Journal of Clinical Epidemiology 110 (2019) 12−22

Journal of Clinical Epidemiology

REVIEW

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]*Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 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 UK*
[c]*Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[d]*Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands*
[e]*Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium*
[f]*Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK*

Accepted 5 February 2019; Published online 11 February 2019

Abstract

Objectives: The objective of this study was to compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling in the literature.

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 of 927 studies. The median sample size was 1,250 (range 72−3,994,872), with 19 predictors considered (range 5−563) and eight events per predictor (range 0.3−6,697). The most common ML methods were classification trees, random forests, artificial neural networks, and support vector machines. In 48 (68%) studies, we observed potential bias in the validation procedures. 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 an 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 (0.20−0.47) higher for ML.

Conclusion: We found no evidence of superior performance of ML over LR. Improvements in methodology and reporting are needed for studies that compare modeling algorithms.

Keywords: Clinical prediction models; Logistic regression; Machine learning; AUC; Calibration; Reporting

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]. Although ML methods have been sporadically used for clinical prediction for some time [3,4], the growing availability of increasingly large, voluminous, and rich data sets such as electronic health records data have 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

* Corresponding author. Tel.: +32-16-377788; fax: +32-16-344205.
E-mail address: ben.vancalster@kuleuven.be (B. Van Calster).

https://doi.org/10.1016/j.jclinepi.2019.02.004
0895-4356/

What is new?

Key findings

- Applied 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 receiver operating characteristic curve (AUC) was almost always provided.
- The AUC of logistic regression and machine learning models for clinical risk prediction were similar when comparisons were at low risk of bias; machine learning (ML) performance was higher in comparisons that were at high risk of bias.

What this adds to what was known?

- ML 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 is the implication and what should change now?

- More attention for calibration performance of regression and ML 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.

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]. By 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 [12]. Some researchers describe the distinction between statistical modeling and ML 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.

Owing 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]. Furthermore, ML models are typically assessed in terms of discrimination performance (e.g., accuracy, area under the receiver operating characteristic [ROC] curve [AUC]), while 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) to base our analysis on recent studies.

2.2. Selection of studies

All abstracts were independently screened by two reviewers (E.C. and J.M.); conflicts were resolved by a third reviewer (B.V.C. or J.Y.V.). The full text of selected abstracts were independently assessed for eligibility by three reviewers (E.C., J.M., B.V.C.), 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 nonhumans,
- 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) before applying LR and ML algorithms,
(3) difference in handling of continuous variables before 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 articles developed several LR and/or ML models. These articles 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 article, we identified all comparisons between LR and ML methods (see Appendix C). We identified multiple comparisons within the same article 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 laboratory 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 articles 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 AUC. We used AUC values in the following order of priority: external validation, internal validation, and 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 article 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 (Fig. 1). Fifty-four 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. Most studies (n = 64, 90%) used existing data, and 27 (38%) used hospital-based multicenter data. The median number of centers was five (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). One hundred and two outcomes were considered in the 71 articles, the median event rate

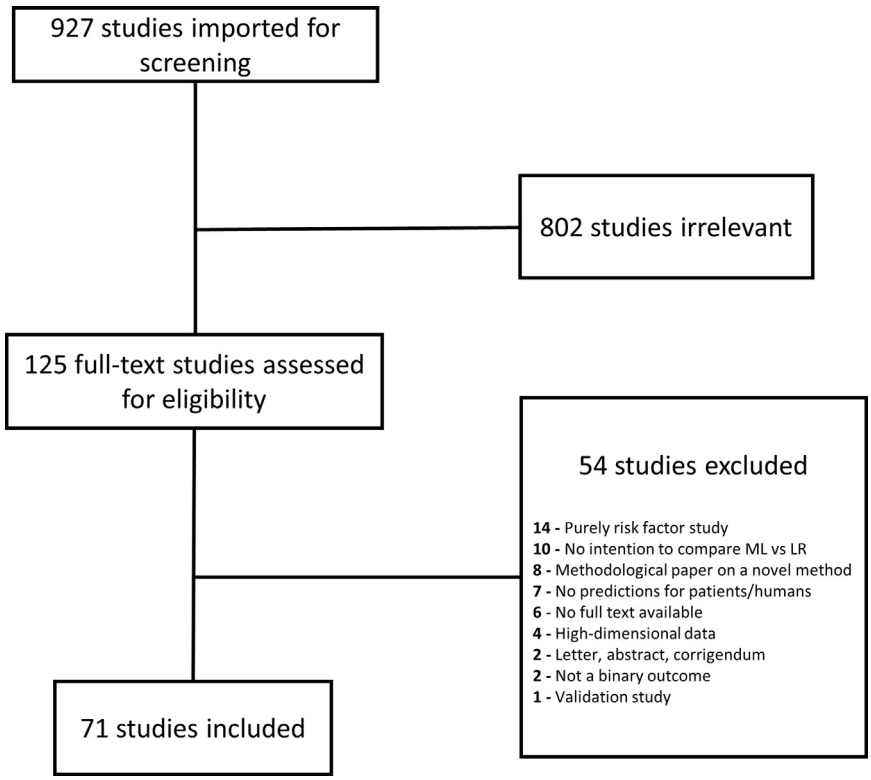

Fig. 1. PRISMA flowchart. PRISMA, preferred reporting items for systematic reviews and meta-analysis.

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) (Fig. 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 Table A.9). Six studies used only penalized LR, and one study used only bagged LR (classified as ML).

Forty-three studies used more than 1 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 one hidden layer, three used multiple hidden layers, and for one 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 (L.R. vs. M.L.), 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, contains hyperparameters that determine the complexity/flexibility of the model. For the most commonly used algorithms,

Table 1. Algorithms used in the studies (n = 71 studies)

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

Counts refer to articles, for example, if one article applies several types of classification trees, this is counted only once.

[a] This excludes simple bagging and boosting.

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

Twenty-nine 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, nine used repeated random splitting, and one 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 performed 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 one study (1%) evaluated performance in terms of clinical utility using decision curve analysis.

In 21 studies, methods were applied to address outcome imbalance, that is, an event rate far from 50% (Table A.16, see Section 4).

3.5. Comparison between performance of LR and ML

The most problematic risk of bias item was an unclear/biased validation procedure (Fig. 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 articles. Of the remaining 13 articles, seven did not report AUCs, three reported AUCs for some algorithms only, one reported AUCs to one decimal, one only applied standard and penalized LR, and one 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 0.12–0.38) (Figs. 3 and 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 (0.20–0.47) 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

Table 2. Overview of methods for model validation at study level (n = 71)[a]

	Validation: risk of bias classification		
Type of validation	**No**	**Unclear/yes**	***N* (%)**
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 data set	1		1
Type depends on algorithm		3	3 (4%)
Total, *n* (%)	23 (32%)	48 (68%)	71

Counts refer to articles. Risk of bias in model validation refers to the first of five bias signaling items that were used in this study.

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.

[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 performed using resampling (repeated data splitting, cross-validation, bootstrapping), it should be clear that all model building steps are repeated in every training data set; ad hoc flaws are documented and tabulated.

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 modeling steps that were used to build the model to estimate performance. Model development on all data was often not performed. 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 nonprobabilistic model outcomes into probabilities [101]. In this article, 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. Furthermore, 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

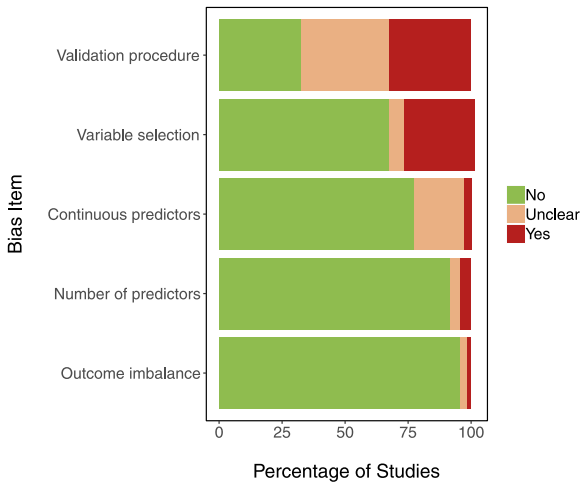

Fig. 2. Summary of the five signaling items at study level (n = 71). No (green): none of the five items were scored as ''unclear'' or ''yes'' in the whole study; unclear (orange): at least one item was scored as ''unclear'' for at least one model; yes (red): at least one item was scored as ''yes'' for at least one model. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

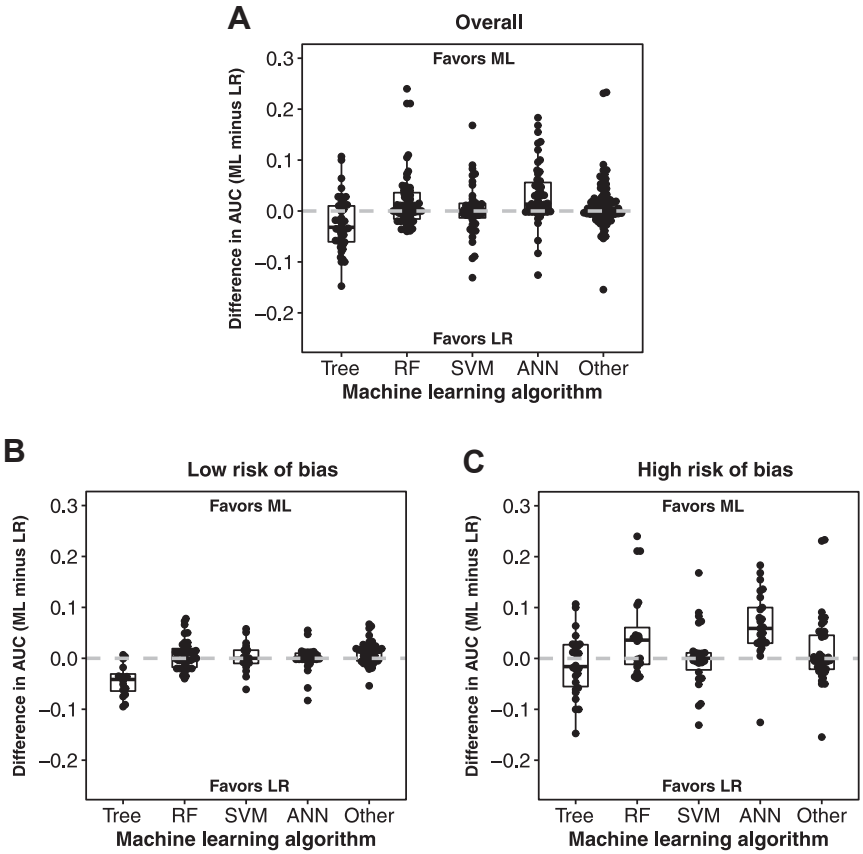

Fig. 3. Beeswarm plots of AUC difference (AUC of ML method minus AUC of LR) for all 282 comparisons by ML category, overall (A) and stratified by risk of bias (B). LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine; ANN, artificial neural network.

common in the ML literature [13]. This is motivated by a dominant focus on classification and overall accuracy based on a 50% risk cutoff. 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 data set 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 GAM, 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], for example, 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 data sets 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 prespecified predictors is considered, and that the advantage of ML lies in better handling a huge amount of

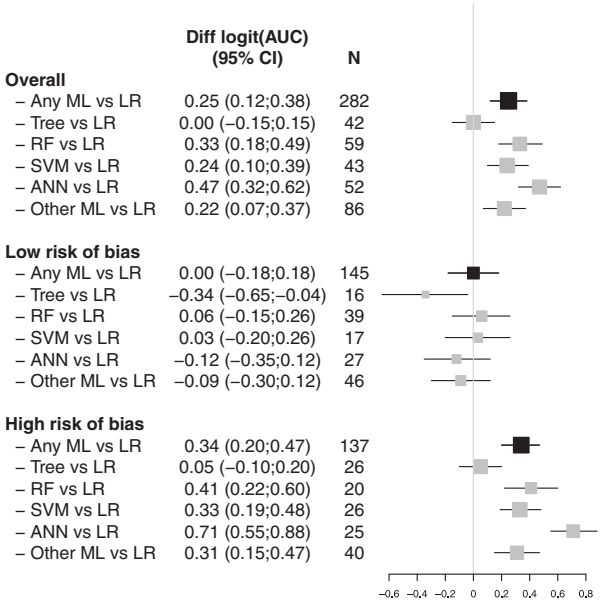

Fig. 4. Differences in discriminative ability between LR and ML models, overall and according to risk of bias (n = 282 comparisons). When LR was compared with traditional statistical methods (discriminant analysis, Poisson regression, generalized estimating equations, generalized additive models), these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

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 aforementioned 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 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 articles 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.

CRediT authorship contribution statement

Evangelia Christodoulou: Conceptualization, Investigation, Formal analysis, Data curation, Writing - original draft. **Jie Ma:** Investigation, Writing - review & editing. **Gary S. Collins:** Conceptualization, Data curation, Writing - review & editing. **Ewout W. Steyerberg:** Conceptualization, Data curation, Writing - review & editing. **Jan Y. Verbakel:** Conceptualization, Investigation, Data curation, Writing - review & editing. **Ben Van Calster:** Conceptualization, Investigation, Formal analysis, Data curation, Writing - original draft.

Acknowledgments

This work was supported 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.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2019.02.004.

References

[1] Steyerberg EW. Clinical prediction models. New York, NY: Springer; 2009.
[2] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009.
[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.
[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.
[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.
[8] Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 2001;16:199–231.
[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, NY: McGraw Hill; 1997.
[11] Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biom J 2014;56:588–93.
[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.
[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.

[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.
[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.
[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.
[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 2018: ArXiv Prepr ArXiv180209596.
[24] Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med 2016;35:4124−35.
[25] Steyerberg EW, Harrell FE Jr, 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.
[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.
[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.
[31] Jahani M, Mahdavi M. Comparison of predictive models for the early diagnosis of diabetes. Healthc Inform Res 2016;22:95−100.
[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.
[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.
[34] Kulkarni P, Smith LD, Woeltje KF. Assessing risk of hospital readmissions for improving medical practice. Health Care Manag Sci 2016;19:291−9.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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 parameters Predicting distant failure in lung SBRT. Radiother Oncol 2016;119:501−4.
[56] Acion L, Kelmansky D, Van der Laan M, Sahker E, Jones DS, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One 2017;12:e0175383.
[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.
[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.
[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 Health Res Perspect 2017;8:195−200.
[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.
[61] Berikol GB, Yildiz O, Özcan T. Diagnosis of acute coronary syndrome with a support vector machine. J Med Syst 2016;40:84.
[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.
[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.
[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.
[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.
[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.
[67] Deng X. Predicting the risk for hospital-acquired pressure ulcers in critical care patients. Crit Care Nurse 2017;37:e1−11.
[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, qin Li W, 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.
[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.
[71] Fei Y, Gao K, Hu J, Tu J, qin Li W, 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.
[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.
[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.
[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.
[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.
[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.
[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.
[78] Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, 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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.
[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.

[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.
[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.
[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 Health (Larchmt) 2017;26:294−302.
[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.
[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. JPEN J Parenter Enteral Nutr 2018;42:132−8.
[94] Souza Filho JB de O, 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.
[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.
[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.
[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.
[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.
[99] Harrell FE Jr. Regression modeling strategies. New York, NY: Springer; 2015.
[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.
[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.
[102] Drummond C, Holte RC. Cost curves: an improved method for visualizing classifier performance. Mach Learn 2006;65:95−130.
[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 2018;. https://doi.org/10.1177/0962280218784726. [Epub ahead of print].
[104] Deo RC. Machine learning in medicine. Circulation 2015;132:1920−30.
[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.
[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.
[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 2018;. https://doi.org/10.1111/bjd.17432. [Epub ahead of print].
[113] Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Comput Biol 2015;11:e1004191.