. Además, hemos propuesto un marco teórico para la identificación de spin en estudios sobre modelos predictivos, SPIN-PM. En general, hemos actualizado la guía actual para la evaluación e interpretación de la calidad de hallazgos en estudios sobre modelos de predicción, lo que podría ayudar a reducir la investigación vaga y sesgada. Podemos anticipar que los futuros estudios sobre modelos de predicción se beneficiarán de estas guías.

Manuscripts published and included in this doctoral thesis

Andaur Navarro CL, Damen JAAG, Takada T, 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(11):1-6. doi:10.1136/bmjopen-2020-038832+

Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021;375:n2281. doi:10.1136/bmj.n2281

Collins GS, Dhiman P, **Andaur Navarro CL**, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open.2021;11(e048008):1-7. doi:10.1136/bmjopen-2020-048008

Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22(1):12. doi:10.1186/s12874-021-01469-6

Dhiman P, Ma J, **Andaur Navarro CL**, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22(1):1-16. doi:10.1186/s12874-021-01469-6

Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2022; doi:10.1016/j.jclinepi.2022.11.015

Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning-based prediction models. *Accepted in Journal of Clinical Epidemiology*

Manuscripts published but not included in this doctoral thesis

Dhiman P, Ma J, **Andaur Navarro CL**, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol. 2021; doi:10.1016/j.jclinepi.2021.06.024

Dhiman P, Ma J, **Andaur Navarro CL**, et al. Risk of bias of prognostic models developed using machine learning: systematic review in oncology. 2022; 6:13. doi: 10.1186/s41512-022-00126-w

Wynants L, van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020; 369:m1328. doi:10.1136/bmj.m1328

Dhiman P, Ma J, **Andaur Navarro CL**, et al. Overinterpreation of findings in machine learning prediction model studies in oncology: an evaluation of 'spin'. J Clin Epidemiol. 2023; doi:10.1016/j.jclinepi.2023.03.012

Twait E, **Andaur Navarro CL**, Gudnason V, et al. Dementia prediction using clinically-accesible variables: a proof-of-concept using machine learning. *Submitted to BMC Medical Informatics and Decision Making*

Manuscripts in preparation

Andaur Navarro CL, Damen JAA, Ghannad M, et al. SPIN-PM: A framework to evaluate presence of spin in studies on prediction models.

Andaur Navarro CL, Damen JAA, Takada T, et al. Estimating risk using machine learnng for clinical prediction models: predicting deep vein thrombosis.

Levis B, Snell KIE, Damen JAA, et al. Risk of bias assessments in individual participant data meta-analyses of test accuracy and prediction models: a review shows improvements are needed.

AUTHORS AFFILIATIONS

Constanza L Andaur Navarro
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Johanna AA Damen
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Karel GM Moons
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Lotty Hooft
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Toshihiko Takada
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Steven WJ Nijman
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Paula Dhiman
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK
NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, NIHR Oxford, Biomedical Research Centre, Oxford, UK

Jie Ma
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK

Gary S Collins
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK
NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, NIHR Oxford, Biomedical Research Centre, Oxford, UK

Ram Bajpai
Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

Richard Riley
Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

Maarten van Smeden
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Johannes Reitsma
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

Mona Ghannad
Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, Netherlands

PhD Portfolio

Name PhD student:	Constanza Lourdes Andaur Navarro
UMC Department:	Julius Center – Epidemiology - Methodology
PhD Period:	September 2019 – September 2022
Promotor:	Prof. Dr. Karel GM Moons Prof. Dr. Lotty Hooft
Co-promotor:	Dr. Johanna AA Damen

1. PhD training

Courses at the MSc Epidemiology	Year	ECTS
Clinical epidemiology	2019	1.5
Prognosis research	2020	1.5
Systematic review and meta-analysis of prognostic studies	2020	1.5
Systematic review of individual patient data	2020	1.5
Machine learning & application in medicine	2020	1.5
Courses at the PhD Graduate School of Life Sciences		
The Art of Presenting in Sciences	2019	1.0
Adobe InDesign – from Dissertation layout to Poster Design	2020	0.6
Writing a scientific paper	2020	1.5
Academic writing in English	2020	1.8
Achieving your goals and performing more successfully during your PhD	2021	1.0
Scientific artwork: Data visualisation and Infographics with Adobe Illustrator	2022	1.4
Specific courses		
Best practices for writing reproducible code	2020	
Quick start to Research Data Management - UU	2021	
Quick start to Research Data Management - UMC	2021	
Courses at other institutions		
Peerspectives – Charite & BMJ	2021	4.0
Publication School – UK EQUATOR Centre	2021	3.0
Conferences and scientific meetings		
MEMTAB – Poster presentation (video 2 min)	2021	1.0
WEON – Oral presentation (video 10 min)	2021	1.0
WEON – Poster presentation	2022	1.0
Teaching		
Bias training for teachers	2021	0.1
Start to Teach	2020	0.6
STARTblock	2020	
STARTblock	2021	
STARTblock	2022	

"A possibilist
Someone who hopes with reason, who sees the progress being made in the world and believe further progress in possible."

— HANS ROSLING, 1948-2017

ABOUT THE AUTHOR

Constanza Lourdes Andaur Navarro was born on October 25th, 1988, in Santiago, Chile. In 2009, she started with her bachelor and master in dentistry at Pontificia Universidad Católica de Chile. She graduated cum laude in 2015. After that, Constanza worked as a dentist in private clinic and deployed a caries preventive program for children in a low-income municipality. In this period, she also obtained a postgraduate certificate in Evidence-based Medicine in Pontificia Universidad Católica de Chile.

In 2017, Constanza received a scholarship from the National Committee for Scientific and technology (Becas Chile) to study Clinical Epidemiology at Erasmus MC in Rotterdam, the Netherlands. Constanza's enthusiasm for research motived her to pursue a PhD in Epidemiology. In 2019, Constanza started working as a PhD student at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, supervised by Prof. Carl Moons, Prof. Lotty Hooft and Dr. Anneke Damen. Constanza participated as fellow for public engagement in the Open Science strategy from Utrecht University.

The results of her PhD research, titled "Quality of machine learning prediction models in healthcare", are presented in this thesis. Constanza currently works as a postdoctoral researcher in Real World Evidence at the department of data science and biostatistics at the Julius Center, UMC Utrecht, the Netherlands.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all people who have contribute and support my PhD journey. First, thanks to my promotors, Carl and Lotty, and my co-promotor, Anneke. Undoubtedly covid times weren't easy for none of us, that's why I appreciate your mentorship, patient, and trust even more. Your guidance and teamwork was fundamental to make this time of my life very enjoyable despite the hard times. Dear Anneke, thank you for leading me and being supportive. I feel truly honored that were you all were part of my PhD supervision team.

Thanks to all the co-authors in this thesis: Paula, Toshi, Steven, Ram, Jie, Mona, Hans, Maarten, Richard, and Gary. I was so lucky to have you all in my projects. You have largely contributed to this thesis but specially to my professional development. Your comments were always so precise that make me rethink why I was doing. Special thanks to Maarten, Richard, and Gary for that. I hope to keep collaborating with you in the future, and finally meet Richard in real life.

Thanks to my paranymphs Paula and Emma, you have been such a great support through this process. Dear Paula, I'm happy to see that our working relationship went beyond screening. I would like to acknowledge your generosity. You were always there to guide me, congratulate me and to share a small chat time to time. My apologies for the excess on WhatsApp stickers. Dear Emma, thank you for your endless encouragement, movies, and for the cups of coffee and wine we shared together.

Thanks to all my colleagues in the Epi-Methods team. You definitely made my work so much fun. Although covid make it more complicate to see each other, I still very much enjoyed our coffie momentjes, lunches, and borreltjes. Special thanks to the PhD ladies: Ilse, Ana, Lieke, Anran, and Julia. Thank you for the writing retreat which lead to this thesis, and I hope to be in all your PhD ceremonies soon. Thanks to my new colleagues in the Real World Evidence team for welcoming and teaching me new things during my post-thesis period. Miriam, thank you for the opportunity to join your team and for the trust you have placed on me. Dear Vjola, thank you for all your support and teaching, I truly admire you and I hope we can continue to work/have fun together. To all, I hope I can keep learning from each of you even further.

Thanks to all my chilean friends in the Netherlands and Belgium. Cabros, que manera de pasarlo bien, gracias por escuchar pacientemente como me quejaba con una chelita en la mano. Paula, gracias por las caminatas, los sustos con los mimos y las infinitas sesiones de salud mental y karaoke que compartimos. Lore and Lucas (y Santi), gracias por acojerme en su casa cada vez que necesito "otro aire" y por el mejor panorama de todos, ver tele. Gracias por los 2kg que subo cada vez que

voy a Leuven. Nicole y Javier, gracias por las sesiones de farándula nacional, los completos, el pelambre, y por querer tanto a mi Chiri. Ine y Marcelo, gracias por recibirme en su casa, la comida gourmet con thermomix, y las infantibles sesiones de terapia alternativa que compartimos. Ven, ninguna weá es imposible, ni una weá.

My dearest Vrimiboat ploeg: Jepke, Monique, Pien, Natasja, Veronique, and Geja. Although we meet because we were slower at learning how to row, it has been an amazing 3 years. Thanks for encouraging me to be out of the house during covid times and my grief. Certainly, the rowing, the cup of coffee, and appeltaart were worth it. Thank you for always stand up to celebrate my achievements. Jullie zijn de aardige vrouwen ter wereld.

Finalmente, quiero agradecer a mi familia en Chile. Gracias mamá, por dejarme partir y nunca dudar que mi camino estaba lejos. Sé que han sido tiempos difíles, pero tu apoyo ha sido esencial para creer en mí y perseverar. Te amo infinito y cada día te admiro aún más. Gracias a mi hermano y los AJ (Crescente y Magnolia), ciertamente su dosis diaria de fotos ha sido una gran motivación para empezar mis días. Espero que pronto puedan venir a visitarme, los amo mucho. A mis tías, tíos, abuelo, y primos, muchas gracias por la preocupación y por recibirme tan bien cuando voy a Chile. A mis amigas Julia, Mariana, Nino, y Camila, gracias infinitas por su apoyo y por las vacaciones en Europa que hemos pasado juntas. Esta tesis está dedicada a mi papá, que aunque no estés aquí y sólo haya silencio, has sido parte fundamental de mi motivación y de ver la vida con otros ojos. Cada parte del diseño está dedicada a ti. Te amo infinito.

Last, but not least my best colleague ever and favorite Dutch. Thank you Chirimoya Alegre, better known as Chiri, my cat. I undoubtedly knew that I need it you in my journey and luckily, you were born the same day I started my PhD and came home soon after. Like a clock you woke me up at 6am and told me to stop working at 18pm while competing many times for my attention against the computer. You were there in my anxiety, stress, and joy. I am excited to see what our next journey would be, you just made my heart bigger. Love you!

Once more, thank you very much to all for your support.