Application of Machine Learning Algorithms to Predict Clinically Meaningful Improvement After Arthroscopic Anterior Cruciate Ligament Reconstruction

Kyle N. Kunze,*[†] MD, Evan M. Polce,[‡] BS, Anil S. Ranawat,[†] MD,
Per-Henrik Randsborg,[†] MD, PhD, Riley J. Williams III,[†] MD, Answorth A. Allen,[†] MD,
Benedict U. Nwachukwu,[†] MD, MBA, and HSS ACL Registry Group[§]
Investigation performed at the Hospital for Special Surgery, New York, New York, USA

Background: Understanding specific risk profiles for each patient and their propensity to experience clinically meaningful improvement after anterior cruciate ligament reconstruction (ACLR) is important for preoperative patient counseling and management of expectations.

Purpose: To develop machine learning algorithms to predict achievement of the minimal clinically important difference (MCID) on the International Knee Documentation Committee (IKDC) score at a minimum 2-year follow-up after ACLR.

Study Design: Case-control study; Level of evidence, 3.

Methods: An ACLR registry of patients from 27 fellowship-trained sports medicine surgeons at a large academic institution was retrospectively analyzed. Thirty-six variables were tested for predictive value. The study population was randomly partitioned into training and independent testing sets using a 70:30 split. Six machine learning algorithms (stochastic gradient boosting, random forest, neural network, support vector machine, adaptive gradient boosting, and elastic-net penalized logistic regression [ENPLR]) were trained using 10-fold cross-validation 3 times and internally validated on the independent set of patients. Algorithm performance was assessed using discrimination, calibration, Brier score, and decision-curve analysis.

Results: A total of 442 patients, of whom 39 (8.8%) did not achieve the MCID, were included. The 5 most predictive features of achieving the MCID were body mass index $\leq$27.4, grade 0 medial collateral ligament examination (compared with other grades), intratunnel femoral tunnel fixation (compared with suspensory), no history of previous contralateral knee surgery, and achieving full knee extension preoperatively. The ENPLR algorithm had the best relative performance (C-statistic, 0.82; calibration intercept, 0.10; calibration slope, 1.15; Brier score, 0.068), demonstrating excellent predictive ability in the study's data set.

Conclusion: Machine learning, specifically the ENPLR algorithm, demonstrated good performance for predicting a patient's propensity to achieve the MCID for the IKDC score after ACLR based on preoperative and intraoperative factors. The femoral tunnel fixation method was the only significant intraoperative variable. Range of motion and medial collateral ligament integrity were found to be important physical examination parameters. Increased body mass index and prior contralateral surgery were also significantly predictive of outcome.

Keywords: anterior cruciate ligament; reconstruction; machine learning; artificial intelligence; clinically meaningful; MCID; IKDC

Implementing value-based health care and shared decision-making models within orthopaedic surgery has challenged clinicians and policy makers to determine which metrics should be considered in determining patient-defined success. Patient-reported outcome measures (PROMs) are subjective metrics that are useful for evaluating a patient's perceived state of health and function

before and after treatment. Psychometric transformations of PROMs, such as defining a minimal clinically important difference (MCID), enhance their value by overcoming the challenge of interpreting raw numeric values and by allowing providers to understand what magnitude of outcome change is perceivable and important to the patient. Not achieving a clinically meaningful improvement may increase the risk of diminished patient satisfaction and suboptimal outcome. Therefore, it is imperative to gain a better understanding of which patients may not experience this level of improvement postoperatively, especially for

common sports medicine procedures where many patients have high preoperative expectations and functional demands.

Many orthopaedic sports medicine subspecialties concerning procedures such as hip arthroscopy and cartilage preservation of the knee have endeavored to determine which patient-specific factors are predictive of clinically meaningful outcome improvement. Various factors such as age at the time of surgery, sex, body mass index (BMI), preoperative outcome scores, and prior surgery have been shown to be associated with outcome.[2,6,25,31] However, a major limitation to these studies is that they provide associations on a global scale and may not accurately represent individual patient risk. This is especially true concerning outcomes after anterior cruciate ligament reconstruction (ACLR), where there is a paucity of literature exploring patient-specific risk and clinically meaningful outcome improvement. Indeed, the International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form is one such PROM frequently used to assess outcomes after ACLR; however, risk factors for not achieving clinically meaningful outcome improvement for the IKDC Subjective Knee Form are not well defined at the global or patient-specific levels.

Machine learning is a subset of artificial intelligence and differs from basic statistical modeling in that the methodology prioritizes making repeatable and accurate predictions over providing interpretability.[3] The application of machine learning has gained recent interest given its robust methods for feature selection and outcome classification, thereby allowing clinicians to better understand risk for events such as complications. Furthermore, machine learning has demonstrated validity in predicting clinically meaningful outcome improvement after common orthopaedic procedures.[15-17,23] This allows for risk prediction at the individual patient level, overcoming the limitations of current sports medicine literature. The purpose of the current study was to develop machine learning algorithms to predict achievement of the MCID on the IKDC score at a minimum 2-year follow-up after ACLR. The authors hypothesized that the best-performing machine learning model would have excellent discriminatory performance (area under the curve, ≥ 0.9) for predicting the MCID.

METHODS

Guidelines

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research were followed for this analysis.[7,21] The TRIPOD guidelines represent a systematic checklist of reporting recommendations to which researchers should adhere when performing predictive modeling and machine learning analyses to optimize reporting clarity and the potential for methodological reproducibility.

Study Population

Institutional review board approval was obtained before performing the query and analysis. Patients were identified from the ACL registry of a large academic tertiary care center that comprises 2111 patients from 27 fellowship-trained sports medicine surgeons. Included patients underwent primary ACLR between 2011 and 2013. Exclusion criteria consisted of (1) revision ACLR cases, (2) missing preoperative (after injury but before surgery) IKDC outcome data, and (3) <2-year follow-up data for the IKDC subjective form. Of the initial 2111 patients, 281 (13.3%) were excluded for undergoing revision ACLR. The final cohort of patients was obtained based on exclusion of patients who did not provide 2-year outcome responses. Analysis of baseline characteristics indicated that these patients did not significantly differ based on age ($P = .16$), BMI ($P = .85$), or sex ($P = .15$).
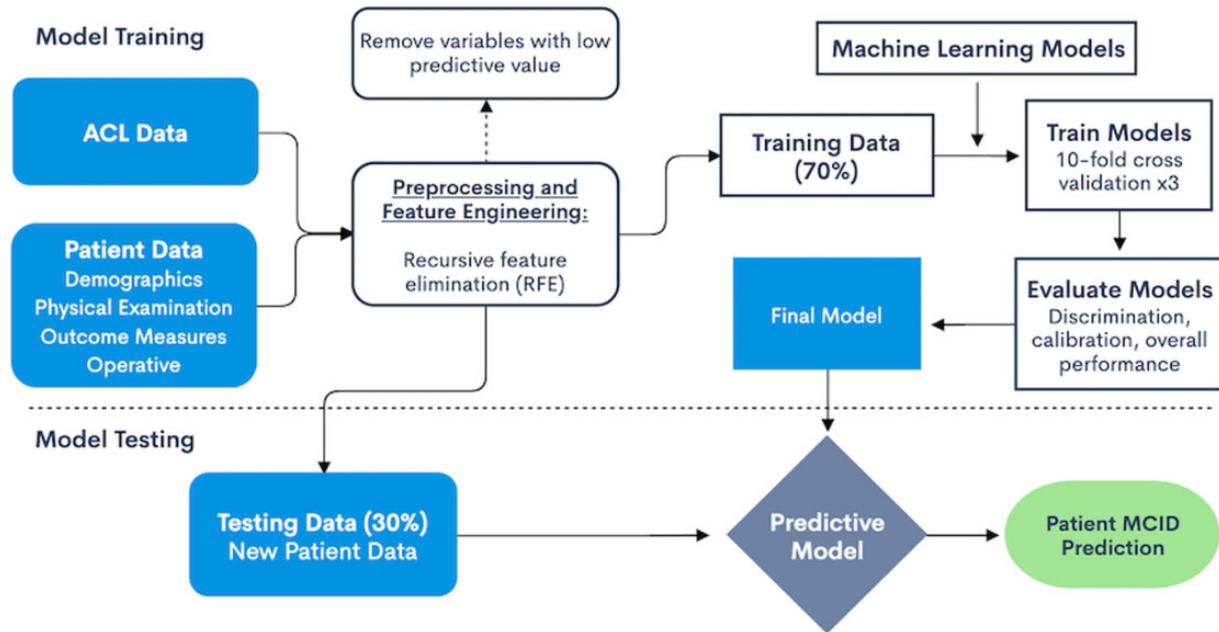
Figure 1. Machine learning algorithm development methodology. ACL, anterior cruciate ligament; MCID, minimal clinically important difference.

Primary Outcome

The primary outcome of interest was the MCID for the IKDC score at a minimum of 2 years postoperatively. The IKDC survey was administered electronically via the Outcomes Based Electronic Research Database platform both preoperatively and at a minimum of 2 years postoperatively. The MCID was calculated using a distribution-based method where the threshold was equal to one-half the standard deviation of the mean change in IKDC outcome scores between 2-year postoperative and preoperative time points.[8] The MCID threshold was determined to be a change of 9.2 points for our specific population.

Covariate Prediction Features and Management of Missing Data

Thirty-six preoperative and intraoperative features routinely collected in the ACL registry were tested for predictive value (Appendix Table A1). All physical examination maneuvers, including the Lachman test, were performed manually. Exploration of the registry revealed that data were missing at random, and therefore multiple imputation was appropriate. No covariates exceeded >30% missing data; therefore, all 36 features were eligible as potential predictors.[18,26] We accounted for missing data using the predictive mean matching method of multiple imputation,[9,19] thereby demonstrating the ability of the machine learning algorithms to address random "missingness" within the registry.

Algorithm Development

Recursive feature elimination with random forest algorithms[12] was applied to determine the covariate features with the highest predictive value (importance). Recursive feature elimination utilizes backward selection by creating a model with all covariates, assigning each variable an importance score and then removing features with the lowest importance scores. After this elimination step, another unique model is built, and the process is repeated until a subset of features that optimizes model performance is selected. These specific variables are used to train the machine learning algorithms.

Algorithm Performance Assessment

The study population of patients was randomly partitioned into training and independent testing (hold-out) sets using a 70:30 split (Figure 1). Six machine learning algorithms (stochastic gradient boosting, random forest, neural network, support vector machine, adaptive gradient boosting, and elastic-net penalized logistic regression [ENPLR]) were trained using 10-fold cross-validation 3 times. Each algorithm uses a different method of optimizing prediction on the training data set based on differences parametricity, assumptions, and methods of "learning." Algorithm performance was then evaluated on an independent testing set of patients (remaining 30%), allowing for internal validation. To determine which model has optimal performance, 4 methods were used to assess each algorithm: (1) discrimination,[29,30] (2) calibration,[29,30] (3) Brier score,[4,14] and (4) decision-curve analysis[29] (Appendix Table A2).

Algorithm Fidelity Assessment

Global variable importance plots and local (patient-specific) interpretable model-agnostic explanations (LIME) were used to assess model fidelity. LIME is a quantitative

TABLE 1
Algorithm Performance in Independent Testing Set (n = 131)[a]

Metric	Stochastic Gradient Boosting	Random Forest	Support Vector Machine	Adaptive Gradient Boosting	Neural Network	Elastic-Net Penalized Logistic Regression
C-statistic	0.70 (0.55 to 0.82)	0.78 (0.65 to 0.92)	0.79 (0.64 to 0.89)	0.79 (0.62 to 0.90)	0.81 (0.68 to 0.90)	0.82 (0.70 to 0.89)
Calibration intercept	0.02 (−0.66 to 0.70)	0.21 (−0.51 to 0.92)	0.19 (−0.43 to 0.81)	0.17 (−0.59 to 0.93)	0.18 (−0.45 to 0.81)	0.10 (−0.56 to 0.75)
Calibration slope	0.49 (0.04 to 0.94)	0.63 (0.21 to 1.06)	5.05 (1.67 to 8.42)	0.49 (0.19 to 0.80)	1.74 (0.66 to 2.83)	1.15 (0.45 to 1.86)
Brier score[b]	0.080 (0.041 to 0.12)	0.083 (0.037 to 0.10)	0.075 (0.038 to 0.11)	0.073 (0.037 to 0.11)	0.069 (0.036 to 0.10)	0.068 (0.035 to 0.10)

[a]Data in parentheses are 95% CIs.
[b]Null model Brier score = 0.077.

visualization technique that provides insight into the decision-making process of complex "black box" machine learning models.[27] Briefly, LIME trains interpretable models to provide numeric and visual representations of the decision the model used to predict the outcome (Appendix Table A2). The best-performing machine learning model (defined as the model with the best discriminatory capability and calibration that had a Brier score less than that of the null Brier score) was subsequently transformed into an open-access application accessible on desktops and smartphones.

RESULTS

Study Population

A total of 442 eligible patients were identified. The median age and BMI were 29.0 years (interquartile range [IQR], 21.0-40.3 years) and 24.2 (IQR, 21.9-26.6), respectively. A total of 231 (52.3%) patients were male. The complete list of preoperative and intraoperative features for the study cohort that were tested for predictive value are listed in Appendix Table A1. The prevalence of patients who achieved the MCID for the IKDC score at a minimum of 2 years postoperatively was 91.2%.

Feature Selection

A combination of the following 8 features optimized algorithm performance: age, BMI, preoperative IKDC score, preoperative Lysholm score, medial collateral ligament (MCL) examination from extension to 30° (grades 0-3), femoral tunnel fixation (intratunnel or suspensory), history of contralateral knee surgery, and preoperative degree of knee extension (recurvatum, neutral, or extension loss). This model did not identify ACL graft type as a feature that optimized algorithm performance.

To determine the relative contribution of the features to the overall predictions, we created and explored a total

of 50 unique cases of LIME with 5000 permutations. Subsequently, preoperative IKDC score >62.1, preoperative Lysholm score between 50 and 64, and BMI >27.4 were associated with not achieving the MCID. Furthermore, use of suspensory femoral fixation, MCL examination grades 2 to 4, previous contralateral knee surgery, knee extension loss or recurvatum, and age >40 or <21 years were consistently feature categories associated with not achieving the MCID.

Relative Algorithm Performance

Performance characteristics of the 6 algorithms are displayed in Table 1. The best-performing algorithm based off of these metrics was the ENPLR model. This model indicated that the 5 most important features for predicting the MCID for the IKDC score were (1) a history of contralateral knee surgery, (2) preoperative knee extension, (3) MCL examination from extension to 30°, (4) method of femoral fixation, and (5) BMI (Figure 2A). This model had a C-statistic of 0.82 (Figure 2B), calibration intercept of 0.10, calibration slope of 1.15 (Figure 3), and Brier score of 0.068. The null model Brier score was 0.077, indicating that this algorithm calibrated predictions appropriately. Decision-curve analysis demonstrated that changes in management based off of the ENPLR model confer the greatest net benefit for optimizing whether a patient would achieve the MCID (Figure 4).

Application Development

The open-source application is available online (http://orthoapps.shinyapps.io/ACLR_IKDC). This application demonstrates how combinations of patient-specific factors can provide risk assessment on a case-by-case basis. An example of the use of this prediction application is shown in Figure 5.
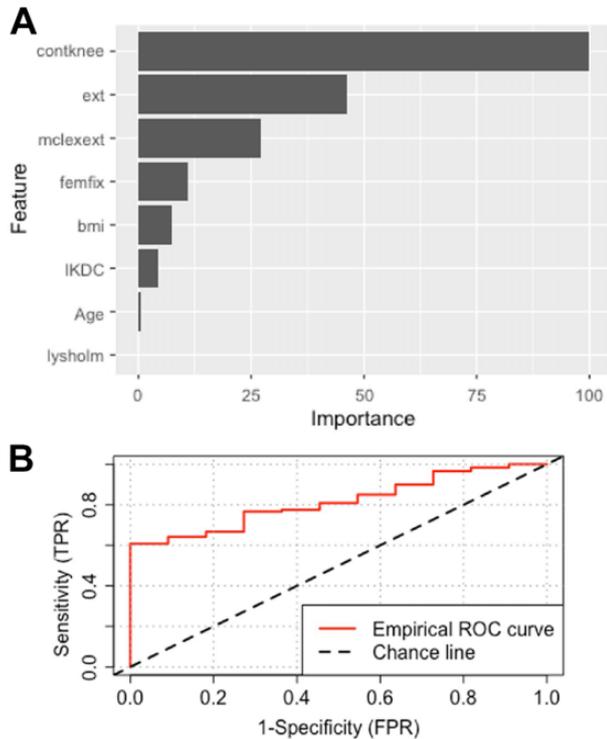
Figure 2. (A) Global variable importance plot and (B) discrimination performance from the elastic-net penalized logistic regression model on the independent testing set. Each predictive weight of each variable is compared among the other 7 variables chosen from recursive feature elimination. The global variable importance plot represents the predictive value of each variable in descending order, with variables having lower predictive value as one moves down the y-axis. This plot indicates that a history of contralateral knee surgery is the most important predictor of achieving the minimal clinically important difference, whereas the importance of the preoperative Lysholm score is negligible. bmi, body mass index; contknee, history of contralateral knee surgery; ext, preoperative knee extension; femfix, femoral tunnel fixation method; FPR, false-positive rate; IKDC, International Knee Documentation Committee; mclexext, medial collateral ligament examination from extension to 30°; ROC, receiver operating characteristic; TPR, true-positive rate.

DISCUSSION

The main findings of the current study are (1) 6 machine learning algorithms were developed, with the ENPLR model demonstrating good ability to predict the MCID for the IKDC score at a minimum of 2 years postoperatively, and (2) the 5 most important features found to predict the MCID for the IKDC score were a history of contralateral knee surgery, preoperative knee extension, MCL examination grade from extension to 30°, method of femoral fixation, and BMI. These findings have important implications for preoperative patient counseling and shared decision-making strategies.
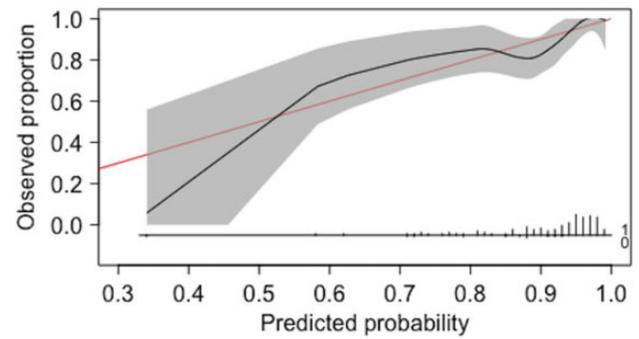

Figure 3. Calibration plot for the elastic net penalized logistic regression (ENPLR) model on the independent testing set of patients. The y-axis displays the true observed proportion of those who achieved the minimal clinically important difference, while the x-axis displays the corresponding predictions made by the ENPLR model. The shaded area indicates the 95% CI of the predicted probabilities. The red line represents perfect prediction.

Machine learning describes statistical processes that exhibit experiential "learning" associated with human intelligence and the capacity to improve via the application and refinement of algorithms.[3] These algorithms learn to make specific decisions based off of this training and can then be modified or enhanced, allowing for the development of a model with powerful ability to transform inputs into an accurate prediction.[3] These predictions are compared against the true outcomes present in the data to determine the accuracy of the algorithms, and models can be modified again to further optimize performance.[13] The current study applied this methodology to predict clinically meaningful outcome improvement after ACLR and potentially enhance the treatment using customized risk predictions.

The 5 most important features for predicting clinically meaningful improvement after ACLR are semimodifiable. For example, in Figure 5, by undergoing a theoretical period of preoperative optimization of knee function, a patient improved the probability of experiencing clinically meaningful improvement by 20% from the previous baseline estimate. Furthermore, it is important that the selected features are clinically plausible. Through multiple permutations of LIME, the current study specifically determined the following 8 preoperative variables as consistently being predictive of not achieving the MCID: IKDC score >62.1, preoperative Lysholm score between 50 and 64, BMI >27.4, use of suspensory femoral fixation, MCL examination grades 2 to 4, previous contralateral knee surgery, knee extension loss or recurvatum, and age >40 or <21 years.

Recent studies have examined all of these factors. Indeed, the importance of the integrity of the MCL as a major restraint to anteromedial instability,[33] the knee extension deficits as a risk factor for poor outcomes and Cyclops syndrome,[10] the potential effect of femoral tunnel fixation methods when used,[32,35] and the associations between characteristic and preoperative PROMs with postoperative outcomes have all been documented.[11]
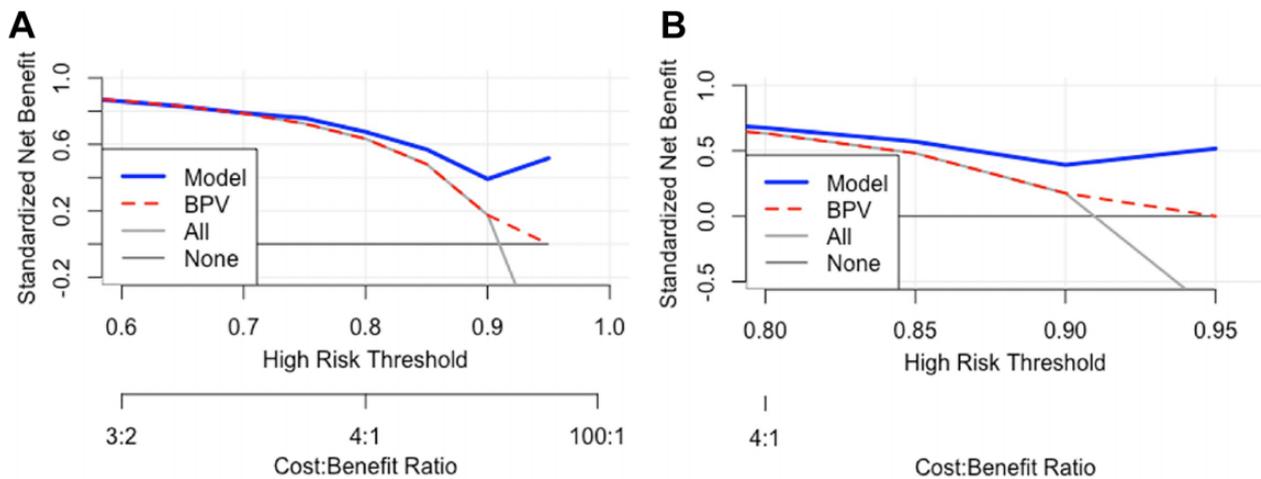
Figure 4 . Decision-curve analysis for the elastic-net penalized logistic regression (ENPLR) model on the independent testing set of patients. The y-axis shows the standardized net benefit of changing management based off of the model (ENPLR), the best-performing variable (BPV; history of contralateral knee surgery), for all patients, and for no patients. The x-axis demonstrates risk thresholds for not achieving the minimal clinically important difference (MCID) as a percentage, as well as the cost to benefit ratio (ratio of false-positive outcomes to true-positive outcomes). (A) View of decision-curve for wide range of risk thresholds. (B) View of decision curves for higher-risk thresholds. When risk is very high (80% likelihood of not achieving MCID), management changes based off of the ENPLR model give greater net benefit (higher likelihood of achieving the MCID) than changing management based on the other decisions.
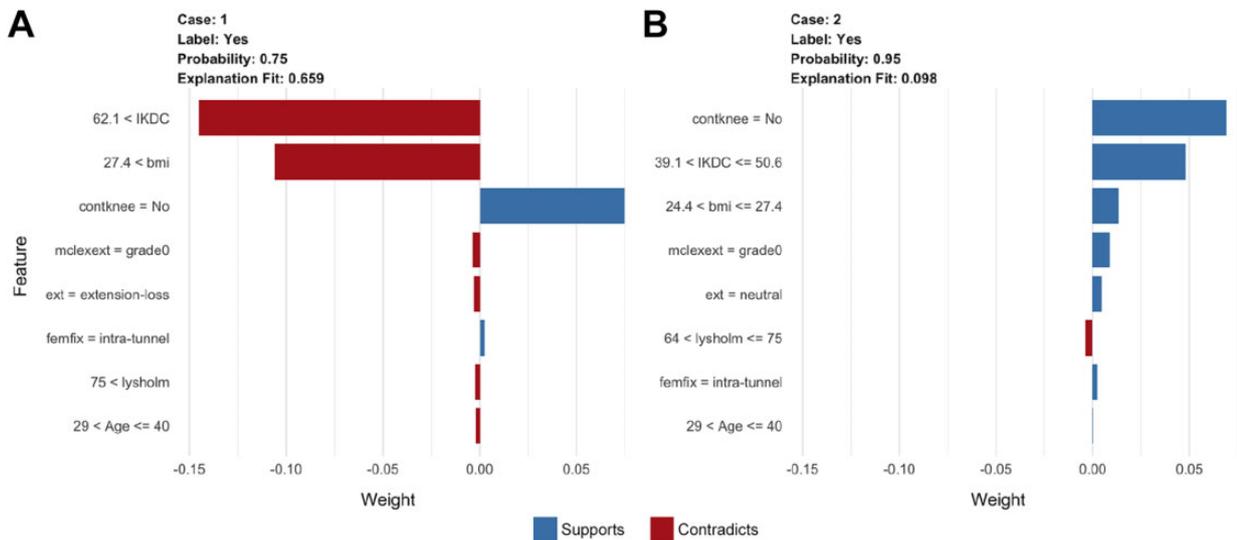
Figure 5.   Demonstration of the clinical effect that application of the clinical decision-making tool derived from the elastic-net penalized logistic regression model can have if applied during the preoperative period. The red bars indicate features that support the probability of achieving the minimal clinically important difference (MCID), and the blue bars indicate features that put the patient at risk of not achieving the MCID. (A) Case 1: A 30-year-old patient with an anterior cruciate ligament tear and body mass index (BMI) of 31 is evaluated at the clinic. The patient has a relatively high level of function (International Knee Documentation Committee [IKDC] score, 75; Lysholm, 80). The patient has never had a contralateral knee surgery. On examination, the patient demonstrates a grade 0 medial collateral ligament examination and has an extension loss; the decision is made to operate using an intratunnel femoral fixation technique. Given this decision, at 2 years postoperatively, there is a 25% chance the patient will not achieve a clinically meaningful improvement in symptoms and function. (B) Case 2: Instead of pursuing surgery, the patient is recommended to first optimize his current health state. The patient is able to decrease BMI into the normal category (BMI, 27) and obtain neutral extension on examination via physical therapy. By using the current algorithm to optimize his health state based off of their specific risk factors, this patient improved the probability of achieving a clinically meaningful improvement in symptoms and function to 95% at 2 years postoperatively. bmi, body mass index; contknee, history of contralateral knee surgery; ext, preoperative knee extension; femfix, femoral tunnel fixation method; mclexext, medial collateral ligament examination from extension to 30°.

Interestingly, although the method of femoral tunnel fixation demonstrated significant predictive value, graft type was not found to optimize algorithm performance in this specific cohort, while previous studies[5,34] have reported associations between graft type and functional outcomes. However, the significant relationship found with fixation may have indirectly been due to graft type. Beyond the scope of the current study, however, it is possible that graft type was not a significant predictor in this cohort given that (1) the majority of patients received autografts, and recent literature has demonstrated inconsistent findings with regard to knee laxity and failure rates among autograft types[28]; and (2) the IKDC score specifically has been demonstrated not to statistically differ among autograft types, suggesting that it may not be sensitive to this specific factor.[1,28,34]

The performance of the ENPLR machine learning model demonstrated excellent discrimination and calibration for predicting which patients will achieve the MCID for the IKDC score at a minimum of 2 years after primary ACLR. Furthermore, the relatively low Brier score of the ENPLR model indicated that the predictions were calibrated well, and decision-curve analysis suggested that patients will experience the greatest benefit from changes in management based off of this model when their risk of not achieving the MCID is high. These findings not only support the validity of the development and performance of the ENPLR model but also the clinical utility that this model confers. The ENPLR model was transformed into an open-access online application that can be used in office-based settings. This type of resource has the potential to enhance shared decision making and improve outcomes for patients undergoing primary ACLR.

A few limitations should be discussed in the context of the current study results. First, although the current study explored a very large number of potential variables, it did not study other variables that may have associations with achieving the MCID for the IKDC score. There remain certain semimodifiable features, such as graft tunnel placement, time from injury to surgery, meniscal integrity, cartilage status, chronicity of MCL laxity, and tibial slope, that have been demonstrated in recent literature[20,22,24] to be associated with outcomes after ACLR and were not routinely collected in the prospective repository used for this study.

Furthermore, in accordance with the purpose of the study and model, which is aimed at allowing for preoperative intervention, we chose features that were modifiable or semimodifiable. This may have also narrowed the potential feature pool. However, we used recursive feature elimination, a powerful statistical tool, to ensure that the variables included in the algorithm development had high predictive value. An additional limitation of the machine learning models in the current study was that they underwent internal validation on patients at a large academic medical center from 27 surgeons but still may not be generalizable to patients in other geographic locations. External validation is required to confirm the performance of these algorithms in heterogeneous populations before using this online tool for active clinical decision making. However, this tool

provides value as an educational aid and demonstrates the value and power of machine learning to integrate individualized patient data to perform clinically useful predictions.

Finally, as this study was not performed prospectively, it is possible that there was heterogeneity in the physical examinations of the 27 surgeons. For example, although knee extension loss and recurvatum were highly predictive variables of not achieving a clinically important outcome, it is theoretically possible that testing specifically for hyperextension was not performed in all patients. However, testing for knee hyperextension is a routine part of the knee examination by sports medicine surgeons at our institution, and the rate of missing data was low for this variable, adding confidence to the knee extension findings and predictive performance of this variable.

CONCLUSION

Machine learning, specifically the ENPLR algorithm, demonstrated good performance for predicting a patient's propensity to achieve the MCID for the IKDC score after ACLR based on preoperative and intraoperative factors. Femoral tunnel fixation method was the only significant intraoperative variable. Range of motion and MCL integrity were found to be important physical examination parameters. Increased BMI and prior contralateral surgery were also significantly predictive of outcome.

AUTHORS

Kyle N. Kunze, MD (Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, USA); Evan M. Polce, BS (University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA); Anil S. Ranawat, MD; Per-Henrik Randsborg, MD, PhD; Riley J. Williams III, MD; Answorth A. Allen, MD; Benedict U. Nwachukwu, MD, MBA (Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, USA); HSS ACL Registry Group (Andrew Pearle, MD; Beth S. Stein, MD; David Dines, MD; Anne Kelly, MD; Bryan Kelly, MD; Howard Rose, MD; Michael Maynard, MD; Sabrina Strickland, MD; Struan Coleman, MD; Jo Hannafin, MD, PhD; John MacGillivray, MD; Robert Marx, MD; Russell Warren, MD; Scott Rodeo, MD; Stephen Fealy, MD; Stephen O'Brien, MD; Thomas Wickiewicz, MD; Joshua S. Dines, MD; Frank Cordasco, MD, MS; and David Altcheck, MD [Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, USA]).

REFERENCES

1. Baverel L, Demey G, Odri GA, Leroy P, Saffarini M, Dejour D. Do outcomes of outpatient ACL reconstruction vary with graft type? *Orthop Traumatol Surg Res*. 2015;101(7):803-806.

2. Beck EC, Nwachukwu BU, Kunze KN, Chahla J, Nho SJ. How can we define clinically important improvement in pain scores after hip arthroscopy for femoroacetabular impingement syndrome? Minimum 2-year follow-up study. *Am J Sports Med*. 2019;47(13):3133-3140.

3. Bini SA. 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):2358-2361.

4. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev*. 1950;78(1):1-3.

5. Cavaignac E, Coulin B, Tscholl P, Nik Mohd Fatmy N, Duthon V, Menetrey J. Is quadriceps tendon autograft a better choice than hamstring autograft for anterior cruciate ligament reconstruction? A comparative study with a mean follow-up of 3.6 years. *Am J Sports Med*. 2017;45(6):1326-1332.

6. Chahla J, Beck EC, Okoroha K, Cancienne JM, Kunze KN, Nho SJ. Prevalence and clinical implications of chondral injuries after hip arthroscopic surgery for femoroacetabular impingement syndrome. *Am J Sports Med*. 2019;47(11):2626-2635.

7. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg*. 2015;102(3):148-158.

8. Copay AG, Eyberg B, Chung AS, Zurcher KS, Chutkan N, Spangehl MJ. Minimum clinically important difference: current trends in the orthopaedic literature, part II. Lower extremity: a systematic review. *JBJS Rev*. 2018;6(9):e2.

9. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study. *BMC Med Res Methodol*. 2019;19(1):14.

10. Delaloye JR, Murar J, Vieira TD, et al. Knee extension deficit in the early postoperative period predisposes to cyclops syndrome after anterior cruciate ligament reconstruction: a risk factor analysis in 3633 patients from the SANTI Study Group database. *Am J Sports Med*. 2020;48(3):565-572.

11. Grindem H, Wellsandt E, Failla M, Snyder-Mackler L, Risberg MA. Anterior cruciate ligament injury—who succeeds without reconstructive surgery? The Delaware-Oslo ACL Cohort Study. *Orthop J Sports Med*. 2018;6(5):2325967118774255.

12. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002;46(1):389-422.

13. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med*. 2020;13(1):69-76.

14. 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(4):E671-E681.

15. Kunze KN, Karhade AV, Sadauskas AJ, Schwab JH, Levine BR. Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. *J Arthroplasty*. 2020;35(8):2119-2123.

16. Kunze KN, Polce EM, Clapp I, Nwachukwu BU, Chahla J, Nho SJ. Machine learning algorithms predict functional improvement after hip arthroscopy for femoroacetabular impingement syndrome in athletes. *J Bone Joint Surg Am*. 2021;103(12):1055-1062.

17. Kunze KN, Polce EM, Rasio J, Nho SJ. Machine learning algorithms predict clinically significant improvements in satisfaction after hip arthroscopy. *Arthroscopy*. 2021;37(4):1143-1151.

18. Lee JH, Huber J Jr. Multiple imputation with large proportions of missing data: how much is too much? Paper presented at: United Kingdom Stata Users' Group Meetings 2011; September 15-16, 2011; London, UK.

19. Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data. *Stat Med*. 2017;36(4):606-617.

20. Lin LJ, Akpinar B, Meislin RJ. Tibial slope and anterior cruciate ligament reconstruction outcomes. *JBJS Rev*. 2020;8(4):e0184.

21. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323.

22. Magnussen RA, Mansour AA, Carey JL, et al. Meniscus status at anterior cruciate ligament reconstruction associated with radiographic signs of osteoarthritis at 5- to 10-year follow-up: a systematic review. *J Knee Surg*. 2009;22(4):347-57.

23. Merali ZG, Witiw CD, Badhiwala JH, Wilson JR, Fehlings MG. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One*. 2019;14(4):e0215133.

24. MOON Knee Group, Spindler KP, Huston LJ, et al. Ten-year outcomes and risk factors after anterior cruciate ligament reconstruction: A MOON longitudinal prospective cohort study. *Am J Sports Med*. 2018;46(4):815-825.

25. Nwachukwu BU, Chang B, Voleti PB, et al. Preoperative Short Form Health Survey score is predictive of return to play and minimal clinically important difference at a minimum 2-year follow-up after anterior cruciate ligament reconstruction. *Am J Sports Med*. 2017;45(12):2784-2790.

26. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649.

27. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Paper presented at: Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA.

28. Samuelsen BT, Webster KE, Johnson NR, Hewett TE, Krych AJ. Hamstring autograft versus patellar tendon autograft for ACL reconstruction: is there a difference in graft failure rate? A meta-analysis of 47,613 patients. *Clin Orthop Relat Res*. 2017;475(10):2459-2468.

29. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.

30. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.

31. Wang D, Chang B, Coxe FR, et al. Clinically meaningful improvement after treatment of cartilage defects of the knee with osteochondral grafts. *Am J Sports Med*. 2019;47(1):71-81.

32. Wang Y, Lei G, Zeng C, et al. Comparative risk-benefit profiles of individual devices for graft fixation in anterior cruciate ligament reconstruction: a systematic review and network meta-analysis. *Arthroscopy*. 2020;36(7):1953-1972.

33. Wierer G, Milinkovic D, Robinson JR, et al. The superficial medial collateral ligament is the major restraint to anteromedial instability of the knee. *Knee Surg Sports Traumatol Arthrosc*. 2021;29(2):405-416.

34. Xie X, Liu X, Chen Z, Yu Y, Peng S, Li Q. A meta-analysis of bone-patellar tendon-bone autograft versus four-strand hamstring tendon autograft for anterior cruciate ligament reconstruction. *Knee*. 2015;22(2):100-110.

35. Zhang S, Liu S, Yang L, Chen S, Chen S, Chen J. Morphological changes of the femoral tunnel and their correlation with hamstring tendon autograft maturation up to 2 years after anterior cruciate ligament reconstruction using femoral cortical suspension. *Am J Sports Med*. 2020;48(3):554-564.

APPENDIX

TABLE A1
Baseline Characteristic and Injury Information for Patients Included in the Final Analysis (N = 442)[a]

Characteristic	Median (IQR) or No. (%)	Missing Data, %	Characteristic	Median (IQR) or No. (%)	Missing Data, %
Age, y	29.0 (21.0-40.3)	0	Anterior drawer endpoint		5.4
Body mass index	24.2 (21.9-26.6)	11.8	A	19 (4.6)	
Male sex	231 (52.3)	0	B	397 (95.4)	
White race	360 (81.4)	0	Posterior sag		6.6
Smoking status		10.2	Normal	407 (98.5)	
Never	320 (80.6)		Flush	3 (0.73)	
Quit >6 mo preoperatively	39 (9.8)		Back	3 (0.73)	
Quit <6 mo preoperatively	14 (3.5)		Posterior drawer endpoint		13.8
Current	24 (6.0)		A	351 (92.6)	
Diabetes mellitus	1 (0.25)	9.7	B	28 (7.4)	
Sports participation	343 (83.7)	7.2	MCL examination		6.6
Contact injury mechanism	112 (28.6)	11.5	Stable	388 (93.0)	
Previous ipsilateral knee surgery	25 (6.3)	0	Loose	29 (7.0)	
Previous contralateral knee surgery	52 (13.2)	6.1	MCL examination: extension to 30°		6.3
Graft source		0	Grade 0	374 (90.3)	
Autograft	316 (71.5)		Grade 1	29 (7.0)	
Allograft	126 (28.5)		Grade 2	6 (1.4)	
Graft configuration		0	Grade 3	5 (1.2)	
Single bundle	426 (96.4)		LCL examination		6.1
Double bundle	16 (3.6)		Stable	410 (98.8)	
Graft type		0	Loose	5 (1.2)	
Bone-patellar tendon-bone	214 (48.4)		LCL examination: extension to 30°		5.9
Hamstring: semitendinosus	38 (8.6)		Grade 0	403 (96.9)	
Hamstring: S+T	72 (16.3)		Grade 1	8 (1.9)	
Quadriceps-bone	5 (1.1)		Grade 2	3 (0.72)	
Iliotibial band	0 (0.0)		Grade 3	2 (0.48)	
Achilles tendon	99 (22.4)		Pivot shift		9.0
Tibialis anterior	13 (2.9)		0	10 (2.5)	
Tibialis posterior	1 (0.23)		1+	100 (24.9)	
Femoral tunnel drilling		4.5	2+	284 (70.6)	
Transtibial	85 (20.2)		3+	8 (2.0)	
Anteromedial	327 (77.9)		Reverse pivot shift		9.0
Outside-in	3 (0.71)		0	392 (97.5)	
Retro-drill	5 (1.2)		1+	5 (1.2)	
Tibial tunnel drilling		6.3	2+	4 (1.0)	
Outside-in	402 (97.1)		3+	1 (0.25)	
Retro-drill	12 (2.9)		SSD: external rotation at 30°		5.7
Femoral/tibial fixation		5.7	Grade 0 (<5°)	409 (98.1)	
Intratunnel	310 (74.0)		Grade 1 (5°-10°)	6 (1.4)	
Suspensory	109 (26.0)		Grade 2 (>10°)	2 (0.48)	
Effusion on examination	89 (21.9)	7.9	SSD: External rotation at 90°		5.7
Preoperative ROM: extension		5.7	Grade 0 (<5°)	406 (97.4)	
Recurvatum	17 (4.1)		Grade 1 (5°-10°)	7 (1.7)	
Neutral	379 (90.9)		Grade 2 (>10°)	4 (0.96)	
Extension loss	21 (5.0)		SSD: Internal rotation at 30°		5.7
Preoperative ROM: flexion		5.9	Grade 0 (<5°)	408 (97.8)	
Symmetric to contralateral side	387 (93.0)		Grade 1 (5°-10°)	6 (1.4)	
Flexion loss	29 (7.0)		Grade 2 (>10°)	3 (0.72)	
Lachman grade		5.7	SSD: Internal rotation at 90°		5.7
0	1 (0.24)		Grade 0 (<5°)	406 (97.4)	
1	12 (2.9)		Grade 1 (5°-10°)	6 (1.4)	
2	400 (96.2)		Grade 2 (>10°)	5 (1.2)	
3	3 (0.72)		Preoperative Lysholm score	64.0 (51.0-76.0)	2.5
			Preoperative IKDC score[b]	50.6 (39.4-61.8)	0
			Preoperative Tegner score	2.0 (1.0-3.0)	0.45

[a]IKDC, International Knee Documentation Committee; IQR, interquartile range; LCL, lateral collateral ligament; MCL, medial collateral ligament; ROM, range of motion; SSD, side-to-side difference; S+T, semitendinosus + gracilis.
[b]At 2-year follow-up, 39 (8.8%) patients did not achieve the MCID for the IKDC score.

TABLE A2
Performance Metric Interpretation Guide[a]

Metric	Description
Discrimination	Assessed through performing ROC analyses and quantifying the AUC (also referred to as the concordance statistic [C-statistic]). The C-statistic is described as the probability that the machine learning model will assign a greater predicted probability to a randomly selected positive case (patient who achieved the MCID) relative to a randomly selected negative case (false-positive case, ie, a patient who did not achieve the MCID).
Calibration	Assesses the agreement between predictions made by the machine learning models and the true observed outcomes. A calibration slope of 1 and calibration intercept of 0 are indicative of perfect prediction by the model. Performance is assessed through quantifying the calibration slope (precision of predictions) and calibration intercept (tendency for model to overestimate or underestimate the observed outcome).
Brier score	A proper scoring function that assesses overall performance and is an extension of calibration and discrimination. The Brier score for each model is equal to the mean squared difference between the true observed outcomes and the model prediction probabilities as a benchmark to quantitatively ensure that the machine learning models are providing valuable predictions and not demonstrating class imbalance; the null model Brier score (Brier score where the predicted probabilities of the null model are equal to the outcome prevalence of the entire study cohort) is calculated. The Brier score of each machine learning model is subsequently compared with this value. In general, lower Brier scores indicate that predictions are better calibrated (with zero being perfect performance and calibration), and Brier scores lower than the null model score indicate model usefulness.
Decision-curve analysis	An analysis that provides insight into potential clinical utility of making changes in patient management based off of the machine learning model and alternative scenarios by comparing the predicted net benefit of using the model at varying risk thresholds. Decision-curve analysis specifically compares changes in management based off of the model, the best-performing predictive variable alone, changes for all patients, and changes for no patients. As the risk threshold probability increases, the cost to benefit ratio (and consequently the weight attributed to false-positive classifications made by the model) increases.
Local interpretable model-agnostic explanations	LIME samples local input variable distributions using a predefined number of permutations and assesses the effect of specific ranges of values for each predictor feature on the primary outcome. The importance of each feature is computed and carried forward based on similarities between the features and the model predictions. LIME then explains model fit (here, how well this local example represents both the global model behavior and its plausibility) and provides a visual explanation of how each feature contributes to the overall predictions, demonstrating how each variable on a case-by-case basis either supports (increases the probability of achieving the MCID) or contradicts (decreases the probability of achieving the MCID) the prediction. A ridge regression model with the Gower distance function and a kernel width of 1.25 was used to optimize LIME in the current study.

[a]AUC, area under the curve; LIME, local interpretable model-agnostic explanations; MCID, minimal clinically important difference; ROC, receiver operating characteristic.