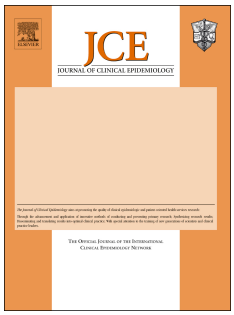
Accepted Manuscript



A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

christodoulou Evangelia, M.A. Jie, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, Ben van Calster

PII: S0895-4356(18)31081-3

DOI: https://doi.org/10.1016/j.jclinepi.2019.02.004

Reference: JCE 9825

To appear in: *Journal of Clinical Epidemiology*

Received Date: 5 December 2018

Revised Date: 18 January 2019

Accepted Date: 5 February 2019

Please cite this article as: Evangelia c, Jie M, 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, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.02.004.

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia CHRISTODOULOU[a], Jie MA[b], Gary S. COLLINS[b,c],

Ewout W. STEYERBERG[d], Jan Y. VERBAKEL[a,e,f], Ben VAN CALSTER[a,d]

a KU Leuven, Department of Development & Regeneration, Herestraat 49 box 805, 3000 Leuven, Belgium; b Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom; c Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom; d Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA Leiden, the Netherlands; e KU Leuven, Department of Public Health & Primary Care, Kapucijnenvoer 33J box 7001, 3000 Leuven, Belgium; f Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom

Word count: 3199

Corresponding author:

Ben Van Calster

KU Leuven, Department of Development and Regeneration

Herestraat 49 box 805

3000 Leuven

Belgium

+32 16 377788

ben.vancalster@kuleuven.be

Abstract

Objective: To compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling. Study design and setting: We conducted a Medline literature search (1/2016 to 8/2017), and extracted comparisons between LR and ML models for binary outcomes. Results: We included 71 out of 927 studies. The median sample size was 1250 (range 72-3,994,872), with 19 predictors considered (range 5-563) and 8 events per predictor (range 0.3-6,697). The most common ML methods were classification trees (30 studies), random forests (28), artificial neural networks (26), and support vector machines (24). Sixty-four (90%) studies used the area under the receiver operating characteristic curve (AUC) to assess discrimination. Calibration was not addressed in 56 (79%) studies. We identified 282 comparisons between a LR and ML model (AUC range, 0.52-0.99). For 145 comparisons at low risk of bias, the difference in logit(AUC) between LR and ML was 0.00 (95% confidence interval, -0.18 to 0.18). For 137 comparisons at high risk of bias, logit(AUC) was 0.34 (0.20 to 0.47) higher for ML. Conclusions: We found no evidence of superior performance of ML over LR for clinical prediction modeling, but improvements in methodology and reporting are needed for studies that compare modeling algorithms.

Key words:

Clinical prediction models; logistic regression; machine learning; AUC; calibration; reporting

What is new

Key Findings

- Studies comparing clinical prediction models based on logistic regression and machine learning algorithms suffered from poor methodology and reporting, in particular with respect to the validation procedure
- The studies rarely assessed whether risk predictions are reliable (calibration), but the area under the ROC curve (AUC) was almost always provided
- The AUC of logistic regression and machine learning models for clinical risk prediction were similar when fair comparisons were made; ML performance was higher in comparisons that were at high risk of bias

What this adds to what is known

- Machine learning models do not automatically lead to improved performance over traditional methods
- Model validation procedures are often not sound or not well reported, which hampers a fair model comparison in real world case studies

What should change now

- More attention for calibration performance of regression and machine learning models is urgently needed
- Model development and validation methodologies should be more carefully designed and reported to avoid research waste
- Research should focus more on identifying which algorithms have optimal performance for different types of prediction problems

1. Introduction

Clinical risk prediction models are ubiquitous in many medical domains. These models aim to predict a clinically relevant outcome using person-level information. The traditional approach to develop these models involves the use of regression models, for example logistic regression (LR) to predict disease presence (diagnosis) or disease outcomes (prognosis) [1]. Machine learning (ML) algorithms are gaining in popularity as an alternative approach for prediction and classification problems. ML methods include artificial neural networks, support vector machines, and random forests [2]. Whilst ML methods have been sporadically used for clinical prediction for some time [3,4], the growing availability of increasingly large, voluminous and rich datasets such as electronic health records data has reignited interest in exploiting these methods [5–7].

Definitions of what constitutes ML and the differences with statistical modeling have been discussed at length in the literature [8], yet the distinction is not clear-cut [9]. The seminal reference on this issue is Breiman's review of the 'two cultures' [8]. Breiman contrasts theory-based models such as regression with empirical algorithms such as decision trees, artificial neural networks, support vector machines, or random forests. A useful definition of ML is that it focuses on models that directly and automatically learn from data [10]. In contrast, regression models are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification. For example, ML performs modeling more automatically than regression regarding the inclusion of nonlinear associations and interaction terms [11]. To do so, ML algorithms are often highly flexible algorithms that require penalization to avoid overfitting, i.e. that predictions generalize poorly to new data [12]. Some researchers describe the distinction between statistical modeling and machine learning as a continuum [5]. Other researchers label any method that deviates from basic regression models as ML [13], such as penalized regression (e.g., LASSO, elastic net) or generalized additive models (GAM). We note that these methods do not belong to ML using the 'automatic learning from data' definition, and did not classify these as ML in this study.

Due to its flexibility, ML is claimed to have better performance over traditional statistical modeling, and to better handle a larger number of potential predictors [5–7,12,14–16]. However, recent research suggested that ML requires more data than LR, which contradicts the above claim [17]. Further, ML models are typically assessed in terms of discrimination performance (e.g. accuracy, area under the ROC curve), whilst the reliability of risk predictions (calibration) is often not assessed [18]. The claim of improved performance in clinical prediction is therefore not established.

The primary objective of this study was to compare the performance of LR with ML algorithms for the development of diagnostic or prognostic clinical prediction models for binary outcomes based on clinical data. Secondary objectives were to describe the characteristics of the studies, the type of ML algorithms that were used, the validation process , the modeling aspects of LR and ML, reporting quality, and risk of bias for comparing performance between regression and ML [19].

2. Materials and methods

The study was registered with PROSPERO (CRD42018068587). We followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement.

2.1. Identification of studies

We searched MEDLINE on August 8th, 2017. We performed a sensitive literature search by using a broad working definition of ML (see the search string in Appendix A). We focused on articles published since 2016 (between January 1st, 2016 and August 8th, 2017) in order to base our analysis on recent studies.

2.2. Selection of studies

All abstracts were independently screened by two reviewers (EC, JM), conflicts were resolved by a third reviewer (BVC or JYV). The full text of selected abstracts were independently assessed for eligibility by three reviewers (EC, JM, BVC), and conflicts were resolved by consensus.

2.3. Inclusion and exclusion criteria

Studies were eligible if the article:

- Described the development of a diagnostic or prognostic prediction model for individualized prediction using two or more predictors
- Compared prediction models based on LR and ML algorithms

Studies were excluded if:

- A new modeling approach was introduced (hence a methodological focus) [20,21]
- Models were developed for non-humans
- The models made predictions for individual images or signals rather than participants
- Models were developed based on high-dimensional data modalities
- The primary interest was assessing risk factors rather than prediction modeling
- They were reviews of the literature
- Studies for which we were unable to obtain the full text

2.4. Data extraction and risk of bias

We focused on methodological issues of model development, and aspects that compromise the comparison of model performance between LR and ML algorithms. The list of extraction items was based on the CHARMS checklist and the QUADAS risk of bias tool and refined after extensive discussion among the authors [9,22]. The extracted items included general study characteristics, applied algorithms and their characteristics, data-driven variable selection, and model performance (Table A.1, Appendix B) [1,2,13,23–25].

From each article, we defined five signaling items to indicate potential bias. We elaborate on these items in Table A.2:

(1) unclear or biased validation of model performance
(2) difference in whether data-driven variable selection was performed (yes/no) prior to applying LR and ML algorithms
(3) difference in handling of continuous variables prior to applying LR and ML algorithms
(4) different predictors considered for LR and ML algorithms
(5) whether corrections for imbalanced outcomes where used only for LR or only for ML algorithms

Most papers developed several LR and/or ML models. These papers contain multiple comparisons between LR and ML algorithms, and we evaluated the signaling items per comparison. Each bias item was scored as no (not present), unclear, or yes (present). We considered a comparison at low risk of bias if the answer was ‘no’ for all five signaling items. If the answer was ‘unclear’ or ‘yes’ for at least one item, we assumed high risk of bias. We also summarized the signaling items for each study as a whole, by noting the worst case (no, unclear, yes) across all comparisons in the study.

2.5. Data analysis

We used descriptive statistics to summarize results. Within each paper, we identified all comparisons between LR and ML methods (see Appendix C). We identified multiple comparisons within the same paper as a result of implementing multiple ML algorithms, developing models for more than one outcome, developing models based on different predictor sets (e.g. once with and once without lab measurements), or developing models for several subgroups separately. Although the search string contrasted standard LR with penalized methods, we consider penalized LR (e.g. lasso, ridge, elastic net) to be LR rather than ML. Some papers contrasted LR with algorithms that are traditional statistical methods, such as discriminant analysis, Poisson regression, generalized estimating equations, and GAM. We did not classify these algorithms as ML. We compared the LR and ML models using the difference in the area under the ROC curve (AUC). We used AUC values in the following order of priority: external validation, internal validation, training data (no validation). Based on the extracted data, we classified ML algorithms into five broad groups: single classification trees, random forests, artificial neural networks, support vector machines, and other algorithms. We analyzed AUC differences for all comparisons, and with stratification for risk of bias. We performed a meta-regression of the difference between logit-transformed AUC using a random effect model to take clustering of comparisons by paper into account, and weighted by the square root of the validation sample size. Logit(AUC) was used to circumvent the bounded nature of the AUC [26].

3. Results

Our search identified 927 articles published since between 1/2016 and 8/2017, of which 802 studies were excluded based on title or abstract (Figure 1). 54 studies were excluded during full text screening. Seventy-one studies met inclusion criteria and came from a wide variety of clinical domains, with oncology and cardiovascular medicine as the most common (Table A.3-4) [27–97]

3.1. General study characteristics

The most common designs were cohort (n=39, 55%) and cross-sectional (n=18, 25%) (Table A.5). Overall, 50 studies (70%) focused on prognostic outcomes, 19 (27%) on diagnostic outcomes, and

two on both. The majority of studies (n=64, 90%) used existing data, and 27 (38%) used hospital-based multicenter data. The median number of centers was 5 (range 2-1,137) (Table A.6).

The median total sample size was 1,250 (range 72-3,994,872), median number of considered predictors was 19 (range, 5-563). 102 outcomes were considered in the 71 articles, the median event rate was 0.18 (range 0.002-0.50). We defined the number of events as the number of participants in the smallest outcome category. Nine articles developed models to predict more than one outcome. The median number of events per predictor in the training data was 8 (range 0.3-6,697) (Figure A.1).

Information on handling of missing data was lacking or unclear in 32 studies (45%) (Tables A.7-8). Sixteen studies (23%) performed a complete case analysis, 14 (20%) relied on ad hoc methods (mean imputation, missing indicator methods, variable deletion), and nine (11%) used single or multiple stochastic imputation, albeit poorly documented.

3.2. Overview of algorithms

Sixty-four studies used standard (maximum likelihood) LR, of which nine also used penalized LR (lasso, ridge, or elastic net) and one also used boosted LR (Table 1 and A.9). Six studies used only penalized LR, and 1 study used only bagged LR (classified as ML).

Forty-three studies used more than one ML algorithm. The most popular algorithms were classification trees (n=30, 42%), random forests (28, 39%), artificial neural networks (26, 37%), and support vector machines (24, 34%). Of 26 studies using artificial neural networks, 22 used 1 hidden layer, 3 used multiple hidden layers, and for 1 study this was unclear (Table A.9). When support vector machines were used, the Gaussian ('radial basis function') kernel was most often used (n=10).

3.3. Model development

Irrespective of algorithm (LR vs ML), 14 studies (20%) were not clear about how continuous variables were handled during model development (Table A.10). Discretization (into two or more categories) was used for some or all algorithms in 18 studies (25%), whereas continuous modeling was observed in 37 studies (52%), although this was often not explicitly stated. Data-driven variable selection before any model fitting was reported for 41 studies (58%).

Specifically for LR, handling of continuous predictors was unclear in 47/71 studies (66%). In 33/47, some or all predictors were kept continuous but it was unclear whether nonlinear associations were examined. For one study, it was clear that continuous variables were assumed to have linear

associations with the outcome. Discretization of some or all continuous predictors was carried out in 20 studies (28%), whereas nonlinearity was investigated in seven studies (10%). Sixty-three studies (89%) did not explicitly mention whether interaction effects were considered for LR models. The remaining eight studies were often unclear on the approach for interaction terms (Table A.11).

Penalized LR as well as many ML algorithms contain hyperparameters that determine the complexity/flexibility of the model. For the most commonly used algorithms, we observed that the approach for determining the hyperparameters was not clear in at least half of the studies (Table A.12). It was either unclear whether hyperparameters were tuned or default settings were used, or hyperparameters were said to be tuned but the tuning procedure was not clear.

3.4. Model validation

29 studies (41%) used a single random split of the data into train-test or train-validate-test parts (Table 2). Twenty-five studies used resampling (35%; 15 used cross-validation, 9 used repeated random splitting, and 1 used bootstrapping). Seven studies (10%) used some form of external validation, most commonly using a chronological split of data into training and test parts. Seven studies (10%) did not validate performance, and for three studies (4%) the approach depended on the algorithm. Importantly, in 48 studies (68%) we observed unclear reporting or potential biases in validation procedures for one or more algorithms. Common reasons were that hyperparameters were tuned or variable selection was done on all data (or this was not clearly specified), or that not all modeling steps were repeated when resampling was used for validation (Table A.13).

The AUC was the most commonly reported performance measure (64 studies, 90%), followed by sensitivity (45, 63%) and specificity (43, 61%) (Table A.14). Calibration performance was not discussed in 56 studies (79%) (Table A.15). Most commonly, calibration was addressed using grouped calibration plots (n=7). Only 1 study (1%) evaluated performance in terms of clinical utility using decision curve analysis.

In 21 studies, methods were applied to address outcome imbalance, i.e. an event rate far from 50% (Table A.16, see Discussion).

3.5. Comparison between performance of LR and ML

The most problematic risk of bias item was an unclear/biased validation procedure (Figure 2, Table A.17).

We identified 282 comparisons between standard/penalized LR (AUC 0.52-0.97) and ML models (AUC 0.58-0.99) in 58 papers. Of the remaining 13 papers, 7 did not report AUCs, 3 reported AUCs for some algorithms only, 1 reported AUCs to one decimal, 1 only applied standard and penalized LR, and 1 only applied bagged LR and random forests. 145 comparisons (51%) were labeled as having low risk of bias. The logit(AUC) was on average 0.25 higher for ML vs LR (95% CI 0.12 to 0.38) (Figures 3-4). However, the logit(AUC) difference was on average 0.00 (-0.18 to 0.18) for comparisons with low risk of bias, and 0.34 higher (0.20 to 0.47) for comparisons with high risk of bias. Trees uniformly had worse performance than other ML algorithms. Otherwise, results for different ML algorithms were similar.

Finally, Table A.18 reports on additional findings on methodology and reporting that could not be discussed in the main text due to space limitations.

4. Discussion

Our systematic review of studies that compare clinical prediction models using LR and ML yielded the following key findings. Reporting of methodology and findings was very often incomplete and unclear, model validation procedures still often were poor. Calibration of risk predictions was seldom examined, and AUC performance of LR and ML was on average no different when comparisons had low risk of bias. The latter finding is in line with the claim that traditional approaches often perform remarkably well [21].

Our findings lead to the following recommendations (Table A.19). First, fully report on all modeling steps and analyses in sufficient detail to maximize transparency and reproducibility. We recommend to adhere to the TRIPOD guidelines [19]. If necessary, include detailed descriptions as supplementary material. For complex procedures, a comprehensive flowchart of the development and validation procedures can be insightful - some studies provided this [53]. Second, if model validation is based on resampling, the model development should be based on all available data, and the resampling should then include all modelling steps that were used to build the model in order to estimate performance. Model development on all data was often not done. In addition, provide all information on these models to allow independent validation. Third, report training and test performance. The difference between these results is informative. Fourth, evaluate model performance in terms of calibration (whether risk estimates are accurate) and clinical utility for

decision making [18]. Preferably, calibration should be investigated using calibration curves, whereas the Hosmer-Lemeshow test should be avoided [18,98,99]. Clinical utility can be assessed using decision curve analysis, which is increasingly used in medical applications [100].

We found several differences between the ML and statistical literature. In the ML literature, calibration often refers to the transformation of non-probabilistic model outcomes into probabilities [101]. In this paper, calibration refers to the evaluation of the reliability of probabilistic (risk) estimates [18]. A transformation of model outcomes into probabilities is part of model development. Further, the ML literature has paid attention to the utility of models. For example cost curves are very similar to decision curve analysis [102]. Finally, the issue of class imbalance is common in the ML literature [13]. This is motivated by a dominant focus on classification and overall accuracy based on a 50% risk cut-off. However, adjusting class imbalance distorts prevalence and yields inadequate risk predictions. This is not acceptable for clinical risk prediction. In particular, downsampling is inefficient because it reduces sample size. Recent research clearly indicated that this increases the risk of overfitting [103].

The comparison of AUC performance between LR and ML depends on how one defines risk of bias and ML. We used five signaling items to consider comparisons as at low or high risk of bias. These items did not address whether LR models were penalized or included nonlinear and/or interaction effects. Regression is sometimes presented as a method that simply assumes linearity and additivity [7,104]. In comparison studies it is usually implemented as such, for example in two recent benchmark studies using dataset repositories [105,106]. Some criticize that assuming linearity and additivity will reduce the performance of regression, although this may depend on sample size. Regarding the definition of ML, we used a broad approach: we focused on alternative algorithms for LR, hereby only excluding classical statistical algorithms (we also excluded GAMs, although some may see this as an ML method). The rationale is that LR has been the standard method for clinical prediction, and more modern approaches are often discussed in relation to LR [6,7,14-17,104,107].

Future research should focus more on delineating the type of predictive problems in which various algorithms have maximal value. For example, the signal-to-noise ratio may be an important aspect in determining how successful ML will be [2,21,107]. ML tends to work well for problems with a strong signal-to-noise ratio [108], e.g. handwriting recognition, gaming, or electric load forecasting. Clinical prediction problems often have a poor signal-to-noise ratio [107] .

A limitation of our study is that it does not investigate which factors influence the difference in performance (e.g. sample size, number of predictors, hyperparameter tuning). We feel that such a study would be relevant, but should be performed by comparing different scenarios on the same

datasets to avoid confounding [106]. Another limitation is that many studies had a fairly limited number of events per considered predictor, a common problem despite repeated warnings [1,17,99,103,109]. This issue urgently needs better consideration. Some researchers claim that ML will not outperform LR when only a limited set of pre-specified predictors is considered, and that the advantage of ML lies in better handling a huge amount of predictors [3,7,12,15,16,104]. Unfortunately, all 23 comparisons that we identified from the seven included studies with >100 predictors were at high risk of bias. Nevertheless, their median AUC difference was -0.005. In contradiction with the above claim, recent research suggests that ML requires more data than LR [17]. A final limitation is that conducting a decent and detailed systematic review on this broad topic was time-consuming. In the meantime, new studies will have been published. Although there is the potential that methodology and reporting has improved, such improvements are slow even when longer time periods are considered [110–112].

In conclusion, evidence is lacking to support the claim that clinical prediction models based on ML lead to better AUCs than clinical prediction models based on LR. Reporting of papers that compare both types of algorithms needs to improve. Correct validation procedures are needed [113], with assessment of calibration and clinical utility in addition to discrimination, to define situations where modern methods have advantages over traditional approaches.

Acknowledgements

Funding source: This work was supported by by the Research Foundation – Flanders (FWO) [grant G0B4716N]; Internal Funds KU Leuven [grant C24/15/037]; Cancer Research UK [grant 5529/A16895]; the NIHR Biomedical Research Centre, Oxford, UK. The funding sources had no role in the conception, design, data collection, analysis, or reporting of this study.

Authors' contributions: E.C. was involved in the conception of the study, data collection, data analysis and interpretation, drafting of the article, and gave her final approval of the version to be published. J.M. was involved in data collection, critical revision of the article, and gave her final approval of the version to be published. G.S.C. was involved in the conception of the study, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. E.W.S. was involved in the conception of the study, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. J.Y.V. was involved in the conception of the study, data collection, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. B.V.C. was involved in the conception of the study, data collection, data analysis and interpretation, drafting of the article, and gave his final approval of the version to be published.

References

[1] Steyerberg EW. Clinical Prediction Models. New York: Springer; 2009. doi:10.1007/978-0-387-77244-8_17.

[2] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (2nd ed). New York: Springer; 2009. doi:https://doi.org/10.1007/978-0-387-84858-7.

[3] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.

[4] Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: A systematic review. Neural Netw 2006;19:408–15. doi:10.1016/j.neunet.2005.10.007.

[5] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317–8.

[6] Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. N Engl J Med 2017;376:2507–9. doi:10.1056/NEJMp1702071.

[7] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. Eur Heart J 2017;38:1805–14. doi:10.1093/eurheartj/ehw302.

[8] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci 2001;16:199–231. doi:10.1214/ss/1009213726.

[9] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11:e1001744.

[10] Mitchell TM. Machine learning. New York: McGraw Hill; 1997.

[11] Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biom J 2014;56:588–93. doi:10.1002/bimj.201300226.

[12] Deo RC, Nallamothu BK. Learning about Machine Learning: The Promise and Pitfalls of Big Data and the Electronic Health Record. Circ Cardiovasc Qual Outcomes 2016;9:618–20. doi:10.1161/CIRCOUTCOMES.116.003308.

[13] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2008;21:1263–84.

[14] Pochet NLMM, Suykens JAK. Support vector machines versus logistic regression: Improving

prospective performance in clinical decision-making. Ultrasound Obstet Gynecol 2006;27:607–8. doi:10.1002/uog.2791.

[15] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. Npj Digit Med 2018;1:1–10. doi:10.1038/s41746-018-0029-1.

[16] 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:e323. doi:10.2196/jmir.5870.

[17] 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;14:137.

[18] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76.

[19] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol 2015;68:134-43.

[20] Boulesteix AL, Lauer S, Eugster MJA. A Plea for Neutral Comparison Studies in Computational Sciences. PLoS One 2013;8:e61562. doi:10.1371/journal.pone.0061562.

[21] Hand DJ. Classifier technology and the illusion of progress. Stat Sci 2006;1:1–14.

[22] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.

[23] Probst P, Bischl B, Boulesteix A-L. Tunability: Importance of hyperparameters of machine learning algorithms. ArXiv Prepr ArXiv180209596 2018.

[24] Collins GS, Ogundimu EO, Cook JA, Manach Y Le, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med 2016;35:4124–35.

[25] Steyerberg EW, Harrell Jr FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.

[26] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003.

[27] Adavi M, Salehi M, Roudbari M. Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes. Med J Islam Repub Iran 2016;30:2–6.

[28] Anderson AE, Kerr WT, Thames A, Li T, Xiao J, Cohen MS. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J Biomed Inform 2016;60:162–8. doi:10.1016/j.jbi.2015.12.006.

[29] Habibi Z, Ertiaei A, Nikdad MS, Mirmohseni AS, Afarideh M, Heidari V, et al. Predicting ventriculoperitoneal shunt infection in children with hydrocephalus using artificial neural network. Childs Nerv Syst 2016;32:2143–51. doi:10.1007/s00381-016-3248-2.

[30] Ichikawa D, Saito T, Ujita W, Oyama H. How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. J Biomed Inform 2016;64:20–4. doi:10.1016/j.jbi.2016.09.012.

[31] Jahani M, Mahdavi M. Comparison of predictive models for the early diagnosis of diabetes. Healthc Inform Res 2016;22:95–100. doi:10.4258/hir.2016.22.2.95.

[32] Kabeshova A, Launay CP, Gromov VA, Fantino B, Levinoff EJ, Allali G, et al. Falling in the elderly: Do statistical models matter for performance criteria of fall prediction? Results from two large population-based studies. Eur J Intern Med 2016;27:48–56. doi:10.1016/j.ejim.2015.11.019.

[33] Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 2016;16:39. doi:10.1186/s12911-016-0277-4.

[34] Kulkarni P, Smith LD, Woeltje KF. Assessing risk of hospital readmissions for improving medical practice. Health Care Manag Sci 2016;19:291–9. doi:10.1007/s10729-015-9323-5.

[35] Lu T, Hu YH, Tsai CF, Liu SP, Chen PL. Applying machine learning techniques to the identification of late-onset hypogonadism in elderly men. Springerplus 2016;5:729. doi:10.1186/s40064-016-2531-8.

[36] Mahajan S, Burman P, Hogarth M. Analyzing 30-day readmission rate for heart failure using different predictive models. Stud Health Technol Inform 2016;225:143–7. doi:10.3233/978-1-

61499-658-3-143.

[37] Malik S, Khadgawat R, Anand S, Gupta S. Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. Springerplus 2016;5 :701. doi:10.1186/s40064-016-2339-6.

[38] Matis GK, Chrysou OI, Silva D, Karanikas MA, Baltsavias G, Lyratzopoulos N, et al. Prediction of lumbar disc herniation patients' satisfaction with the aid of an artificial neural network. Turk Neurosurg 2016;26:253–9. doi:10.5137/1019-5149.JTN.8492-13.0.

[39] Belliveau T, Jette AM, Seetharama S, Axt J, Rosenblum D, Larose D, et al. Developing Artificial Neural Network Models to Predict Functioning One Year After Traumatic Spinal Cord Injury. Arch Phys Med Rehabil 2016;97:1663–1668.e3. doi:10.1016/j.apmr.2016.04.014.

[40] Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. Circ Cardiovasc Qual Outcomes 2016;9:629–40. doi:10.1161/CIRCOUTCOMES.116.003039.

[41] Nakas CT, Schütz N, Werners M, Leichtle ABL. Accuracy and Calibration of Computational Approaches for Inpatient Mortality Predictive Modeling. PLoS One 2016;11: e0159046. doi:10.1371/journal.pone.0159046.

[42] Ratliff JK, Balise R, Veeravagu A, Cole TS, Cheng I, Olshen RA, et al. Predicting occurrence of spine surgery complications using big data modeling of an administrative claims database. J Bone Joint Surg Am 2016;98:824–34. doi:10.2106/JBJS.15.00301.

[43] Rau HH, Hsu CY, Lin YA, Atique S, Fuad A, Wei LM, et al. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Comput Methods Programs Biomed 2016;125:58–65. doi:10.1016/j.cmpb.2015.11.009.

[44] Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J Vasc Surg 2016;64:1515–1522.e3. doi:10.1016/j.jvs.2016.04.026.

[45] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med 2016;23:269–78. doi:10.1111/acem.12876.

[46] Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PLoS One 2016;11: e0155705.

doi:10.1371/journal.pone.0155705.

[47] Tong L, Erdmann C, Daldalian M, Li J, Esposito T. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. BMC Med Res Methodol 2016;16: 26. doi:10.1186/s12874-016-0128-0.

[48] van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. J Clin Epidemiol 2016;78:83–9. doi:10.1016/j.jclinepi.2016.03.002.

[49] Wang HY, Hsieh CH, Wen CN, Wen YH, Chen CH, Lu JJ. Cancers screening in an asymptomatic population by using multiple tumour markers. PLoS One 2016;11: e0158285. doi:10.1371/journal.pone.0158285.

[50] Berchialla P, Scarinzi C, Snidero S, Gregori D, Lawson AB, Lee D, et al. Comparing models for quantitative risk assessment: An application to the European Registry of foreign body injuries in children. Stat Methods Med Res 2016;25:1244–59. doi:10.1177/0962280213476167.

[51] Wang Z, Wen X, Lu Y, Yao Y, Zhao H. Exploiting machine learning for predicting skeletal-related events in cancer patients with bone metastases. Oncotarget 2016;7:12612–22. doi:10.1063/1.1781034.

[52] Wu HY, Gong CSA, Lin SP, Chang KY, Tsou MY, Ting CK. Predicting postoperative vomiting among orthopedic patients receiving patient-controlled epidural analgesia using SVM and LR. Sci Rep 2016;6:1–7. doi:10.1038/srep27041.

[53] Yahya N, Ebert MA, Bulsara M, House MJ, Kennedy A, Joseph DJ, et al. Statistical-learning strategies generate only modestly performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: A comparison of conventional and machine-learning methods. Med Phys 2016;43:2040. doi:10.1118/1.4944738.

[54] Zhang Y-D, Wang J, Wu C-J, Bao M-L, Li H, Wang X-N, et al. An imaging-based approach predicts clinical outcomes in prostate cancer through a novel support vector machine classification. Oncotarget 2016;7:78140. doi:10.18632/oncotarget.11293.

[55] Zhou Z, Folkert M, Cannon N, Iyengar P, Westover K, Zhang Y, et al. Predicting distant failure in early stage NSCLC treated with SBRT using clinical parametersPredicting distant failure in lung SBRT. Radiother Oncol 2016;119:501–4. doi:10.1016/j.radonc.2016.04.029.

[56] Acion L, Kelmansky D, Laan MD Van, Sahker E, Jones DS, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One 2017;12:

e0175383. doi:10.1371/journal.pone.0175383.

[57] Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS One 2017;12: e0179805. doi:10.1371/journal.pone.0179805.

[58] Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis. PLoS One 2017;12:e0169772. doi:10.1371/journal.pone.0169772.

[59] Amini P, Maroufizadeh S, Samani RO, Hamidi O, Sepidarkish M. Prevalence and Determinants of Preterm Birth in Tehran, Iran: A Comparison between Logistic Regression and Decision Tree Methods. Osong Public Heal Res Perspect 2017;8:195–200. doi:10.24171/j.phrp.2017.8.3.06.

[60] Asaoka R, Hirasawa K, Iwase A, Fujino Y, Murata H, Shoji N, et al. Validating the Usefulness of the "Random Forests" Classifier to Diagnose Early Glaucoma With Optical Coherence Tomography. Am J Ophthalmol 2017;174:95–103. doi:10.1016/j.ajo.2016.11.001.

[61] Berikol GB, Yildiz O, Özcan T. Diagnosis of Acute Coronary Syndrome with a Support Vector Machine. J Med Syst 2016;40:84. doi:10.1007/s10916-016-0432-6.

[62] Batterham M, Neale E, Martin A, Tapsell L. Data mining: Potential applications in research on nutrition and health. Nutr Diet 2017;74:3–10. doi:10.1111/1747-0080.12337.

[63] Batterham M, Tapsell L, Charlton K, O'Shea J, Thorne R. Using data mining to predict success in a weight loss trial. J Hum Nutr Diet 2017;30:471–8. doi:10.1111/jhn.12448.

[64] Cheng FW, Gao X, Bao L, Mitchell DC, Wood C, Sliwinski MJ, et al. Obesity as a risk factor for developing functional limitation among older adults: A conditional inference tree analysis. Obesity 2017;25:1263–9. doi:10.1002/oby.21861.

[65] Chiriac AM, Wang Y, Schrijvers R, Bousquet PJ, Mura T, Molinari N, et al. Designing Predictive Models for Beta-Lactam Allergy Using the Drug Allergy and Hypersensitivity Database. J Allergy Clin Immunol Pract 2018;6:139–148.e2. doi:10.1016/j.jaip.2017.04.045.

[66] Dean JA, Welsh LC, Wong KH, Aleksic A, Dunne E, Islam MR, et al. Normal Tissue Complication Probability (NTCP) Modelling of Severe Acute Mucositis using a Novel Oral Mucosal Surface Organ at Risk. Clin Oncol 2017;29:263–73. doi:10.1016/j.clon.2016.12.001.

[67] Deng X. Predicting the Risk for Hospital-Acquired Pressure Ulcers in Critical Care Patients. Crit Care Nurse 2017;37:e1–11. doi:10.4037/ccn2017548.

[68] Ebell MH, Hansen JG. Proposed clinical decision rules to diagnose acute rhinosinusitis among adults in primary care. Ann Fam Med 2017;15:347-54.

[69] Fei Y, Hu J, Gao K, Tu J, Li W qin, Wang W. Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. J Crit Care 2017;39:115–23. doi:10.1016/j.jcrc.2017.02.032.

[70] Fei Y, Hu J, Li WQ, Wang W, Zong GQ. Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. J Thromb Haemost 2017;15:439–45. doi:10.1111/jth.13588.

[71] Fei Y, Gao K, Hu J, Tu J, Li W qin, Wang W, et al. Predicting the incidence of portosplenomesenteric vein thrombosis in patients with acute pancreatitis using classification and regression tree algorithm. J Crit Care 2017;39:124–30. doi:10.1016/j.jcrc.2017.02.019.

[72] Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of incident diabetes in the jackson heart study using high-dimensional machine learning. PLoS One 2016;11: e0163942. doi:10.1371/journal.pone.0163942.

[73] Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. JAMA Cardiol 2017;2:204–9. doi:10.1001/jamacardio.2016.3956.

[74] Hettige NC, Nguyen TB, Yuan C, Rajakulendran T, Baddour J, Bhagwat N, et al. Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach. Gen Hosp Psychiatry 2017;47:20–8. doi:10.1016/j.genhosppsych.2017.03.001.

[75] Hu YH, Tai CT, Chen SCC, Lee HW, Sung SF. Predicting return visits to the emergency department for pediatric patients: Applying supervised learning techniques to the Taiwan National Health Insurance Research Database. Comput Methods Programs Biomed 2017;144:105–12. doi:10.1016/j.cmpb.2017.03.022.

[76] Huang SH, Loh JK, Tsai JT, Houg MF, Shi HY. Predictive model for 5-year mortality after breast cancer surgery in Taiwan residents. Chin J Cancer 2017;36:23. doi:10.1186/s40880-017-0192-9.

[77] Imai S, Yamada T, Kasashi K, Kobayashi M, Iseki K. Usefulness of a decision tree model for the analysis of adverse drug reactions: Evaluation of a risk prediction model of vancomycin-

associated nephrotoxicity constructed using a data mining procedure. J Eval Clin Pract 2017;23:1240–6. doi:10.1111/jep.12767.

[78] Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova M V., Rosellini AJ, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res 2017;26: e1575. doi:10.1002/mpr.1575.

[79] Kim SM, Kim Y, Jeong K, Jeong H, Kim J. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. Ultrasonography 2018;37:36–42. doi:10.14366/usg.16045.

[80] Luo Y, Li Z, Guo H, Cao H, Song C, Guo X, et al. Predicting congenital heart defects: A comparison of three data mining methods. PLoS One 2017;12:e0177811. doi:10.1371/journal.pone.0177811.

[81] Nuutinen M, Leskelä RL, Suojalehto E, Tirronen A, Komssi V. Development and validation of classifiers and variable subsets for predicting nursing home admission. BMC Med Inform Decis Mak 2017;17: e0177811. doi:10.1186/s12911-017-0442-4.

[82] Shi KQ, Zhou YY, Yan HD, Li H, Wu FL, Xie YY, et al. Classification and regression tree analysis of acute-on-chronic hepatitis B liver failure: Seeing the forest for the trees. J Viral Hepat 2017;24:132–40. doi:10.1111/jvh.12617.

[83] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med 2016;44:368–74. doi:10.1097/CCM.0000000000001571.

[84] Shneider BL, Moore J, Kerkar N, Magee JC, Ye W, Karpen SJ, et al. Initial assessment of the infant with neonatal cholestasis-Is this biliary atresia? PLoS One 2017;12: e0176275. doi:10.1371/journal.pone.0176275.

[85] Tighe DF, Thomas AJ, Sassoon I, Kinsman R, McGurk M. Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma. Head Neck 2017;39:1357–63. doi:10.1002/hed.24769.

[86] Wallert J, Tomasoni M, Madison G, Held C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. BMC Med Inform Decis Mak 2017;17:99. doi:10.1186/s12911-017-0500-y.

[87] 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:e0174944.

[88] Yip TCF, Ma AJ, Wong VWS, Tse YK, Chan HLY, Yuen PC, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther 2017;46:447–56. doi:10.1111/apt.14172.

[89] Zhang C, Garrard L, Keighley J, Carlson S, Gajewski B. Subgroup identification of early preterm birth (ePTB): Informing a future prospective enrichment clinical trial design. BMC Pregnancy Childbirth 2017;17:18. doi:10.1186/s12884-016-1189-0.

[90] Zhao Y, Healy BC, Rotstein D, Guttmann CRG, Bakshi R, Weiner HL, et al. Exploration of machine learning techniques in predicting multiple sclerosis disease course. PLoS One 2017;12: e0174866. doi:10.1371/journal.pone.0174866.

[91] Zhao Y, Xiong P, McCullough LE, Miller EE, Li H, Huang Y, et al. Comparison of Breast Cancer Risk Predictive Models and Screening Strategies for Chinese Women. J Womens Heal 2017;26:294–302. doi:10.1089/jwh.2015.5692.

[92] Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. Comput Methods Programs Biomed 2016;130:87–92. doi:10.1016/j.cmpb.2016.03.022.

[93] Chen W, Sun C, Wei R, Zhang Y, Ye H, Chi R, et al. Establishing Decision Trees for Predicting Successful Postpyloric Nasoenteric Tube Placement in Critically Ill Patients. J Parenter Enter Nutr 2018;42:132-8. doi:10.1177/0148607116667282.

[94] de O. Souza Filho JB, de Seixas JM, Galliez R, de Bragança Pereira B, de Q Mello FC, dos Santos AM, et al. A screening system for smear-negative pulmonary tuberculosis using artificial neural networks. Int J Infect Dis 2016;49:33–9. doi:https://doi.org/10.1016/j.ijid.2016.05.019.

[95] Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo A, Barreto SM, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil: Accuracy study Sao Paulo Med J 2017; 135:234-46. doi:10.1590/1516-3180.2016.0309010217.

[96] Dean JA, Wong KH, Welsh LC, Jones AB, Schick U, Newbold KL, et al. Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. Radiother Oncol 2016;120:21–7. doi:10.1016/j.radonc.2016.05.015.

[97] Eigentler T, Assi Z, Hassel JC, Heinzerling L, Starz H, Berneburg M, et al. Which melanoma

patient carries a BRAF-mutation? A comparison of predictive models. Oncotarget 2016;7:36130. doi:10.18632/oncotarget.9143.

[98] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014;33:517–35. doi:10.1002/sim.5941.

[99] Harrell, FE Jr. Regression Modeling Strategies. New York: Springer; 2015. doi:10.1007/978-3-319-19425-7.

[100] 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;74:796–804. doi:10.1016/j.eururo.2018.08.038.

[101] Chen W, Sahiner B, Samuelson F, Pezeshk A, Petrick N. Calibration of medical diagnostic classifier scores to the probability of disease. Stat Methods Med Res 2016;27:1394–409. doi:10.1177/0962280216661371.

[102] Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. Mach Learn 2006;65:95–130. doi:10.1007/s10994-006-8199-5.

[103] van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res, in press. doi: 10.1177/0962280218784726.

[104] Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30. doi:10.1161/CIRCULATIONAHA.115.001593.

[105] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15:3133–81.

[106] Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics 2018;19:270. doi:10.1186/s12859-018-2264-5.

[107] Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17:2501–8.

[108] Mitchell T. Does machine learning really work? AI Mag 1997;18:11. doi:10.1609/aimag.v18i3.1303.

[109] Steyerberg EW, Uno H, Ioannidis JPA, Van Calster B. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol 2018;98:133–43.

[110] Pouwels KB, Widyakusuma NN, Groenwold RHH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. J Clin Epidemiol 2016;69:217–24.

[111] Michelessi M, Lucenteforte E, Miele A, Oddone F, Crescioli G, Fameli V, et al. Diagnostic accuracy research in glaucoma is still incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. PLoS One 2017;12:e0189716.

[112] Kim DY, Park HS, Cho S, Yoon HS. The quality of reporting randomized controlled trials in the dermatology literature in an era where the CONSORT statement is a standard. Br J Dermatol, in press. doi: 10.1111/bjd.17432.

[113] Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Comput Biol 2015;11:e1004191. doi:10.1371/journal.pcbi.1004191.

Table 1. Algorithms used in the studies (n=71 studies). Counts refer to papers, e.g. if one paper applies several types of classification trees, this is counted only once.

Type of algorithm	**N (%)**
Logistic regression (LR) methods	71 (100%)
Standard LR only	54
Standard and penalized LR	9
Penalized LR only	6
Standard LR and boosted LR	1
Bagged LR	1
Alternative machine learning methods	
Classification tree (e.g. CART, C4.5)	30 (42%)
Random forest (RF)	28 (39%)
Support vector machine (SVM)	24 (34%)
Artificial neural network (ANN)	26 (37%)
Other algorithms	30 (42%)
Boosted tree methods (e.g. gradient boosting machines)	16
Naïve Bayes	9
Ensemble of methods [a]	4
K nearest neighbors (KNN)	3
Multivariate adaptive regression splines (MARS)	3
Bayesian Network	2
Bagged classification trees	1
Bayesian additive regression trees (BART)	1
Genetic algorithm	1
RF combined with LR	1
RF combined with SVM	1
Fuzzy logic	1
Logistic model tree	1
Naïve Bayes tree	1
Tree-augmented naïve Bayes	1
Alternative traditional statistical methods	5 (7%)
Generalized additive models (GAM)	2
Discriminant analysis	1
Poisson regression	1
Generalized estimating equations (GEE)	1

a This excludes simple bagging and boosting

Table 2. Overview of methods for model validation at study level (n=71). Counts refer to papers. Risk of bias in model validation refers to the first of five bias signaling items that were used in this study.[a] No risk of bias: the item was scored as 'no' for all models in the study; unclear: the item was scored as 'unclear' for at least one model; yes: the item was scored as 'yes' (bias present) for at least one model.

Type of validation	**Validation: risk of bias classification**		**N (%)**
	No	**Unclear/yes**	
None		7	7 (10%)
Single random split	10	19	29 (41%)
Resampling	6	19	25 (35%)
Repeated random splits	3	6	9
Cross-validation	3	12	15
Bootstrapping		1	1
External	7		7 (10%)
Chronological split	4		4
Split by center	1		1
Internal-external CV	1		1
Different dataset	1		1
Type depends on algorithm		3	3 (4%)
Total, n (%)	23 (32%)	48 (68%)	71

[a] Table A.2 describes the five bias items. For bias in model validation, we repeat the description here: We discern two general criteria to assess the validation: first, it should be clear that models are developed using training data only; second, if validation is done using resampling (repeated data splitting, cross-validation, bootstrapping), it should be clear that all model building steps are repeated in every training dataset; ad hoc flaws are documented and tabulated.

Figure 1. PRISMA flowchart

Figure 2. Summary of the five signaling items at study level (n=71). No (green): none of the five items were scored as 'unclear' or 'yes' in the whole study; unclear (orange): at least one item was scored as 'unclear' for at least one model; yes (red): at least one item was scored as 'yes' for at least one model.

Figure 3. Beeswarm plots of AUC difference (AUC of ML method minus AUC of LR) for all 282 comparisons by ML category, overall (panel A) and stratified by risk of bias (panel B). Abbreviations: LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine; ANN, artificial neural network.

Figure 4. Differences in discriminative ability between LR and ML models, overall and according to risk of bias (n=282 comparisons). LR, logistic regression; RF, random forest; SVM, support vector machine; ANN, artificial neural network. When LR was compared with traditional statistical methods (discriminant analysis, Poisson regression, generalized estimating equations, generalized additive models), these methods were not included as 'Other ML methods' and were thus excluded from this plot.

Supplementary material: PRISMA checklist

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2, (abstract had a 200 word limit though; registration number on p5)
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	5
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6, 7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix A
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6, Appendix B, Appendix C
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6, Appendix B, Appendix C
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Table A.1
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6, 7, Table A.2
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6, 7

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9, 10

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6,7,Table A.2, Table A.13
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	7 (no subgroup analyses done)
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7, Figure 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7-9, 25-26, Table A.2-Table A.6, 16-23
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Table A.3
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Not relevant
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	9-10, Figure 4
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	9-10, Figure 3, Figure 4
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	9-10, Figure 4
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	10-12
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	11-12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	12
FUNDING			

Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	13

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

What is new

Key Findings

- Studies comparing clinical prediction models based on logistic regression and machine learning algorithms suffered from poor methodology and reporting, in particular with respect to the validation procedure
- The studies rarely assessed whether risk predictions are reliable (calibration), but the area under the ROC curve (AUC) was almost always provided
- The AUC of logistic regression and machine learning models for clinical risk prediction were similar when fair comparisons were made; ML performance was higher in comparisons that were at high risk of bias

What this adds to what is known

- Machine learning models do not automatically lead to improved performance over traditional methods
- Model validation procedures are often not sound or not well reported, which hampers a fair model comparison in real world case studies

What should change now

- More attention for calibration performance of regression and machine learning models is urgently needed
- Model development and validation methodologies should be more carefully designed and reported to avoid research waste
- Research should focus more on identifying which algorithms have optimal performance for different types of prediction problems

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia CHRISTODOULOU, Jie MA, Gary S. COLLINS,

Ewout W. STEYERBERG, Jan Y. VERBAKEL, Ben VAN CALSTER

CONFLICTS OF INTEREST

Corresponding author:

Ben Van Calster

KU Leuven, Department of Development and Regeneration

Herestraat 49 box 805

3000 Leuven

Belgium

+32 16 377788

ben.vancalster@kuleuven.be

Declaration of interests: none.

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia CHRISTODOULOU, Jie MA, Gary S. COLLINS,

Ewout W. STEYERBERG, Jan Y. VERBAKEL, Ben VAN CALSTER

SUPPLEMENTARY MATERIAL

Corresponding author:

Ben Van Calster

KU Leuven, Department of Development and Regeneration

Herestraat 49 box 805

3000 Leuven

Belgium

+32 16 377788

ben.vancalster@kuleuven.be

Appendix A: Search string

```
(machine learning[MeSH Terms] OR support vector machine[MeSH Major Topic] OR neural
networks[MeSH Major Topic] OR support vector machine OR multilayer perceptron OR neural
network OR random forest OR lasso OR ridge OR kernel OR bayesian network OR classification
tree OR regression tree OR relevance vector machine OR nearest neighbor OR probability
estimation tree OR elastic net OR ensemble OR penalized OR regularized OR bagging OR boosting
OR fuzzy OR Naive bayes OR deep learning OR genetic algorithms) AND (logistic models[MeSH
Terms] OR multinomial logistic regression OR ordinal logistic regression OR logistic
regression OR proportional odds regression).
```

Appendix B: Further details on data extraction

- Categorical predictors with >2 levels are counted as one predictor.
- Missing indicator variables (binary variable indicating whether a predictor is missing or not) are not counted as predictor variables
- We extracted detailed information on tuning of hyperparameters, and classified these later on into four categories: unclear, use of default values, values are tuned but procedure is unclear, and values are tuned (procedure clear).
- Several papers also analyze 'variable importance' of the included predictors, we did not extract information on this.
- As an additional analysis, some studies also investigated the impact of sample size on performance based on smaller subsamples of the full dataset. In that case, we did not extract data from subsample models. We only extracted data for modeling based on the full dataset.

Appendix C: Criteria for identifying comparisons

Criteria for identifying comparisons between logistic regression (LR) and machine learning (ML) within a paper were:

- Comparisons involve standard/penalized LR vs a ML method
- When a paper compared LR with traditional statistical models, these are not identified as LR vs ML comparisons. We regarded discriminant analysis, Poisson regression, generalized estimating equations, and generalized additive models as traditional statistical methods
- If a paper develops models for more than one outcome, comparisons involve LR and ML models for the same outcome
- If a paper develops models for multiple subgroups: comparisons involve LR and ML models for the same subgroup
- If a paper develops models with different predictor sets that are clearly described (e.g. only clinical variables vs clinical variables and lab values): comparisons involve LR and ML models within the same predictor set
- The AUC must be available for both models, with at least 2 decimals

Table A.1. Overview of extracted items for each study.

Extracted item	**Comments**
Journal	Name of the journal in which the study was published
Impact factor	Impact factor in year of publication
Data collection	Retrospective vs prospective. If data collection was prospective, but the aim to build prediction models arose after data collection, we classified the study as retrospective.
Study design	E.g. cohort, cross-sectional, pooled data from interventional studies
Outcome type	Diagnostic vs prognostic
Predicted outcome(s)	Actual outcome(s) predicted in the study; if multiple outcomes are predicted, multiple rows are used in the extraction sheet
Sampling procedure	Population-based (registries, administrative or claims databases, recruitment from general population outside medical sites) vs hospital-based
Number of centers	In case of multicenter hospital-based study
Sample size	Sample size used in modeling, including training and validation data. E.g. if complete case analysis is used, it is the number of complete cases used in modeling. If prediction was performed in several subgroups, we recorded sample size per subgroup, and used a different row for each subgroup
Train-test split ratio and sample size for training and test sets	If only ratio (e.g. 80:20) or only sample size per dataset was given, we calculated the other one; we also recorded the sample sizes if cross-validation was used (e.g. if 10-fold cross-validation was used, training sample size was 90% of total sample size)
Number of outcome events (overall, in training)	Defined as number of participants in the smallest outcome category; if the exact number of events for the training data was not reported, we approximated the number of events based on the overall event rate (assuming equal distribution)
Missing data statements	The specific statements on amount of missing data
Method(s) to deal with missing data	E.g. complete case analysis, variable deletion, multiple imputation
Applied algorithms	We recorded every algorithm that was fitted, each algorithm was entered on a different row in the extraction sheet
Considered predictors	This is the number of predictors considered prior to data-driven selection (if done); Nominal predictors are counted as 1; extracted per algorithm
Predictors included in final model	This information is recorded per algorithm
Interaction terms for LR	Whether interaction terms were considered for LR, and which approach was used for this (e.g. all pairwise, prespecified terms)
Hyperparameter tuning	Which hyperparameters were tuned, and the method used for tuning these hyperparameters; extracted per algorithm
Type of data-driven variable selection	Whether data-driven selection was performed prior to model development, and which method was used; extracted per algorithm
Handling of continuous covariates	Whether continuous variables were kept continuous, or whether some or all continuous variables were categorized or dichotomized; for LR we also extracted information about investigation of nonlinear effects; extracted per algorithm
Type of validation	E.g., none, (repeated) train-test splitting, 10-fold cross-validation, external validation (type of external validation added; extracted per algorithm
Validation risk of bias	Whether validation of model performance was clearly described and did not have a potential for bias; extracted per algorithm
Validation issues	If risk of bias was observed, the specific issue(s) are stated here; extracted per algorithm
AUC	AUC result per algorithm; we recorded one value in this order of priority: external validation, internal validation; training data
Calibration information	Whether calibration of risk predictions was examined, and which method(s) was/were used; extracted per algorithm

Other reported performance measures	Other performance measures are listed here (not the values, only the measures); extracted per algorithm
Method to deal with class imbalance	Whether class imbalance was addressed, and which method was used; extracted per algorithm
Type of predictors	A list of the broad type of predictors that were used in the study (e.g. demographic)

Table A.2. Description of the five risk of bias items.

Risk of bias item	**Description**
Unclear or biased validation of performance	We discern two general criteria to assess the validation: first, it should be clear that models are developed using training data only; second, if validation is done using resampling (repeated data splitting, cross-validation, bootstrapping), it should be clear that all model building steps are repeated in every training dataset [26]; ad hoc flaws are documented and tabulated.
Difference in use of data-driven variable selection	This item refers to the situation where the LR model was preceded by data-driven variable selection but the ML model was not, or vice versa. This item did not refer to the use of different methods for data-driven selection, or inherent differences in selection between algorithms (e.g. LASSO automatically includes variable selection).
Difference in handling of continuous variables	This item refers to the situation where the LR model uses categorized versions of continuous variables as predictors, but the ML model kept these variables continuous, or vice versa. This item did not refer to inherent differences in handling of continuous variables between algorithms (e.g. CART) automatically dichotomizes continuous variables during model development).
Difference in considered predictors	This item refers to whether both models considered the same predictors or not.
Difference in methods for class imbalance	As discussed elsewhere in this report, some studies used methods to correct imbalance in the outcome (i.e. event rate far away from 50%). This item refers to the situation where such methods were used for the LR model but not for ML model, or vice versa.

Table A.3. List of 71 papers [28–98].

Paper	Research Field	Sample size	Predictors	Bias item 1	Bias item 2	Bias item 3	Bias item 4	Bias item 5
Acion 2017	Psychiatry	99,013	28	No	No	No	No	No
Alghamdi 2017	Endocrinology	32,555	26	Yes	No	No	No	Unc
Allyn 2017	Cardiology	6,520	66	Unc	No	Unc	No	No
Amini 2017	Preterm Birth	4,415	14	Unc	Yes	No	No	No
Asaoka 2017	Ophthalmology	374	84	No	Yes	Unc	No	No
Batterham 2017	Nutrition & DIet	295	23	Unc	No	No	No	No
Batterham 2017b	Nutrition & DIet	76	5	Yes	No	No	No	No
Cheng 2017	Geriatrics	1,951	11	No	Yes	No	No	No
Chiriac 2017	Allergy & Immunology	2,191	9	No	Yes	No	No	No
Dean 2017	Oncology	179	32	No	No	No	No	No
Deng 2017	Critical Care	417	28	Yes	No	No	No	No
Ebell 2017	Primary Care	175	17	Yes	Yes	Yes	Unc	No
Fei 2017	Critical Care	353	11	Unc	Yes	No	No	No
Fei 2017b	Critical Care	353	11	Unc	Yes	No	No	No
Fei 2017c	Critical Care	72	11	Unc	No	Unc	No	Unc
Frizzell 2017	Cardiology	56,477	83	Unc	Yes	No	No	No
Hettige 2017	Psychiatry	345	27	Unc	No	No	No	No
Hu 2017	Health care services	125,940	35	No	No	No	No	No
Huang 2017	Oncology	3,632	11	Unc	No	No	No	No
Imai 2017	Allergy & Immunology	592	11	Yes	Yes	No	Yes	No
Kessler 2017	Psychiatry	2,114,855	381	No	Unc	No	No	No
Kim 2017	Oncology	139		Un	No	No	No	No
Luo 2017	Cardiology	33,831	9	Unc	No	No	No	Yes
Nuutinen 2017	Geriatrics	3,056	97	Yes	No	No	No	No
Olivera 2017	Endocrinology	12,447	27	No	No	No	No	No
Shi 2017	Hepatology	777	22	No	No	No	No	No
Shneider 2017	Neonatology	660	22	Yes	Yes	No	No	No
Tighe 2017	Oncology	979	10	Unc	Unc	Unc	No	No
Wallert 2017	Cardiology	51,943	28	No	No	No	No	No
Weng 2017	Cardiology	378,256	30	No	No	No	No	No
Yip 2017	Hepatology	922	23	Yes	No	Unc	No	No
Zhang 2017	ObGyn	3,994,872	14	No	Yes	No	No	No
Zhao 2017	Phys. Med. & Rehab.	1,331	35	Unc	No	No	No	No
Zhao 2017b	Oncology	13,355	10	No	No	Yes	No	No
Adavi 2016	Endocrinology	12,000	7	Unc	No	No	No	No
Anderson 2016	Endocrinology	9,948	298	Yes	No	No	No	No
Arslan 2016	Cardiology	190	17	Yes	No	No	No	No
Belliveau 2016	Phys. Med. & Rehab.	3,142		Yes	No	No	Unc	No
Berchialla 2016	Health care services	7,296	12	Unc	Yes	No	No	No
Berikol 2016	Cardiology	228	7	Unc	No	No	No	No
Casanova 2016	Endocrinology	3,363	93	No	No	No	No	No
Chen 2016	Critical care	939	10	Yes	Yes	No	No	No
Churpek 2016	Critical care	269,999	29	No	No	No	No	No
De Souza Filho 2016	Infectious diseases	136	12	Yes	No	No	No	No
Dean 2016	Oncology	183	32	No	No	No	No	No
Eigentler 2016	Oncology	1,170	7	Unc	No	No	No	No
Habibi 2016	Neonatology	148	19	Yes	No	Unc	No	No
Ichikawa 2016	Primary Care	61,313	12	No	No	No	No	No
Jahani 2016	Endocrinology	545	5	Yes	No	No	No	No
Kabeshova 2016	Geriatrics	3,525	17	No	No	No	No	No
Kate 2016	Hepatology	25,521	42	Unc	No	Unc	No	No
Kulkarni 2016	Health care services	112,749	8	Yes	No	No	No	No
Lu 2016	Geriatrics	772	16	Unc	No	Unc	No	No
Mahajan 2016	Cardiology	1,037	48	Yes	Yes	Unc	No	No
Malik 2016	Endocrinology	175	7	Unc	No	No	No	No
Matis 2016	Health care services	145	13	No	No	Unc	No	No
Mortazavi 2016	Cardiology	1,004	236	Yes	Yes	No	No	No
Nakas 2016	Health care services	106,688	25	Unc	No	Unc	No	No
Ratliff 2016	Surgery		18	Yes	No	No	No	No
Rau 2016	Endocrinology	65,871		Unc	Unc	No	Unc	No
Ross 2016	Cardiology	1,047	130	Yes	Yes	Unc	No	No
Taylor 2016	Critical care	5,278	563	No	No	No	Yes	No
Thottakkara 2016	Hepatology	50,318	285	Yes	No	No	No	No
Tong 2016	Critical care	162,466	273	Yes	Yes	Unc	No	No

van der Ploeg 2016	Neurology	11,026	10	No	No	No	No	No
Wang 2016	Oncology	20,696	7	Unc	No	No	No	No
Wang 2016b	Oncology	1,143	19	Unc	Yes	No	No	No
Wu 2016	Surgery	195	9	No	Yes	No	No	No
Yahya 2016	Oncology	754	28	No	Yes	No	No	No
Zhang 2016	Oncology	205	11	Yes	No	No	Yes	No
Zhou 2016	Oncology	81	18	Unc	No	Unc	No	No

Table A.4. List of domains (n=71 studies).

Clinical discipline	**N**
Oncology	12 (17%)
Cardiovascular medicine	10 (14%)
Critical care	8 (11%)
Endocrinology	8 (11%)
Health care services	5 (7%)
Geriatrics	4 (6%)
Hepatology	4 (6%)
Psychiatry	3 (4%)
Allergy & Immunology	2 (3%)
Neonatology	2 (3%)
Nutrition	2 (3%)
Obstetrics & Gynecology	2 (3%)
Physical medicine & rehabilitation	2 (3%)
Primary care	2 (3%)
Surgery	2 (3%)
Infectious diseases	1 (1%)
Neurology	1 (1%)
Ophthalmology	1 (1%)

Table A.5. Overview of study characteristics.

Study characteristic	**N (%)**
Study design	
Unclear	3 (4%)
Cohort study	39 (55%)
Cross-sectional study	18 (25%)
Pooled data from interventional studies	6 (8%)
(Nested) case-control	2 (3%)
Pooled data from cohort and interventional studies	2 (3%)
Mix of cross-sectional and cohort data	1 (1%)
Type of outcome	
Prognostic only	50 (70%)
Diagnostic only	19 (27%)
Prognostic and diagnostic outcomes	2 (3%)
Study timing	
Unclear	4 (6%)
Retrospective	64 (90%)
Prospective	3 (4%)
Participant sampling	
Unclear	3 (4%)
Hospital-based multicenter	27 (38%)
Hospital-based single center	22 (31%)
Population-based	19 (27%)

Table A.6. Descriptive statistics, of papers and study characteristics.

Variable	N	Unknown or NA	Median	Interquartile range	Range
Journal impact factor	71	6	2.8	2.5-4.2	0.6-10.1
Number of centers if multicenter	27	10	5	4-15	2-1,137
Total sample size [a]	71	1	1,250	353-188,861	72-3,994,872
Number of predictors [b]	71	3	19	11-32	5-563
Event rate [c]	102	14	0.18	0.09-0.35	0.002-0.50
Events per predictor, training data [d]	128	26	8	4-34	0.3-6,697

[a] Some studies included an assessment of performance by sample size by also developing models for different subsamples of the full dataset. Here, we recorded information on the core analysis using the full dataset.

[b] In some cases, the number of predictors was not mentioned explicitly but could be reasonably derived from a table.

[c] Event rate: in total 102 outcomes are predicted (62 papers predicted 1 outcome, 9 predicted multiple outcomes; event is defined as the smallest outcome group.

[d] Events per predictor: papers can predict outcomes in multiple subgroups/cohorts, or with multiple predictor sets, or for multiple outcomes; in total 128 settings were identified in 71 papers. The size of and number of events in the training data was recorded exactly where possible. In some papers, size of the training data was approximated based on the reported train-test split ratio or number of folds if cross-validation was used, and number of events was approximated based on event rate. If this information was also absent, we could not derive the number of events per predictor (this happened in 26 settings).

Table A.7. Approaches to deal with missing data (n=71 studies).

Missing data approach	**N (%)**
Unclear / no information	32 (45%)
Complete case analysis (CCA)	16 (23%)
Ad hoc methods	14 (20%)
Replacement with fixed value (FV), e.g. mean imputation	4
Mixture of CCA and Missing indicator methods	3
Missing indicator methods only	1
Mixture of FV and missing indicator methods	1
Mixture of variable deletion and FV	1
Mixture of CCA and variable deletion	1
Mixture of variable deletion and missing indicator methods	1
Mixture of CCA and linear interpolation	1
Mixture of missing indicator methods and an unclear method	1
Single/Multiple stochastic imputation – see table S7	9 (13%)

Table A.8. Descriptions in papers where single or multiple imputation was used (n=9 studies)

Description in paper	**N**
Multiple imputation, no further information	2
Complete case analysis, multiple (5) imputation using propensity score method as sensitivity analysis	1
Participants with less than 75% complete information were omitted, multiple (25) information using fully conditional specification for clinically important variables	1
Less important predictors with >5% missing values were removed, important predictors with >15% missing values were removed, then complete cases were used. As a sensitivity analysis, multiple (5) imputation was done using multivariable imputation through chained equations and predictive mean matching	1
Single imputation using sequential regression imputation, no further information	1
Single imputation with knnImpute with k=5 in caret R package, no further information	1
Imputation using multivariate imputation by chained equations (mice), no further information (unclear whether single or multiple imputation)	1
Single imputation based on correlations between predictors, no further information	1

Table A.9. Detailed information about methods that were used for penalized regression, classification trees, support vector machines, and artificial neural networks. Some studies used multiple methods, therefore numbers within an algorithm category may not sum to the subtotal.

Algorithm category	**N studies**
Penalized logistic regression	15
Lasso	8
Elastic net	5
Ridge	4
Lasso or ridge used as tuning parameter	2
Classification trees	30
Classification and Regression Trees (CART)	20
C4.5	5
Chi-square Automatic Interaction Detection (CHAID)	4
Conditional inference tree	1
Unclear	2
Artificial neural networks	26
1 hidden layer	22
>1 hidden layer	3
# hidden layers unclear	1
Support vector machine	24
Radial basis function (RBF) kernel	10
Kernel unclear	7
Linear kernel	5
Kernel part of tuning process	5
Polynomial kernel	2

Table A.10. Approaches to deal with predictors (n=71 studies). Counts refer to papers.

Issue	**N (%)**
Continuous variables: general approach	
Unclear	14 (20%)
Kept continuous	37 (52%)
Categorized (i.e. >2 categories) [a]	10 (14%)
Dichotomized (2 categories) [b]	8 (11%)
Depends on algorithm	2 (3%)
Continuous variables: approach for logistic regression	
Unclear	14 (20%)
Continuous, nonlinearity unclear	29 (41%)
Discretized (2 or more categories) all variables	16 (23%)
Continuous, nonlinearity investigated	7 (10%)
Generalized additive modes used as alternative	2
Unclear, piecewise effects noted in results	2
Restricted cubic splines	1
Penalized spline functions	1
BMI categorized because nonlinearity expected	1
Discretized some variables, unclear for others	4 (6%)
Continuous, with linear effect	1 (1%)
Data-driven variable selection [c]	
Unclear	2 (3%)
No (i.e. a priori prespecification)	28 (39%)
For some algorithms	22 (31%)
For all algorithms [d]	19 (27%)
Interaction terms for logistic regression modeling	
Not explicitly mentioned	63 (89%)
Interaction terms were considered [e]	8 (11%)

a 2 studies categorized some variables, but not all

b 1 study dichotomized some variables, and categorized others; 3 studies dichotomized some variables, but not all.

c This refers to data-driven variable selection before applying the algorithms, not to variable selection that is inherent in algorithms (e.g. as in CART or lasso)

d 5 studies applied the algorithms both with and without data-driven variable selection

e The description of what was done was often unclear.

Table A.11. Descriptions in papers where interaction terms were examined (n=8 studies)

Description in paper	**N**
All two-way interactions were included	1
All two- and three-way interactions considered for LASSO model, no interactions for standard model	1
All two-way interactions screened	1
Interactions were tested	1
Models were tested for significant interactions	1
A number of interactions between socio-demographic features are included	1
Potential interactions detected through the CART model were considered	1
Interactions were checked using a backward method	1

Table A.12. Summary of hyperparameter tuning for the most common algorithms. Counts refer to papers.

	Penalized LR (N=15)	**Tree (N=30)**	**RF (N=28)**	**SVM (N=24)**	**ANN (N=26)**
Tuning approach	**n (%)**	**n (%)**	**n (%)**	**n (%)**	**n (%)**
Unclear	3 (20%)	10 (33%)	6 (21%)	7 (29%)	5 (19%)
Default setting	1 (7%)	5 (17%)	7 (25%)	2 (8%)	4 (15%)
Tuned, unclear approach	7 (47%)	11 (37%)	8 (29%)	11 (46%)	13 (50%)
Tuned	4 (27%)	4 (13%)	7 (25%)	4 (17%)	4 (15%)

LR, logistic regression; RF, random forest; SVM, support vector machine; ANN, artificial neural network.

Table A.13. Reasons for labeling a validation approach as unclear or biased (n=71 studies). Multiple reasons may apply to the same study.

Biased validation approach	**N**
Yes	
No validation of model performance	10
Model optimized using test data	5
Variable selection not repeated during resampling	4
Selective reporting of ML performance (only for the best ones)	3
Variable selection done on all data, then train-test split	2
Resampling used to tune and validate at the same time	1
Recoding of categorical predictors using the outcome	1
Performance calculated for all data despite validation procedure	1
Tuning based also on test data	1
Unclear	
Not clear on which data the hyperparameters were tuned	27
Not clear on which data variable selection was done	4
Resampling may have been used to tune and validate at the same time	4
Unclear whether tuning repeated during resampling	3
Paper states the model 'was fitted to the test sample'	1
Unclear whether variable selection repeated during resampling	1
Unclear whether all procedures repeated during resampling	1
No information on how the bootstrap validation was done	1
No information at all, except that the algorithm was used	1

Table A.14. Measures used to assess model performance (n=71 studies)

Performance criterion	**N (%)**
Area under the ROC curve (AUC)	64 (90%)
Sensitivity	45 (63%)
Specificity	43 (61%)
Positive predictive value	31 (44%)
Overall accuracy	29 (41%)
Negative predictive value	25 (35%)
Positive likelihood ratio (LR+)	4 (6%)
Negative likelihood ration (LR-)	4 (6%)
F1 score	4 (6%)
Brier	4 (6%)
Youden index	4 (6%)
Misclassification rate / overall error rate	4 (6%)
Kappa	3 (4%)
R-squared information	3 (4%)
False positive rate	3 (4%)
False negative rate	3 (4%)
Logloss / entropy	2 (3%)
Balanced accuracy	1 (1%)
Weighted accuracy	1 (1%)
Balanced error rate	1 (1%)
G mean	1 (1%)
Net reclassification improvement	1 (1%)
Matthews correlation coefficient	1 (1%)
Gini coefficient	1 (1%)
Pearson correlation	1 (1%)
Root mean squared error (RMSE)	1 (1%)
Avg absolute error	1 (1%)
Max absolute error	1 (1%)
Relative risk reduction	1 (1%)
Absolute risk reduction	1 (1%)
Absolute risk increase	1 (1%)

Table A.15. Approaches used to assess the accuracy of risk estimates (calibration) (n=71 studies).

Method	**N (%)**
Calibration not discussed	56 (79%)
Calibration discussed [a]	15 (21%)
Grouped calibration plot (or table)	8
Calibration intercept and slope	3
Hosmer-Lemeshow test on training data only	3
Hosmer-Lemeshow test on validation data	3
Calibration slope	1
Smoothed calibration plot	1
Overall predicted vs observed events	1

a Some papers used more than one method, hence numbers per method do not sum to 15.

Table A.16. Methods used for imbalanced outcome (event rate far from 50%) (71 studies)

Methods for imbalance used	**N (%)**
No	50 (70%)
Yes	21 (30%)
Undersampling	8
Weighting approach	5
Several methods tried [a]	3
Unclear	3
Synthetic minority oversampling (SMOTE) technique	1
Sampling a balanced training set	1

a Two papers tried undersampling and SMOTE, one paper tried undersampling, oversampling, and SMOTE

Table A.17. Overview of the bias items for the 71 studies and the 282 comparisons. The table indicates for how many studies/comparisons the bias item was present or was unclear.

	Item unclear or bias present, n (%)	
Bias item	**Study level (N=71)**	**Comparison level (N=282)**
Validation procedure	48 (68)	119 (42)
Variable selection	24 (32)	39 (14)
Continuous predictors	16 (23)	44 (16)
Number of predictors	6 (8)	14 (5)
Outcome imbalance	3 (4)	5 (2)

Table A.18. Overview of further anecdotal observations in the included studies [28–98].

Number	Description
General observations	
1	The measurement scale of predictors was often lacking
2	The number of predictors selected in the final model was often lacking
3	The exact type of data-driven variable selection was often unclear
4	Several extractions were implicit, by checking tables, figures or footnotes, but without clear explicit statements
Anecdotal observations	
1	We observed selective reporting of performance in some studies. It happened that several ML algorithms were applied but only results for the best were shown (1 study), or that results were shown only for ML algorithms that performed better than LR (1 study).
2	Regarding prognostic outcomes: - We observed several studies where a prognostic outcome was predicted without taking into account the time horizon. The outcome was defined as the occurrence of the condition within the available follow-up time, which could be different for each participant. - One prognostic study predicted functional limitations in the elderly, but excluded participants who died irrespective of the reason. - One study aimed to make a prognostic model based on cross-sectional data: a model to predict who is at risk of developing the condition was made by distinguishing between participants who already had or had not experienced the condition.
3	In one study, a split into train-validate-test parts was reported. The models were developed on the training set using default values, and performance was reported for the validation set. There was no further mention of the test set.
4	One paper reported an AUC of 0.52 for logistic regression, but with a sensitivity of 84% and a specificity of 87%.
5	Some papers present ROC plots showing binary ROC curves, i.e. ROC curves that are not based on the absolute risk predictions but rather on the classification after applying a cut-off.
6	One paper included a sensitivity analysis where models were training on 50% of the data, and then validated on all data.
7	One study mentions very high AUCs for two ML algorithms in the abstract and discussion, but without any mention in the results section.
8	One study matched participants with and without the outcome condition on age and gender, and then used these variables as predictors for the outcome.
9	One paper deletes the top and bottom 1% of values for continuous predictors to avoid a large influence of outliers, but then imputes these values using mean imputation.
10	One paper gives numerical values to different levels of nominal predictors based on the association of each level with the outcome that is predicted.
11	One paper deletes nearly all data in order to obtain a ‘balanced’ data set (i.e. 50% event rate). The observed event rate is 1%, such that nearly all non-events had to be excluded.

Table A.19. Overview of recommendations with the rationale and further explanation.

Recommendation	Rationale and further explanation
Fully report on all modeling steps	Incomplete reporting makes it impossible to judge on the likely robustness and validity of a model. Full reporting includes for example clear information of sample size and number of outcome events in the dataset and in dataset splits if appropriate, an unambiguous overview of all predictors that were considered in data-driven modeling and how these were selected, hyperparameter tuning, explicit statements of how continuous variables were addressed in logistic regression models, explicit statements of whether and how interaction terms were used in logistic regression models, whether and what kind of data-driven variable selection was performed, and a clear description of how modeling was done in each resampled dataset.
If resampling is used for internal validation, also develop and report the models on the full dataset	When the aim of a study is to develop clinical prediction models for use in medical practice, these models should be fully reported and available to allow external validation studies. When a study uses a single train-test split, the development data is the training set. The model based on this set is applied to the test set, and should be available for further external validation. When models are internally validated using resampling, test performance is based on multiple training and testing datasets generated from the total study sample. This means that the development data is the total study sample, and the model based on all data should be available for validation.
Report training and validation performance	Often, performance on the development data is not provided because it tends to be optimistic. However, the difference in performance with the internally validated performance (whether based on a single test set or on resampling) is informative of the amount of optimism or overfitting.
Assess calibration of the risk predictions	In clinical medicine, risk predictions are important for making decisions for individuals. Therefore, discrimination performance of a model is not sufficient. The calibration of the predicted risks should be evaluated as well. This informs on the likely over- or underestimation of the predicted risks. For example, overfitted models tend to underestimate low risks and overestimate high risks. Poor calibration reduces the utility of a model.

Figure A.1. Scatter plot of the number of considered predictors by the number of events in the training data for all 71 studies. The plot contains >71 data points: some studies predicted multiple outcomes, made predictions for different subgroups, or considered multiple predictor sets.

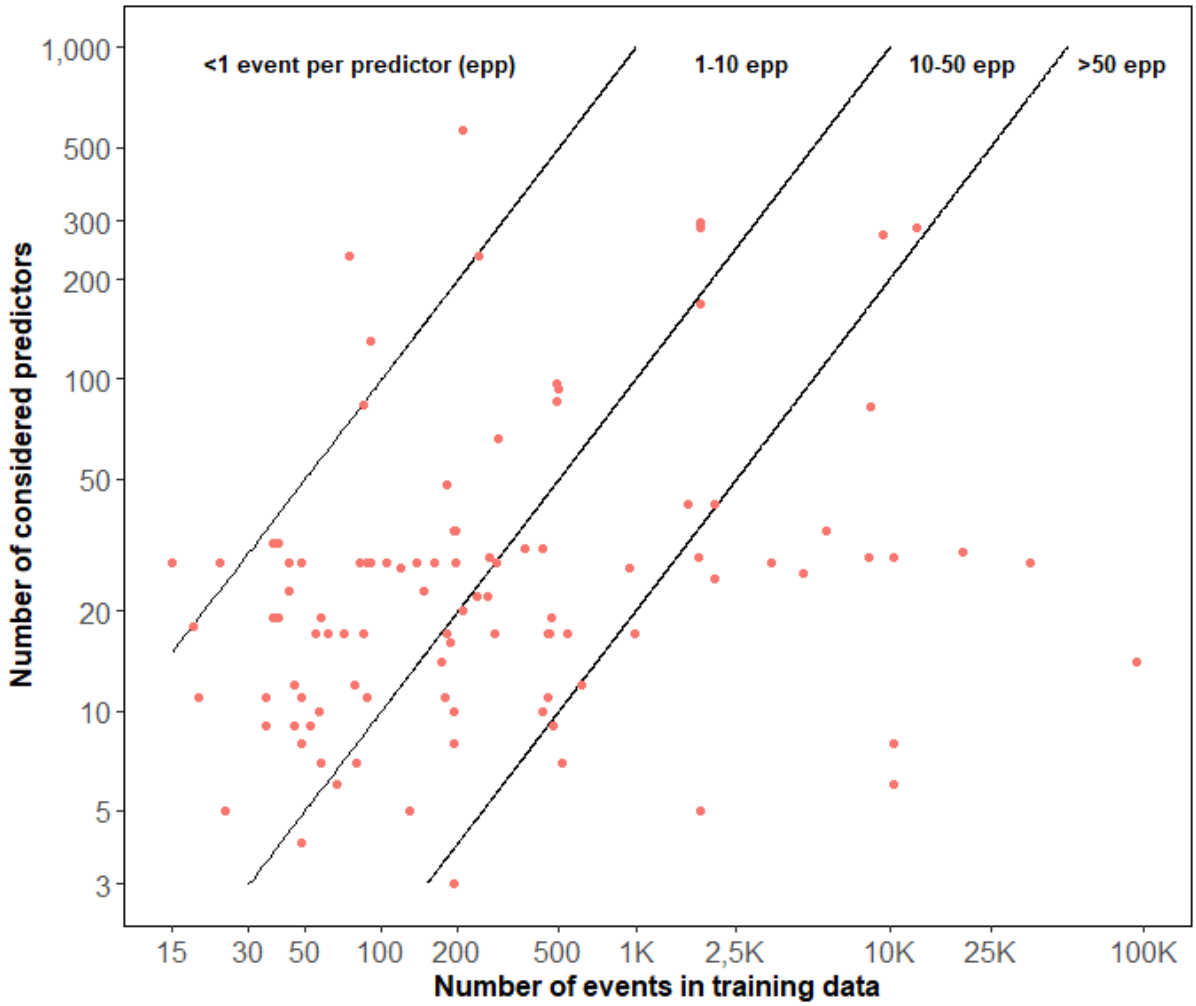
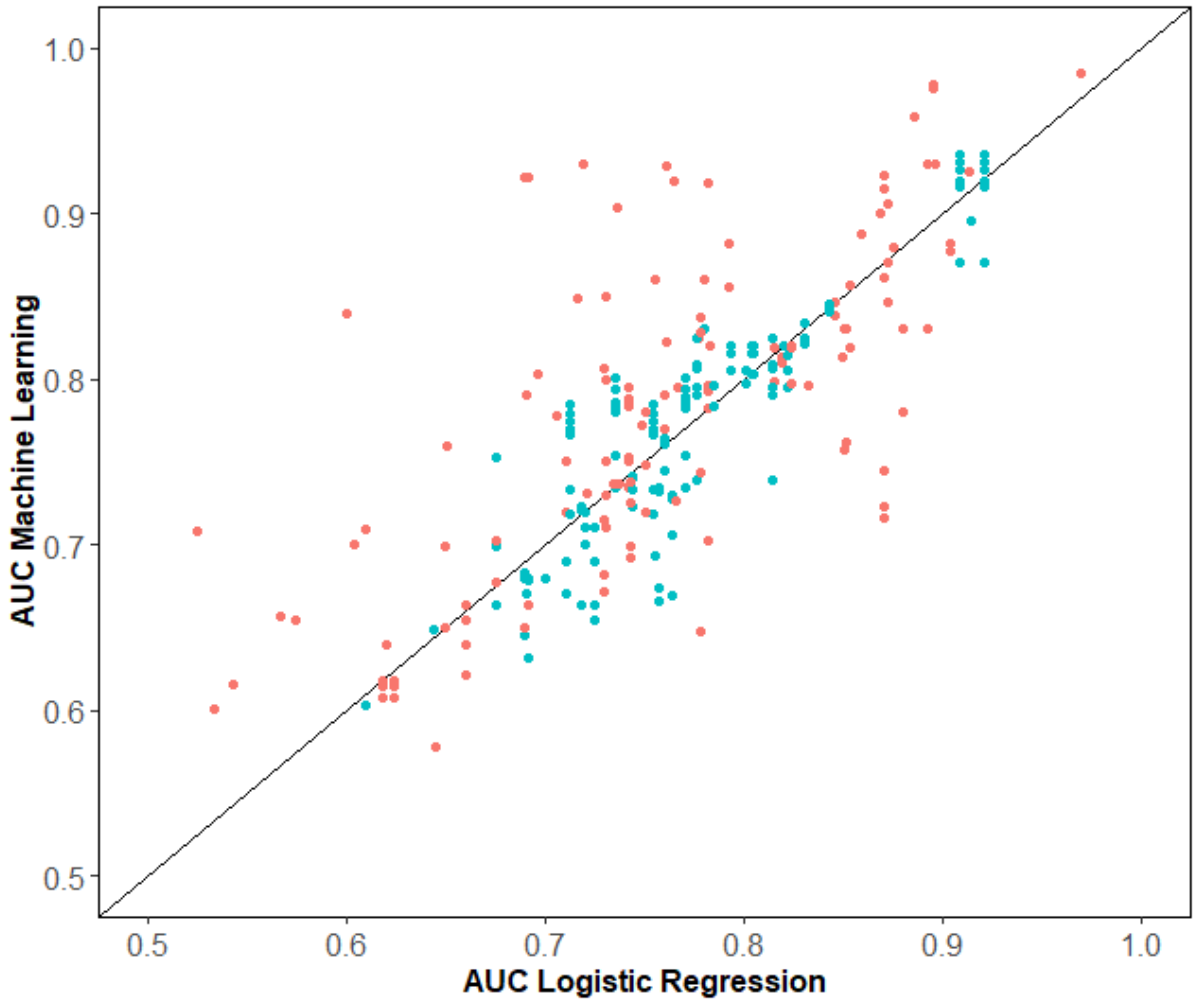
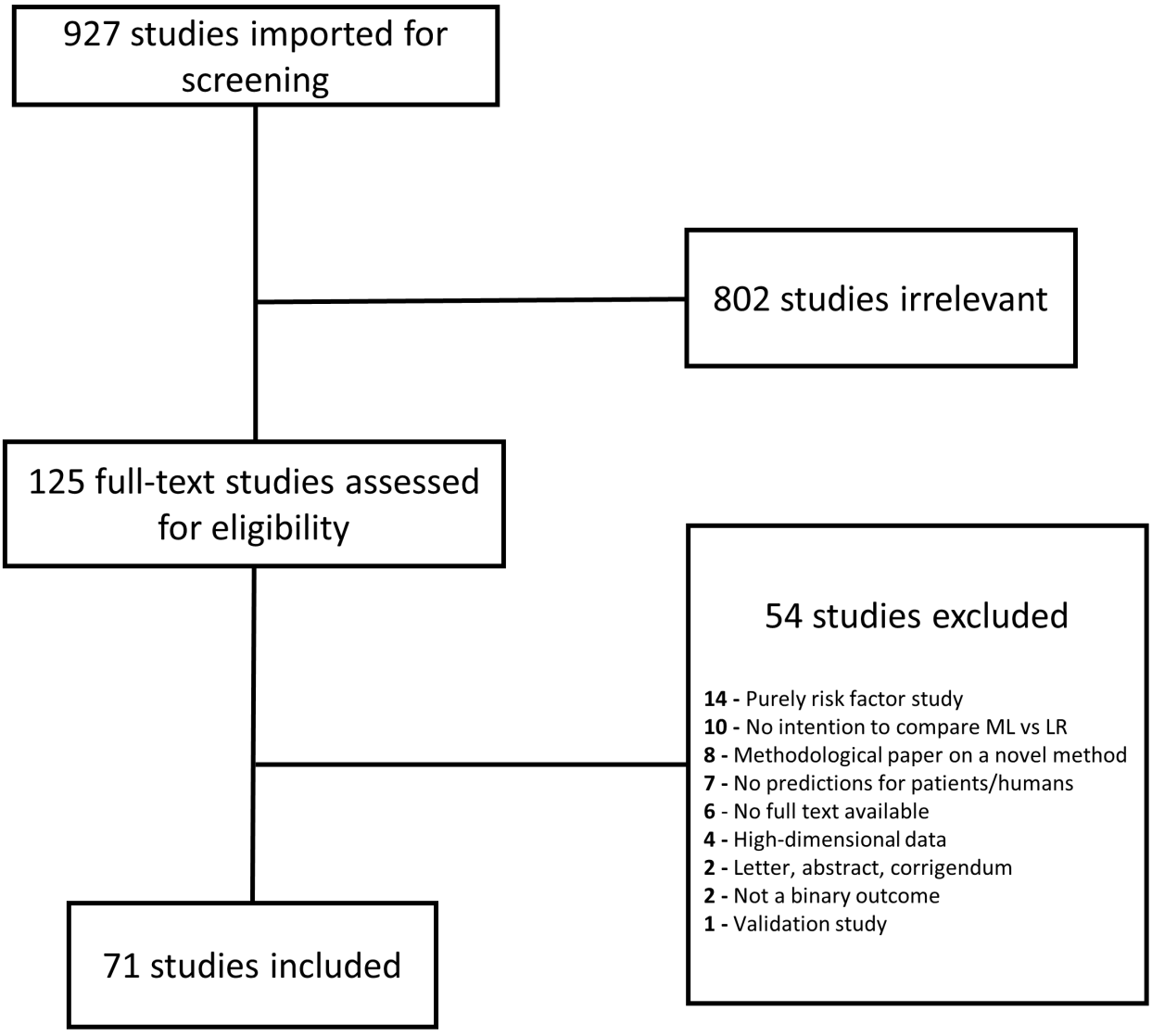
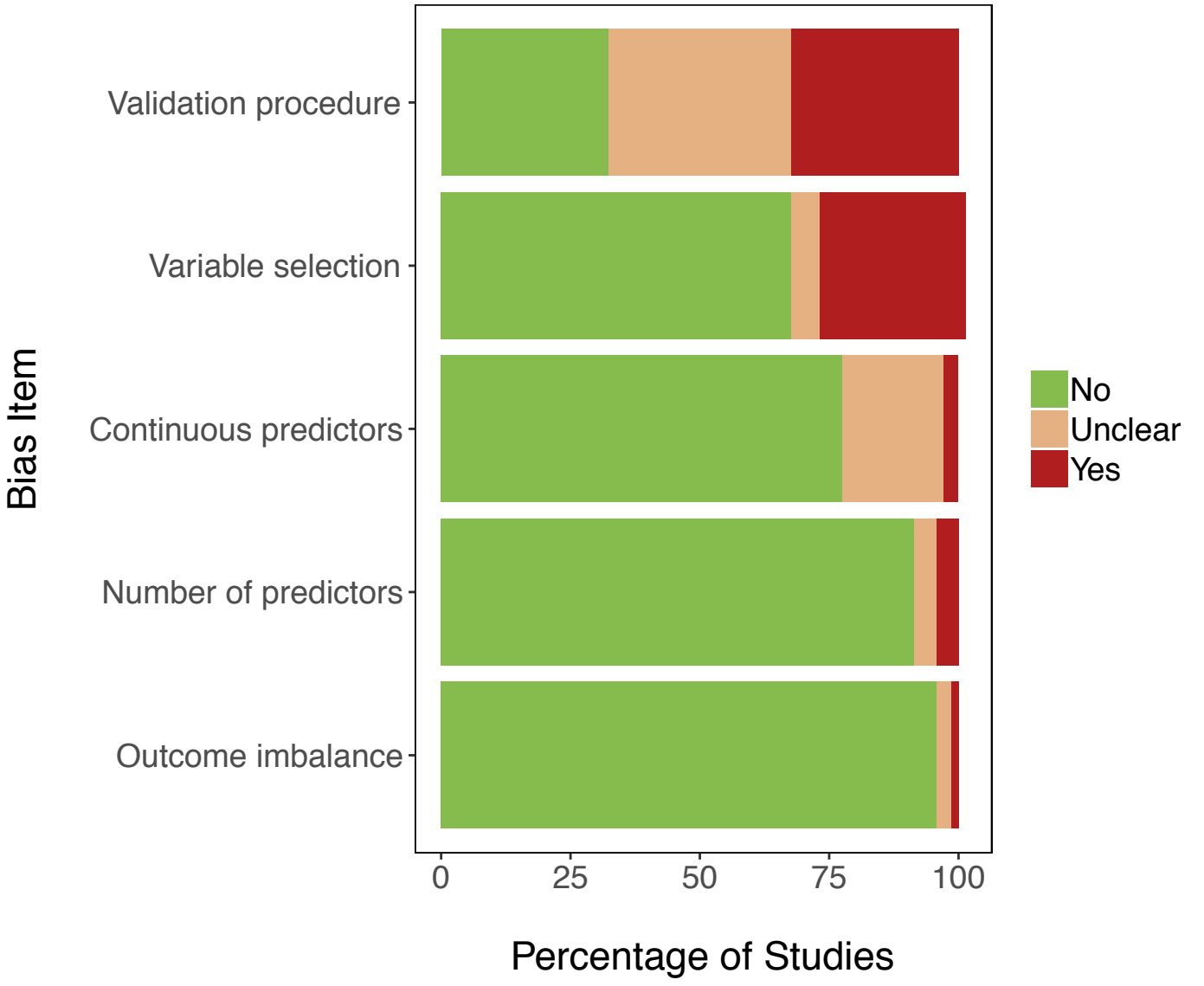
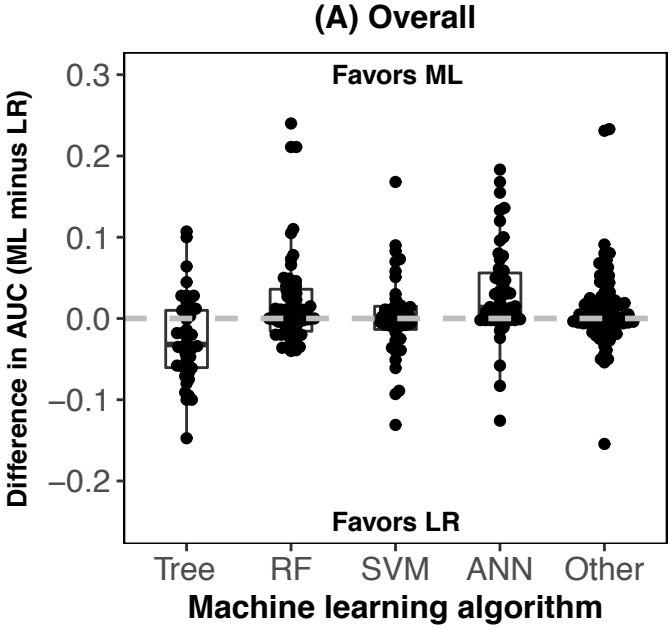


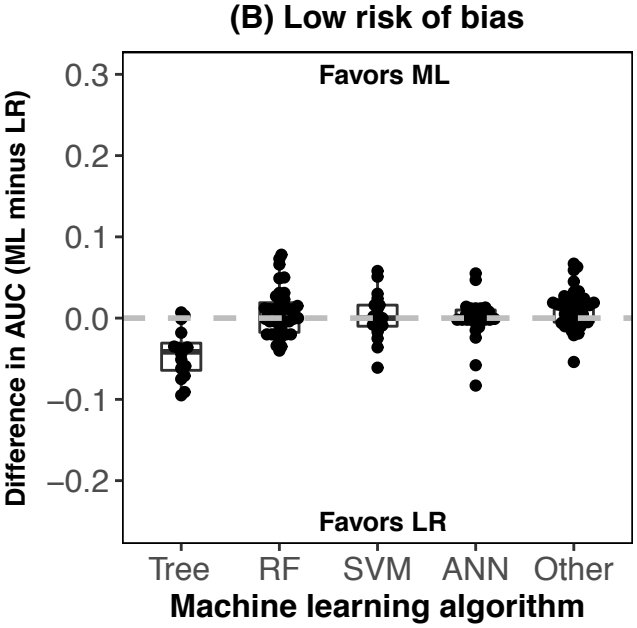
Figure A.2. Scatter plot of the area under the ROC curve (AUC) for LR vs ML for all 282 comparisons. Comparisons with low risk of bias are shown in green, comparisons with high risk of bias in red.

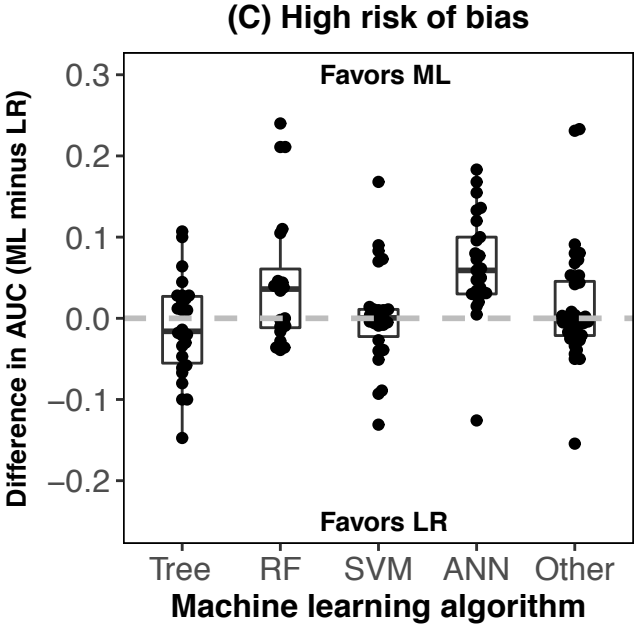


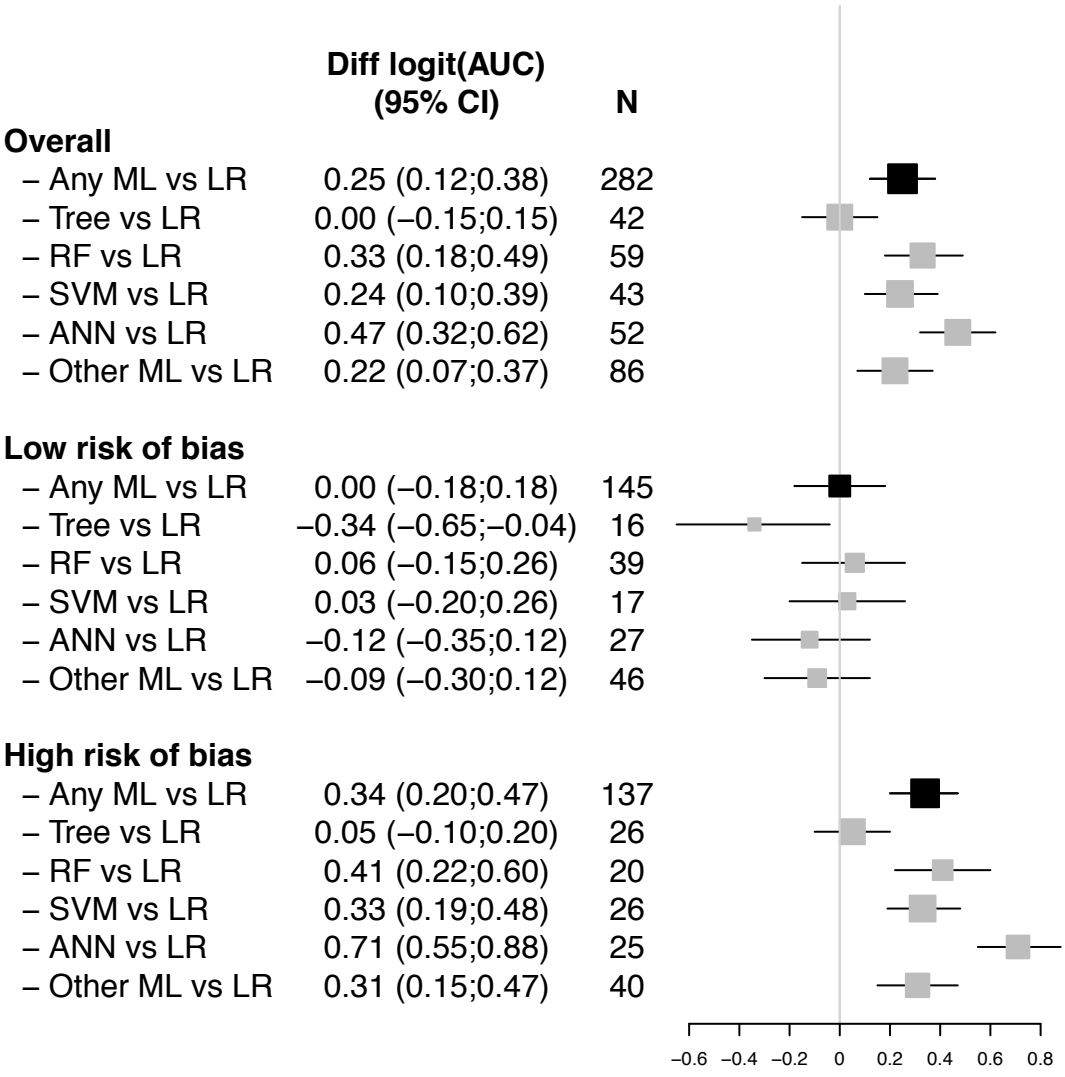

927 studies imported for screening
802 studies irrelevant
125 full-text studies assessed for eligibility
54 studies excluded
14 - Purely risk factor study
10 - No intention to compare ML vs LR
8 - Methodological paper on a novel method
7 - No predictions for patients/humans
6 - No full text available
4 - High-dimensional data
2 - Letter, abstract, corrigendum
2 - Not a binary outcome
1 - Validation study
71 studies included




(A) Overall
Favors ML
Favors LR
Difference in AUC (ML minus LR)
0.3
0.2
0.1
0.0
−0.1
−0.2
Tree
RF
SVM
ANN
Other
Machine learning algorithm


(B) Low risk of bias
Favors ML
Favors LR
Difference in AUC (ML minus LR)
0.3
0.2
0.1
0.0
−0.1
−0.2
Tree
RF
SVM
ANN
Other
Machine learning algorithm


(C) High risk of bias
Favors ML
Favors LR
Difference in AUC (ML minus LR)
0.3
0.2
0.1
0.0
−0.1
−0.2
Tree
RF
SVM
ANN
Other
Machine learning algorithm


Diff logit(AUC) (95% CI)
N
Overall
– Any ML vs LR 0.25 (0.12;0.38) 282
– Tree vs LR 0.00 (–0.15;0.15) 42
– RF vs LR 0.33 (0.18;0.49) 59
– SVM vs LR 0.24 (0.10;0.39) 43
– ANN vs LR 0.47 (0.32;0.62) 52
– Other ML vs LR 0.22 (0.07;0.37) 86
Low risk of bias
– Any ML vs LR 0.00 (–0.18;0.18) 145
– Tree vs LR –0.34 (–0.65;–0.04) 16
– RF vs LR 0.06 (–0.15;0.26) 39
– SVM vs LR 0.03 (–0.20;0.26) 17
– ANN vs LR –0.12 (–0.35;0.12) 27
– Other ML vs LR –0.09 (–0.30;0.12) 46
High risk of bias
– Any ML vs LR 0.34 (0.20;0.47) 137
– Tree vs LR 0.05 (–0.10;0.20) 26
– RF vs LR 0.41 (0.22;0.60) 20
– SVM vs LR 0.33 (0.19;0.48) 26
– ANN vs LR 0.71 (0.55;0.88) 25
– Other ML vs LR 0.31 (0.15;0.47) 40
–0.6 –0.4 –0.2 0 0.2 0.4 0.6 0.8

Authors' contributions

E.C. was involved in the conception of the study, data collection, data analysis and interpretation, drafting of the article, and gave her final approval of the version to be published. J.M. was involved in data collection, critical revision of the article, and gave her final approval of the version to be published. G.S.C. was involved in the conception of the study, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. E.W.S. was involved in the conception of the study, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. J.Y.V. was involved in the conception of the study, data collection, interpretation of the data, the critical revision of the article, and gave his final approval of the version to be published. B.V.C. was involved in the conception of the study, data collection, data analysis and interpretation, drafting of the article, and gave his final approval of the version to be published.