Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

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

Lertkong Nitiwarangkul, MD.

From Journal of Clinical Epidemiology 2019

Introduction

- Clinical risk prediction models aim to predict a clinically relevant outcome
- Traditional approach
  - Use of regression, e.g., logistic regression (LR)
  - To predict for diagnosis or prognosis
- Machine learning (ML): Alternative approach
  - Artificial neural networks
  - Support vector machines
  - Random forests

Introduction

- Useful definition of ML
  - Focuses on models directly
  - Automatically learn from data
- By contrast, regression models
  - Based on theory and assumptions
- For example, ML performs modeling
  - More automatically than regression
    - Regarding the inclusion of nonlinear associations and interaction terms
  - More flexible but require penalization to avoid overfitting

Introduction

- Primary objective
- Compare the performance of LR with ML algorithms
- For the development of diagnostic or prognostic clinical prediction models of binary outcomes
- Secondary objectives
- Describe the characteristics of the studies
- Type of ML algorithms
- Validation process
- Modeling aspects of LR and ML
- Reporting quality
- Risk of bias

Materials and methods

- Identification of studies
- Search from Medline by using a broad working definition of ML
- Since 2016 to August 2017
- Selection of studies
- Screened by two reviewers
- Conflicts were resolved by a third reviewer
- Full text of selected abstracts were independently assessed for eligibility by three reviewers

Materials and methods

- Inclusion criteria
- Development of a diagnostic or prognostic prediction model
- Compared prediction models based on LR and ML algorithms
- Exclusion criteria
- New modeling approach was introduced
- Models were developed for **nonhumans**
- Models made predictions for **images or signals**
- Models were developed based on **high-dimensional data**
- Primary interest was assessing risk factors
- Reviews of the literature
- Unable to obtain the full text

Materials and methods

- Data extraction and risk of bias
  - List of extraction items → CHARMS check list
  - Risk of bias tool → QUADAS risk
- Extracted items included
  - General study characteristics
  - Applied algorithms
  - Their characteristics
  - Data-driven variable selection
  - Model performance

CHARMS check list

Table 1. Key items to guide the framing of the review aim, search strategy, and study inclusion and exclusion criteria.

Item	Comments and examples
1. Prognostic versus diagnostic prediction model	Define whether the aim is to review models to predict:
	• Future events: prognostic prediction models
	• Current (disease) status: diagnostic prediction models
2. Intended scope of the review	Define intended scope of the review and intended purpose of the models reviewed in it. Examples:
	• Models to inform physicians' therapeutic decision making
	• Models to inform referral to or withholding from invasive diagnostic testing
3. Type of prediction modelling studies (see also Box 1)	Define the type of prediction modelling studies to include. Examples of study types (Box 1):
	• Prediction model development without external validation in independent data
	• Prediction model development with external validation in independent data
	• External model validation, possibly with model updating
4. Target population to whom the prediction model applies	Define the target population relevant to the review scope. Examples:
	• Women with diagnosed breast cancer
	• Healthy adult men in the general population
5. Outcome to be predicted	Define the outcome of interest to be predicted:
	• Specific future event, such as a fatal or non-fatal coronary heart disease
	• Specific diagnostic target disease, such as presence of lung embolism
6. Time span of prediction	Define over what specific time period the outcome is predicted (prognostic models only). Example:
	• Event within a specific time interval, such as event within 3 months, 1 year, or 10 years
7. Intended moment of using the model	The systematic review may focus on models to be used at a specific moment in time. Examples:
	• Models to be used at the moment of diagnosis of a particular disease
	• Models to be used preoperatively to predict the risk of postoperative complications
	• Models to be used in asymptomatic adults to detect undiagnosed type 2 diabetes mellitus

QUADAS risk

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Study 1	☺	☺	☺	☺	☹	☺	☺
Study 2	☺	☺	☺	☺	☹	☺	☺
Study 3	☹	☹	☺	☺	☹	☺	☺
Study 4	☹	☹	☺	☺	☹	☺	☺
Study 5	☹	?	☺	☺	☹	☺	☺
Study 6	☹	?	☺	☺	☹	?	☺
Study 7	☹	?	☺	☺	☹	☺	☺
Study 8	☹	?	☺	☺	☹	?	☺
Study 9	☹	?	☺	☺	☹	☺	☺
Study 10	☹	?	☺	☹	☹	☺	☺
Study 11	☺	?	☺	☹	☺	☺	☺

☺ Low Risk ☹ High Risk ? Unclear Risk

Materials and methods

- Five signaling items to indicate potential bias
- Unclear **validation procedure**
- Difference in whether **data-driven variable selection** was performed (yes/no) before applying LR and ML algorithms
- Difference in **handling of continuous variables** before applying LR and ML algorithms,
- **Different predictors** considered for LR and ML algorithms (Number of predictor)
- Whether corrections for **imbalanced outcomes** where used only for LR or only for ML algorithms

Materials and methods

- Rating for potential bias
- Each bias item was scored as
  - No (not present)
  - Unclear
  - Yes (present)
- Low risk of bias → ‘‘no’’ for all five signaling items
- High risk of bias → ‘‘unclear’’ or ‘‘yes’’ for at least one item

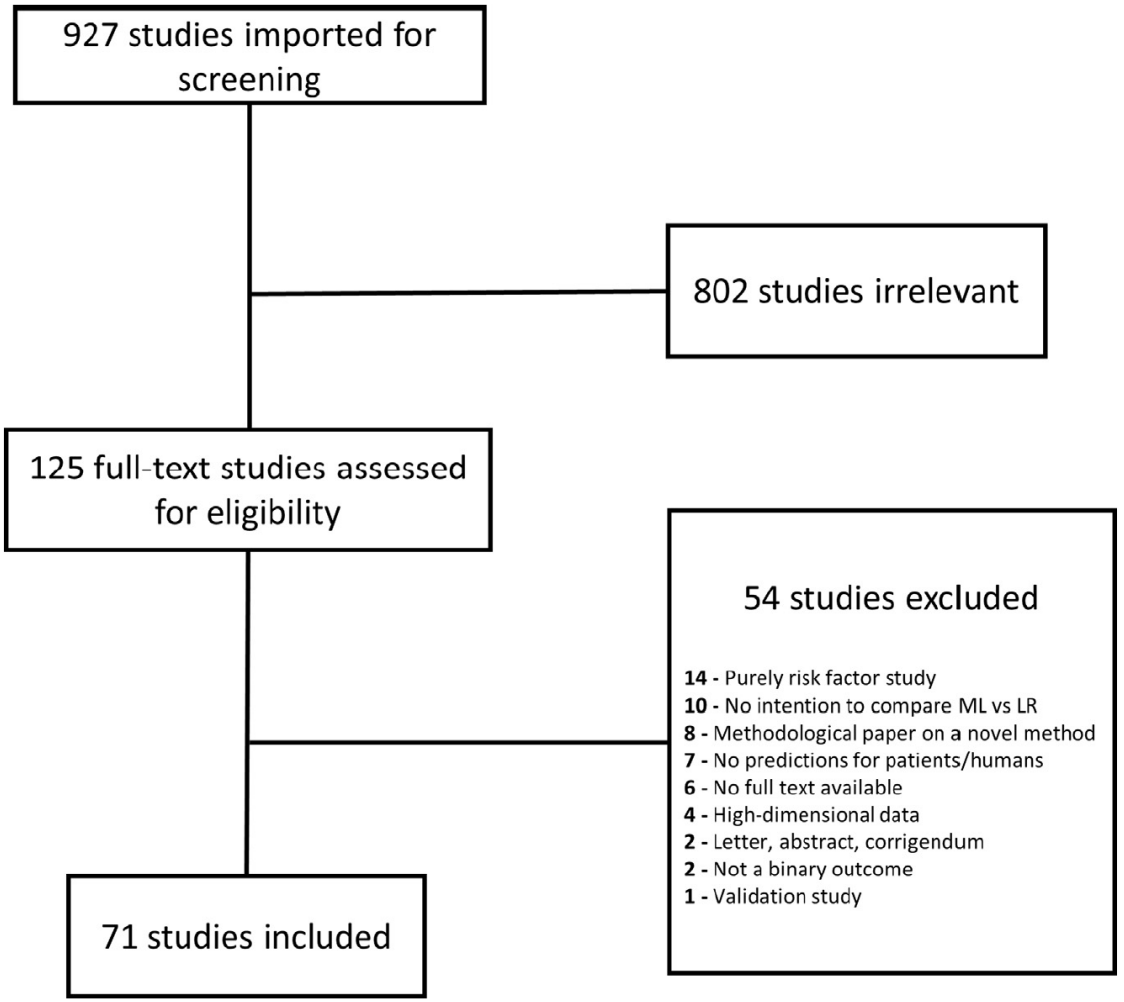
Materials and methods

- Data analysis
- We compared the LR and ML models using the difference in the AUC
  - External validation, internal validation, and training data
- ML algorithms classification into 5 groups
  - Single classification trees
  - Random forests
  - Artificial neural networks
  - Support vector machines
  - Other algorithms

Materials and methods

- We analyzed AUC differences with stratification for risk of bias
- Meta-regression of the difference between logit AUC using a random effect model
- Weighted by the square root of the validation sample size
- Logit(AUC) was used to avoid the bounded nature of the AUC

Result


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

General study characteristics

- 39 studies (59%) were cohort (M/C)
- 18 studies (25%) were cross-sectional

- 50 studies (70%) focused on prognostic outcomes
- 19 studies (27%) on diagnostic outcomes
- 2 studies (3%) on both

- 64 studies (90%) used existing data
- 27 studies (38%) used hospital-based multicenter data
  - Median number of centers was 5 (range 2-1,137)

General study characteristics

- 102 outcomes were considered
- 9 articles → Models to predict more than one outcome

- *** Report in Median ***
- Total sample size: 1,250 (72-3,994,872)
- Number of predictors: 19 (5-563)
- Event rate was 0.18 (0.002-0.50)
- Number of events per predictor was 8 (0.3-6,697)

General study characteristics

- Missing data
- 32 studies (45%) → Information on handling was unclear
- 16 studies (23%) performed a complete case analysis
- 14 studies (20%) relied on ad hoc methods (mean imputation, missing indicator methods, variable deletion)
- 9 studies (11%) used single or multiple stochastic imputation

Overview of algorithms

- 64 studies used standard LR
  - 9 also used penalized LR (LASSO, ridge or elastic net)
  - 1 also used boosted LR
- 6 studies used only penalized LR
- 1 study used only bagged LR (classified as ML)

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

Overview of algorithms

- 43 studies used more than 1 ML algorithm
- Most popular algorithms were
  - Classification trees (n = 30, 42%)
  - Random forests (n = 28, 39%)
  - Artificial neural networks (n = 26, 37%)
  - Support vector machines (n = 24, 34%)
- 26 studies using artificial neural networks
  - 22 used one hidden layer
  - 3 used multiple hidden layers
  - 1 study was unclear
- When support vector machines were used → the Gaussian (‘‘radial basis function’’) kernel was most often used (n = 10)

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

Model development

- 14 studies (20%) were not clear about how continuous variables were handled
- 18 studies (25%) used discretization (into 2 or more categories)
- 41 studies (58%) reported about Data-driven variable selection

Model development

- 47/71 studies (66%) of LR was unclear in handling of <u>continuous</u> predictors
  - In 33/47, → unclear whether nonlinear associations were examined
  - 1 study, clear that continuous variables have linear associations with the outcome
  - 20 studies (28%) used discretization
  - 7 studies (10%) investigated nonlinearity
- <u>Interaction</u>
  - 63 studies (89%) did not mention of interaction effects
  - 8 studies (11%) → Unclear: Approach for interaction terms

Model validation

- 29 studies (41%) used a single random split
- 25 studies used resampling
  - 15 studies used cross validation
  - 9 studies used repeated random splitting
  - 1 study used bootstrapping
- 7 studies (10%) used some form of external validation
- 7 studies (10%) did not validate performance
- 3 studies (4%), the approach depended on the algorithm

Type of validation	Validation: risk of bias classification		
	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

Model validation

- 48 studies (68%), unclear reporting or potential biases in validation procedures
- The AUC was the most commonly reported performance measure (64 studies, 90%)
- Sensitivity (45 studies, 63%)
- Specificity (43 studies, 61%)
- Most of the studies (56 studies, 79%) not discussed about calibration performance

Comparison between performance of LR and ML

- The most problematic risk of bias item was an unclear validation procedure
- 282 comparisons identified
- 145 comparisons → low risk of bias
  - Logit (AUC) difference was on average 0.00 (-0.18 to 0.18) for comparisons
- 137 comparisons → high risk of bias
  - Logit (AUC) 0.34 higher in ML (0.20 to 0.47) for comparisons
- Results for different ML algorithms were similar except for Trees uniformly had worse performance than others

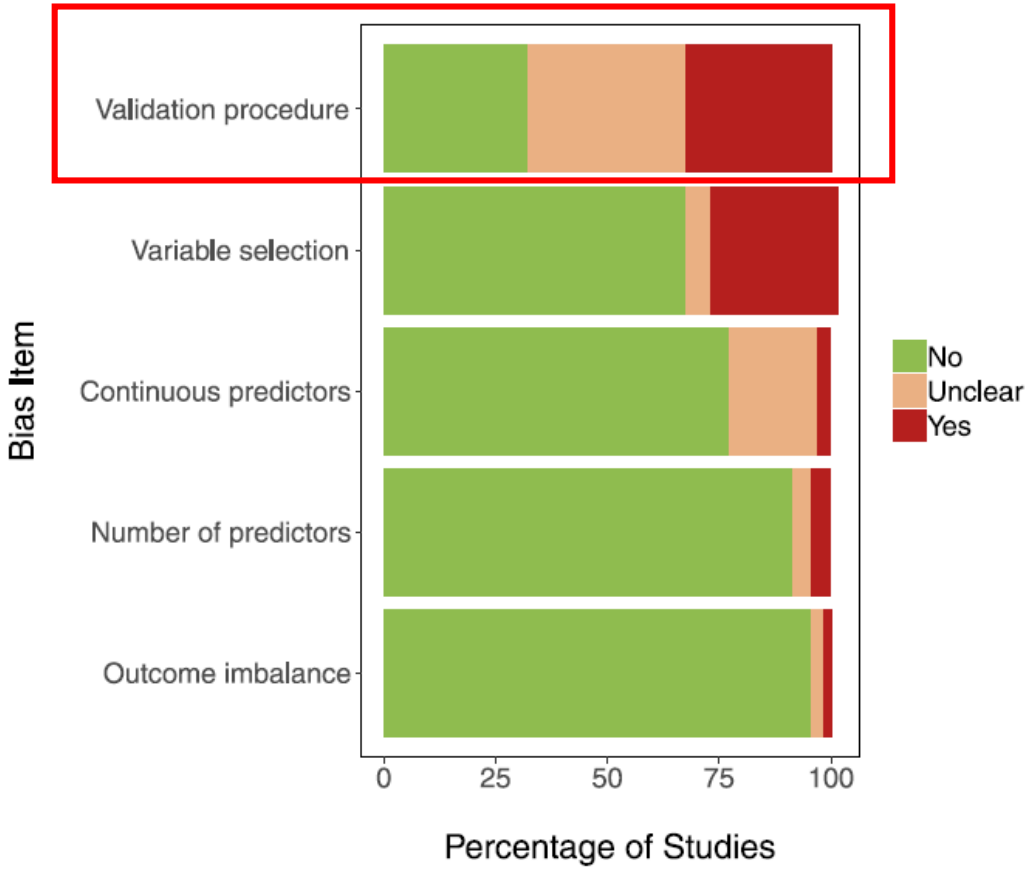
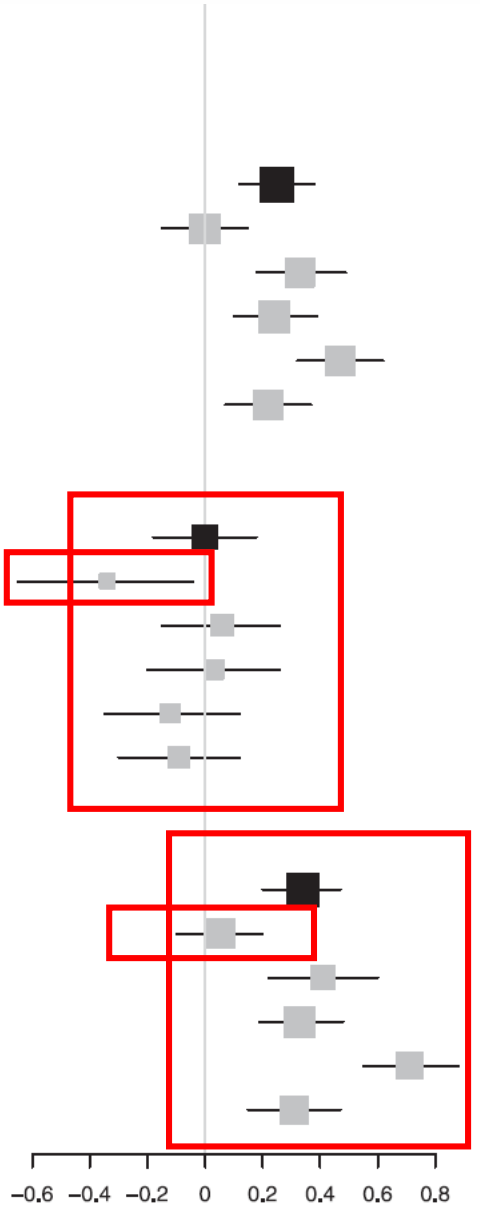
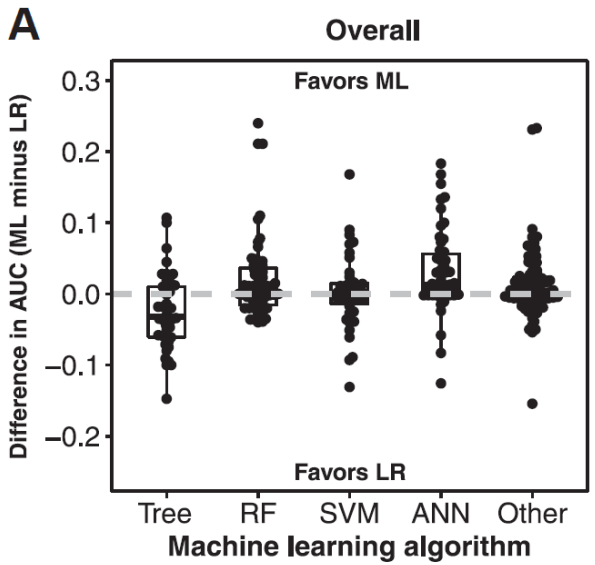
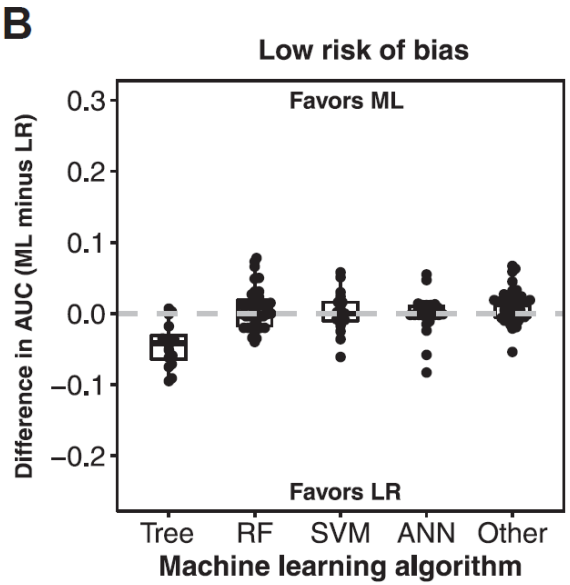


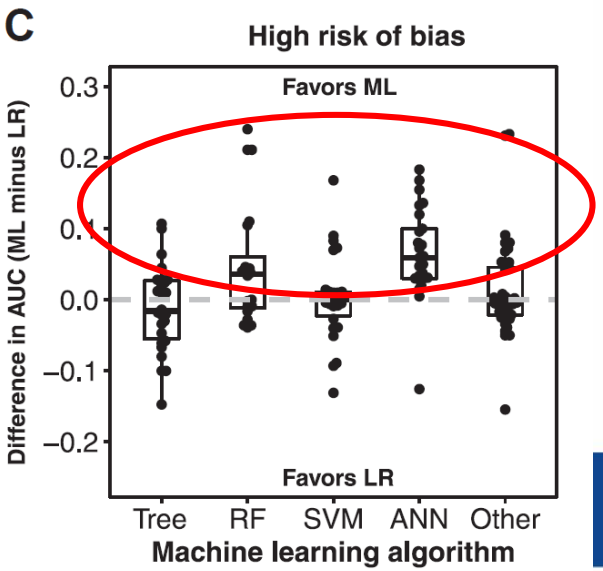
Fig. 2. Summary of the five signaling items at study level ($n = 71$). No (green): none of the five items were scored as ''unclear'' or ''yes'' in the whole study; unclear (orange): at least one item was scored as ''unclear'' for at least one model; yes (red): at least one item was scored as ''yes''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

	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




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
0.0
−0.1
−0.2
Tree
RF
SVM
ANN
Other
Machine learning algorithm

Discussion

- Summary from most of studies
- Reporting of methodology → often incomplete and unclear
- Model validation procedures were poor
- Calibration of risk predictions was seldom examined
- AUC performance of LR and ML was on average no different when comparisons had low risk of bias

Recommendation

- 1st
- Fully report the steps and analyses → maximize transparency and reproducibility
  - Adhere to the TRIPOD guidelines
- For complex procedures
  - Use flowchart of the development and validation procedures can be insightful

TRIPOD Checklist: Prediction Model Development

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	Describe eligibility criteria for participants.	
	5c	Give details of treatments received, if relevant.	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	Explain how the study size was arrived at.	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
Risk groups	11	Provide details on how risk groups were created, if done.	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
Model development	14a	Specify the number of participants and outcome events in each analysis.	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	Explain how to the use the prediction model.	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	Give the source of funding and the role of the funders for the present study.	

TRIPOD Checklist: Prediction Model Validation

TRIPOD

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	Describe eligibility criteria for participants.	
	5c	Give details of treatments received, if relevant.	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	Explain how the study size was arrived at.	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10c	For validation, describe how the predictions were calculated.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	Provide details on how risk groups were created, if done.	
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	Give the source of funding and the role of the funders for the present study.	

TRIPOD Checklist: Prediction Model Development and Validation

TRIPOD

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

Recommendation

- 2nd if model validation is based on resampling
  - Model development should be based on all available data
  - Resampling should then include all modeling steps that were used to build the model to estimate performance
  - In addition, provide all information on these models to allow independent validation

Recommendation

- 3rd
- Report training and test performance
- The difference between these results is informative

- 4th
- Evaluate model performance in terms of calibration and clinical utility for decision-making

- Comparison of AUC performance between LR and ML
- Depends on how one defines risk of bias and ML
- Five signaling items to consider comparisons as at low or high risk of bias
  - Did not address whether LR models were penalized or included nonlinear or interaction effects
- Regression is sometimes assumes linearity and additivity
  - Some criticize say that this may reduce the performance of regression, although this may depend on sample size

Future research

- Should focus more on explaning the type of predictive problems
- For example, the signal-to-noise ratio
  - may be an important aspect in determining how successful ML will be
  - ML tends to work well for problems with a strong signal-to-noise ratio
  - For example, handwriting recognition, gaming, or electric load forecasting.
- But → Clinical prediction problems often have a poor signal-to noise ratio

Limitation

- Does not investigate which factors influence the difference in performance
  - Sample size
  - Number of predictors
  - Hyperparameter tuning
- Should be performed by comparing different scenarios on the same data sets to avoid confounding
- Limited number of events per predictor
- 23 comparisons with 100 predictors were at high risk of bias

Conclusion

- Evidence is lacking to claim that ML lead to better AUCs than clinical prediction models based on LR
- Reporting of articles → needs to improve
- Validation procedures should add **calibration** and **clinical utility** to improve discrimination
  - To define situations where modern methods have advantages over traditional approaches

Thank you for your attention