Journal of Clinical Epidemiology 154 (2023) 8−22

Journal of Clinical Epidemiology

REVIEW ARTICLE

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

Constanza L. Andaur Navarro, Doctoral Student[a,b,*], Johanna A.A. Damen, Assistant Professor[a,b], Maarten van Smeden, Associate Professor[a], Toshihiko Takada, Assistant Professor[a], Steven W.J. Nijman, Doctoral Student[a], Paula Dhiman, Research Fellow[c,d], Jie Ma, Medical Statistician[c], Gary S. Collins, Professor[c,d], Ram Bajpai, Research Fellow[e], Richard D. Riley, Professor[e], Karel G.M. Moons, Professor[a,b], Lotty Hooft, Professor[a,b]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[c]*Center for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[d]*NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[e]*Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK*

Accepted 22 November 2022; Published online 25 November 2022

Abstract

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

Methods: 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%

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. 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.

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

Competing interests: There are no conflicts of interest to declare.

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.

Ethical approval: Not required for this work.

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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; 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.

* Corresponding author. Julius Centre for Health Sciences and Primary Care, Universiteitsweg 100, P.O. Box 85500, 3508 GA Utrecht, The Netherlands.

E-mail address: c.l.andaurnavarro@umcutrecht.nl (C.L. Andaur Navarro).

https://doi.org/10.1016/j.jclinepi.2022.11.015

[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]).

Conclusion: 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.

Keywords: Predictive algorithm; Risk prediction; Diagnosis; Prognosis; Development; Validation

1. 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 nonlinear 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, nonlinear associations, and interactions. This model can be later applied in other but related individuals to predict their (yet unknown) outcome. Support vector machines (SVMs), random forests (RFs), and neural networks (NNs) 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 summarize the characteristics on study design, modelling steps, and performance measures reported in studies of prediction models using supervised machine learning.

2. Methods

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

2.1. 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 or validation of at least one 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 nonregression 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 (for example, ordinary least squares or maximum likelihood logistic regression), or penalized regression (for example, 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

What is new?

Key findings

- Design and methodological conduct of studies on clinical prediction models based on machine learning vary substantially.

What this adds to what was known?

- Studies on clinical prediction models based on machine learning suffered from poor methodology and reporting similar to studies using regression approaches.

What is the implication and what should change now?

- Methodologies for model development and validation should be more carefully designed and reported to avoid research waste.
- More attention is needed to missing data, internal validation procedures, and calibration.
- Methodological guidance for studies on prediction models based on machine learning techniques is urgently needed.

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].

2.2. 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.

2.3. 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 (for example, cohort, case-control, randomized trial) and data source (for example, 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 (for example, train-validation-test), method for internal validation (for example., 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 (for example, discrimination and calibration).

Items were recorded by two independent reviewers. One reviewer (CLAN) recorded all items, while 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.

2.4. 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 were 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 resampling methods were used; ''corrected performance'' when studies reported model performance assessed in test dataset and/or using resampling 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

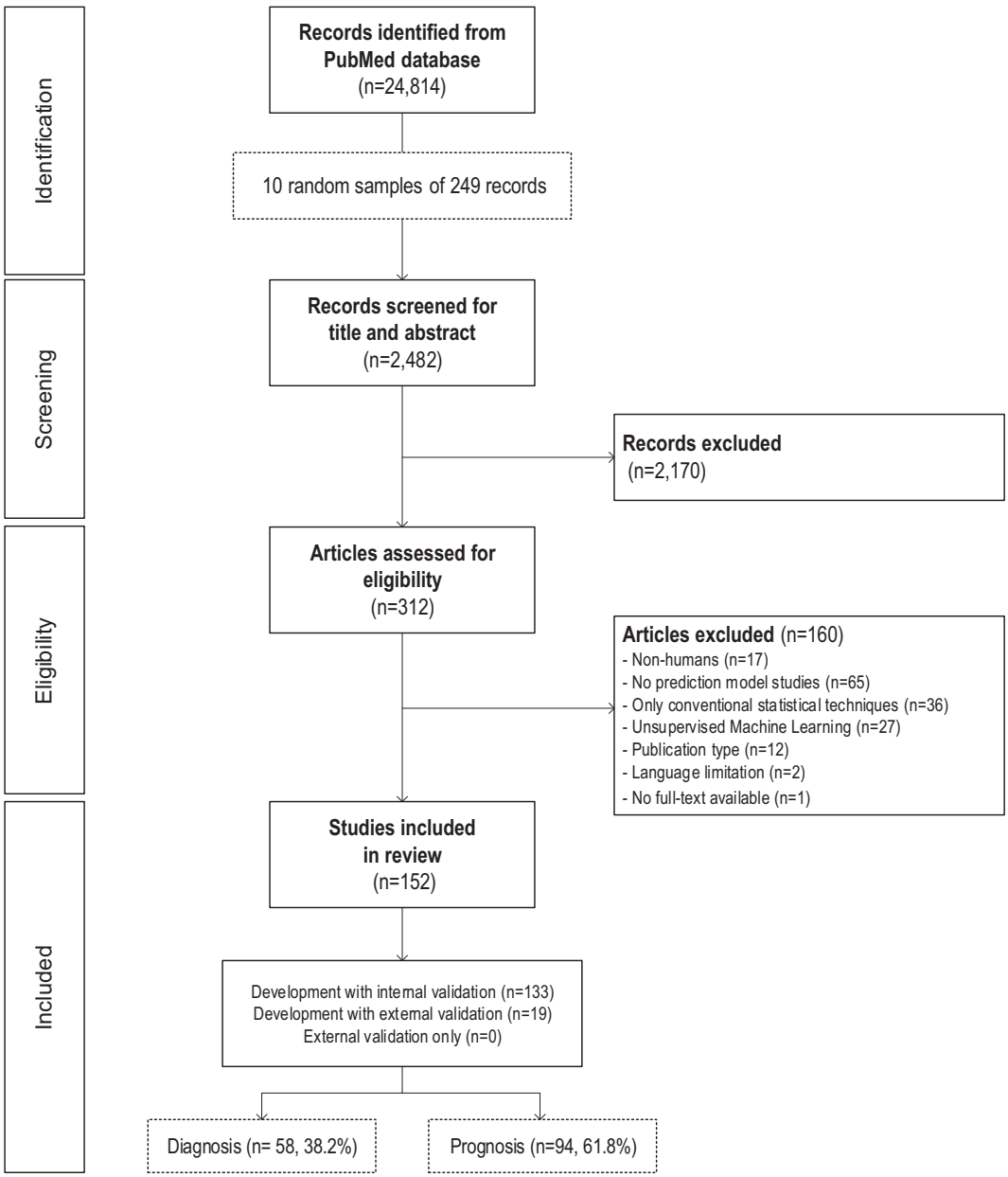

Fig. 1. Flowchart of included studies.

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 (that is, 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).

3. Results

Among the 24,814 articles retrieved, we drew a random sample of 2,482 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% confidence interval (CI) 53.9–69.2]) prognostic and 58 (38.2% [95% CI 30.8–46.1]) diagnostic prediction model studies (Fig. 1). Detailed description of the included articles is provided in Supplemental File 2.

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

Table 1. General characteristics of included studies

Key characteristics	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[a]	
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[b]	20 (13.2) [8.7–19.5]
Mentioning of reporting guidelines[a]	
TRIPOD	8 (5.3) [2.7–10]
STROBE	3 (2.0) [0.7–5.6]
Other[c]	5 (3.3) [1.4–7.5]
None	139 (91.4) [85.9–94.9]
Model availability[a]	
Repository for data	18 (11.8) [7.6–17.9]
Repository for code	13 (8.6) [5.1–14.1]
Model presentation[d]	31 (20.4) [14.8–27.5]
None	121 (79.6) [72.5–85.2]

[a] 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.

[b] 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.

[c] Guidelines for developing and reporting machine learning models in biomedical research (n = 2), STARD (n = 2), BRISQ (n = 1).

[d] This includes simplified scoring rule, chart, nomogram, online calculator, or worked examples.

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 (for example, CART)[b]	46 (8.8) [6.6–11.7]
Random forest[c]	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[d]	34 (6.5) [4.6–9.1]
XGBoost	11 (2.1) [1.1–3.9]
Neural Network (incl. deep learning)[e]	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]

Abbreviations: 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] This includes conditional inference tree (n = 3), optimal tree (n = 1).

[c] This includes Random Survival Forest (n = 2).

[d] This includes lightGBM (n = 1), adaBoost (n = 8), catBoost (n = 1), logitboost (n = 1), RUSBoost (n = 1), and stochastic (n = 1).

[e] Multilayer perceptron, denseNet, convolutional, recurrent, and Bayesian neural networks were extracted as neural networks.

[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).

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

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

Key items	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,c]			
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 51,920	1 [1–3], 1 to 712	1 [1–10], 1 to 51,920
Follow-up (mo)[b]	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 (mo)[b]	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[a]			
Split sample with test set	86 (56.6) [48.6–64.2]	NA	NA
(Random) split	49 (57) [46.4–66.9]		
(Nonrandom) split	9 (10.5) [5.6–18.7]		
Split[d]	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[a]			
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]

[a] 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.
[b] We collected the longest follow-up and longest prediction horizon, both in months.
[c] Data sources also included surveys (n = 2), cross-sectional studies (n = 2).
[d] Unclear whether split sample was performed random or nonrandom.

with a clinical affiliation (n = 85/152, 56% [95% CI 48–63.6]). Other characteristics are shown in Table 1.

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 (that is, AUC).

3.1. 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).

3.2. 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).

3.3. 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).

3.4. 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 4. Predictors in included studies

Key items	Total (n = 152) n (%) [95% CI]
Type of candidate predictors[a]	
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[b]	
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[a,b]	
Linear (no change)	13 (8.6) [5.1–14.1]
Nonlinear (planned)	2 (1.3) [0.4–4.7]
Nonlinear (unplanned)	4 (2.6) [1–6.6]
Categorized (some)	16 (10.5) [6.6–16.4]
Categorized (all)	18 (11.8) [7.6–17.9]
Unclear	104 (68.4) [60.7–75.3]
Categorization of continuous predictors[b]	
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]

[a] 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.
[b] As data preparation.

3.5. Sample size

Studies had a median sample size of 587 participants (IQR 172–6,328). 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).

3.6. Missing values

Missing values were an explicit exclusion criterion of participants in 56 studies (n = 56/152, 36.8% [95% CI

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

Key items	Total (n = 152)	
	n (%)	Median [IQR], range
Initial sample size	93 (61.2)	999 [272–24,522], 8 to 1,093,177
External validation[a]	13 (68.4)	318 [90–682], 19 to 1,113,656
Final sample size	151 (99.3)	587 [172–6,328], 8 to 594,751
Model development	83 (54.6)	641 [226–10,512], 5 to 392,536
Internal validation[b]	83 (54.6)	230 [75–2,892], 2 to 202,215
External validation[a]	18 (94.7)	293 [71–1,688], 19 to 59,738
Initial number of events	10 (6.6)	66 [15–207], 15 to 4,370
External validation[a]	1 (5.3)	107
Final number of events	37 (24.3)	106 [5–-364], 15 to 7,543
Model development	19 (13.2)	156 [47–353], 10 to 5,054
Internal validation[b]	19 (13.2)	35 [26–109], 4 to 2,489
External validation[a]	4 (21.1)	250 [121–990], 107 to 2,834
Number of candidate predictors	119 (78.3)	24 [13–112], 2 to 39,212
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, for example, split sample, cross-validation, bootstrapping.
[c] For model development.

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.

3.7. 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].

3.8. 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.

3.9. 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.

3.10. 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]).

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

	Total (n = 152)	Development only (n = 133)	Development with external validation (n = 19)
Key items	***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 [19–4,209], (1 to 627,180)	224 [16–4,699], (1 to 627,180)	0
Methods of handling missing data[a]			
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

[a] 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.

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).

3.11. 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 (that is, 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]. We also found overall metrics such as classification accuracy (n = 324/522, 62.1% [95% CI 57.8–66.2]) and F1-score (n = 79/522, 15.1% [95% CI 12.2–18.6]).

3.12. Uncertainty quantification

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

3.13. Predictive performance

Most models achieved discriminative ability better than chance (that is, 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 7. Machine learning aspects in the included studies

Key items	Total (n = 152) n (%) [95% CI]
Data preparation[a]	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[b]	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[a]	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[c]	7 (25.9) [13.2–44.7]
Strategy for hyperparameter optimization[a]	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]
Predefined 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]

Abbreviations: CART, classification and regression tree.

[a] 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.

[b] This includes matching, augmentation, noise filtering, merging, splitting, binning.

[c] This includes matching, resampling, class weighting, inverse class probability.

Table 8. Model building of all included studies

Key items	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]
Penalized 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.

3.14. 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 (nonrandom) 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 1,000 iterations and one did not report the number of iterations. For further details see Table 3.

3.15. 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

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

	All extracted models (n = 522)	
	n (%) [95% CI]	
Key items	**DEV**	**VAL**
Calibration[a]		
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[a]		
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[a]		
Decision Curve Analysis	2 (0.4) [0.01–1.5]	0
IDI	1 (0.2) [0.01–1.2]	0
Overall[a]		
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[b]	160 (30.7) [26.8–34.8]	10 (1.9) [1–3.6]
AUC difference	2 (0.4) [0.01–1.5]	0
Accuracy[c]	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

Abbreviations: 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.

[a] 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.

[b] This includes models reporting positive predictive value as precision.

[c] This includes models reporting balance accuracy.

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.

3.16. 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.

4. Discussion

4.1. Principal findings

In this study, we evaluated the study design, data sources, modelling steps, and performance measures in studies

Table 10. Predictive performance of all extracted models[a]

	All extracted models (n = 522)					
		Apparent performance		Corrected performance[b]		Externally validated performance
Key items	Reported, n (%)	Median [IQR], range	Reported, n (%)	Median [IQR], range	Reported, n (%)	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 to −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

[a] 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 prespecified.

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

on clinical prediction models using machine learning. 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 subsets 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]. Several studies also applied regression-based methods (LR in particular) as benchmark to compare against the predictive performance of machine learning-based models.

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 ML remains relatively low. Furthermore, splitting datasets persists as a method for internal validation (that is, 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, 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 (that is, data preprocessing) 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 (that is, 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.

4.2. Comparison to previous studies

Although regression methods were not our focus (as we did not define them to be machine learning methods), 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 5 years [10]. Methods to correct for class imbalance in datasets concerning EHR increased from 7% to 13% [41].

4.3. 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 3 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 (that is, 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 ML-based models as well.

4.4. 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 (for example, 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 (that is, 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].

5. Conclusions

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.

Acknowledgments

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.

Supplementary data

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

References

[1] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:1317−20.
[2] van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142−5.
[3] Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3(1):126.
[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):281.
[5] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383:101−4.
[6] Jong YD, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, Diepen MV. 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;26:939−47.
[7] Damen JAAG, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353: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.
[9] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375: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.
[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):30.
[12] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022;22:1−16.
[13] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
[14] Andaur Navarro CL, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. BMJ Open 2020;10(11):1−6.
[15] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744.
[16] Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, 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.
[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;162:55.
[18] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323.
[19] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321−57.
[20] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern Syst Hum 2010;40(1):185−97.
[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, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for investigators. Eur Urol 2018;74(6):796−804.
[23] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022;22:12.
[24] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317−8.
[25] Hastie T, Tibshirani R, Friedman J. In: The Elements of Statstical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer-Verlag; 2009.
[26] Breiman L. Random forests. California; Mach Learn 2001;45(1):5−32.
[27] Scholkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Massachusetts: MIT Press; 2001.
[28] Ploeg TVD, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137.
[29] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans RM, Vergouwe Y, Habbema J. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2007;42:774−81.
[30] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245−7.
[31] Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):1−7.
[32] Goorbergh RVD, 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 Inform Assoc 2022;29(9):1525−34.

[33] Pajouheshnia R, Damen JAAG, Groenwold RHH, Moons KGM, Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. Diagn Progn Res 2017;1(1):1−10.
[34] Nijman SWJ, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs J, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol 2022;142:218−29.
[35] Dhiman P, Ma J, Andaur Navarro C, Speich B, Bullock G, Damen JA, 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.
[36] Heus P, Reitsma JB, Collins GS, Damen JAAG, Scholten RJPM, Altman DG, et al. Transparent reporting of multivariable prediction models in journal and conference abstracts: TRIPOD for abstracts. Ann Intern Med 2022;173(1):42−8.
[37] Collins GS, De Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.
[38] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.
[39] Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, 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.
[40] Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1−12.
[41] Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc 2022;29(5):983−9.
[42] Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, Heus P, Hooft L, Moons KGM, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. BMJ Open 2020;10(9):e041537.
[43] Kim DW, Jang HY, Ko Y, Ko Y, Son JH, Kim PH, 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;15:1−10.
[44] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008.
[45] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.
[46] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51−8.
[47] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1−33.
[48] Damen JAAG, Debray TPA, Pajouheshnia R, Reitsma JB, Scholten RJPM, Moons KGM, 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.
[49] Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577−9.
[50] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453−73.
[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 Inform Assoc 2019;26(12):1651−4.