Level V Evidence

A Guide for the Application of Statistics in Biomedical Studies Concerning Machine Learning and Artificial Intelligence

Evan M. Polce, B.S., and Kyle N. Kunze, M.D.

Abstract: With the plethora of machine learning (ML) analyses published in the orthopaedic literature within the last 5 years, several attempts have been made to enhance our understanding of what exactly ML means and how it is used. At its most fundamental level, ML comprises a branch of artificial intelligence that uses algorithms to analyze and learn from patterns in data without explicit programming or human intervention. On the other hand, traditional statistics require a user to specifically choose variables of interest to create a model capable of predicting an outcome, the output of which (1) may be falsely influenced by the variables chosen to be included by the user and (2) does not allow for optimization of performance. Early publications have served as succinct editorials or reviews intended to ease audiences unfamiliar with ML into the complexities that accompany the subject. Most commonly, the focus of these studies concerns the terminology and concepts surrounding ML because it is important to understand the rationale behind performing such studies. Unfortunately, these publications only touch on the most basic aspects of ML and are too frequently repetitive. Indeed, the conclusion of these articles reiterate that the potential clinical utility of these algorithms remains tangential at best in their current form and caution against premature adoption without external validation. By doing so, our perspective and ability to draw our own conclusions from these studies have not advanced, and we are left concluding with each subsequent study that a new algorithm is published for an outcome of interest that cannot be used until further validation. What readers now need is to regress to embrace the principles of the scientific method that they have used to critically assess vast numbers of publications before this wave of newly applied statistical methodology—a guide to interpret results such that their own conclusions can be drawn. **Level of Evidence:** Level V, expert opinion.

See commentary on page 159

What is Machine Learning?

Although a complete, in-depth discussion of machine learning (ML) terminology is outside of the scope of the present work, several key concepts and definitions frequently encountered in ML research are important to understand. At its most fundamental level, ML comprises a branch of artificial intelligence that uses algorithms to analyze and learn from patterns in data without explicit programming or human intervention. Generally, there are 2 basic approaches that ML algorithms use to "learn" from data: supervised and unsupervised learning. In supervised learning (e.g., classification and regression), labeled datasets are provided to the ML algorithm such that the relationship between input variables and the predicted dependent variable can be learned with increasing accuracy over time. In unsupervised learning (e.g., clustering and dimensionality reduction), unlabeled datasets are provided to the ML algorithm that subsequently cluster subjects based on previously unknown patterns within the data. ML offers several benefits over more traditional statistical methods, such as simple regression. First, ML algorithms are better able to model complex, nonlinear associations between many input variables in large datasets with a high degree of efficiency. Second, ML can use knowledge gained from a prior task and apply it to a similar, but novel, problem in a process

From the Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, U.S.A.

The authors report that they have no conflicts of interest in the authorship and publication of this article. Full ICMJE author disclosure forms are available for this article online, as supplementary material.

Received November 10, 2021; accepted April 19, 2022.

Address correspondence to Kyle N. Kunze, M.D., Hospital for Special Surgery, Department of Orthopedic Surgery, 535 E. 70th St., New York, NY 10021, U.S.A. E-mail: kylekunze7@gmail.com

0749-8063/211635/$36.00

https://doi.org/10.1016/j.arthro.2022.04.016

commonly referred to as transfer learning. However, it is important to note that traditional statistics may continue to be useful in situations where interpretation rather than overall model accuracy is preferred.

A subset of ML that has increased in popularity as of late due to improvements in computational power and speed is deep learning (DL). DL leverages the power of specific ML algorithms referred to as artificial neural networks (ANN), which are designed to emulate the neuronal circuitry of the brain. Briefly, ANNs are comprised of an input layer of nodes (analogous to neurons) connected to a series of multiple hidden layers by weights which are periodically updated as the ANN learns. The weighted sum of the hidden layers culminates in an output layer, representing the final model prediction. Besides the use of ANNs, DL also is distinct with respect to the type of data used. Whereas other ML algorithms require data to be in a structured format, DL can analyze unstructured data, such as images and text. For example, an analysist must first mask and segment properties within an image to establish a ground truth for algorithms to learn. DL algorithms then leverage the ability to recognize patterns in the structure of pixels in images based on this segmentation and by doing so can be applied to identify the presence or absence of pathology, identify specific types of implants, and automate measurements.

What are Machine Learning Studies Actually Reporting?

It is far too easy to become lost in the nuances and messages portrayed by ML studies that develop novel prediction tools and laud the ways in which they may someday change clinical practice given the substantial increase in ML literature in the past 5 years (Fig 1). Despite their potential clinical utility, they should be interpreted with caution because the current literature remains in the proof-of-concept phases.[1,2] Regardless, the potential of ML is far-reaching and here to stay, and it would be in opposition to the quintessential principle we as clinician-scientists pride and nurture if we were to simply take all of the conclusions from these studies at face value—question everything and understand why. Without being able to fully understand what we are reading, we limit our ability to draw meaningful conclusions and critically evaluate this new and important literature. This puts the reader at risk of accepting overstatements and misstatements, a common limitation of observational studies.[3]

We must "level the playing field" so to speak to allow readers to draw their own conclusions from such studies. As mentioned, although several well-written introductions to ML and associated analyses have been published in recent years,[4-7] they unfortunately fall short with regard to providing a guide to readers for the interpretation of commonly reported statistics in ML. The following review will provide a comprehensive yet digestible introduction to interpreting these statistics to allow for a better understanding of algorithm performance and use.

Classification Analysis

The Receiver-Operator-Curve (ROC) Analysis

The ROC analysis will likely be familiar to many readers because it is commonly implicated in diagnostic studies. This is also known as discrimination, or the ability for a diagnostic or predictive model to distinguish between true-positive and false-positive cases (Fig 2). In a hypothetical scenario, this analysis could be used to describe the probability that the ML model will assign a greater predicted probability to a randomly selected positive case (patient who experienced an ACL graft failure) relative to a randomly selected negative case (false-positive case, i.e., a patient who did not have an ACL graft failure).

The quantitative metric used to describe the performance of the model is the area under the curve, also known as the concordance statistic (C-statistic). In a ROC analysis, a c-statistic of 0.5 is equivalent to the model performing no better than chance, whereas a c-statistic of 1.0 indicates perfect discrimination.[8,9] Therefore the greater the c-statistic, the better a model is at discriminating between cases.

The Calibration Analysis

Calibration is an analysis similar to that of ROC analysis; however, the major difference between ROC and calibration analyses is that unlike ROC analysis where the prediction is a binary label (i.e., ACL graft failure or no failure), calibration leverages the ability to extract probabilities from these binary class labels. In other words, while discrimination assesses the accuracy of correctly predicted class labels (true positives vs false positives), calibration extracts the probabilities of those labels being correct. This analysis is extremely important in clinical practice as it is more pragmatic in terms of prognosis. Specifically, it is impossible to know with 100% certainty that an ACL graft will fail before it actually fails, which is what discrimination describes; however, in practice, we speak of probabilities of an event occurring based on data and experience, which is what calibration represents. As such, it is possible for an ML algorithm to systematically overestimate or underestimate the likelihood of ACL graft failure for all patients (i.e., poor calibration) while still adequately identifying those that ultimately experience the event or not (i.e., good discrimination). Thus calibration analyses can evaluate model risk estimates and determine how well they are in accordance with the true observed outcomes. For example, among patients receiving a

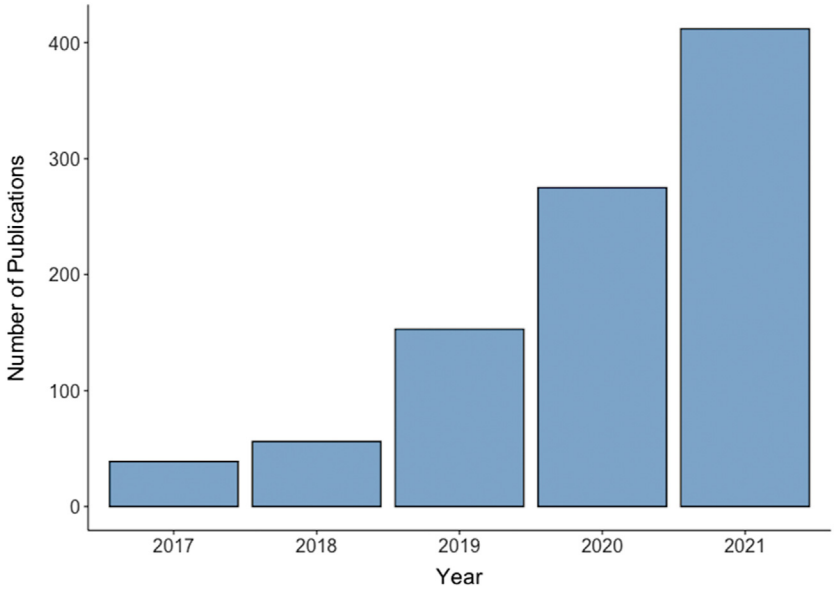

Fig 1. Bar plot depicting the publication of ML-related research (articles per year) since 2017 based on the PubMed search term "machine learning orthopedic." This graph depicts an approximate 10-times increase in the amount of ML literature since 2017 per year.

predicted risk of 5% from a perfectly calibrated model, 5 out of 100 experience ACL graft failure.

The quantitative metrics used to describe calibration analyses are the calibration intercept and calibration slope. The calibration intercept represents the propensity of the model on average to overestimate or underestimate the observed outcome prevalence, whereas the calibration slope reflects whether predictions were too moderate or too extreme (Fig 3).[8,9] A calibration intercept = 0 and a calibration slope = 1 indicate perfect calibration.

The Brier Score

The Brier score is an extension of calibration analysis and describes overall model performance. In other words, this score tells us how accurate the predictions were from the calibration model. The score is calculated by determining the mean squared difference between true outcomes and the corresponding predicted probabilities from the model (calculated from calibration analysis).[10,11] The best possible Brier score is equal to 0, which indicates total accuracy. The worst possible Brier score is 1, indicating the predictions were completely inaccurate. Therefore lower scores on the scale indicate better prediction accuracy. Below is the formal equation, although the Brier score is automatically calculated when performing calibration:

$$\text{Brier Score} = 1 \Big/ N \sum_{t=1}^{N} (f_t - o_t)^2$$

Where N = the number of items being calculated, f_t is the predicted probability of the ML algorithm, o_t is the outcome (ACL failure or no failure), and $\sum$ is the summation symbol.

It is also recommended to compare the Brier score to a correlative function called the Null model Brier score. The Null model Brier score is the Brier score for which the predicted probabilities are simply equivalent to the outcome prevalence in the study population and represents a benchmark value for overall comparison of model performance. ML model Brier scores less than or equal to the null model Brier score indicate appropriate performance because predictions are more informed.

The Decision-Curve Analysis

The decision-curve analysis (DCA) allows us to understand how using the ML model in practice to make decisions for treatment compares to other methods of clinical decision making.[9,12] Though theoretical, it provides greater insight, and more validity, as to whether the ML model may actually be useful.[9] In contrast to discrimination and calibration, net benefit reflects whether the prediction model *should* or *should not* be used in the clinical setting by weighing the consequences of model decisions and potential misclassifications (Fig 4).

The net benefit of a model is how well it correctly identifies patients that do and do not have the outcome of interest (i.e., true positive predictions confer benefit and false positives result in cost/harm). The net benefit of the model (blue line) is shown relative to default strategies of changing management for patients in the absence of a prediction model ("all" for all patients, "none" for no patients, or the highest weighted variable [i.e., the variable demonstrated to have the most importance for predicting the outcome of interest]). Changing management for "all" and "none" are included for comparison because these are reasonable strategies in most clinical contexts. The "all" line slopes

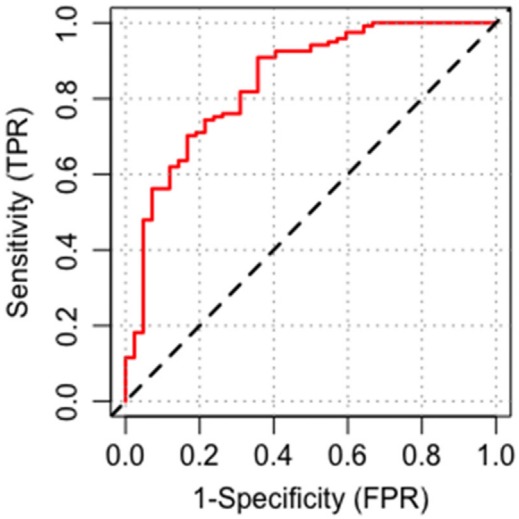

Fig 2. Receiver operative curve. The Y axis represents the sensitivity, or true-positive rate (TPR), of model predictions, whereas the X axis represents 1 minus specificity, or false-positive rate (FPR), of model predictions. Reprinted with permission from Kunze et al.[23]

down because at a threshold of zero, false-positives are given no weight relative to true positives; as the threshold increases, false positives gain increased weight relative to true positives and changing management for all patients results in decreasing net benefit. The horizontal line ("none") represents the default strategy of changing management for no patients and therefore the net benefit is zero at all thresholds.

$$Net\,benefit = \frac{True\,Positives - Cost:Benefit\,Ratio(False\,Positives)}{Total\,Number\,of\,Patients}$$

Knowing this, we see that the line showing the net benefit of using the ML model always lies superior (at a greater net benefit) than the other theoretical treatment strategies, and therefore the DCA indicates the model would be clinically useful in this patient population.

In contrast, the x-axis represents preference; that is, how one may value different outcomes for a patient in a given scenario, which can vary from physician to physician. As we move along the x-axis, the high-risk threshold (i.e., risk level above which management would be altered), and cost/benefit ratio increases. The cost/benefit ratio is equivalent to the relative weight of false-positive to true-positive predictions and is calculated by

$$Cost:Benefit\,Ratio = \frac{risk\,threshold\,probability}{1 - risk\,threshold\,probability}$$

We can imagine this trend as beginning to encounter the primary ACL patients who have several risk factors for ACL graft failure, and we know that treating these patients becomes riskier. Therefore a risk threshold of 1 indicates the highest probability of graft failure, whereas the risk threshold of 0 indicates no risk. It is reasonable to assume that if a given patient's risk of ACL graft failure was 1% that both the patient and physician would be comfortable proceeding with surgery; however, if the risk for that same patient was 99%, additional assessment and clinical judgment would be necessary before proceeding with treatment. In practice, the model predictions of ACL graft failure (based on patient risk factors) and the risk threshold at which additional/different clinical management is necessary (based on physician experience and preference) is often somewhere between these two extremes. Thus DCA assists the physician and patient with making an informed treatment decision.

Explanation of Model Predictions with Global and Local Behavior

ML models are often referred to as "black boxes" because of their sophisticated computations and lower interpretability relative to simpler algorithms (e.g., linear regression).[13] However, several techniques have been developed to increase the interpretation of *how* ML models reach their predictive decisions to increase both accountability and transparency.

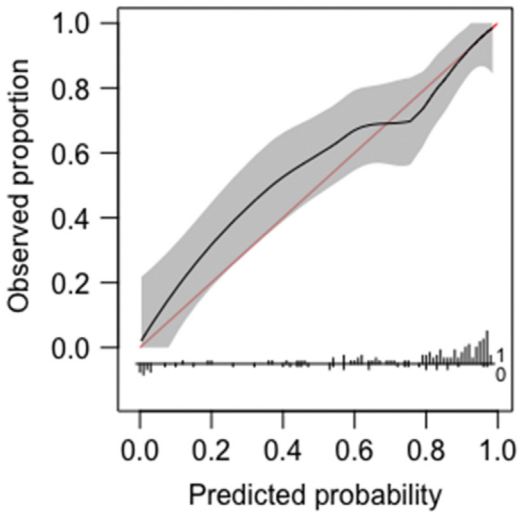

Fig 3. Calibration plot. The Y axis represents the true proportion of the outcome within the dataset, whereas the X axis represents the predicted probability of the outcome per the model. The *red line* represents perfect calibration whereas the *solid black line* is the predicted risk from the model. With perfect calibration, the *solid black line* would exactly match the *red line* (e.g., among patients with a predicted risk of 30%, 30/100 ultimately experienced ACL graft failure). The shaded areas represent the 95% confidence interval of the predictions. In this example, the model tends to underestimate the true observed outcome when the prevalence of the outcome is 0% to 70%, after which it predicts the true outcome precisely. Reprinted with permission from Kunze et al.[23]

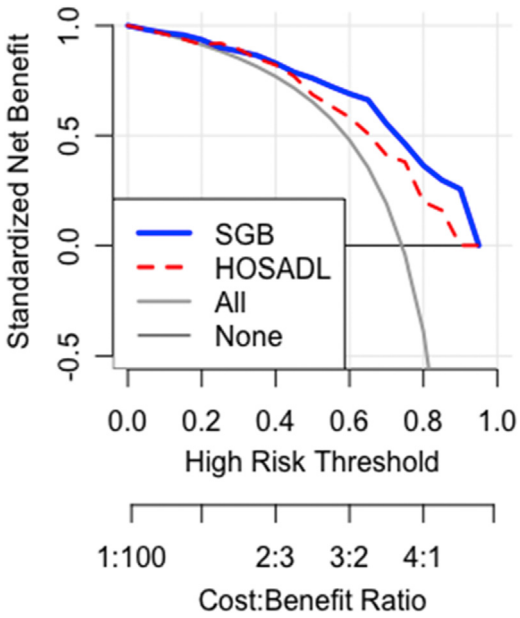

Fig 4. Decision-curve analysis output. The Y axis demonstrates the net benefit of decision-making based off of each scenario, whereas the X axis depicts both the cost-to-benefit (cost:benefit) ratio and risk thresholds. A net benefit of 1 is ideal, whereas the net benefit of 0 is the worst possible outcome. In this example, the net benefit of using the machine-learning algorithm to make treatment decisions is greater than the best performing variable (preoperative hip outcome score—activities of daily living [HOSADL]), treating all patients as if they were to achieve the outcome of interest, and treating no patients as if they were to achieve the outcome of interest. Specifically, we see that the algorithm outperforms the predictive abilities of the HOSADL at higher risk thresholds (0.6 and above), suggesting that using a model with multiple contributing variables is more useful than considering a variable in isolation. SGB, stochastic gradient boosting. Reprinted with permission from Kunze et al.[23]

Global model behavior corresponds to how each input variable influences model predictions at the overall population level. The graphical depiction of global model behavior is often demonstrated visually with global variable importance plots (Fig 5). In these plots, the Y-axis lists the predictor variables included in the model, whereas the X-axis is a scale of importance (i.e., how strongly each variable influenced model predictions in a global sense).

In contrast, local model behavior refers to how the model computed individual predictions at the patient level. Several techniques for explaining local behavior have been proposed including local agnostic model interpretations, which involves using a separate, interpretable algorithm to explain individual predictions made by more complex models.[13] The qualitative output of local agnostic model interpretations provides a visual explanation of how each input feature contributes to the model output in the instance being predicted. Specifically, it demonstrates how each input variable either supports (e.g., increases the likelihood of ACL graft failure) or contradicts (e.g., decreases the likelihood of ACL graft failure) the model prediction (Fig 6). It is important to bear in mind that while possessing intuitive appeal, these explanations represent approximations of complex model behavior mapped to individual predictions and, as such, carry important limitations.[14] As is the case when evaluating all novel interventions and devices, the reliability of and trust in machine learning models should be determined through rigorous experimentation in different patient populations and geographic locations.

Similarly, the impact of individual predictor variables on model output can be understood using partial dependence plots.[15,16] For instance, as the symptom duration prior to ACL reconstruction increases, the model output decreases (Fig 7). These plots may subsequently inform researchers and clinicians about how specific thresholds or values of a variable influence model predictions.

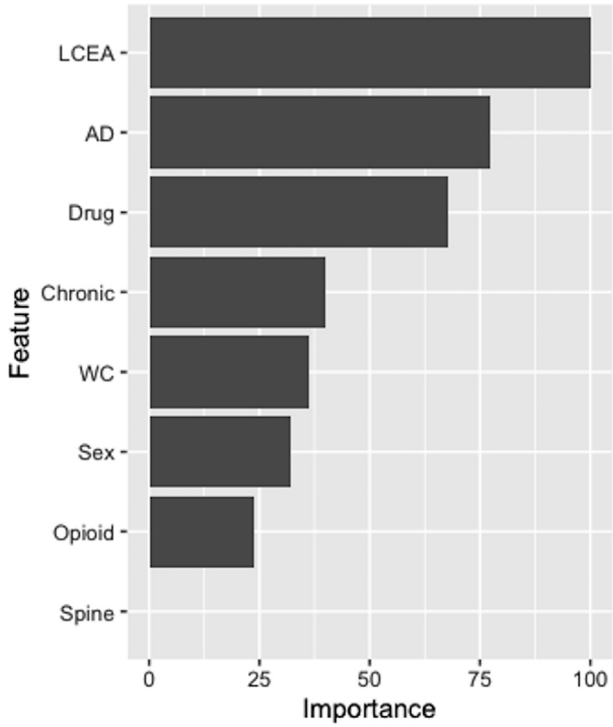

Fig 5. Global variable importance plot. In this example, the lateral center edge angle of the hip (LCEA) confers the greatest contribution to predicting the outcome of interest, followed by a history of anxiety and depression (AD), drug allergies (drug), chronic symptom duration (chronic), workers' compensation status (WC), sex, and chronic opioid use (opioid). A history of spine surgery is the least important variable in the model, contributing to less than 1% of the model prediction. This information is useful in terms of understanding model behavior, as well as during subsequent fine tuning of algorithm performance. Reprinted with permission from Kunze et al.[24]

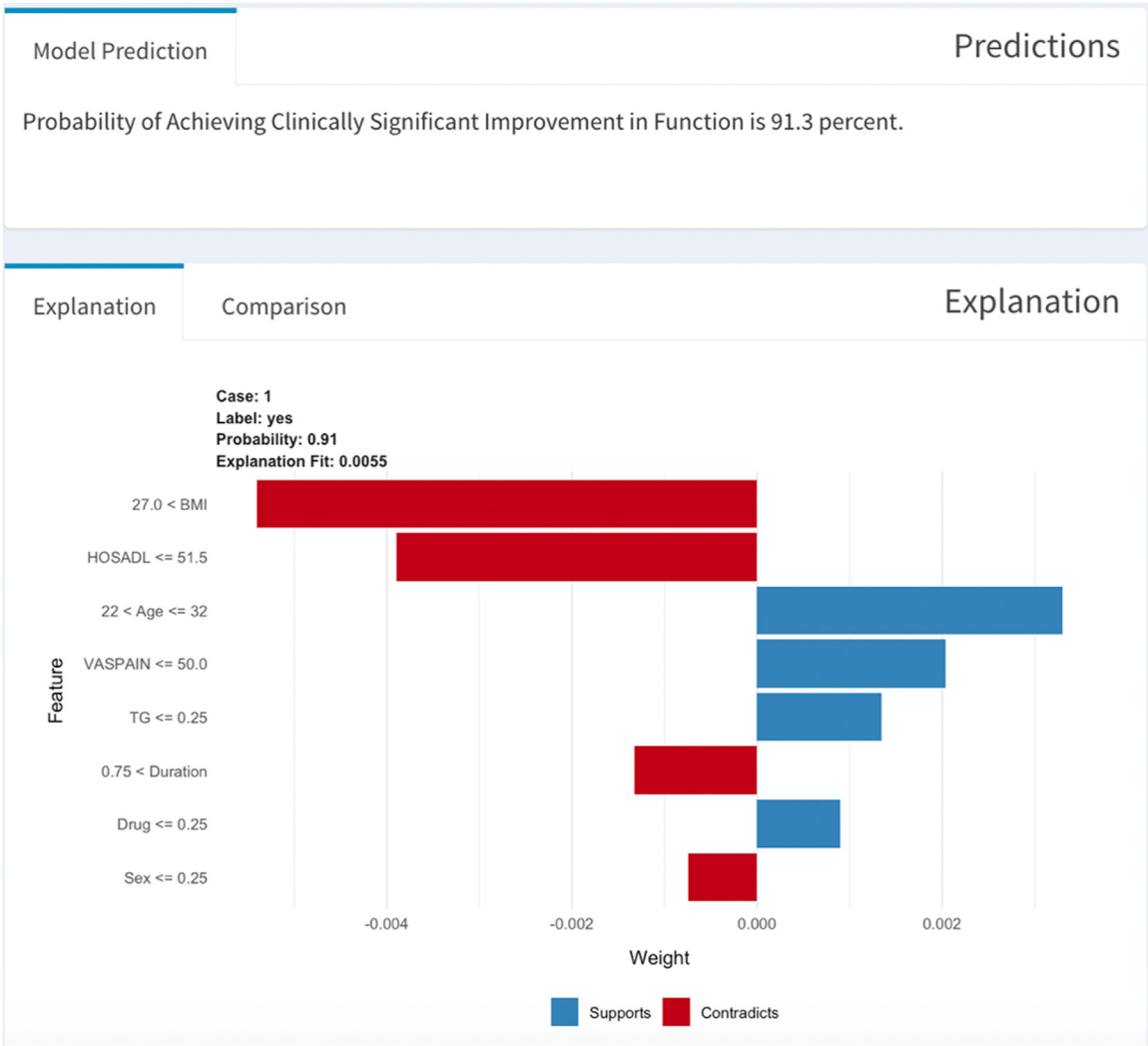

Fig 6. Local interpretable model agnostic plot derived from an online prediction tool generated from machine learning algorithms. The X axis represents weight of each variable, or their importance to the individual patient prediction, whereas the Y axis lists features included in the model. The longer the bar, the more weight the variable contributes to the prediction. The colors represent whether the variable supports (*blue*: positively influence) or contradicts (*red*: negatively influence) the prediction. BMI, body mass index; HOSDAL, hip outcome score—activities of daily living; TG, tonnis grade; VASPAIN, visual analog scale for pain. Reprinted with permission from Kunze et al.[23]

Regression Analysis

In contrast to classification, regression analysis refers to the use of statistical models to predict a continuous outcome. Clinical questions that would be approached using regression analysis include: How many months will it take for my patient to return to sport after an ACL reconstruction? How many dollars is an inpatient surgical intervention likely to cost given patient demographic and operative factors? In both cases, the predicted output is a numeric value (i.e., number of months or dollars, respectively). Unlike in classification analysis, the predictions made by regression models are often evaluated using error metrics rather than accuracy. Error metrics provide an estimate of how close the prediction was to the actual (i.e., ground truth) value, which is likely to be more clinically useful than exact accuracy. Although not an exhaustive list, there are 3 metrics frequently used for the evaluation of regression models with which clinicians should be familiar: R squared, root mean squared error, and mean absolute error.

R Squared (R^2)

R^2, also referred to as the coefficient of determination, is a measure of the degree of variability in the dependent variable that is accounted for or explained by the regression model.[17] It is calculated by dividing the sum of predictive errors (i.e., explained variance) by the total variance and subtracting this from 1. The value of R^2 is always between 0 and 1, with values closer to 1 indicating greater fidelity between the predicted and actual values.

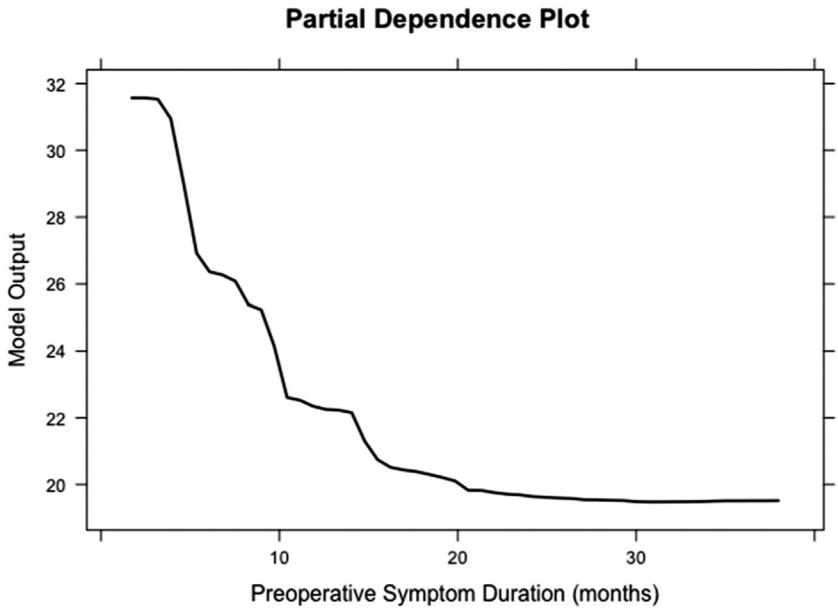

Fig 7. Partial dependence plot demonstrating how different levels of a hypothetical input variable (e.g., preoperative symptom duration) impact model output. As the preoperative symptom duration increases, the corresponding model output decreases. The rate of change in model output is greatest from 0 to 10 months, meaning these levels of the variable have the most influence on the model output. Partial dependence plots are useful for visually depicting the relationship between an input variable and predicted dependent variable. Furthermore, the relationship between variables (e.g., linear, parabolic, complex) can be determined over the entire range of a given input variable.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Although R^2 is useful for determining the model fit, it does not account for overfitting. In situations where many independent variables are used to create the predictive model, the model may be fit too well to the training data and result in poor generalization to the testing data. To combat this, the adjusted R^2 is used that introduces a penalty for adding additional independent variables to the model.

Root Mean Squared Error (RMSE)

Whereas R^2 provides a *relative* measure of model fit, RMSE represents an *absolute* measure of model fit.[17,18] The RMSE is obtained by first calculating the mean squared error (MSE) between model predictions and actual values. Because of the MSE being squared, the square root of the MSE (i.e., RMSE) is calculated to return the prediction error in the original units and augment interpretability. Although a perfect value of RMSE is theoretically 0 (i.e., the predictions and actual values matched exactly), this often never occurs in practice. Rather, a good RMSE value is dataset- or task-dependent and can be used to choose among several candidate models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE)

Like the RMSE, the MAE provides an *absolute* measure of model fit and is easily interpretably because its value matches the units of the predicted output (i.e., the MAE is in dollars or months).[17,18] Similar to RMSE, a perfect MAE value of 0 is rarely achieved in practice, and instead the MAE is used as a comparator between several candidate regression models. However, unlike RMSE, the MAE does not assign greater weight to larger predictive errors (because of the squared term in the RMSE calculation), and thus the MAE changes linearly with the magnitude of prediction error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Limitations of ML

First, although ML as an analytical tool has demonstrated value when applied to diagnostic and prognostic problems in orthopaedic surgery,[4,19-22] it is important to acknowledge several current limitations of this methodology. Specifically, consumers of ML research should be aware of deficiencies in data (i.e., quality and quantity) and model interpretability. The validity of the results of predictive modeling is heavily reliant on the data used to train algorithms. Although not a unique problem to ML, the application of ML to small databases with substantial degrees of missing and inaccurate data are unlikely to produce meaningful or useful results. Second and perhaps more consequential, especially in the field of medicine, is the issue of model interpretability. Owing to their technical complexity, the precise associations between variables and computations of ML models are often unknown, even to their human programmers. Understanding the reasoning leading to specific predictions is essential for building trust in ML,[13] which has far reaching implications in terms of

model deployment and integration into clinical decision making. Both accuracy and interpretability must be taken into consideration when assessing the potential impact of ML research. Finally, a considerable proportion of ML models now published in the literature remain limited to development and internal validation, and the paucity of models that have been successfully validated in external (i.e., independent) populations preclude their use in clinical settings and integration into clinical workflow. Therefore a greater effort is needed to externally validate models to expedite the transition of existing models from proof-of-concept and research practices to clinically useful entities that can be applied to enhance patient care.

Conclusion

The application of applied statistics and specifically ML is continuing to expand among published literature within orthopaedic surgery. As such, it is imperative to understand how to interpret the results of these studies to critically assess reported data and avoid misinterpretation. The current review introduces the interpretation of several commonly reported metrics used in machine learning analyses to promote independent assessments and research transparency.

References

1. Fontana MA. CORR Insights(R): Can machine-learning algorithms predict early revision TKA in the Danish Knee Arthroplasty Registry? *Clin Orthop Relat Res* 2020;478:2102-2104.
2. Leopold SS. Editor's Spotlight/Take 5: Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin Orthop Relat Res* 2019;477:1262-1266.
3. Varady NH, Feroe AG, Fontana MA, Chen AF. Causal language in observational orthopaedic research. *J Bone Joint Surg Am* 2021;103:e76.
4. Ramkumar PN, Kunze KN, Haeberle HS, et al. Clinical and research medical applications of artificial intelligence. *Arthroscopy* 2021;37:1694-1697.
5. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: Definitions, applications, and future directions. *Curr Rev Musculoskelet Med* 2020;13:69-76.
6. Makhni EC, Makhni S, Ramkumar PN. Artificial intelligence for the orthopaedic surgeon: An overview of potential benefits, limitations, and clinical applications. *J Am Acad Orthop Surg* 2021;29:235-243.
7. Myers TG, Ramkumar PN, Ricciardi BF, Urish KL, Kipper J, Ketonis C. Artificial intelligence and orthopaedics: An introduction for clinicians. *J Bone Joint Surg Am* 2020;102:830-840.
8. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010;21:128-138.
9. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925-1931.
10. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: Development and internal validation. *Neurosurgery* 2019;85:E671-E681.
11. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev* 1950;78:1-3.
12. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.
13. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. *Proc 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016:1135-1144.
14. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;3:e745-e750.
15. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001:1189-1232.
16. Greenwell BM. pdf: An R package for constructing partial dependence plots. *R Journal* 2017;9:421-436.
17. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer J Comput Sci* 2021;7:e623.
18. Botchkarev A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *ArXiv* 2018:abs/1809.03006.
19. Kunze KN, Krivicich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: A systematic review [published online December 27, 2021]. *Arthroscopy*. https://doi.org/10.1016/j.arthro.2021.12.030.
20. Kunze KN, Orr M, Krebs V, Bhandari M, Piuzzi NS. Potential benefits, unintended consequences, and future roles of artificial intelligence in orthopaedic surgery research: A call to emphasize data quality and indications. *Bone Jt Open* 2022;3:93-97.
21. Polce EM, Kunze KN, Dooley MS, Piuzzi NS, Boettner F, Sculco PK. Efficacy and applications of artificial intelligence and machine learning analyses in total joint arthroplasty: A call for improved reporting. *J Bone Joint Surg Am* 2022:10-2106.
22. Kunze KN, Rossi DM, White GM, et al. Diagnostic performance of artificial intelligence for detection of anterior cruciate ligament and meniscus tears: A systematic review. *Arthroscopy* 2021;37:771-781.
23. Kunze KN, Polce EM, Nwachukwu BU, Chahla J, Nho SJ. Development and internal validation of supervised machine learning algorithms for predicting clinically significant functional improvement in a mixed population of primary hip arthroscopy. *Arthroscopy* 2021;37:1488-1497.
24. Kunze KN, Polce EM, Rasio J, Nho SJ. Machine learning algorithms predict clinically significant improvements in satisfaction after hip arthroscopy. *Arthroscopy* 2021;37:1143-1151.