Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review

Paul T Ogink, Olivier Q Groot, Aditya V Karhade, Michiel E R Bongers, F Cumhur Oner, Jorrit-Jan Verlaan & Joseph H Schwab

Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review

Paul T OGINK [1], Olivier Q GROOT [2], Aditya V KARHADE [2], Michiel E R BONGERS [2], F Cumhur ONER [1], Jorrit-Jan VERLAAN [1], and Joseph H SCHWAB [2]

[1] Department of Orthopedic Surgery, University Medical Center Utrecht – Utrecht University, Utrecht, The Netherlands; [2] Department of Orthopedic Surgery, Orthopedic Oncology Service, Massachusetts General Hospital – Harvard Medical School, Boston, USA
Correspondence: ptogink@gmail.com

Background and purpose — Advancements in software and hardware have enabled the rise of clinical prediction models based on machine learning (ML) in orthopedic surgery. Given their growing popularity and their likely implementation in clinical practice we evaluated which outcomes these new models have focused on and what methodologies are being employed.

Material and methods — We performed a systematic search in PubMed, Embase, and Cochrane Library for studies published up to June 18, 2020. Studies reporting on non-ML prediction models or non-orthopedic outcomes were excluded. After screening 7,138 studies, 59 studies reporting on 77 prediction models were included. We extracted data regarding outcome, study design, and reported performance metrics.

Results — Of the 77 identified ML prediction models the most commonly reported outcome domain was medical management (17/77). Spinal surgery was the most commonly involved orthopedic subspecialty (28/77). The most frequently employed algorithm was neural networks (42/77). Median size of datasets was 5,507 (IQR 635–26,364). The median area under the curve (AUC) was 0.80 (IQR 0.73–0.86). Calibration was reported for 26 of the models and 14 provided decision-curve analysis.

Interpretation — ML prediction models have been developed for a wide variety of topics in orthopedics. Topics regarding medical management were the most commonly studied. Heterogeneity between studies is based on study size, algorithm, and time-point of outcome. Calibration and decision-curve analysis were generally poorly reported.

Surgical decision-making in orthopedic surgery involves weighing the benefits of an intervention against its inherent risks. Prognostic scoring tools have been devised to individualize risk prediction and thus improve surgical decision-making (Janssen et al. 2015, Pereira et al. 2016, Shah et al. 2018). Although clinical prediction models are not new, recent advancements in artificial intelligence have created a host of prediction models based on machine learning (ML) (Cabitza et al. 2018).

ML is a branch of artificial intelligence that enables computer algorithms to learn from experience from large datasets without explicit programming. Figure 1 shows 3 commonly employed algorithms. Existing reviews of machine learning studies have provided a broad overview of applications ranging from vision to natural language processing and predictive analytics (Cabitza et al. 2018). To our knowledge, there is no study that has critically assessed the body of studies focused on ML prediction models for surgical outcome in orthopedics. These types of prediction models are most likely the first branch of artificial intelligence to be employed in clinical practice (Staartjes et al. 2020). Therefore, familiarizing practicing orthopedic surgeons with ML's concepts and the topics these new methods have focused on can optimize their implementation in clinic.

As such, the purpose of this systematic review is to (1) evaluate which surgical outcomes orthopedic clinical prediction models have focused on, and (2) determine which techniques current prediction models use for development and validation.

Material and methods

Systematic literature search

Adhering to the 2009 PRISMA guidelines a systematic search was performed in PubMed, Embase, and the Cochrane Library for articles published up to June 18, 2020. 2 different domains
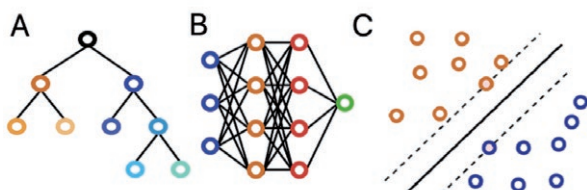
Figure 1. (A) Decision trees are hierarchical structures in which each node performs a test on the input value with the subsequent branches representing the outcomes. Their graphical representation as seen here makes them easy to understand and interpret. However, they are prone to overfitting. (B) Neural networks are based on interconnected nodes. The input features are represented by the first (blue) layer. The designated outcome is represented by the final (green) layer. The middle, hidden layers (blue and orange) base their output on the input they get from prior layers. Neural networks have been around for a long time and offer good discriminative abilities, but interpretation of the relationships between the different layers remains difficult. (C) Support vector machines (SVMs) perform classification by determining the optimal separating hyperplane between datapoints, which maximizes the distance between the 2 closest points of either group. They can be used for both linear and nonlinear relationships. While they remain effective in data with a great number of features, they do not work well in larger datasets.

Figure 2. Flowchart of study inclusions and exclusions.

of medical subject headings (MeSH) terms and keywords were combined with "AND" and within the 2 domains the terms were combined with "OR." The 1st domain included words related to ML and the second domain related to possible orthopedic specialties (Appendix 1, see Supplementary data). Terms were restricted to MeSH, title, abstract, and keywords. Two reviewers (PTO, OQG) independently screened all titles and abstracts for eligible articles based on predefined criteria. Eligible full-text articles were evaluated and cross-referenced for potentially relevant articles not identified by the initial search (Figure 2). Discrepancies between the 2 reviewers were adjudicated by the senior author (JHS).

Eligibility criteria
Studies reporting on ML-based prediction models addressing orthopedic surgical outcomes were included, as were all intraoperative and postoperative outcomes. The surgical orthopedic population was defined as disorders of the bones, joints, ligaments, tendons, or muscles treated by any type of operation. Excluded were studies (1) that did not include at least 1 ML-based prediction models for surgical outcome (e.g., logistic regression-based models), (2) non-English studies, (3) lack of full text, and (4) non-relevant study types such as animal studies, letters to the editors, and case reports.

Assessment of methodological quality
Quality assessment was performed based on a modified nine-item Methodological Index for Non-Randomized Studies (MINORS) checklist (Slim et al. 2003). We made it applicable to our systematic review by including disclosure, study aim, input feature, output feature, validation method, dataset dis-
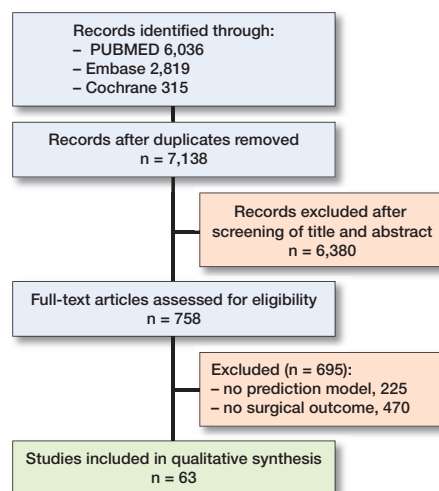
tribution, performance metric, and explanation of the used AI model (Langerhuizen et al. 2019). These 9 items were scored on a binary scale: 0 (not reported or unclear) and 1 (reported and adequate).

Data extraction
Table 1 lists the data we extracted from each study. For this review, 6 main orthopedic surgical outcome domains were identified, consisting of (1) intraoperative complications (e.g., blood transfusion, prolonged operative time), (2) postoperative complications (e.g., venous thromboembolism), (3) survival, (4) patient reported outcome measures (PROMs), (5) medical management (e.g., hospitalization), and (6) other. For studies reporting the performance of multiple ML models, the best performing ML model was used. 13 studies provided multiple models for multiple surgical outcomes; these were extracted separately resulting in more ML models than studies. Only the 2 performance measures AUC and accuracy were extracted as they were most the commonly reported results.

Study characteristics
After screening of titles and abstracts, 758 full-text articles were assessed for eligibility and ultimately 59 articles were included reporting on 77 ML prediction models (Table 1). Median sample size was 5,818 (IQR 635–26,869). Using the MINORS criteria, all 59 articles were found to be of similar quality. All included a minimum of 8 out of 9 appraisal items (Appendix 2, see Supplementary data).

Statistics
AUC scores and accuracies in tables are expressed as they were originally reported. For studies that reported multiple results within a single outcome domain (e.g., multiple different postoperative PROMs, each with an independent AUC) averages were taken. The sizes of the training, validation, and

Table 1. Data extracted from each study

1	Year of publication
2	First author
3	Disease condition
4	Type of surgery
5	Input feature
6	Number of features in final model
7	Type of outcome
8	Time points of outcome
9	Number of output classes
10	ML algorithm used
11	Number of patients
12	Distribution between training, validation, and test set
13	Validation method
14	AUC and accuracy of model
15	Reporting of calibration and Brier score
16	Decision-curve analysis
17	Digital application of the model

Table 3. Characteristics of studies (n = 77). Values are count (%) unless otherwise specified

Sample size, median (IQR)	5,818 (635–26,364)
Predictors included in final model, median (IQR) [a]	10 (8–15)
Outcome domain	
Medical management	17 (22)
Survival	16 (21)
Complication	15 (19)
PROMs	12 (16)
Intraoperative complication	3 (3.9)
Other	14 (18)
Orthopedic subspecialty	
Spine	28 (36)
Arthroplasty	21 (27)
Trauma	13 (17)
Oncology	6 (7.8)
Other	9 (12)
National/Registry database [b]	42 (55)
Split sample	
70:30	22 (29)
80:20	19 (25)
Other	36 (46)
ML algorithm [c]	
Neural network	42 (55)
Single layer	36 (47)
Deep learning	6 (8)
Convolutional	1 (1)
Random forest	30 (39)
Support vector machine	19 (25)
Naive Bayes	11 (14)
Stochastic gradient boosting	10 (13)
Performance metric [c]	
AUC	74 (96)
Accuracy	39 (51)
Brier score	23 (30)
Calibration	26 (34)
Model explanation	
Global	34 (44)
Local	17 (22)
Decision curve analysis	14 (18)
Digital application available	18 (23)

AUC = area under the curve, IQR = interquartile range, ML = machine learning, PROM = patient reported outcome measure.
[a] Amount of predictors that were included in the final, best performing machine learning algorithm. In 16% (13/81) this could not be extracted from the study or was unclear.
[b] This includes databases such as Surveillance, Epidemiology, and End Results (SEER) or American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).
[c] Not mutually exclusive.

test sets are reported as percentages of the total dataset. No meta-analysis was performed because of obvious heterogeneity between studies and in orthopedic applications. However, to summarize the findings in some quantitative form, the median AUC and accuracy of the prediction performance were calculated for all studies.

We used Microsoft Excel (Version 16.31; Microsoft Inc, Redmond, WA, USA) for standardized forms for data extraction and quality assessment, and Mendeley as reference management software.

Ethics, funding, and potential conflicts of interests

Institutional review board approval was not required for this systematic review. No external funding was received. The authors have no conflicts of interest to declare.

Results

Study design

Table 2 lists the characteristics of all included studies. More than half of the 77 models were developed with data from national databases or registries (42) (Table 3). The median number of predictor variables used in the ML model was 10 (IQR 8–15). Models using national data did not include more variables: 10 (IQR 8–13). 68 of the models had a binary distribution of the outcome variable. Most frequently employed algorithms were neural networks (42) and random forests (30). 36 of the neural networks were single-layer, 5 deep learning, and 1 convolutional. The median number of patients used was 5,507 (IQR 635–26,364). Median AUC was 0.80 (IQR 0.73–0.86) and median accuracy was 79% (IQR 75–88). Calibration was reported for 26 of the models and 23 provided Brier scores. Decision-curve analysis was employed in 14 studies. 18 provided a digital application for their prediction model.

Outcome

The most commonly reported outcome domains were medical management (17) and survival (16). Medical management mostly focused on discharge destination (7) and hospitalization (4). The studies on survival all addressed patient survival. 6 survival studies were in orthopedic oncology and 5 in orthopedic trauma. Both medical management and survival had a higher median AUC (0.82 and 0.84 than overall median AUC). Spinal surgery was the most commonly involved subspecialty (28).

Discussion

Recent years have seen an increasing interest in artificial intelligence and ML in orthopedics (Bini 2018, Jayakumar et al. 2019). With this systematic review we aimed to provide an introduction to the main concepts of developing ML models for orthopedic surgeons and analyze the current application and design of these models in orthopedic surgery. We found a wide range of potential applications ranging from predicting survival in spinal metastases, clinical outcome after shoulder arthroplasty, and hospitalization after hip fracture surgery.

This systematic review has a number of limitations. 1st, due to the relative novelty of this field of research in orthopedic surgery, the variety in study designs renders comparisons and comprehensive quantitative analysis difficult. We therefore opted to perform a qualitative analysis of the current publications. Hopefully, the increasing familiarity with these types of studies will lead to better reporting and open up the possibility to perform quantitative analyses. 2nd, this review is likely influenced by publication bias. ML prediction models with good performance are more likely to be published than models with mediocre or poor performance. This positive publication bias has been shown both in medicine and computational sciences (Boulesteix et al. 2015). The performance measures presented here were therefore likely to be more favorable than those of all developed models. 3rd, despite our efforts to perform a search across multiple online libraries, we have missed a number of studies reporting ML prediction models. Whilst unfortunate, we do no not think these omissions will significantly alter our findings on research topics or most utilized methodology as this review included nearly 60 studies.

This systematic review shows that ML models have been developed for a wide variety of topics across all subspecialties within orthopedics. Perhaps surprisingly, medical management was the most studied domain with the majority of models focusing on readmissions and discharge placement. Both readmissions and discharge delays impose a heavy burden on healthcare costs (Wan et al. 2016). Healthcare expenditure has risen steadily throughout the developed world in recent decades (OECD 2019). While there is enormous variation in healthcare systems, government institutions in virtually all countries have looked at improving medical management to help curb costs (Schwierz 2016). Papanicolas et al. (2018) found activities relating to planning, regulating, and managing health services was a major factor in the difference in healthcare expenditure between the United States and 10 other high-income countries. Shrank et al. (2019) concluded failure of care coordination, leading to unnecessary readmissions among other things, amounts to $78 billion of waste in the United States. To address this problem the Centers for Medicare and Medicaid Services started the Hospital Readmissions Reduction Program in 2012, incentivizing hospitals to lower readmission rates. Knowing in advance

which patients are at risk of being readmitted within 30 days after discharge is crucial, which is a possible explanation as to why so many prediction models focus on this topic. Similarly, knowing in advance where patients are likely to be discharged to makes preventing delayed discharge a lot easier than the other interventions tried over the years (Bryan 2010, Ou et al. 2011). Furthermore, the databases available in the studies on medical management appear to be larger, enabling researchers to include more variables and create better performing prediction models. These models are more likely to be published as evidenced by the higher AUC for medical management compared to overall AUC.

Survival was the other commonly studied outcome domain. Accurately estimating remaining life-expectancy is an important feature in medical decision-making in orthopedic oncology (Pereira et al. 2016). In a patient group with only limited life-span remaining, the aim of treatment is to preserve quality of life. Accurate survival estimations can guide decision-making on whether or not to perform surgery and if so, which operative treatment should be opted for (Quinn et al. 2014). With an ageing population and cancer patients surviving longer, the incidence of bone metastases will continue to rise and prediction models will likely play an increasing role in this field (Quinn et al. 2014).

The AAOS Census 2018 showed only 8.3% of orthopedic surgeons' primary specialty area was the spine, while one-third of the prediction models were linked to spinal surgery (AAOS Department of Clinical Quality and Value 2019). Cost reduction may also be the driving factor in the overrepresentation of spinal surgery prediction models; the economic cost of spinal surgery is large and growing with spinal fusions alone costing $30 billion annually in the United States (Johnson and Seifi 2018). Prediction models could play a role in curbing costs by improving patient selection and surgical decision-making, although this could be said for all other subspecialties. Another possible explanation for the disproportionate number is the overlap with neurosurgery. The neurosurgical field was relatively quicker to use ML to develop prediction models and had developed several models in spinal surgery earlier on (Senders et al. 2018). Finally, the field of prediction models is expanding but still small. A significant proportion of the prediction models are developed by a few research groups that happen to focus on spine surgery. With the field expanding as fast as it is with new prediction models being published every month, we expect the overrepresentation of spine surgery to be temporary in a field in its infancy.

While there is wide variation in study design, certain study design elements are fairly similar across most studies. The most common designs comprise binary outcomes; either a 70:30 or 80:20 split between training and test set; and 10-FCV as method of internal validation. Wide variety exists in study size, time-point of outcome, and choice of ML algorithms. Study size is mostly defined by whether a national database or registry was used for model development. These quality improvement

databases offer a large number of datapoints with a variety of variables of a diverse group of hospitals, enabling the creation of prediction models. However, these databases are sometimes flawed by errors and their generalizability is also yet to be assessed (Rolston et al. 2017). External validation remains crucial considering generalizability outside the geographical origin of the database is not ensured (Janssen et al. 2018). Institutional databases offer the advantage of more veracious data, for instance including PROM data, which can extend over longer periods of time, but often lack adequate size.

Which ML algorithm is chosen seems highly random. While studies do list the pros and cons of certain algorithms, no study elaborates on why those algorithms were specifically chosen. A potential reason neural networks and random forests are selected so often is the familiarity of these algorithms. Neural networks have been around for decades, but were limited by lagging computational power (Hopfield 1988). The increase in computational power has led to a significant expansion of what neural networks can process and scientists have been able to build on the work of previous decades (Schmidhuber 2015). Future research should report on multiple ML algorithms and provide the performance measures of all models, thus enabling comparison between different approaches.

Despite the importance of performance metrics, a mere one-third of prediction models included information on calibration, similar to prior studies assessing prediction models in multiple medical domains (Bouwmeester et al. 2012, Heus et al. 2018). Calibration is important to evaluate wehther the model is under- or overestimating the risk regardless of the discriminative abilities. Systematically underestimating risk can lead to undertreatment, while overestimating risk can cause overtreatment (Van Calster and Vickers 2015, Van Calster et al. 2019). To improve the quality of reporting of clinical prediction models, Collins et al. (2015) published the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement. While not tailored for ML prediction models this guideline can provide a framework for researchers to use during development. Hopefully, a more widespread adaptation of the TRIPOD statement can lead to less variation in study designs and better reporting of performance metrics.

Only one-fifth of prediction models have a digital application available. The purpose of prediction models is to aid clinicians and patients in decision-making, which can be achieved only if the models are available for use. Otherwise, predictive analytics based on ML will remain a mere theoretical exercise. Furthermore, researchers should be encouraged to not only provide a digital application of their prediction model, but share their code as well. With a field in its infancy, providing code of more experienced researchers can guide beginning research groups in their endeavors. Additionally, this can greatly increase the small number of external validation studies being performed.

In conclusion, ML prediction models have been developed for a wide variety of topics in orthopedic surgery. Topics regarding medical management and survival were the most commonly studied and spine surgery was the most involved subspecialty. Heterogeneity between studies is mostly based on study size, choice of ML algorithm, and time-point of outcome. Most published prediction models showed fair to good discriminative abilities, while calibration was poorly reported. Future studies should preferably include more multi-institutional, prospective databases and develop multiple models enabling comparison between different ML approaches. Also, important performance measures such as calibration should be reported to evaluate the prediction model accurately.

Supplementary data

Table 2 and appendices 1 and 2 are available as supplementary data in the online version of this article, http://dx.doi.org/10.1080/17453674.2021.1932928

AAOS Department of Clinical Quality and Value. Orthopaedic Practice in the US 2018. 2019 (January): 1-68.

Bini S A. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? J Arthroplasty 2018; 33(8): 2358-61.

Boulesteix A L, Stierle V, Hapfelmeier A. Publication bias in methodological computational research. Cancer Inform 2015; 14(Suppl. 5): 11-19.

Bouwmeester W, Zuithoff N P A, Mallett S, Geerlings M I, Vergouwe Y, Steyerberg E W, Altman D G, Moons K G M. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012; 9(5): e1001221. https://doi.org/10.1371/journal.pmed.

Bryan K. Policies for reducing delayed discharge from hospital. Br Med Bull 2010; 95(1): 33-46.

Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 2018; 6: 75.

Collins G S, Reitsma J B, Altman D G, Moons K G M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Eur Urol 2015; 67(6): 1142-51.

Heus P, Damen J A A G, Pajouheshnia R, Scholten R J P M, Reitsma J B, Collins G S, Altman D G, Moons K G M, Hooft L. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 2018; 16(1): 120.

Hopfield J J. Artificial neural networks. IEEE Circuits Devices 1988; 4(5): 3-10.

Janssen S J, van der Heijden A S, van Dijke M, Ready J E, Raskin K A, Ferrone M L, Hornicek F J, Schwab J H. 2015 Marshall Urist Young Investigator Award: Prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? Clin Orthop Relat Res 2015; 473(10): 3112-21.

Janssen D M C, van Kuijk S M J, D'Aumerie B B, Willems P C. External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. J Orthop Surg Res 2018; 13(1): 114.

Jayakumar P, Moore M L G, Bozic K J. Value-based healthcare: can artificial intelligence provide value in orthopaedic surgery? Clin Orthop Relat Res 2019; 477(8): 1777-80.

Johnson W C, Seifi A. Trends of the neurosurgical economy in the United States. J Clin Neurosci 2018; 53(2018): 20-6.

Langerhuizen D W G, Janssen S J, Mallee W H, Van Den Bekerom M P J, Ring D, Kerkhoffs G M M J, Jaarsma R L, Doornberg J N. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. Clin Orthop Relat Res 2019; 477(11): 2482-91.

OECD 2019. Health at a Glance 2019. Available at: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2019_4dd50c09-en.

Ou L, Chen J, Young L, Santiano N, Baramy L-S, Hillman K. Effective discharge planning: timely assignment of an estimated date of discharge. Aust Heal Rev 2011; 35(3): 357.

Papanicolas I, Woskie L R, Jha A K. Health care spending in the United States and other high-income countries. JAMA 2018; 319(10): 1024-39.

Pereira N R P, Janssen S J, Van Dijk E, Harris M B, Hornicek F J, Ferrone M L, Schwab J H. Development of a prognostic survival algorithm for patients with metastatic spine disease. J Bone Joint Surg Am 2016; 98(21): 1767-76.

Quinn R H, Randall R L, Benevenia J, Berven S H, Raskin K A. Contemporary management of metastatic bone disease: tips and tools of the trade for general practitioners. Instr Course Lect 2014; 63: 431-41.

Rolston J D, Han S J, Chang E F. Systemic inaccuracies in the National Surgical Quality Improvement Program database: implications for accuracy and validity for neurosurgery outcomes research. J Clin Neurosci 2017; 37(2017): 44-7.

Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks 2015: 85-117.

Schwierz C. Cost-containment in the European Union 2016. Https://ec.europa.eu/info/publications/economy-finance/cost-containment-policies-hospital-expenditure-european-union_en

Senders J T, Staples P C, Karhade A V, Zaki M M, Gormley W B, Broekman M L D, Smith T R, Arnaout O. Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg 2018; 109: 476-486.e1.

Shah A A, Ogink P T, Nelson S B, Harris M B, Schwab J H. Nonoperative management of spinal epidural abscess: development of a predictive algorithm for failure. J Bone Joint Surg Am 2018; 100(7): 546-55.

Shrank W H, Rogstad T L, Parekh N. Waste in the US health care system: estimated costs and potential for savings. JAMA 2019; 322(15): 1501-9.

Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (Minors): development and validation of a new instrument. ANZ J Surg 2003; 73(9): 712-16.

Staartjes V E, Stumpo V, Kernbach J M, Klukowska A M, Gadjradj P S, Schröder M L, Veeravagu A, Stienen M N, van Niftrik C H B, Serra C, Regli L. Machine learning in neurosurgery: a global survey. Acta Neurochir (Wien) 2020; 162(12): 3081-91.

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

Van Calster B, McLernon D J, Van Smeden M, Wynants L, Steyerberg E W, Bossuyt P, Collins G S, MacAskill P, McLernon D J, Moons K G M, Steyerberg E W, Van Calster B, Van Smeden M, Vickers A J. Calibration: the Achilles heel of predictive analytics. BMC Med 2019; 17(1): 1-7.

Wan H, Zhang L, Witz S, Musselman K J, Yi F, Mullen C J, Benneyan J C, Zayas-Castro J L, Rico F, Cure L N, Martinez D A. A literature review of preventable hospital readmissions: [receding the Readmissions Reduction Act. IIE Trans Healthc Syst Eng 2016; 6(4): 193-211.

Supplementary data

APPENDIX 1: Search syntaxes for the PubMed, Embase, and Cochrane databases

PubMed: June 18, 2020—6,036 hits

(("Foot"[Mesh] OR "Ankle"[Mesh] OR "Knee Joint"[Mesh] OR "Knee"[Mesh] OR "Ankle Joint"[Mesh] OR "Hip"[Mesh] OR "Hip Joint"[Mesh] OR "Hip Prosthesis"[Mesh] OR "Hip Fractures"[Mesh] OR "Shoulder Joint"[Mesh] OR "Shoulder"[Mesh] OR "Shoulder Fractures"[Mesh] OR "Shoulder Dislocation"[Mesh] OR "Elbow"[Mesh] OR "Elbow Joint"[Mesh] OR "Wrist Joint"[Mesh] OR "Spine"[Mesh] OR "Intervertebral Disc Degeneration"[Mesh] OR "Bone Neoplasms"[Mesh] OR "Arthroplasty"[Mesh] OR "Fractures, Bone"[Mesh] OR "Orthopedics"[Mesh] OR "Foot"[Tiab] OR "Ankle"[Tiab] OR Knee[Tiab] OR Hip[Tiab] OR "Shoulder"[Tiab] OR Elbow[Tiab] OR Wrist[Tiab] OR Spina*[Tiab] OR Spine*[tiab] OR "degenerative disc"[Tiab] OR "Bone Neoplasms"[Tiab] OR Arthroplast*[Tiab] OR Fractur*[Tiab] OR Orthop*[Tiab])) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Neural Networks Computer"[Mesh] OR "Deep Learning"[Mesh] OR "support vector machine"[MeSH Terms] OR "support vector machine"[All Fields] OR "Support Vector Machine"[Mesh] OR naive bayes[tiab] OR "bayesian learning"[tiab] OR neural network*[tiab] OR "support vector"[tiab] OR support vectors[tiab] OR random forest[tiab] OR "deep learning"[tiab] OR "machine prediction"[tiab] OR "machine intelligence"[tiab] OR "computational intelligence"[tiab] OR "computational learning"[tiab] OR "computer reasoning"[tiab] OR "machine learning"[tiab] OR convolutional network*[tiab] OR "artificial intelligence"[tiab])

Embase: June 18, 2020—2,819 hits

('foot'/exp/mj OR 'ankle'/exp/mj OR 'knee'/exp/mj OR 'hip'/exp/mj OR 'hip prosthesis'/exp/mj OR 'hip fracture'/exp/mj OR 'shoulder'/exp/mj OR 'shoulder fracture'/exp/mj OR 'shoulder dislocation'/exp/mj OR 'elbow'/exp/mj OR 'wrist'/exp/mj OR 'spine'/exp/mj OR 'intervertebral disk disease'/exp/mj OR 'bone tumor'/exp/mj OR 'arthroplasty'/exp/mj OR 'fracture'/exp/mj OR 'orthopedic surgery'/exp/mj OR foot:ab,ti OR ankle:ab,ti OR knee:ab,ti OR hip:ab,ti OR shoulder:ab,ti OR spine:ab,ti OR 'degenerative disc':ab,ti OR elbow:ab,ti OR wrist:ab,ti OR 'bone tumor':ab,ti OR arthroplasty:ab,ti OR fractur:ab,ti OR orthop:ab,ti) AND ('artificial intelligence'/exp/mj OR 'machine learning'/exp/mj OR 'supervised machine learning'/exp/mj OR 'artificial neural network'/exp/mj OR 'deep learning'/exp/mj OR 'support vector machine'/exp/mj OR 'bayesian learning'/exp/mj OR 'neural network':ab,ti OR 'naive bayes':ab,ti OR 'beyesian learning':ab,ti OR 'support vector':ab,ti OR 'support vectorts':ab,ti OR 'random forest':ab,ti OR 'deep learning':ab,ti OR 'machine prediction':ab,ti OR 'machine intelligence':ab,ti OR 'computational intelligence':ab,ti OR 'computer learning':ab,ti OR 'computer reasoning':ab,ti OR 'machine learning':ab,ti OR 'convolutional network':ab,ti OR 'artificial intelligence':ab,ti)

Cochrane: June 18, 2020—315 hits

([mh Foot] OR [mh Knee] OR [mh "Knee Joint"] OR [mh "Ankle Joint"] OR [mh Hip] OR [mh "Hip Joint"] OR [mh "Hip Prosthesis"] OR [mh "Hip Fractures"] OR [mh "Shoulder Dislocation"] OR [mh Elbow] OR [mh "Elbow Joint"] OR [mh "Wrist Joint"] OR [mh Spine] OR [mh "Intervertebral Disk Degeneration"] OR [mh "Bone Neoplasms"] OR [mh Arthroplasty] OR [mh "Fractures, Bone"] OR [mh Orthopedics] OR ((Foot OR Ankle OR Knee OR Hip OR Shoulder OR Elbow OR Wrist OR Spine OR Spina* OR "degenerative disk" OR "Bone Neoplasms" OR Arthroplast* OR Fractur* OR Orthop*):ti,ab,kw)) AND (([mh "Artificial Intelligence"] OR [mh "Machine Learning"] OR [mh "Supervised Machine Learning"] OR [mh "Neural Networks (Computer)"] OR [mh "Deep Learning"] OR [mh "Support Vector Machine"] OR (("naive bayes" OR "bayesian learning" OR "neural network*" OR "support vector" OR "support vectors" OR "random forest" OR "deep learning" OR "machine prediction" OR "machine intelligence" OR "computational intelligence" OR "computational learning" OR "computer reasoning" OR "machine learning" OR "convolutional network*" OR "artificial intelligence"):ti,ab,kw)))

Appendix 2. Critical appraisal of included studies

Author, year	Disclosure	Study aim	Input feature	Output feature	Validation method	Dataset distribution	Performance metric	AI model
Anderson, 2020	1	1	1	1	1	1	1	1
Arvind, 2018	1	1	1	1	1	1	1	1
Azimi, 2014	1	1	1	1	1	1	1	1
Azimi, 2015	1	1	1	1	1	1	1	1
Bevevino, 2014	1	1	1	1	1	1	1	1
Chen, 2020	1	1	1	1	1	1	1	1
Durand, 2018	1	1	1	1	1	1	1	1
Fatima, 2020	1	1	1	1	1	1	1	1
Fontana, 2018	1	1	1	1	1	1	1	1
Forsberg, 2011	1	1	1	1	1	1	1	1
Gabriel, 2019	1	1	1	1	1	1	1	1
Gowd, 2019	1	1	1	1	1	1	1	1
Goyal, 2019	1	1	1	1	1	1	1	1
Han, 2019	1	1	1	1	1	1	1	1
Harris, 2018	1	1	1	1	1	1	1	1
Harris, 2019	1	1	1	1	1	1	1	1
Hopkins, 2020a	0	1	1	1	1	1	1	1
Hopkins, 2020b	1	1	1	1	1	1	1	1
Huang, 2018	1	1	1	0	1	1	1	1
Huber, 2018	1	1	0	1	1	1	1	1
Kalagara, 2018	0	1	1	1	1	1	1	1
Karhade, 2018a	1	1	1	1	1	1	1	1
Karhade, 2018b	1	1	1	1	1	1	1	1
Karhade, 2018c	1	1	1	1	1	1	1	1
Karhade, 2019a	1	1	1	1	1	1	1	1
Karhade, 2019b	1	1	1	1	1	1	1	1
Karhade, 2019c	1	1	1	1	1	1	1	1
Karhade, 2019d	1	1	1	1	1	1	1	1
Karhade, 2020a	1	1	1	1	1	1	1	1
Karhade, 2020b	1	1	1	1	1	1	1	1
Karnuta, 2019	1	1	1	1	1	1	1	1
Karnuta, 2020	1	1	1	1	1	1	1	1
Katakam, 2020	1	1	1	1	1	1	1	1
Khan,2019	1	1	1	1	1	1	1	1
Kim, 2018a	1	1	1	1	1	1	1	1
Kim, 2018b	1	1	1	1	1	1	1	1
Kukar, 1996	0	1	1	1	1	1	1	1
Kumar, 2020	1	1	1	1	0	1	1	1
Kunze, 2020	1	1	1	1	1	1	1	1
Lin, 2010	1	1	1	1	1	1	1	1
Lungu, 2015	1	1	1	1	1	1	1	1
Martini, 2020	1	1	1	1	1	1	1	1
Merali, 2019	1	1	1	1	1	1	1	1
Merrill, 2018	1	1	1	1	1	1	1	1
Nwachukwu, 2020	1	1	1	1	1	1	1	1
Ogink, 2019a	1	1	1	1	1	1	1	1
Ogink, 2019b	1	1	1	1	1	1	1	1
Ottenbacher, 2004	1	1	1	1	1	1	1	1
Paulino Pereira, 2016	1	1	1	1	1	1	1	1
Pua, 2019	1	1	1	1	1	1	1	1
Ramkumar, 2019	1	1	1	1	1	1	1	1
Scheer, 2017	1	1	1	1	1	1	1	1
Schwartz, 1997	0	1	1	1	1	1	1	1
Shi, 2013	1	1	1	1	0	1	1	1
Siccoli, 2019	1	1	1	1	1	1	1	1
Thio, 2020	1	1	1	1	1	1	1	1
Wu, 2016	1	1	1	1	1	1	1	1
Zhang, 2020a	1	1	1	1	0	1	1	1
Zhang, 2020b	1	1	1	1	1	1	1	1

Table 2. Studies evaluating ML models for orthopedic surgical outcome prediction

A B C	D	E	F		G	H	I	J	K	L	M	N	O	P
Intraoperative complications														
Durand, 2018	SpD	NOS	C, S, H	4	Intraop.	3 d	2	**RF**, DT	1,029	80	10-FCV	20	0.85	
Huang, 2018	NA	THA, TKA	C, S	7	Intraop.	NA	2	**RF**, LR	15,187	100	5-FCV	NA	0.84	
Siccoli, 2019	SpS	Decomp.	C	15	Intraop.	45 min	2	**RF**, XGB, BDT, KNN, ANN, GLM, BGLM	635	70	NA	30	0.54	78
Postoperative complications														
Arvind, 2018	NA	ACDF	C		Compl.	NA	2	**ANN**, SVM, RF	20,879	70	5-FCV	30	0.65	
Fatima, 2020	Degen. SO	NOS	C, S	10	Compl.	1 m	2	**LR**, LASSO	80,610	70	10-FCV	30	0.70	
Gowd, 2019	Shoulder arthritis	TSA	C, S		Compl.	1 m	2	**LR**, GBM, RF, KNN, DT, NB	17,119	80	CV (nos)	20	0.71	95
Han, 2019	SpP	Sp surg.	C, S	274	Compl.	1 m	2	**LR**, LASSO	11,04233	70	10-FCV	30	0.70	
Harris, 2018	OA	THA, TKA	C, S	13	Compl.	1 m	2	**BR**, LASSO	70,569	100	10-FCV	NA	0.70	
Harris, 2019	Nonemergent primary	THA, TKA	C		Compl.	1 m	2	**LASSO**	10,7792	100	10-FCV	NA	0.64	
Hopkins, 2020a	SpP	Posterior fusion	C, S, H		Compl.	NA		**NN**	4,046	75	CV (nos)	25	0.79	
Karhade, 2020a	SpP	ALIF	C, S	6	Compl.	intra-op.		**EPLR**, SGB, RF, SVM, NN	1,035	75	CV (nos)	25	0.73	
Kim, 2018a	SpD	NOS	C	12	Compl.	NA	2	**ANN**, LR	5,818	70	5-FCV	30	0.64	
Kim, 2018b	Degen. SpP	PLIF	C	12	Compl.	NA	2	**ANN**, LR	22,629	70	NA	30	0.63	
Kukar, 1996	Femur fracture	NOS	C	17	Compl.	24 m	2	**Backpropagation** ANN, NB, KNN, LFC, DT	151	70	10-FCV	30		71
				17			5	**Semi NB**, ANN, NB, KNN, LFC, DT	151	70	10-FCV	30		67
Pua, 2019	Knee OA	TKA	C		Compl.	6 m	2	**LR**, RF, GBM	4,026	70	ICVL	30	0.75	
Scheer, 2017	Adult SpD	NOS	C, S, R	20	Compl.	1.5 m	2	**RT**	557	70	NA	30	0.89	88
Wu, 2016	Lower extremities (NOS)	NOS (including PCEA)	C, S	9	Compl.	NA	2	**SVM**, LR	195	75	CV (nos)	25	0.93	88
Medical management														
Gabriel, 2019	OA	THA	C	9	Hosp.	≤ 3 d	2	**RR**, LASSO, RF, MLR	960	67	NA	33	0.76	
Goyal, 2019	SpP	Spinal fusion	C		Non-HD	1 m	2	**GLM**, NB, ANN, RF, GBM, LDA	59,145	100	10-FCV	NA	0.87	79
Gowd, 2019	Shoulder arthritis	TSA	C, S		Extended LOS	1 m	2	**GBM**, RF, KNN, DT, NB, LR	17,119	80	CV (nos)	20	0.68	82
Karhade, 2018b	LDDD	NOS	C	10	Non-HD	NA	2	**NN**, BPM, BDT, SVM	26,364	80	10-FCV	20	0.82	
Karnuta, 2019	Hip fracture	NOS	C	7	Hosp.	NA	4	**NB**	98,562	90	10-FCV	10	0.88	77
		NOS	C	7	Cost	NA	3	**NB**	98,562	90	10-FCV	10	0.89	79
Karnuta, 2020	SpP	Sp. fusionl	C	8	Cost	NA	3	**NB**	38,070	100	10-FCV	NA	0.88	80
		Sp. fusionl	C	8	LOS	NA	3	**NB**	38,070	100	10-FCV	NA	0.94	87
		Sp. fusionl	C	8	Non-HD	NA	3	**NB**	38,070	100	10-FCV	NA	0.91	88
Merrill, 2018	Ankle fracture	ORIF	C	9	Hosp.	3 d	2	**Bo**, LR	16,501	70	CV (nos)	30	0.76	72
Ogink, 2019b	SpS	Surgery	C	10	Non-HD	NA	2	**ANN**, SVM, BPM, BDT	28,600	80	10-FCV	20	0.74	
Ogink, 2019a	Degen. SO	Surgery	C	10	Non-HD	NA	2	**BPM**, ANN, SVM, BDT	9,338	80	10-FCV	20	0.75	
Ottenbacher, 2004	Hip fracture	NOS	C, R	6	Non-HD	80 d	2	**ANN**, LR	3,708	67	3-FCV	33	0.73	
Ramkumar, 2019	OA	THA	C, H	15	LOS	NA	2	**ANN**	78,335	100	10-FCV		0.82	75
		THA	C, H	15	Charges	NA	2	**ANN**	78,335	100	10-FCV		0.83	76
		THA	C, H	15	Non-HD	NA	2	**ANN**	78,335	100	10-FCV		0.79	72
Siccoli, 2019	SpS	Decomp.	C	15	Hosp.	28 h	2	**XGB**, RF, BDT KNN, ANN, GLM, BGLM	635	70	NA	30	0.58	77
PROMs														
Azimi, 2014	Lumbar SpS	NOS	C	7	PROM	24 m	2	**ANN**, LR	168	50	25	25	0.80	97
Fontana, 2018	OA	THA, TKA	C, S, H		PROM	24 m	2	**LASSO**, RF, SVM	13,719	80	5-FCV	20	0.80	
Huber, 2018	OA	THA, TKA	C		PROM	NA	2	**XGB**, ANN, KNN, NB , RF, MSAENET, LM, LB	66,356	97	5-FCV	3	0.81	75
Khan, 2019	DCM	NOS	C	28	PROM	12 m	2	**MARS**, CT, SVM, PLS, GBoM, GAM, RF, LR	193	75	10-FCV	25	0.78	71

A B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Kumar, 2020	Shoulder pathology	aTSA	C, S	291	PROM	1 y, 2–3, 3–5, > 5 y	2	**NN**, LM, DT	4,782	67	NA	33	0.86	91
		rTSA	C, S	291	PROM	1 y, 2–3, 3–5, > 5 y	2	**NN**, LM, DT	4,782	67	NA	33	0.88	94
Kunze, 2020	OA	THA	C	8	PROM	24 m	2	RF, SGB, SVM NN, EPLR	616	80	CV (nos)	20	0.97	
Lungu, 2015	OA	THA	C	6	PROM	12 m, 24 m	2	**RF**	265	100	bootstrap resamping	NA		89
Merali, 2019	DCM	Decomp.	C, S	5	PROM	6 m, 12, 24 m	2	**RF**	605	70	10-FCV	30	0.72	71
Nwachukwu, 2020	FAI	Hip arthroscopy	C	5	PROM	24 m	2	**LR**	1,103	100	10-FCV	NA	0.86	
Siccoli, 2019	SpS	Decomp.	C	15	PROM	1.5 m, 3 m	2	**BDT**, RF, XGB, KNN, ANN, GLM, BGLM	635	70	NA	30	0.86	76
Schwartz, 1997	OA	THA	C	14	PROM	12 m	2	**ANN**, LR	221	95	LOOCV	5	0.79	
Survival														
Arvind, 2018	NA	ACDF	C		Survival	NA	2	**ANN**, SVM, RF	20,879	70	5-FCV	30	0.98	
Chen, 2020	Hip fracture	Nos	C, H	11	Survival	NA	2	**ANN**	10,534	70	15	15	0.93	93
Forsberg, 2011	Bone metastases	Nos	C		Survival	3 m, 12 m	2	**BNN**	189	90	10-FCV	10	0.84	
Harris, 2018	OA	THA, TKA	C, S	13	Survival	1 m	2	**BR**, LASSO	70,569	100	10-FCV	NA	0.73	
Harris, 2019	Elective PA	THA, TKA	C		Survival	1 m	2	**LASSO**	10,7792	100	10-FCV	NA	0.73	
Karhade, 2018c	Spine metastasis	NOS	C	7	Survival	1 m	2	**BPM**, NN, DT, SVM	1,790	80	10-FCV	20	0.78	
Karhade, 2018a	Spinal chordoma	NOS	C, S	5	Survival	60 m	2	**BPM**, BDT, SVM, ANN	265	100	10-FCV	0	0.80	
Karhade, 2019d	Spine metastasis	NOS	C	17	Survival	3 m, 12 m	2	**SGB**, PLR, RF, NN, SVM	732	80	10-FCV	20	0.86	
Kim, 2018a	SpD	NOS	C	12	Survival	NA	2	**ANN**, LR	5,818	70	5-FCV	30	0.84	69
Kim, 2018b	Various degen. diseases	PLIF	C	12	Survival	NA	2	**ANN**, LR	22,629	70	NA	30	0.70	60
Lin, 2010	Femur fracture	Various	C, R	11	Survival	12 m	2	**ANN**	286	70	NA	30	0.95	96
Merrill, 2018	Ankle fracture	ORIF	C	9	Survival	NA	2	**Bo**, LR	16,501	70	CV (nos)	30	0.74	85
Paulino Pereira, 2016	Spine metastasis	Various	C	9	Survival	1 m, 3, 12 m	2	**Nomogram**, Bo	649	80	5-FCV	20	0.74	75
Shi, 2013	Femur fracture	DHS	C	9	Survival	12 m	2	**ANN**, LR	2,150	67	NA	33	0.87	86
Thio, 2020	Extremity metastatsis	NOS	C	15	Survival	3 m, 12 m	2	**SGB**, RF,SVN, NN, PLR	1,090	80	10-FCV	20	0.86	
Zhang, 2020b	Pertrochanteric fracture	PFNA	C, H	14	Survival	12 m	2	**BNN**	448	100	10-FCV	NA	0.85	
Other														
Anderson, 2020	ACL rupture	ACL reconstruction	C		Sustained opioid use	3 m	2	**GBM**, LR, BNN, RF	10,919	80	CV (nos)	20	0.77	
Azimi, 2015	LDH	MicroDE	C	14	Recurrence	NA	2	**ANN**, LR	402	50	NA	25	0.83	94
Bevevino, 2014	Calcaneus fracture	Limb salvage	C, R	8	Amputation	NA	2	**ANN**, LR	155	100	10-FCV	NA	0.80	79
Hopkins, 2020b	SpP	Posterior fusion	C, S, H	177	Readmission	1 m	2	**ANN**	23,264	75	Cv (nos)	25	0.81	79
Kalagara, 2018	NA	Lumbar laminectomy	C, S, H	13	Readmission	1 m	2	**GBM**	26,869	85	10-FCV	15	0.81	95
Karhade, 2019a	Cervical pathology	ACDF	C, S	10	Sustained opioid use	3m	2	**SGB**, RF, NN, SVM, EPLR	2,737	80	10-FCV	20	0.81	
Karhade, 2019b	Hip arthritis	THA	C	7	Sustained opioid use	3 m	2	**EPLR**, SGB, RF, SVM, ANN	5,507	80	10-FCV	20	0.77	
Karhade, 2019c	LDH	NOS	C	9	Sustained opioid use	6 m	2	**EPLR**, RF, SGB, ANN, SVM	5,413	80	10-FCV	20	0.81	
Karhade, 2020b	LDH, SpS, SO	Decomp. and/or fusion	C, S	6	Sustained opioid use	3 m	2	**EPLR**, SGB, RF, SVM, ANN	8,435	80	10-FCV	20	0.70	
Katakam, 2020	Knee OA	TKA	C	9	Sustained opioid use	6 m	2	**SGB**, RF, SVM, ANN, EPLR	12,542	80	CV (nos)	20	0.76	
Martini, 2020	Degen. SpP	NOS	C, S	30	Readm.	1 m	2	**RF**	11,150	75	5-FCV	25	0.75	
Merrill, 2018	Ankle fracture	ORIF	C	9	Readm.	1 m	2	**Bo**, LR	33,504	70	CV (nos)	30	0.70	85
Siccoli, 2019	SpS	Decomp.	C	15	Reoperations	NA	2	**XGB**, RF, BDT, KNN, ANN, GLM, BGLM	635	70	NA	30	0.66	69
Zhang, 2020a	Low back and lower extremity pain	Thoracic or lumbar surgery	C, S	9	Sustained opioid use	12 m	2	**LR**, RF, SGB, SVM, NN	19,317	80	NA	20	0.85	

NA = not available
NOS = not otherwise specified
A. Output category
B. First author, year of publication
C. Disease/condition
    DCM = degenerative cervical myelopathy
    FAI = femoroacetabular impingement
    LDDD = lumbar degenerative disc disease
    LDH = lumbar disc herniation
    SO = spondylolisthesis
    SpD = spinal deformity
    SpP = spinal patholgy
    SpS = spinal stenosis
D. Operation
    ACDF = anterior cervical discectomy and fusion
    ALIF = anterior lumbar spine fusion
    Decomp. = decompression
    DHS = dynamic hip screws
    MicroDE = microdiscectomy
    ORIF = open reduction and internal fixation
    PA = primary arthroplasty
    PCEA = patient-controlled epidural analgesia
    PLIF = posterior lumbar spine fusion
    PFNA = proximal femoral nail antirotation
    THA = total hip arthroplasty
    TKA = total knee arthroplasty
    TSA = total shoulder arthroplasty (a = anatomic, r = reverse)
E. Input features
    C = clinical
    H = hospital-related factors (surgeon volume, hospital volume)
    S = surgical
F. Number of features
G. Output
    Hosp. = hospitalization
    LOS = length of stay
    Non-HD = Non-home discharge
    Readm. = readmission
H. Output: time points
I. Number of classes
J. Machine learning model. Best performing ML model is in bold.
    ANN = artificial neural network
    BDT = boosted decision tree
    BGLM = Batesian generalized linear models
    BNN = Bayesian belief network

Bo = boosting
BPM = Bayes point machine
BR = boosting regression
CHAID = chi-square automatic interaction dector
CT = classification tree
DT = decision tree
EPLR = elastic-net penalized logistic regression
FCM = fuzzy C-means
FIS = fuzzy inference system
GAM = generalized additive models
GBM = gradient boosting machine
GboM = generalized boosted models
GLM = generalized linear models
KNN = K-nearest neighbors
LASSO = least absolute shrinkage and selection operator
LB = logistic boost
LDA = linear discriminant analysis
LFC = lookahead feature construction
LM = linear model
LR = logistic regression
MARS = multivariable adaptive regression splines
MLR = multivariable logistic regression
MSAENET = multi-step elastic-net
NB = naive Bayes
PCA = principal component analysis
PLR = penalized logistic regression
PLS = partial least squares
RF = random forests
RT = random trees
RR = ridge regression
SGB = stochastic gradient boosting
SVM = support vector model
SVR = support vector regression
XGB = extreme gradient boosting
K. Number of patients
L. Size training set (%)
M. Validation method/size
    CV (nos) = cross-validation not otherwise specified
    FCV = fold cross validation
    ICVL = Inner cross-validation loop
    LOOCV = leave-one-out cross validation,
N. Size test set (%)
O. Area under the curve (AUC)
P. Accuracy

Anderson A B, Grazal C F, Balazs G C, Potter B K, Dickens J F, Forsberg J A. Can predictive modeling tools identify patients at high risk of prolonged opioid use after ACL reconstruction? Clin Orthop Relat Res 2020; 478(7): 00-1618.

Arvind V, Kim J S, Oermann E K, Kaji D, Cho S K. Predicting surgical complications in adult patients undergoing anterior cervical discectomy and fusion using machine learning. Neurospine 2018; 15(4): 329-37.

Azimi P, Benzel E C, Shahzadi S, Azhari S, Mohammadi H R. Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis. J Neurosurg Spine 2014; 20(3): 300-5.

Azimi P, Mohammadi H R, Benzel E C, Shahzadi S, Azhari S. Use of artificial neural networks to predict recurrent lumbar disk herniation. J Spinal Disord Tech 2015; 28(3): E161-5.

Bevevino A J, Dickens J F, Potter B K, Dworak T, Gordon W, Forsberg J A. A model to predict limb salvage in severe combat-related open calcaneus fractures. Clin Orthop Relat Res 2014; 472(10): 3002–9.

Chen C Y, Chen Y F, Chen H Y, Hung C T, Shi H Y. Artificial neural network and cox regression models for predicting mortality after hip fracture surgery: A population-based comparison. Medicina (Kaunas) 2020; 56(5): 243.

Durand W M, DePasse J M, Daniels A H. Predictive modeling for blood transfusion after adult spinal deformity surgery: A tree-based machine learning approach. Spine (Phila Pa 1976) 2018; 43(15): 1058-66.

Fatima N, Zheng H, Massaad E, Hadzipasic M, Shankar G M, Shin J H. Development and validation of machine learning algorithms for predicting adverse events after surgery for lumbar degenerative spondylolisthesis. World Neurosurg 2020; 140: 627-41.

Fontana M A, Lyman S, Sarker G K, Padgett D E, MacLean C H. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clin Orthop Relat Res 2019; 477(6): 1267-79.

Forsberg J A, Eberhardt J, Boland P J, Wedin R, Healey J H. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. PLoS One 2011; 6(5): e19956.

Gabriel R A, Sharma B S, Doan C N, Jiang X, Schmidt U H, Vaida F. A predictive model for determining patients not requiring prolonged hospital length of stay after elective primary total hip arthroplasty. Anesth Analg 2019; 129(1): 43-50.

Gowd A K, Agarwalla A, Amin N H, Romeo A A, Nicholson G P, Verma N N, Liu J N. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. J Shoulder Elbow Surg 2019; 28(12): e410-e421.

Goyal A, Ngufor C, Kerezoudis P, McCutcheon B, Storlie C, Bydon M. Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? Analysis of a national surgical registry. J Neurosurg Spine 2019: 1-11. Online ahead of print.

Han S S, Azad T D, Suarez P A, Ratliff J K. A machine learning approach for predictive models of adverse events following spine surgery. Spine J 2019; 19(11): 1772-81.

Harris A H, Kuo A C, Bowe T, Gupta S, Nordin D, Giori N J. Prediction models for 30-day mortality and complications after total knee and hip arthroplasties for Veteran Health Administration patients with osteoarthritis. J Arthroplasty 2018; 33(5): 1539-45.

Harris A H S, Kuo A C, Weng Y, Trickey A W, Bowe T, Giori N J. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin Orthop Relat Res 2019; 477(2): 452-60.

Hopkins B S, Mazmudar A, Driscoll C, Svet M, Goergen J, Kelsten M, Shlobin NA, Kesavabhotla K, Smith Z A, Dahdaleh N S. Using artificial intelligence (AI) to predict postoperative surgical site infection: A retrospective cohort of 4046 posterior spinal fusions. Clin Neurol Neurosurg 2020a; 192: 105718.

Hopkins B S, Yamaguchi J T, Garcia R, Kesavabhotla K, Weiss H, Hsu W K, Smith Z A, Dahdaleh N S. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: An NSQIP study involving 23,264 patients. J Neurosurg Spine 2020b; 32(3): 399-406.

Huang Z, Huang C, Xie J, Ma J, Cao G, Huang Q, Shen B, Byers Kraus V, Pei F. Analysis of a large data set to identify predictors of blood transfusion in primary total hip and knee arthroplasty. Transfusion 2018; 58(8): 1855-62.

Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC Med Inform Decis Mak 2019;1 9(1): 3.

Kalagara S, Eltorai A E M, Durand W M, DePasse J M, Daniels A H. Machine learning modeling for predicting hospital readmission following lumbar laminectomy. J Neurosurg Spine 2018; 30(3): 344-52.

Karhade A V, Thio Q, Ogink P, Kim J, Lozano-Calderon S, Raskin K, Schwab J H. Development of machine learning algorithms for prediction of 5-year spinal chordoma survival. World Neurosurg 2018a; 119(0): e842–7.

Karhade A V, Ogink P, Thio Q, Broekman M, Cha T, Gormley W B, Hershman S, Peul W C, Bono C M, Schwab J H. Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders. Neurosurg Focus 2018b; 45(5): E6.

Karhade A V, Thio Q C B S, Ogink P T, Shah A A, Bono C M, Oh K S, Saylor P J, Schoenfeld A J, Shin J H, Harris M B, Schwab J H. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 2018c; 85(1): E83-E91.

Karhade A V, Ogink P T, Thio Q C B S, Broekman M L D, Cha T D, Hershman S H, Mao J, Peul W C, Schoenfeld A J, Bono C M, Schwab J H. Machine learning for prediction of sustained opioid prescription after anterior cervical discectomy and fusion. Spine J 2019a; 19(6): 976-83.

Karhade A V., Schwab J H, Bedair H S. Development of machine learning algorithms for prediction of sustained postoperative opioid prescriptions after total hip arthroplasty. J Arthroplasty 2019b; 34(10): 2272-7.e1.

Karhade A V, Ogink P T, Thio Q C B S, Cha T D, Gormley W B, Hershman S H, Smith T R, Mao J, Schoenfeld A J, Bono C M, Schwab J H. Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation. Spine J 2019c; 19(11): 1764-71.

Karhade A V, Thio Q C B S, Ogink P T, Bono C M, Ferrone M L, Oh K S, Saylor P J, Schoenfeld A J, Shin J H, Harris M B, Schwab J H. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. Neurosurgery 2019d; 85(4): E671-81.

Karhade A V, Bongers M E R, Groot O Q, Cha T D, Doorly T P, Fogel H A, Hershman S H, Tobert D G, Srivastava S D, Bono C M, Kang J D, Harris M B, Schwab J H. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. Spine J 2020a; S1529-9430(20)30135-2. Online ahead of print.

Karhade A V, Cha T D, Fogel H A, Hershman S H, Tobert D G, Schoenfeld A J, Bono C M, Schwab J H. Predicting prolonged opioid prescriptions in opioid-naïve lumbar spine surgery patients. Spine J 2020b; 20(6): 888-95.

Karnuta J M, Navarro S M, Haeberle H S, Billow D G, Krebs V E, Ramkumar P N. Bundled care for hip fractures: a machine learning approach to an untenable patient-specific payment model. J Orthop Trauma 2019; 33(7): 324-30.

Karnuta J M, Golubovsky J L, Haeberle H S, Rajan P V, Navarro S M, Kamath A F, Schaffer J L, Krebs V E, Pelle D W, Ramkumar P N. Can a machine learning model accurately predict patient resource utilization following lumbar spinal fusion? Spine J 2020; 20(3): 329-36.

Katakam A, Karhade A V, Schwab J H, Chen A F, Bedair H S. Development and validation of machine learning algorithms for postoperative opioid prescriptions after TKA. J Orthop 2020; 22: 95-9.

Khan O, Badhiwala J H, Witiw C D, Wilson J R, Fehlings M G. Machine learning algorithms for prediction of health-related quality-of-life after surgery for mild degenerative cervical myelopathy. Spine J 2020: S1529-9430(20)30047-4. Online ahead of print.

Kim J S, Arvind V, Oermann E K, Kaji D, Ranson W, Ukogu C, Hussain A K, Caridi J, Cho S K. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. Spine Deform 2018a; 6(6): 762-70.

Kim J S, Merrill R K, Arvind V, Kaji D, Pasik S D, Nwachukwu C C, Vargas L, Osman N S, Oermann E K, Caridi J M, Cho S K. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine (Phila Pa 1976) 2018b; 43(12): 853-60.

Kukar M, Kononenko I, Silvester T. Machine learning in prognosis of the femoral neck fracture recovery. Artif Intell Med 1996; 8(5): 431-51.

Kumar V, Roche C, Overman S, Simovitch R, Flurin P H, Wright T, Zuckerman J, Routman H, Teredesai A. What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? Clin Orthop Relat Res 2020; 478(10): 2351-63.

Kunze K N, Karhade A V. Sadauskas A J, Schwab J H, Levine B R. Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. J Arthroplasty 2020;3 5(8): 2119-23.

Lin C-C, Ou Y-K, Chen S-H, Liu Y-C, Lin J. Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. Injury 2010; 41(8): 869-73.

Lungu E, Vendittoli P A, Desmeules F. Identification of patients with suboptimal results after hip arthroplasty: development of a preliminary prediction algorithm. BMC Musculoskelet Disord. BMC Musculoskeletal Disorders; 2015; 16(1): 1-10.

Martini M L, Neifert S N, Oermann E K, Gal J, Rajan K, Nistal D A, Caridi J M. Machine learning with feature domains elucidates candidate drivers of hospital readmission following spine surgery in a large single-center patient cohort. Neurosurgery 2020; 87(4): E500-10.

Merali Z G, Witiw C D, Badhiwala J H, Wilson J R, Fehlings M G. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. PLoS One 2019; 14(4): e0215133.

Merrill R K, Ferrandino R M, Hoffman R, Shaffer G W, Ndu A. Machine learning accurately predicts short-term outcomes following open reduction and internal fixation of ankle fractures. J Foot Ankle Surg 2019; 58(3): 410-6.

Nwachukwu B U, Beck E C, Lee E K, Cancienne J M, Waterman B R, Paul K, Nho S J. Application of machine learning for predicting clinically meaningful outcome after arthroscopic femoroacetabular impingement surgery. Am J Sports Med 2020; 48(2): 415–23.

Ogink P T, Karhade A V, Thio Q C B S, Hershman S H, Cha T D, Bono C M, Schwab J H. Development of a machine learning algorithm predicting discharge placement after surgery for spondylolisthesis. Eur Spine J 2019a; 28(8): 1775-82.

Ogink P T, Karhade A V, Thio Q C B S, Gormley W B, Oner F C, Verlaan J J, Schwab J H. Predicting discharge placement after elective surgery for lumbar spinal stenosis using machine learning methods. Eur Spine J 2019b; 28(6): 1433-40.

Ottenbacher K J, Linn R T, Smith P M, Illig S B, Mancuso M, Granger C V. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Ann Epidemiol 2004; 14(8): 551-9.

Paulino Pereira N R, Janssen S J, van Dijk E, Harris M B, Hornicek F J, Ferrone M L, Schwab J H. Development of a prognostic survival algorithm for patients with metastatic spine disease. J Bone Joint Surg Am 2016; 98(21): 1767-76.

Pua Y H, Kang H, Thumboo J, Clark R A, Chew E SX , Poon C L L, Chong H C, Yeo S J. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. Knee Surg Sport Traumatol Arthrosc 2020; 28(10): 3207-16.

Ramkumar P N, Navarro S M, Haeberle H S, Karnuta J M, Mont M A, Iannotti J P, Patterson B M, Krebs V E. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. J Arthroplasty 2019; 34(4): 632-7.

Scheer J K, Smith J S, Schwab F, Lafage V, Shaffrey C I, Bess S, Daniels A H, Hart R A, Protopsaltis T S, Mundis G MJ , Sciubba D M, Ailon T, Burton D C, Klineberg E, Ames C P. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J Neurosurg Spine 2017; 26(6): 736-43.

Schwartz M H, Ward R E, Macwilliam C, Verner J J. Using neural networks to identify patients unlikely to achieve a reduction in bodily pain after total hip replacement surgery. Med Care 1997; 35(10): 1020-30.

Shi L, Wang X C, Wang Y S. Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. Braz J Med Biol Res 2013; 46(11): 993-9.

Siccoli A, de Wispelaere M P, Schröder M L, Staartjes V E. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. Neurosurg Focus 2019; 46(5): E5.

Thio Q C B S, Karhade A V, Ogink P T, Bramer J AM , Ferrone M L, Calderón S L, Raskin K A, Schwab J H. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. Clin Orthop Relat Res 2020; 478(2): 322-33.

Wu H-Y, Gong C-S A, Lin S-P, Chang K-Y, Tsou M-Y, Ting C-K. Predicting postoperative vomiting among orthopedic patients receiving patient-controlled epidural analgesia using SVM and LR. Sci Rep 2016; 6: 27041.

Zhang Y, Fatemi P, Medress Z, Azad T D, Veeravagu A, Desai A, Ratliff J K. A predictive-modeling based screening tool for prolonged opioid use after surgical management of low back and lower extremity pain. Spine J 2020a; 20(8): 1184-95.

Zhang Y, Huang L, Liu Y, Chen Q, Li X, Hu J. Prediction of mortality at one year after surgery for pertrochanteric fracture in the elderly via a Bayesian belief network. Injury 2020b; 51(2): 407-13.