Knee Surgery, Sports Traumatology, Arthroscopy (2022) 30:3917–3923
https://doi.org/10.1007/s00167-022-07155-4

EDITORIAL

The development and deployment of machine learning models

James A. Pruneski[1] · Riley J. Williams III[2] · Benedict U. Nwachukwu[2] · Prem N. Ramkumar[2] · Ata M. Kiapour[1] · R. Kyle Martin[3] · Jón Karlsson[4] · Ayoosh Pareek[2]

Received: 23 August 2022 / Accepted: 31 August 2022 / Published online: 9 September 2022

Abstract
Applications of artificial intelligence, specifically machine learning, are becoming increasingly popular in Orthopaedic Surgery, and medicine as a whole. This growing interest is shared by data scientists and physicians alike. However, there is an asymmetry of understanding of the developmental process and potential applications of machine learning. As new technology will undoubtedly affect clinical practice in the coming years, it is important for physicians to understand how these processes work. The purpose of this paper is to provide clarity and a general framework for building and assessing machine learning models.

Introduction

In the last decade, increasing access to powerful computers and larger data sources from electronic medical records, fitness trackers, genetic testing, and other pools have escalated the practicality and utility of powerful machine learning algorithms [2, 26]. Recent studies have demonstrated the ability of complex algorithms to compliment, and even outperform physicians in some diagnostic tasks [11, 21]. The drastic increase in machine learning applications in orthopaedics has been well described, with almost 200 studies related to artificial intelligence in orthopaedics published between 2018 and 2021 [19].

The terms "artificial intelligence" and "machine learning" are often incorrectly used interchangeably [19]. Simply put, machine learning is a subset of artificial intelligence, in which algorithms learn from input data to make predictions and identify patterns. With this seemingly vague definition, it is often beneficial to break up the algorithms into a spectrum, between either fully human-guided or fully machine-guided analyses [2]. An overview of machine learning and its applications in orthopaedic surgery was the subject of a previous study and will not be addressed in this work [16]. The purpose of this study is to provide an overview of the process of development and deployment of machine learning models for musculoskeletal health professionals. In addition, further studies will enable a deeper dive into the nuances of machine learning algorithms, common issues encountered, and costs associated with machine learning model maintenance.

Table 1 provides an overview of key terms discussed in this paper while Table 2 introduces some essential components of machine learning model creation.

Problem identification

As with any scientific study, the first step in developing a machine learning model is to define a clinical problem and ask an appropriate question. Most often, the goal of this endeavor should not be solely to develop the model, but to successfully deploy the model and provide real-time value. It is important to consider not only what the problem is, but also why the problem needs to be solved. Researchers must have a firm understanding of what is currently being done for the problem, and why previously implemented solutions have been insufficient. This will not only provide a reference for performance but may also provide insight on

✉ Ayoosh Pareek
ayooshp@gmail.com

[1] Department of Orthopedic Surgery, Boston Children's Hospital, Boston, MA, USA

[2] Sports Medicine and Shoulder Service, Department of Orthopedic Surgery, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA

[3] Department of Orthopedic Surgery, University of Minnesota, Minneapolis, MN, USA

[4] Orthopaedic Research Department, Göteborg University, Göteborg, SE, Sweden

Table 1 Important definitions

Term	Definition
Machine learning (ML)	The process by which algorithms learn from input data to make predictions or identify patterns. ML exists on a spectrum between fully human-guided and fully machine-guided analyses
Training set	The subset of the data (generally 70–80%) used to train the initial model(s)
Testing set	The remaining subset of the data that the model is blind to; these data are used only to evaluate the final model and report performance metrics
Data splitting/sampling	The process by which data are separated into training and testing sets; can be simple (random), stratified random (by outcome), or convenience (time-dependent)
Brier score	The mean squared error calculation between predicted and expected values. Used to assess the performance of regression models. Ranges from 0 to infinity, where lower values indicate higher performance
Area under the curve (AUC) of the receiver operating characteristic (ROC) curve	Assesses a classification models' ability to discriminate between positive and negative cases. The AUC considers the classifier as a whole (not just the optimized cut-off point). Ranges from 0 to 1, where higher values indicate higher performance
F1-Score	A single metric that combines (through a harmonic mean) the precision and recall of a classification model at an optimized cut-off point. Ranges from 0 to 1, where higher values indicate higher performance
k-fold cross validation	Model validation technique used to assess how a model will perform on an independent/external dataset. The data are split into *k* (e.g., 10) subsets, and the model is trained on *k*-1 (e.g., 9) subsets and evaluated on 1 subset. This is repeated *k* times, and the error is averaged and reported as a single value
External validation	The process of assessing a model's generalizability by testing a model on independent, unseen data (e.g., patients from a different institution)

Table 2 Key takeaways

Building a successful machine learning model requires a clearly identified problem and a deliberate plan from the beginning.
A model's performance will be limited by the quality and quantity of data used to train it. Care must be taken to collect or acquire high-quality data capable of representing the target population
It is important to consider the distribution of positive and negative cases within the data set. If there is low prevalence of positive cases, or they distributed in a time-dependent manner, simple random sampling may not yield reproducible results.
It is good practice to start with simple models and progress to more complicated models to reach the desired performance. A "good" model balances underfitting and overfitting.
It is important to have a plan for what happens once the model is evaluated internally. It is important for models to be validated externally before being used for clinical decision-making, and models should be monitored and retrained with new data continuously to maintain high-level performance.

whether the new model should target improvements in terms of accuracy, efficiency, cost, or another realm. In general, narrowly posed questions with discrete answers are more effectively addressed with available data compared to more general questions.

Once the clinical problem is defined and the question is proposed, a task must be provided to the computer. One must ask, "HOW will we answer this question?" It is imperative to be explicit at this stage. For example, the process of developing a model to predict which patients will tear their anterior cruciate ligament (ACL) graft is different from the process for predicting the time-to-failure of the ACL graft or identifying subgroups of patients that may be at increased risk of ACL graft tear. The framing of this task is dependent on a variety of questions. Will this model be supervised, meaning the goal is to predict labels based on labeled training data, or unsupervised, where the goal is to identify structure or patterns in an unlabeled dataset? If the model is supervised, is it intending to classify and predict the labels of two or more categories, or anticipating the use of regression to predict continuous labels? If the data are unlabeled, is the objective to use unsupervised learning to cluster and identify distinct groups within the data, or reduce dimensions and detect lower level patterns and structure within the large dataset? The types of machine learning algorithms best suited for each question type is beyond the scope of this paper and will be discussed in upcoming papers (Fig. 1).

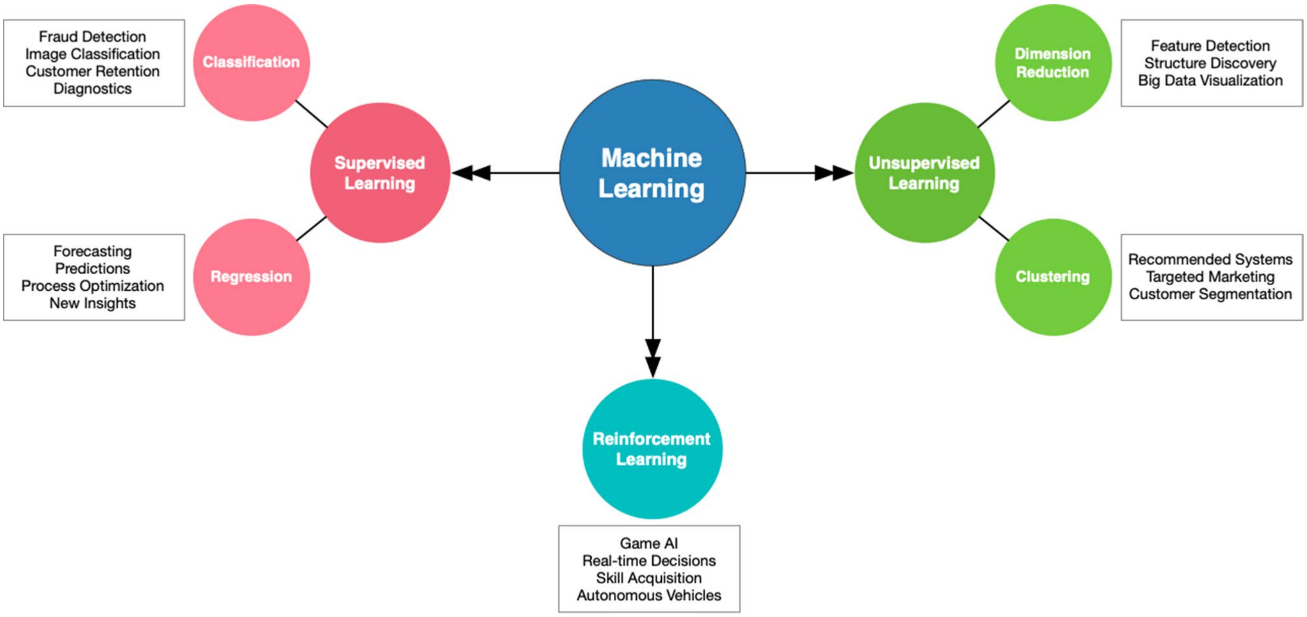

Fig. 1 Machine learning can be divided into supervised, unsupervised, and reinforcement learning. These three sub-fields each have their own applications

Data collection and analysis

The data acquisition stage of a project varies on the prospective versus retrospective nature of the study. Whichever the case, emphasis must be placed on acquiring complete, high-quality data through standardized processes. Most often, cohorts are identified with billing codes, and their demographic and clinical data are manually documented. However, researchers have recently demonstrated success in using machine learning to extract data from clinical documentation using a variety of techniques including natural language processing (NLP) [22, 23, 27]. If the clinical and statistical roles within a project are split between members of the team (which is typical), communication at this stage is critical to ensure that sufficient, usable, and clinically relevant data are available to accomplish the goal of the study. The usability of data can be thought of as the degree to which data are sufficiently accurate, complete, relevant, and timely to allow for improved clinical decision-making [4].

Ideally, a large quantity of high-quality data should be available, however, this is rarely the case. While power analyses provide useful insight about the sample size needed for statistical analyses, there is no defined equation guiding the quantity of data needed to train an effective machine learning model. Overall, the goal is to expose the model to enough diverse training data points to represent the testing population. A closer look at the heterogeneity of the data are completed in the data exploration stage.

Splitting the data

It is important to split the available data into unique training and testing sets. Intuitively, the larger “training” set is used to train and tune the model and compare different models before selecting the most optimal model. The “testing” set is only to be used to test the final model; if the test data are seen at any time during the training of the model, the model will learn from the data it will be evaluated on, introducing bias and nullifying the results. In this sense, the model should be “blind” to the testing data until final evaluation is done to avoid the potential for significantly overestimating model performance.

There are no formal rules for how to split the data as it is dependent on the quality and quantity of the data available. Most commonly, data are split with 70–80% used for training, and the remaining 20–30% used for evaluation. Within the training set, a 10–20% subset is dedicated for the tuning and validation of the model (discussed below). Simple random splitting is the most common approach to splitting, in which each sample has an equal probability of selection. This leads to low bias of the model performance, but can lead to high variance in complex, non-uniform datasets (e.g., unequal distribution of positive and negative cases - such as in rare diseases or conditions) [15, 20]. Other approaches include stratified random sampling, where samples from each cluster (label) are selected with uniform probability to better distribute positive and negative cases between the training and test sets, and convenience sampling, which is where the dataset is split into discrete time intervals before sampling. The latter is helpful in time series to decrease recency bias and account for changes or trends over time [5, 20].

One good “use-case” example of stratified sampling is when the outcome of interest is imbalanced or rare in the dataset: such as in patients with graft tear after ACL reconstruction. Since this occurrence is low (in this case, approximately 5–10%), when the data are split an approximately equal number of ACL graft tears in both the training and testing data sets is desirable. These situations often benefit from stratified random sampling instead of simple random splitting or sampling, which may lead to a disproportionate rate of graft tears in either the training or testing set.

Once the data are split, it can be helpful to think of the subsequent steps as a “pipeline.” A series of functions will be applied to the training data to prepare it for the model, and these will eventually be used on the testing set before evaluating the performance of the final model.

Data exploration and manipulation

Once the problem has been identified and framed, and the data have been split, the next step is to explore and understand the data. Fundamentally, it is important to understand what data features, or predictors, are available, and whether these will allow for accurate predictions. As previously mentioned, data must be sufficiently heterogeneous, in that the spectrum of variations within the predictors of the training set closely represents what will be seen in the testing set and in the real-world. Implied in this is that the training set has a sufficient distribution of “positive” and “negative” cases or labels, which can be addressed through alterations in the splitting methods as mentioned above.

Subsequently, the data can be explored, looking for correlations and redundancies within the features. For example, having a patient’s height, weight, and BMI as features could be redundant; it is not necessary to remove any predictors at this point, but it may prove important if the model is eventually found to overfit. It is also important to understand the number of data points, or instances available, as well as the amount of missingness from each instance. A number of machine learning models will not accept instances with significant missing features, and the researcher can decide to remove the instances with missing features or fill them using imputation. It is also important to characterize the types of data features available; common examples include categorical, binary, and continuous variables. Different models will accept different classes of features, but it is at this point in the process one would transform features such as “Autograft BTB” or “Tibialis Allograft” into categorical numeric values (categorization or binning: example 0 or 1).

Other important transformations improve normalization or standardization of the features, as many models perform better with similar scales (ranges, minima, maxima) so their comparison or correlation is assessed with a similar magnitude in mind. Again, for efficiency and reproducibility, it is reasonable to add the imputations and transformations in this stage of the model building process into the pipeline mentioned in the previous step. Data exploration and manipulation is largely the most investigative and *time-consuming* portion of the model creation process.

Model development and selection

The output defined in the problem identification stage is important, as the process required to develop a model that predicts an exact tumor size (regression) differs substantially from a model that classifies the tumor as “small,” medium,” or “large” (categorical: nominal or ordinal). With the data split, explored, and prepared, one can now start to build the models with the training set.

One tenet to be aware of at the onset is the *bias-variance tradeoff*. This idea describes that a model should balance underfitting and overfitting; it should be powerful enough to characterize the structure within the data without fitting/recreating false patterns [3]. Models with near 100% accuracy are often excellent at predicting the outcome of interest in the training data, but perform much more poorly with the testing data or during external validation with outside data sources. This is likely because they lack the generalizability to be accurate when data differs from the data provided in training sets as is often the case in testing sets. This is the most common problem encountered by novice machine learning practitioners.

Considering the delicate bias-variance tradeoff, it is common to start with simple models (using simpler or more easily understood algorithms) before progressing along the spectrum. If the initial model is underperforming (underfitting), one can either (1) select a more powerful model (i.e. support vector machines, random forests, neural networks), or (2) provide different features/data to the model. If the model is overfitting, options include (1) simplifying the model, or (2) provide more data along with other options (such as stopping model training early) which are outside the scope of this study [10].

To assess for model fit or performance, quantitative evaluations are used based on the type of model selected and the question to be answered. While there are a variety of performance parameters, for regression models, root mean squared error or Brier score are often used to quantify performance and compare models [6]. Notably, these scores provide a single value from 0 up to infinity, so there remains utility in reporting R^2 or adjusted-R^2 values for interpretability and to characterize the amount of variance in the predicted outcome that is accounted for by the features provided to the model [7].

For classification problems or problems assessing categorical variables, the area under (AUC) the receiver operating characteristic (ROC) curve is often used [9]. This value characterizes the classifier as a whole (at all cut-off points), and may be adversely affected by skewed data, which is why the F1-score is commonly used as it evaluates the classifier's performance at an optimized cut-off point [12]. These metrics incorporate elements of the model's accuracy, sensitivity (recall), specificity, and precision (positive predictive value), which can be selectively optimized depending on the goal of the model (e.g., to maximize true positives and/or minimize false negatives).

These are commonly used in conjunction with *k*-fold cross validation, in which the training set is randomly divided into *k* "folds," or subsets, where *k* is usually ten. From here, the model is trained on nine of the folds, and evaluated on one fold; the result is ten evaluations of the model's performance (Fig. 2). From here, small tweaks to model parameters can be explored with, and the desired models can be used in subsequent stages.

Model tuning and validation

The tuning and validation stage is continuous with the previous training and model selection stage. It is normal to go between the two when exploring and deciding which models to select. While small changes to the model parameters were made in the previous step, more discreet approaches to tuning model hyperparameters are made once the most promising models are identified. Simply defined, hyperparameters are actually parameters of the machine learning algorithm itself, instead of parameters related to the data. These hyperparameters essentially dictate how the model learns from the data based on its programming and associated mathematical results [8]. While it is possible to manually alter the hyperparameters and re-assess performance, there are more efficient ways to optimize the model, including grid and random searches which automate this process, often trying tens or hundreds of possibilities in an automated manner [10]. Before proceeding to the evaluation stage, it is often beneficial to take a closer look at the errors your model of choice is making. This allows assessment of the feature importance, a quantitative measure of the individual predictor's impact. Small changes at this stage may improve (or worsen) the final evaluation, and since this is still largely a manual process, it is important to carefully assess why the model is good at predicting certain outcomes and poor at predicting others. Before proceeding to model evaluation using the testing data, there should be a reasonable level of confidence that the model is capable of completing the task as intended.

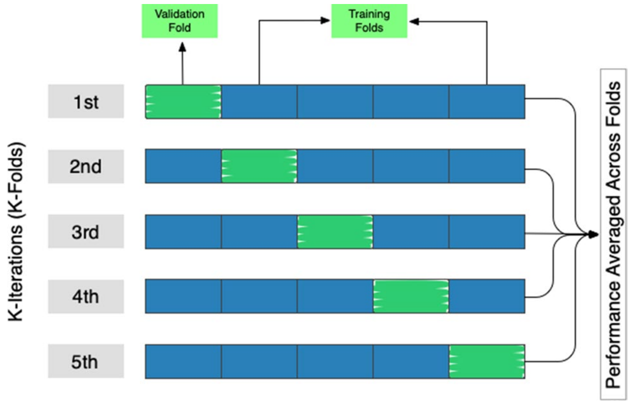

Fig. 2 K-fold cross validation is used to divide the training data into *k* "folds" or subsets. The model performance is an average of the performance on all the subsets. In the figure below, we have used $k=5$ for a fivefold cross-validation

Model evaluation

After making the above changes to the model, the most straightforward stage of the process is to evaluate the model using the dedicated testing set. For this, the testing set needs to be prepared and transformed exactly as the training set (following the pipeline mentioned above). The features are input into the model, and predictions are generated. The metrics described above can be used to assess the model's performance/quantify the error per the discretion of the team creating the model and the task at hand. For instance, in a classification problem, although accuracy is commonly used (percent of predictions that are correct), it may be more desirable to maximize sensitivity instead, accepting false positives if the objective is to analyze a diagnostic test where one does not want to miss any true positives [1, 25].

Model deployment and monitoring

The final step in the model building process is the deployment and monitoring of the model. This step is highly dependent on the purpose of the model. If the model was built for research purposes, it can be published and shared on online repositories if future collaboration is desired. This is especially important for models used in clinical decision-making, to prevent ill-informed predictions or prematurely made decisions. Traditionally, research has been focused on maximizing internal validity in the development of institution-based models and data repositories [24]. However,

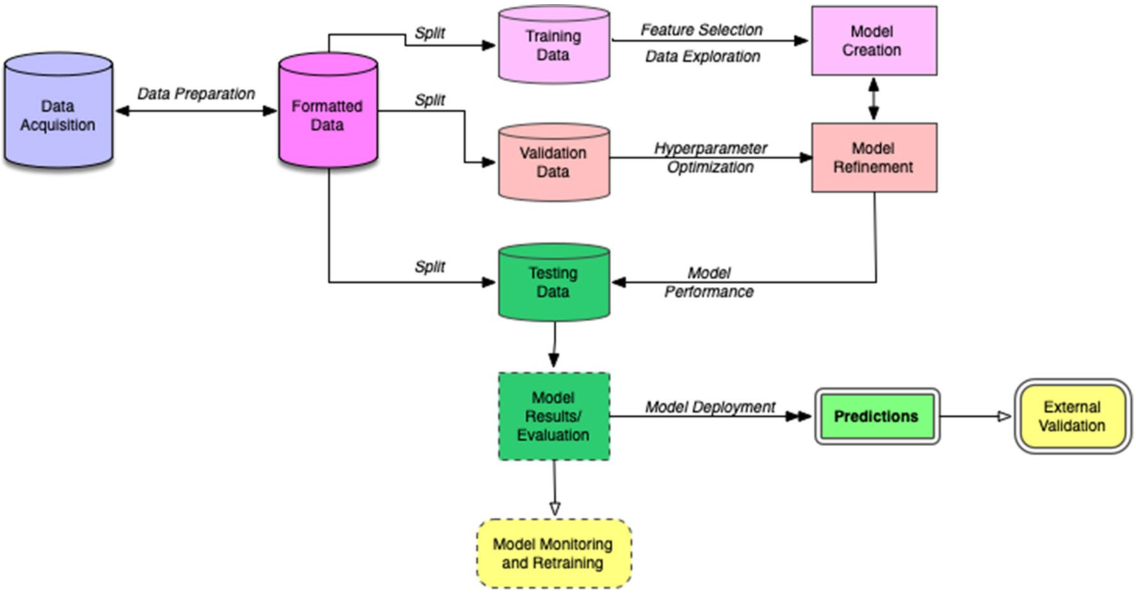

Fig. 3 A typical workflow for machine learning model creation, evaluation, and deployment. Once prepared, data are typically split into training, validation, and test sets. Training data are typically used to create the model and choose the algorithm that performs best, whereas validation sets are used for hyperparameter selection for model refinement. The model should then be evaluated on test data sets, which have been blinded to model creation before deployment into a usable model for predictions. The model should lastly be monitored and retrained for maintenance with the option of deployment to another site for external validation

this failure to account for external validity (generalizability) can lead to models underperforming at outside institutions, potentially due to variations in data collection, documentation, or baseline patient characteristics. When such models are used in real-time to affect a patient's care, external validation is critical and should be highly emphasized, and there is a growing trend in this direction within the orthopaedic literature [13, 14, 17, 18].

If the model was built for business purposes or task automation, it can be launched using web and cloud services. Whichever the case, if high performance is desired after the point of deployment, it is important to regularly assess the model, and update it with new training data. This requires a team to monitor the incoming data, incorporate these data into the model, and at times create a new model that can maintain or exceed previous model performance to optimize the result of the model or the task of interest. A flowchart summarizing the workflow in this editorial is shown in Fig. 3.

Conclusion

This editorial presents an overview of the process required to build a machine learning model. The primary stages involve problem identification, data collection and exploration, data splitting, model development, model validation, model evaluation, model deployment, and model maintenance. While an in-depth tutorial on how to program models for a specific application is beyond the scope, the goal of this review is to provide the structure for how clinicians can think about the stepwise process behind machine learning methodology and associated model creation.

Funding No funding was obtained for this study.

Declarations

Conflict of interest Authors declare that they have no conflict of interest with regards to this publication.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Altman DG, Bland JM (1994) Diagnostic tests. 1: sensitivity and specificity. BMJ 308:1552
2. Beam AL, Kohane IS (2018) Big data and machine learning in health care. JAMA 319:1317–1318
3. Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias-variance tradeoff. Proc Natl Acad Sci U S A 116:15849–15854

4. Bloland P, MacNeil A (2019) Defining & assessing the quality, usability, and utilization of immunization data. BMC Public Health. https://doi.org/10.1186/s12889-019-6709-1
5. Bowden GJ, Maier HR, Dandy GC (2002) Optimal division of data for neural network models in water resources applications. Water Resour Res 38:2–1
6. Brier GW (1950) Verification of forecasts in terms of probability. Mon Weather Rev 78:1–3
7. Chicco D, Warrens MJ, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7:e623
8. Claesen M, de Moor B (2015) Hyperparameter Search in Machine Learning. Paper presented at the 11th metaheuristics international conference, Katholieke Universiteit Leuven, 7–10 June 2015
9. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874
10. Géron A (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media Inc, Sebastopol
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402–2410
12. Jeni LA, Cohn JF, de La Torre F (2013) Facing Imbalanced Data Recommendations for the Use of Performance Metrics. In: Paper presented at the international conference on affective computing and intelligent interaction, Carnegie Mellon University, 2–5 September 2013.
13. Karhade AV, Ahmed AK, Pennington Z, Chara A, Schilling A, Thio QCBS, Ogink PT, Sciubba DM, Schwab JH (2020) External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. Spine J 20:14–21
14. Karnuta JM, Haeberle HS, Luu BC, Roth AL, Molloy RM, Nystrom LM, Piuzzi NS, Schaffer JL, Chen AF, Iorio R, Krebs VE, Ramkumar PN (2021) Artificial intelligence to identify arthroplasty implants from radiographs of the hip. J Arthroplasty 36:S290-S294.e1
15. Lohr SL (2021) Sampling: design and analysis. Chapman and Hall, London
16. Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R (2022) Artificial intelligence and machine learning: an introduction for orthopaedic surgeons. Knee Surg Sports Traumatol Arthrosc 30:361–364
17. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L (2022) Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. Knee Surg Sports Traumatol Arthrosc 30:368–375
18. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Iorio R, Mont MA, Patterson BM, Krebs VE (2019) Preoperative prediction of value metrics and a patient-specific payment model for primary total hip arthroplasty: development and validation of a deep learning model. J Arthroplasty 34:2228-2234.e1
19. Ramkumar PN, Pang M, Polisetty T, Helm JM, Karnuta JM (2022) Meaningless applications and misguided methodologies in artificial intelligence–related orthopaedic research propagates hype over hope. Arthroscopy. https://doi.org/10.1016/j.arthro.2022.04.014
20. Reitermanova Z (2010) Data splitting. In: Šafránková J, Pavlů J (eds) WDS'10 Proceedings of contributed papers, part I, Prague, 2010
21. Richens JG, Lee CM, Johri S (2020) Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun 11:3923
22. Sagheb E, Ramazanian T, Tafti AP, Fu S, Kremers WK, Berry DJ, Lewallen DG, Sohn S, Maradit Kremers H (2021) Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. J Arthroplasty 36:922–926
23. Shah RF, Bini S, Vail T (2020) Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. Bone Joint J 102-B:99–104
24. Steckler A, McLeroy KR (2008) The importance of external validity. Am J Public Health 98:9–10
25. Trevethan R (2017) Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 5:307
26. Weber GM, Mandl KD, Kohane IS (2014) Finding the missing link for big biomedical data. JAMA 311:2479–2480
27. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, Berry DJ, Lewallen DG, Maradit-Kremers H (2019) Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. J Bone Joint Surg Am 101:1931–1938

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.