Feasibility of machine learning and logistic regression algorithms to predict outcome in orthopaedic yrauma surgery

Oosterhoff, J.H.F.; Gravesteijn, B.Y.; Karhade, A.V.; Jaarsma, R.L.; Kerkhoffs, G.M.M.J.; Ring, D.V.; ... ; Machine Learning Consortium

Feasibility of Machine Learning and Logistic Regression Algorithms to Predict Outcome in Orthopaedic Trauma Surgery

Jacobien H.F. Oosterhoff, MD, Benjamin Y. Gravesteijn, MSc, Aditya V. Karhade, MD, MBA, Ruurd L. Jaarsma, MD, PhD, FRACS, FAOrthA, Gino M.M.J. Kerkhoffs, MD, PhD, David Ring, MD, PhD, Joseph H. Schwab, MD, MS, Ewout W. Steyerberg, MSc, PhD, Job N. Doornberg, MD, PhD, and the Machine Learning Consortium*

Investigation performed at Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands

Background: Statistical models using machine learning (ML) have the potential for more accurate estimates of the probability of binary events than logistic regression. The present study used existing data sets from large musculoskeletal trauma trials to address the following study questions: (1) Do ML models produce better probability estimates than logistic regression models? (2) Are ML models influenced by different variables than logistic regression models?

Methods: We created ML and logistic regression models that estimated the probability of a specific fracture (posterior malleolar involvement in distal spiral tibial shaft and ankle fractures, scaphoid fracture, and distal radial fracture) or adverse event (subsequent surgery [after distal biceps repair or tibial shaft fracture], surgical site infection, and postoperative delirium) using 9 data sets from published musculoskeletal trauma studies. Each data set was split into training (80%) and test (20%) subsets. Fivefold cross-validation of the training set was used to develop the ML models. The best-performing model was then assessed in the independent testing data. Performance was assessed by (1) discrimination (c-statistic), (2) calibration (slope and intercept), and (3) overall performance (Brier score).

Results: The mean c-statistic was 0.01 higher for the logistic regression models compared with the best ML models for each data set (range, −0.01 to 0.06). There were fewer variables strongly associated with variation in the ML models, and many were dissimilar from those in the logistic regression models.

Conclusions: The observation that ML models produce probability estimates comparable with logistic regression models for binary events in musculoskeletal trauma suggests that their benefit may be limited in this context.

Several statistical approaches can produce estimates of the probability of binary events such as diagnosis of fracture, death, infection, and reoperation[1,2]. The most used method is logistic regression analysis. The estimates from statistical models have the potential to inform patient and surgeon decision-making.

In a relatively new method, commonly referred to as machine learning (ML), the computer trains an existing human-created algorithm to recognize patterns in the data and iteratively alters the algorithm for optimal performance. In musculoskeletal trauma, the most common ML algorithms used to date are decision-tree-based, support vector machine, neural network, a Bayesian method, and penalized logistic regression[3-7].

There is growing interest in ML statistical methods for estimating probabilities of binary events such as diagnosis of a fracture (posterior malleolar involvement in distal spiral tibial shaft and ankle fractures, scaphoid fracture, and distal radial fracture) or occurrence of a specific adverse event (mortality, treatment failure, reoperation, infection, or sustained opioid use). It is not clear that ML methods provide better estimates. A recent systematic review comparing ML and logistic regression models for binary events among 71 studies found no benefit to ML, with studies showing a benefit rated as more prone to bias[8].

The aim of this study was to investigate the performance of ML compared with standard regression modeling for estimating probabilities of binary events after musculoskeletal

*A list of the Machine Learning Consortium members is included in a note at the end of the article.

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJS/G831).

trauma using the data sets from 9 published studies. Our study questions were: (1) Do ML models produce better probability estimates than logistic regression models? (2) Are ML models influenced by different variables than logistic regression models?

Materials and Methods

Guidelines

This study was conducted according to the *Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View* and the *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement*[9,10].

Study Design and Participants (Data Sources) (Table I)

The data sets from 1 published randomized controlled trial[11] and 8 published cohort studies[12-19] in musculoskeletal trauma were obtained from individual authors who are members of the Machine Learning Consortium[20] (Table I). The sample size of these data sets ranged from 263 to 28,207 patients. Four studies focused on adverse events (surgical site infection after operative fracture care[12], adverse events after distal biceps tendon surgery[13], subsequent surgery in tibial shaft fractures[11], and postoperative delirium in elderly hip fracture patients[17]), and 5 focused on diagnostic events (scaphoid fracture[14,16], posterior malleolar involvement in distal spiral tibial and ankle fractures[15,18], and distal radial fracture[19]). We merged the raw data of 2 studies on the diagnosis of posterior malleolar involvement in distal spiral tibial shaft and ankle fractures[15,18] and scaphoid fracture[14,16], resulting in 7 data sets for comparison of logistic regression and ML (Table I).

Data Analysis

Logistic Regression

Given the rule of thumb in logistic regression to enter no more than 1 explanatory variable for every 10 response-variable events[2], we developed parsimonious logistic regression models using an initial bivariate analysis. Associations with continuous independent variables were measured with the Student t test for parametric variables and the Mann-Whitney U test for nonparametric variables. Associations with dichotomous independent variables were measured with the Fisher exact test. The chi-square test was used for analysis of ordinal data. Associations with a p value of <0.05 were considered significant. After variable selection, missing values for variables with <30% missing data were imputed using multiple imputation by chained equations (MICE)[21,22].

Seven backward stepwise logistic regression models were created to estimate the probabilities of a fracture (posterior malleolar involvement in distal spiral tibial shaft and ankle fractures, scaphoid fracture, or distal radial fracture) or a specific adverse event (subsequent surgery [after distal biceps or tibial shaft fracture], surgical site infection, and postoperative delirium).

Machine Learning

As a first step in the ML models, variables potentially associated with each outcome were identified using least absolute shrinkage and selection operator (LASSO) penalized logistic regression, Boruta, and random forest recursive selection algorithms (see Appendix). After variable selection, missing values for variables with <30% missing data were imputed using MICE[21,22].

Following convention, we trained and internally validated several different ML algorithms on each data set to choose the best-performing algorithm[23]: penalized logistic regression, support vector machine, decision tree algorithms (gradient boosting machine, random forest, and boosting decision tree), neural network, and Bayes classifiers (Bayes point machine and naïve Bayes) (Fig. 1). For each analysis, the data sets were split randomly into training (80%) and test subsets (20%). We trained each ML algorithm using fivefold cross-validation on the training set. The best-performing model was selected and used for performance assessment in the test set.

Model Performance

Model performance was evaluated according to a proposed framework for evaluation of a clinical prediction model[24] that included (1) discrimination with the c-statistic, (2) calibration with the calibration slope and intercept (in line with the method described by Cox[25]), and (3) overall performance with the Brier score.

The c-statistic (area under the curve of a receiver operating characteristic curve) is a score ranging from 0.50 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 indicating the lowest. The higher the discrimination score, the better the model's ability to distinguish between classes (i.e., patients who had the outcome from those who had not)[26].

A calibration plot plots the estimated versus the observed probabilities for the primary outcome. A perfect calibration plot has an intercept of 0 (<0 reflects overestimation, and >0 reflects underestimation of the probability of the outcome) and a slope of 1 (the model is performing similarly in training and test sets)[24,27]. In a small data set, the slope is often <1, reflecting model overfitting; probabilities are too extreme (low probabilities that are too low, and high probabilities that are too high)[26].

The Brier score calculates a composite of discrimination and calibration, with 0 indicating perfect prediction and 1, the poorest prediction[24].

We visualized model performance comparisons in a beeswarm plot—a scatterplot of the differences in c-statistics of each ML and logistic regression algorithm pair.

Software

Data preprocessing and analysis were performed using R version 5.3[28] and RStudio version 1.2.1335 (R Foundation for Statistical Computing). Packages used were rms, pROC, randomForest, caret, gbm, nnet, mice, kernlab, dplyr, mice, and beeswarm. Hyperparameter tuning was performed as recommended in the vignettes.

Source of Funding

No external funding was received for any aspect of this work.

TABLE I Characteristics of Included Studies

Study	Title	Journal*	No. of Patients	Outcome	No. (%) of Patients with Outcome
Sobol et al.[18] (2018) and Hendrickx et al.[15] (2019)	The incidence of posterior malleolar involvement in distal spiral tibia fractures: Is it higher than we think? and Incidence, predictors, and fracture mapping of (occult) posterior malleolar fractures associated with tibial shaft fractures	Both from JOT	263†	Posterior malleolar fracture	75 (28.5)
Beks et al.[13] (2016)	Factors associated with adverse events after distal biceps tendon repair or reconstruction	JSES	373	Adverse events	82 (22.0)
Duckworth et al.[14] (2012) and Mallee et al.[16] (2020)	Predictors of fracture following suspected injury to the scaphoid; and Detecting scaphoid fractures in wrist injury: a clinical decision rule	JBJS-Br and AOTS, respectively	420†	Scaphoid fracture	117 (27.7)
Walenkamp et al.[19] (2015)	The Amsterdam wrist rules: the multicenter prospective derivation and external validation of a clinical decision rule for the use of radiography in acute wrist trauma	BMC MD	854	Distal radial fracture	376 (44.0)
SPRINT Investigators[11] (2008)	Randomized trial of reamed and unreamed intramedullary nailing of tibial shaft fractures	JBJS-A	1,198	Subsequent surgery	214 (17.9)
Bachoura et al.[12] (2011)	Infirmity and injury complexity are risk factors for surgical-site infection after operative fracture care	CORR	2,000	Surgical site infection	90 (4.5)
Oosterhoff et al.[17] (in press)	Development of machine learning algorithms for prediction of postoperative delirium in elderly hip fracture patients	GOS	28,207	Postop. delirium	8,030 (28.5)

*JOT = Journal of Orthopaedic Trauma; JSES = Journal of Shoulder and Elbow Surgery; JBJS-Br = Journal of Bone and Joint Surgery, British Volume; AOTS = Archives of Orthopaedic and Trauma Surgery; BMC MD = BMC Musculoskeletal Disorders; JBJS-A = Journal of Bone and Joint Surgery, American Volume; CORR = Clinical Orthopaedic Related Research; and GOS = Geriatric Orthopaedic Surgery & Rehabilitation. †Number of patients in combined dataset.

Results

Comparison of Model Performance

The c-statistic was on average 0.01 higher for regression than for ML, with a range from −0.01 to 0.06, indicating that ML models and logistic regression models produce comparable probability estimates (Fig. 2, Table II). Among the 7 ML analyses, boosting decision tree, support vector machine, and penalized logistic regression were each the best-performing algorithm twice, and the Bayes point machine performed best once.

The calibration slopes and intercepts and the Brier scores also produced comparable estimates, showing little or no advantage to ML over logistic regression models (Table II).

Comparison of Variable Selection (Table III)

Logistic regression models included between 2 and 17 variables, compared with 4 to 8 included variables in the ML model (Table III). Several key variables were included in both models, with variation in the lesser variables.

Findings for Specific Binary Events

As an example, a clinical decision rule based on regression modeling estimating the probability of a distal radial fracture in acute wrist trauma to aid decision-making for the use of radiography included 8 variables in the logistic regression model with a c-statistic of 0.86[19] and 4 variables in the ML model with an identical c-statistic of 0.86.

As another example, a probability calculator for the risk of a surgical site infection after operative fracture care included 6 variables in the logistic regression model with a c-statistic of 0.76[12] and 6 variables in the ML model with a c-statistic of 0.75. Of the variables included in the models, 3 were identical.

Discussion

There is a growing interest in the use of ML for predicting probability estimates in many fields, and it is increasingly used in studies of musculoskeletal trauma. This study compared logistic regression and ML methodology using data sets from published studies of musculoskeletal trauma and found little or no benefit to ML. This is consistent with findings from other fields and brings into question the benefit of ML models for small, relatively simple data sets. Conversely, the ability of ML models to make comparable probability estimates using fewer and somewhat different variables might prove useful.

Limitations

We acknowledge several limitations of this study. First, statisticians consider logistic regression to be a relatively simplistic supervised ML algorithm, so our grouping of the other 5 approaches into a

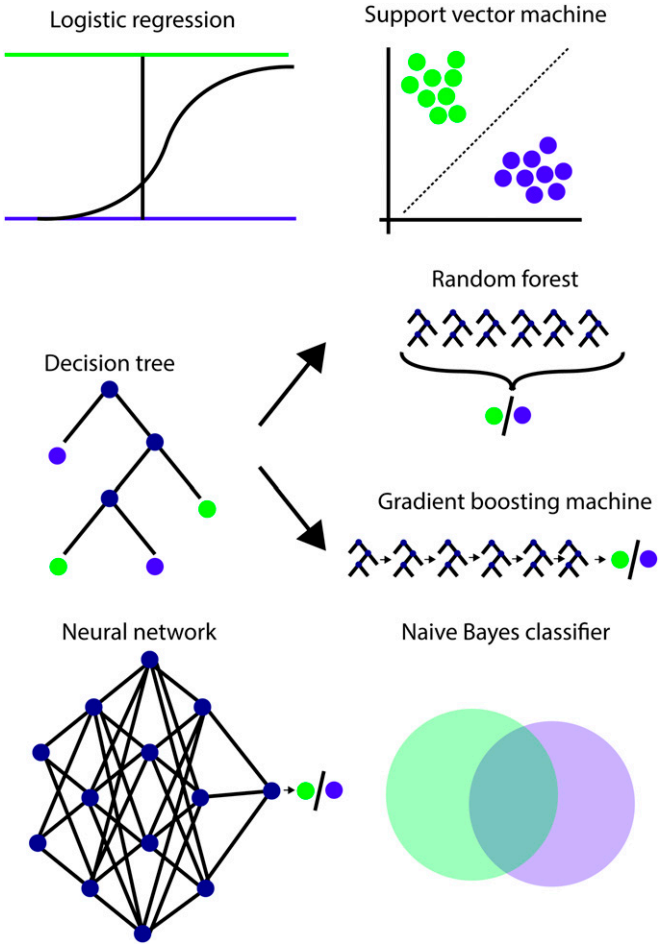

Fig. 1

Classification algorithms. Logistic regression is a calculation used to estimate the probability of binary events and involves fitting an S-shaped probability curve. Support vector machines are based on the idea of finding a hyperplane in a 3-dimensional (kernel) scatterplot that divides a data set into 2 classes and works well on smaller data sets. It is more difficult to draw 1 line in more complex data sets. Decision trees (e.g., random forest and gradient boosting machines) use flowchart-like structures to make decisions, which can be readily understood and visualized. Data points are split into similar categories at particular times (at each "branch from the tree," so-called split points). Neural networks are layers of complex regressions that are interrelated, analogous to the biological neural networks in the human brain. Neural networks benefit from large amounts of data. A naïve Bayes classifier is a product of probabilities, which works well with smaller data sets and is best with categorical events rather than continuous variables.

"machine learning" group may be a bit misleading. Readers can interpret the comparison as "traditional" versus newer methods (Fig. 1). Second, some readers might consider categorizing age, but we prefer not to categorize continuous variables. Categories assume that people below and above a certain threshold are intrinsically different, which can introduce bias and lead to inaccuracies. Categorization may also create false dichotomies with the potential to reinforce stigma[29,30]. We believe strongly that a variable measured on a continuum should be analyzed on its continuum so that information is not lost. One of the advantages of ML methods is the ability to model nonlinear associations, which are common among natural phenomena. Analyzing age on its continuum is also consistent with the original papers. Third, we only studied model development and evaluated model performance according to the discriminative ability of the algorithm. In practice, there would be an implementation phase in which the probability estimates are assessed using decision-curve analysis (a measure of the ability to make better decisions with a model than without)[26]. Fourth, there is no consensus on the quantity of data needed to develop a well-performing algorithm; the minimum sample size needed often depends on the magnitude of the association of the available variables with the event under study. Fifth, the use of a p value threshold in bivariate analysis as a basis for selecting variables for multivariable analysis is debated because p values may not represent the clinical importance of a variable[31]. An alternative approach is to preselect variables according to the Akaike information criterion (AIC)[32]. Finally, ML methods benefit from large amounts of data to capture complex nonlinear and interaction effects and may have advantages only in very large data sets[33]. Finally, in our study, odds ratios were used as a result of logistic regression analysis as our focus was on the prediction of absolute risk. When other outcomes are concerned, we note that, rather than odds ratios, relative risk estimates may be preferred for easier interpretation.

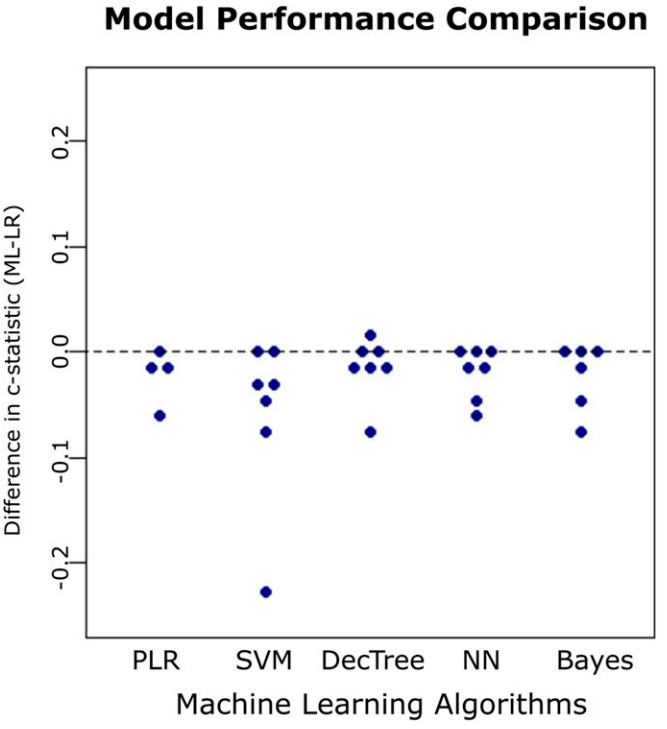

Fig. 2

A bee-swarm plot of model performance c-statistic differences (Δ_{ML-LR}). ML = machine learning; LR = logistic regression; PLR = penalized logistic regression; SVM = support vector machine; DecTree = gradient boosting machine, random forest, and boosting decision tree; NN = neural network; and Bayes = Bayes point machine and naïve Bayes.

TABLE II Logistic Regression and ML: Model Performance

Study	Algorithm	C-statistic	Calibration Slope	Calibration Intercept*	Brier*
Sobol et al.[18] (2018) and Hendrickx et al.[15] (2019)	Logistic regression	0.89	0.84	−0.28	0.12
	Bayes point machine†	0.89	1.02	−0.06	0.11
	Boosting decision tree	0.81	1.02	0.01	0.11
	Neural network	0.89	1.26	0.03	0.11
	Support vector machine	0.89	0.94	−0.02	0.11
Beks et al.[13] (2016)	Logistic regression	0.64	1.13	0.07	0.16
	Boosting decision tree	0.62	0.13	−0.82	0.23
	Neural network	0.59	6.80	7.29	0.17
	Bayes point machine	0.57	0.45	−0.70	0.17
	Penalized logistic regression	0.58	0.55	−0.56	0.17
	Support vector machine†	0.59	0.80	−0.24	0.17
Duckworth et al.[14] (2012) and Mallee et al.[16] (2020)	Logistic regression	0.76	0.93	0.13	0.17
	Boosting decision tree†	0.77	0.84	−0.01	0.16
	Penalized logistic regression	0.74	0.99	0.00	0.17
	Neural network	0.76	0.88	−0.05	0.16
	Support vector machine	0.73	0.86	−0.01	0.17
	Bayes point machine	0.72	0.92	−0.03	0.17
Walenkamp et al.[19] (2015)	Logistic regression	0.86	1.07	NA	NA
	Support vector machine†	0.86	0.85	−0.05	0.16
	Bayes point machine	0.86	0.84	−0.13	0.16
	Boosting decision tree	0.86	0.72	−0.11	0.16
	Neural network	0.86	0.80	−0.21	0.16
SPRINT Investigators[11] (2008)	Logistic regression	0.80	1.01	−0.01	0.10
	Penalized logistic regression†	0.80	0.97	0.00	0.13
	Bayes point machine	0.80	0.09	−0.02	0.13
	Boosting decision tree	0.80	0.92	0.00	0.13
	Neural network	0.79	0.77	0.12	0.14
	Support vector machine	0.77	0.89	0.00	0.14
Bachoura et al.[12] (2011)	Logistic regression	0.76	1.19	0.49	0.04
	Gradient boosting machine†	0.75	0.86	−0.41	0.04
	Support vector machine	0.54	1.77	2.31	0.04
	Neural network	0.70	0.46	−1.50	0.05
	Naïve Bayes	0.75	0.11	−2.77	0.10
	Random forest	0.66	0.32	−1.20	0.05
Oosterhoff et al.[17] (in press)	Logistic regression	0.78	0.98	−0.01	0.15
	Penalized logistic regression†	0.77	1.08	0.01	0.16
	Stochastic gradient boosting	0.77	1.04	0.00	0.16
	Random forest	0.75	0.55	0.21	0.17
	Support vector machine	0.71	1.00	0.01	0.17
	Neural network	0.77	0.97	0.02	0.16

*NA = not available. †Best-performing algorithm.

Findings

The observation that ML and logistic regression-derived probability estimates were comparable suggests that ML provides little advantage in musculoskeletal trauma. That finding is consistent with previous evidence comparing ML and logistic regression models for binary events. On the other hand, the

TABLE III Logistic Regression and ML: Variable Selection

Study	Logistic Regression*		ML*	
	No.	Included Variables	No.	Included Variables
Sobol et al.[18] (2018) and Hendrickx et al.[15] (2019)	4	Trauma mechanism Age Type of tibial fracture Location of tibial fracture	6	Trauma mechanism Age Type of tibial fracture Location of tibial fracture Type of fibular fracture† Location of fibular fracture†
Beks et al.[13] (2016)	2	Obesity Single-incision anterior approach†	5	Obesity Age† Time until surgery† Experience of surgeon† Side†
Duckworth et al.[14] (2012) and Mallee et al.[16] (2020)	4	Sex ASB pain on ulnar deviation Sports injury† Scaphoid tubercle tenderness at 2 wk†	4	Sex ASB pain on ulnar deviation Age† Trauma mechanism†
Walenkamp et al.[19] (2015)	8	Age Wrist swelling Visible deformation Tender on palpation of distal radius Sex† ASB swelling† Radial deviation pain† Thumb axial compression pain†	4	Age Wrist swelling Visible deformation Tender on palpation of distal radius
SPRINT Investigators[11] (2008)	6	Trauma mechanism Gustilo-Anderson classification Tscherne classification AO/OTA fracture classification Sex† Smoking†	7	Trauma mechanism Gustilo-Anderson classification Tscherne classification AO/OTA fracture classification Location of tibial fracture† Polytrauma† Postop. fracture gap†
Bachoura et al.[12] (2011)	6	Drain No. of operations Diabetes Congestive heart failure† Tibial shaft and/or plateau fracture† Elbow fracture†	6	Drain No. of operations Diabetes Wound classification† Preop. hospital stay† Previous external fixation†
Oosterhoff et al.[17] (in press)	17	Age BMI ASA Functional status Preop. delirium Preop. dementia Preop. mobility aid Bleeding disorder† Diabetes† Dyspnea Sex† Medical comanagement† Preop. bone medication† Preop. hematocrit† Preop. platelets† Systemic inflammatory response syndrome† Wound infection†	8	Age BMI ASA Functional status Preop. delirium Preop. dementia Preop. mobility aid Preop. creatinine level†

*ASB = anatomic snuff box, BMI = body mass index, and ASA = American Society of Anesthesiologists classification. †Variability between the predictive variables included in the logistic regression and ML algorithm.

mean difference in the c-statistic of 0.01 is modest but was considered relevant in studies that evaluated the incremental value of biomarkers for probability estimates[34,35].

The observation that different variables were included in the ML and logistic regression algorithms is likely a demonstration of the complexity of probability estimation and may also be interpreted as cautioning against overreliance on specific variables. These findings support the use of principles, experience, and judgment to select variables thought to be clinically meaningful. Starting with a more limited set of variables limits the potential for overfitting[36,37]. The strategy of thoughtful preselection of variables may increase generalizability of the probabilities estimates, may be easier to interpret, and may seem more clinically relevant, thereby balancing model fit and applicability in clinical practice. One principle for preselection might be to favor factors that can be modified either before or after surgery.

Conclusions

We found that the accuracy of more complex statistical ML models was comparable with that of logistic regression for binary events, but fewer and somewhat different variables were used. In our opinion, this supports a pragmatic approach favoring preselection of clinically relevant variables (perhaps modifiable health factors) in the development of clinical prediction models in orthopaedic surgery. Probability estimates also need validation in different time periods and settings. Validated algorithms for probability estimation could be a software add-on to an electronic health record, with automatic calculation and recording for decision support[38].

Appendix

eA Supporting material provided by the authors is posted with the online version of this article as a data supplement at jbjs.org (http://links.lww.com/JBJS/G832). ■

NOTE: The Machine Learning Consortium includes Mohit Bhandari, MD, PhD, FRCSC; Reinier Beks, MD; Anne Eva Bulstra, MD; Sofia Bzovsky, MSc; J. Carel Goslings, MD, PhD; Gordon Guyatt, MD, FRCP(C), OC; Laurent Hendrickx, MD; David Langerhuizen, MD; Wouter H. Mallee, MD, PhD; Rob Nelissen, MD, PhD; Rudolf Poolman, MD, PhD; Vincent Stirler, MD, PhD; Paul Tornetta III, MD; Emil H. Schemitsch, MD, FRCSC; Inger B. Schipper, MD, PhD; Marc Swiontkowski, MD, FAOA; David Sanders, MD; Sheila Sprague, PhD; and Stephen D. Walter, PhD.

Jacobien H.F. Oosterhoff, MD[1,2,3]
Benjamin Y. Gravesteijn, MSc[4]
Aditya V. Karhade, MD, MBA[1]
Ruurd L. Jaarsma, MD, PhD, FRACS, FAOrthA[3]
Gino M.M.J. Kerkhoffs, MD, PhD[2]
David Ring, MD, PhD[5]
Joseph H. Schwab, MD, MS[1]
Ewout W. Steyerberg, MSc, PhD[4]
Job N. Doornberg, MD, PhD[3,6]

[1]Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

[2]Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands

[3]Department of Orthopaedic & Trauma Surgery, Flinders Medical Centre, Flinders University, Adelaide, South Australia, Australia

[4]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

[5]Department of Surgery and Perioperative Care, Dell Medical School, University of Texas, Austin, Texas

[6]Department of Orthopaedic Surgery, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands

Email for corresponding author: david.ring@austin.utexas.edu

References

1. Harrell FE. Regression Modeling Strategies. Springer; 2006.
2. Steyerberg EW. Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer; 2019. Study design for prediction modeling;p; p 37-56.
3. Liu NT, Salinas J. Machine Learning for Predicting Outcomes in Trauma. Shock. 2017 Nov;48(5):504-10.
4. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min Knowl Discov. 1998;2(2):121-67.
5. Jain A, Mao K, Mohiuddin K. Artificial neural networks: a tutorial. IEEE Comput Soc. 1996;29(3):31-44.
6. Afanador N, Smolinska A, Tran T, Blanchet L. Unsupervised random forest: a tutorial with case studies. J Chemometr. 2016;30(5):232-41.
7. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013 Dec 4;7:21.
8. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019 Jun; 110:12-22.
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.
10. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec 16;18(12):e323.
11. Study to Prospectively Evaluate Reamed Intramedullary Nails in Patients with Tibial Fractures (SPRINT) Investigators. Randomized trial of reamed and unreamed intramedullary nailing of tibial shaft fractures. J Bone Joint Surg Am. 2008 Dec; 90(12):2567-78.
12. Bachoura A, Guitton TG, Smith RM, Vrahas MS, Zurakowski D, Ring D. Infirmity and injury complexity are risk factors for surgical-site infection after operative fracture care. Clin Orthop Relat Res. 2011 Sep;469(9):2621-30.
13. Beks RB, Claessen FMAP, Oh LS, Ring D, Chen NC. Factors associated with adverse events after distal biceps tendon repair or reconstruction. J Shoulder Elbow Surg. 2016 Aug;25(8):1229-34.
14. Duckworth AD, Buijze GA, Moran M, Gray A, Court-Brown CM, Ring D, McQueen MM. Predictors of fracture following suspected injury to the scaphoid. J Bone Joint Surg Br. 2012 Jul;94(7):961-8.
15. Hendrickx LAM, Cain ME, Sierevelt IN, Jadav B, Kerkhoffs GMMJ, Jaarsma RL, Doornberg JN. Incidence, Predictors, and Fracture Mapping of (Occult) Posterior Malleolar Fractures Associated with Tibial Shaft Fractures. J Orthop Trauma. 2019 Dec;33(12):e452-8.
16. Mallee WH, Walenkamp MMJ, Mulders MAM, Goslings JC, Schep NWL. Detecting scaphoid fractures in wrist injury: a clinical decision rule. Arch Orthop Trauma Surg. 2020 Apr;140(4):575-81.
17. Oosterhoff JHF, Karhade AV, Oberai T, Doornberg JN, Schwab JH. Development of machine learning algorithms for prediction of postoperative delirium in elderly hip fracture patients. Geriatr Orthop Surg Rehabil. 2021. In press.
18. Sobol GL, Shaath MK, Reilly MC, Adams MR, Sirkin MS. The Incidence of Posterior Malleolar Involvement in Distal Spiral Tibia Fractures: Is it Higher than We Think? J Orthop Trauma. 2018 Nov;32(11):543-7.

19. Walenkamp MMJ, Bentohami A, Slaar A, Beerekamp MS, Maas M, Jager LC, Sosef NL, van Velde R, Ultee JM, Steyerberg EW, Goslings JC, Schep NW. The Amsterdam wrist rules: the multicenter prospective derivation and external validation of a clinical decision rule for the use of radiography in acute wrist trauma. BMC Musculoskelet Disord. 2015 Dec 18;16(1):389.

20. Machine Learning Consortium, on behalf of the SPRINT and FLOW Investigators. A Machine Learning Algorithm to Identify Patients with Tibial Shaft Fractures at Risk for Infection After Operative Treatment. J Bone Joint Surg Am. 2021 Mar 17;103(6): 532-540.

21. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Software. 2011;45(3):1-67.

22. Gravesteijn BY, Sewalt CA, Venema E, Nieboer D, Steyerberg EW; CENTER-TBI Collaborators. Missing Data in Prediction Research: A Five-Step Approach for Multiple Imputation, Illustrated in the CENTER-TBI Study. J Neurotrauma. 2021 Jun 1; 38(13):1842-57.

23. Oosterhoff JHF, Doornberg JN; Machine Learning Consortium. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. EFORT Open Rev. 2020 Oct 26;5(10):593-603.

24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010 Jan;21(1):128-38.

25. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958;45(3-4):562-5.

26. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014 Aug 1;35(29): 1925-31.

27. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2015 Feb;35(2):162-9.

28. R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation; 2013.

29. Seann S. Dichotomania: An Obsessive Compulsive Disorder That Is Badly Affecting the Quality of Analysis of Pharmaceutical Trials. In: Proceedings of the 55th Session of the International Statistical Institute. Sydney, Australia, 2005. International Statistical Institute; 2006.

30. Harrell FE Jr. How To Do Bad Biomarker Research. Presented at the Vanderbilt Center for Quantitative Sciences Workshop, Nashville, Tennessee, 2015 Jan 16.

31. Rodgers JL. The epistemology of mathematical and statistical modeling: a quiet methodological revolution. Am Psychol. 2010 Jan;65(1):1-12.

32. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected Papers of Hirotugu Akaike. Springer; 1998. p 199-213.

33. 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 Dec 22;14(1):137.

34. Retel Helmrich IRA, Lingsma HF, Turgeon AF, Yamal JM, Steyerberg EW. Prognostic Research in Traumatic Brain Injury: Markers, Modeling, and Methodological Principles. J Neurotrauma. 2021 Sep 15;38(18):2502-13.

35. Baker SG, Schuit E, Steyerberg EW, Pencina MJ, Vickers A, Moons KG, Mol BW, Lindeman KS. How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through. Stat Med. 2014 Sep 28; 33(22):3946-59.

36. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res. 2020 Nov;29(11):3166-78.

37. Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, Collins GS. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. J Clin Epidemiol. 2021 Apr;132:88-96.

38. Karhade AV, Schwab JH, Del Fiol G, Kawamoto K. SMART on FHIR in spine: integrating clinical prediction models into electronic health records for precision medicine at the point of care. Spine J. 2021 Oct;21(10):1649-51.