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.

