Clin Orthop Relat Res (2019) 477:1267-1279
DOI 10.1097/CORR.0000000000000687

Clinical Orthopaedics
and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Selected Proceedings from the 7th International Congress of Arthroplasty Registries

Can Machine Learning Algorithms Predict Which Patients Will Achieve Minimally Clinically Important Differences From Total Joint Arthroplasty?

Mark Alan Fontana PhD, Stephen Lyman PhD, Gourab K. Sarker BA, Douglas E. Padgett MD, Catherine H. MacLean MD, PhD

Received: 14 September 2018 / Accepted: 30 January 2019 / Published online: 15 May 2019

Abstract

Background Identifying patients at risk of not achieving meaningful gains in long-term postsurgical patient-reported outcome measures (PROMs) is important for improving patient monitoring and facilitating presurgical decision support. Machine learning may help automatically select and weigh many predictors to create models that maximize predictive power. However, these techniques are underused among studies of total joint arthroplasty (TJA) patients, particularly those exploring changes in postsurgical PROMs.

Question/purposes (1) To evaluate whether machine learning algorithms, applied to hospital registry data, could predict patients who would not achieve a minimally clinically important difference (MCID) in four PROMs 2 years after TJA; (2) to explore how predictive ability changes as more information is included in modeling; and (3) to identify which variables drive the predictive power of these models.

Methods Data from a single, high-volume institution's TJA registry were used for this study. We identified 7239

The institution of one or more of the authors (MF, GS, CM, DP, SL) has received funding from National Institute of Arthritis, Musculoskeletal, and Skin Disease (NIAMS) and the Agency for Healthcare Research and Quality (AHRQ).
One of the authors certifies that he (SL), or a member of his immediate family, has received or may receive payments or benefits, during the study period, in an amount less than USD 10,000 from the Japanese Orthopaedic Society of Knee, Arthroscopy, and Sports Medicine (JOSKAS), (Tokyo, Japan); in an amount less than USD 10,000 from Omni Inc (Raynham, MA, USA); in an amount of USD 10,000 to USD 100,000 from Universal Research Solutions LLC (Columbia, MO, USA).
One of the authors certifies that he (DP) is a consultant for and has received or may receive payments or benefits, during the study period, in an amount of USD 10,000 to USD 100,000 from DJO Global (Vista, CA, USA).
Clinical Orthopaedics and Related Research® neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA approval status, of any drug or device before clinical use.
Each author certifies that his or her institution approved the human protocol for this investigation and that all investigations were conducted in conformity with ethical principles of research.
This work was performed at Hospital for Special Surgery, New York, NY, USA.

M. A. Fontana, S. Lyman, G. K. Sarker, D. E. Padgett, C. H. MacLean, Hospital for Special Surgery, Center for the Advancement of Value in Musculoskeletal Care, New York, NY, USA

M. A. Fontana, S. Lyman, Weill Cornell Medical College, Department of Healthcare Policy and Research, New York, NY, USA

M. A. Fontana (✉), Center for the Advancement of Value in Musculoskeletal Care, Hospital for Special Surgery, 535 E 70th St., New York, NY 10021 USA, Email: fontanam@hss.edu

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

Wolters Kluwer

hip and 6480 knee TJAs between 2007 and 2012, which, for at least one PROM, patients had completed both baseline and 2-year followup surveys (among 19,187 TJAs in our registry and 43,313 total TJAs). In all, 12,203 registry TJAs had valid SF-36 physical component scores (PCS) and mental component scores (MCS) at baseline and 2 years; 7085 and 6205 had valid Hip and Knee Disability and Osteoarthritis Outcome Scores for joint replacement (HOOS JR and KOOS JR scores), respectively. Supervised machine learning refers to a class of algorithms that links a mapping of inputs to an output based on many input-output examples. We trained three of the most popular such algorithms (logistic least absolute shrinkage and selection operator (LASSO), random forest, and linear support vector machine) to predict 2-year postsurgical MCIDs. We incrementally considered predictors available at four time points: (1) before the decision to have surgery, (2) before surgery, (3) before discharge, and (4) immediately after discharge. We evaluated the performance of each model using area under the receiver operating characteristic (AUROC) statistics on a validation sample composed of a random 20% subsample of TJAs excluded from modeling. We also considered abbreviated models that only used baseline PROMs and procedure as predictors (to isolate their predictive power). We further directly evaluated which variables were ranked by each model as most predictive of 2-year MCIDs.

Results The three machine learning algorithms performed in the poor-to-good range for predicting 2-year MCIDs, with AUROCs ranging from 0.60 to 0.89. They performed virtually identically for a given PROM and time point. AUROCs for the logistic LASSO models for predicting SF-36 PCS 2-year MCIDs at the four time points were: 0.69, 0.78, 0.78, and 0.78, respectively; for SF-36 MCS 2-year MCIDs, AUROCs were: 0.63, 0.89, 0.89, and 0.88; for HOOS JR 2-year MCIDs: 0.67, 0.78, 0.77, and 0.77; for KOOS JR 2-year MCIDs: 0.61, 0.75, 0.75, and 0.75. Before-surgery models performed in the fair-to-good range and consistently ranked the associated baseline PROM as among the most important predictors. Abbreviated LASSO models performed worse than the full before-surgery models, though they retained much of the predictive power of the full before-surgery models.

Conclusions Machine learning has the potential to improve clinical decision-making and patient care by helping to prioritize resources for postsurgical monitoring and informing presurgical discussions of likely outcomes of TJA. Applied to presurgical registry data, such models can predict, with fair-to-good ability, 2-year postsurgical MCIDs. Although we report all parameters of our best-performing models, they cannot simply be applied off-the-shelf without proper testing. Our analyses indicate that machine learning holds much promise for predicting orthopaedic outcomes.

Level of Evidence Level III, diagnostic study.

Introduction

Patient-reported outcome measures (PROMs) are increasingly collected as a means of measuring healthcare quality and value before and after elective total joint arthroplasty (TJA) [10, 34]. While measuring PROMs is an important step toward more patient-centered care, the mere determination of whether a patient's score goes up or down after an intervention is insufficient to determine whether that intervention was effective. What really matters is whether a patient's score changed by a sufficiently large margin, that is, whether an improved score constitutes a minimally clinically important difference (MCID, sometimes called a "minimally clinically important change" or "minimally clinically important improvement") [27, 30]. The MCID is defined as the minimum change in PROM scores that patients perceive as beneficial or clinically meaningful [3, 4, 15, 25, 37]. Identifying patients at risk of not achieving a PROM MCID, particularly before surgery, is important for allocating resources toward better monitoring patients and may aid in presurgical decision support. Many papers have explored predicting PROMs, for example, predicting pain and function after spine surgery [26] and total knee replacement [41], predicting quality of life after total hip replacement [35], predicting satisfaction after TJA [7, 20, 42, 45], and predicting whether a patient undergoing foot and ankle surgery will achieve a MCID [23].

At the same time, the use of machine learning, a subfield of artificial intelligence at the intersection of computer science and statistics that uses data-driven approaches to "teach" computer algorithms to perform specific tasks (for example, prediction), has seen more common use. With sustained improvements in processing power, the rise of cloud-based computing, and ever-larger datasets, machine learning's application to healthcare has involved tasks as diverse as classifying early detection of heart failure onset [11] (using electronic health records) and classifying skin cancer [16] (using images). Outcomes more directly relevant for orthopaedics include predicting mortality [12, 22, 38, 40]; readmissions [17, 21, 38, 44]; complications [12, 22, 47], such as sepsis [18, 24]; and prolonged length of stay [38].

Although machine learning algorithms are often rebranded classical statistical techniques, there are deeper methodological distinctions that only a subset of the orthopaedics prediction evidence implements. For example, machine learning approaches typically do not define a priori exactly which variables will be predictive and how; instead, the algorithm is allowed to perform variable selection and weighting among all available variables. Moreover, many papers report in-sample statistics, evaluating model performance on the same patients used to generate the model. By definition, a model generated using specific data explains those same data well; the true test of predictive ability is whether the model performs well on a validation sample of data from other patients. Although some of the papers referenced above predicted orthopaedic outcomes with proper machine learning techniques, and others explored factors associated with postsurgical PROMs with more traditional techniques, no studies have attempted to predict postsurgical PROMs with proper machine learning techniques among TJA patients, particularly in terms of whether a patient is likely to achieve an MCID.

The purposes of this study, therefore, were (1) to evaluate whether machine learning algorithms, applied to hospital registry data, could predict patients who would not achieve a MCID in four PROMs 2 years after TJA; (2) to explore how predictive ability changes as more information is included in modeling; and (3) to identify which variables drive the predictive power of these models.

Patients and Methods

We conducted a retrospective study using data from a single, high-volume institution's hip and knee replacement registry. We identified 7239 adult hip and 6480 adult knee patients who underwent elective total joint arthroplasty (TJA) between May 2007 and April 2012, and, for at least one of four PROMs, completed both baseline and 2-year followup PROMs. Most patients in our machine learning sample underwent primary unilateral replacements, although both revisions and bilateral arthroplasties were also included (Table 1). Medicare was the primary payor for more than half of patients. Women were also more than half of patients (3764 of 7239 [52%] of hips and 3953 of 6480 [61%] of knees). Hip patients were aged 63.0 ± 11.6 years (mean ± SD) and knee patients were aged 66.9 ± 9.7 years. Summaries of ASA scores, years of education, BMI, total operation time in minutes, length of stay in days, and number of final procedure and diagnosis codes were recorded (Table 1).

During this period, there were 43,313 total TJAs eligible for registry inclusion, among which 19,187 (44.3%) joined the registry by completing a baseline survey. Among those, 13,719 (71.5%) are included in our sample; these patients, compared to the 29,594 who were not included, are less likely to be female

Table 1. THA and TKA registry study sample characteristics

Variable	THA (n = 7239) Number (%)				TKA (n = 6480) Number (%)			
Binary								
Unilateral primary	6370 (88%)				5314 (82%)			
Unilateral revision	507 (7%)				389 (6%)			
Bilateral	362 (5%)				778 (12%)			
Medicare	3185 (44%)				3758 (58%)			
Female	3764 (52%)				3953 (61%)			
Ordinal								
ASA score = 1	636 (9%)				199 (3%)			
= 2	5380 (74%)				4917 (76%)			
= 3	1212 (17%)				1359 (21%)			
= 4	8 (0.1%)				2 (0.03%)			
= Missing	3 (0.04%)				3 (0.05%)			
Continuous	**Mean**	**SD**	**Minimum**	**Maximum**	**Mean**	**SD**	**Minimum**	**Maximum**
Age (years)	63	12	18	102	67	10	22	96
Years of education	16	3	10	19	16	3	10	19
BMI (kg/m^2)	28	5	15	70	30	6	16	63
Total time in OR (minutes)	134	43	41	508	142	38	66	391
Length of stay (days)	5	2	2	58	5	2	1	54
Number of procedure codes	2	1	1	12	2	1	1	13
Number of diagnosis codes	6	3	1	28	7	4	1	28

ASA = American Society of Anesthesiologists; OR = operating room.

(p = 0.065), likely to have a lower body mass index (BMI, p = 0.008), likely to be American Society of Anesthesiologists (ASA) class 2-4 (as opposed to ASA class 1, p = 0.006, 0.006, 0.043, respectively), likely to have a shorter length of stay (p < 0.001), and likely to have fewer diagnoses (p < 0.001). However, overall, these variables explained less than 1.2% of the variation in cohort inclusion (pseudo R^2 from multivariable logistic regression).

The four PROMs we focused on were the SF-36 physical component score (PCS), the SF-36 mental component score (MCS), the Hip Disability and Osteoarthritis Outcome Score for joint replacement (HOOS JR), and the Knee Disability and Osteoarthritis Outcome Score for joint replacement (KOOS JR). The SF-36 was collected on all patients; we included the PCS and MCS given their plenary focus on a patients' overall health. The HOOS JR and KOOS JR were collected on only hip and knee patients, respectively; we included these short-form PROMs because of their focus on the particular joint and procedure of interest, as well as the fact that the Centers for Medicare & Medicaid Services has adopted them for its bundle payment programs.

Our primary outcome of interest was whether a patient achieved an MCID between preoperative baseline to 2 years after surgery for each PROM. Where possible, we favored published, anchor-based MCIDs, given their superior construct and face validity compared with distribution-based MCIDs (for example, a value of 17.7 for the HOOS JR and a value of 13.6 for the KOOS JR) [28]. Given a lack of published anchor-based MCIDs for the SF-36 PCS and MCS for TJA patients, we relied on the distribution-based heuristic of one-half SD and, therefore, used a value of 5.0 for both (given the component scores are calibrated to have a SD of 10.0). This choice is also corroborated by analyses of the highly similar SF-12 [5, 6, 43].

In all, 6465 patients who underwent hip and 5738 patients who underwent knee replacement had valid SF-36 PCS and MCS scores at baseline and 2 years. Further, 7085 patients who underwent THA had valid HOOS JR scores, and 6205 patients who underwent TKA had valid KOOS JR scores. We determined the percent of TJAs who reached the 2-year MCID for each PROM, who did not reach the MCID, or had such a high baseline score that they could not mathematically reach the MCID (Table 2). More than 75% of patients achieved an MCID for the PCS, HOOS JR, and KOOS JR; this number is only 39% for the MCS. We did not include the subset of patients with a sufficiently high preoperative baseline score such that it is mathematically impossible for them to achieve an MCID (for example, a patient with a 90 of 100 baseline HOOS JR could not possibly achieve a 17.7 point improvement; we can perfectly predict they will not achieve an MCID). There were 18 (0.3%) such patients for the KOOS JR and 114 (1.6%) for the HOOSJR (and none for the two SF-36 PROMs) (Table 2).

We next defined the predictors or features (using machine learning terminology) and categorized all registry features into four buckets based on when each was available in our sample (Table 3). These were: (1) before the decision to have surgery, such as demographics and self-reported medical, surgical, and medication history (see Appendix, Supplemental Digital Content 1, http://links.lww.com/CORR/A141); (2) before surgery, such as baseline PROM scores, primary operating surgeon, and other procedures conducted before arthroplasty during the index inpatient stay (see Appendix, Supplemental Digital Content 2, http://links.lww.com/CORR/A142); (3) before

Table 2. Percentage of patients who reached, did not reach, or could not reach the MCID for each PROM

		THA				TKA			
PROM	MCID	Number	Percent reached MCID	Percent did not reach MCID	Percent could not reach MCID*	Number	Percent reached MCID	Percent did not reach MCID	Percent could not reach MCID*
SF-36 PCS	5.0	6465	85%	15%	0%	5738	75%	25%	0%
SF-36 MCS	5.0	6465	39%	61%	0%	5738	35%	65%	0%
HOOS JR†	17.7	7085	88%	11%	2%				
KOOS JR†	13.6					6205	82%	18%	0.3%

*Patients who could not reach the MCID were those who had such a high baseline score that it was mathematically impossible possible for them to improve enough to achieve an MCID; for example, someone with a baseline KOOS JR score of 99 of 100 could only improve by one point, which is much lower than the MCID.

†percentages for HOOS JR and KOOS JR do not add to 100% because of rounding; MCID = minimally clinically important difference; PROM = patient-reported outcome measure; PCS = physical component score; MCS = mental component score; HOOS JR = Hip Disability and Osteoarthritis Outcome Score for joint replacement; KOOS JR = Knee Disability and Osteoarthritis Outcome Score for joint replacement.

Table 3. Numerical and categorical features used in each machine learning model

Model	Numerical features	Categorical features
Before decision	Age; BMI; years of education; years of pain medication use	Knee versus hip; race; Hispanic or not; sex; laterality (left, right, bilateral); primary versus revision; zip code; primary and secondary payor; cohabitation; prior shoulder, hip, or knee replacements (same-side or contralateral); prior spinal surgery; ever taken bisphosphonates; ever had cortisone injection on operated joint; self-reported disease history
Before surgery	Baseline PROM summary scores for: SF-36 (PCS, MCS, and eight domains), HOOS/KOOS (five domains and JR), LEAS, WOMAC (three domains), VAS (pain, fatigue, general health), EQ-5D, expectations; number of procedures before index surgery date during inpatient stay; all earlier numerical features	Day of week of baseline survey; month of year of baseline survey; doctor; surgeon; day of week of surgery; month of year of surgery; whether attended pre-surgery class; CCS procedure codes before surgery date during inpatient stay; ASA score; all earlier categorical features
Before discharge	Total time in OR; number of nonindex procedures on day of surgery; all earlier numerical features	Hour admitted; anesthesia type; CCS nonindex procedure codes on surgery date during inpatient stay; all earlier categorical features
After discharge	Length of stay; total number of procedure codes during inpatient stay; number of procedure codes after surgery date during inpatient stay; total number of diagnosis codes during inpatient stay; Charlson comorbidity index; all earlier numerical features	Discharge disposition; CCS nonindex procedure codes after surgery date during inpatient stay; CCS diagnosis codes during inpatient stay; Elixhauser and Charlson comorbidity indicators; all earlier categorical features
Abbreviated 1	PROM summary scores for outcome of interest (SF-36 PCS, SF-36 MCS, HOOS JR, or KOOS JR)	Knee versus hip
Abbreviated 2	PROM summary scores for all baseline PROMs (SF-36 PCS, SF-36 MCS, HOOS JR, and KOOS JR)	Knee versus hip

ASA = American Society of Anesthesiologists; OR = operating room; PROM = patient-reported outcome measure; PCS = physical component score; MCS = mental component score; HOOS = Hip Disability and Osteoarthritis Outcome score; HOOS JR = Hip Disability and Osteoarthritis Outcome Score for joint replacement; KOOS = Knee Disability and Osteoarthritis Outcome Score; KOOS JR = Knee Disability and Osteoarthritis Outcome Score for joint replacement; LEAS = Lower Extremity Activity Scale; EQ-5D = 5-domain EuroQol quality of life survey; CCS = clinical classifications software.

hospital discharge, such as total operation time, hour admitted, anesthesia type, and other procedures (see Appendix, Supplemental Digital Content 3, http://links.lww.com/CORR/A143); and (4) after hospital discharge, such as length of stay, claims-based diagnosis codes, and discharge disposition (see Appendix, Supplemental Digital Content 3, http://links.lww.com/CORR/A143). Categorical variables with "n" categories were transformed into n-1 binary variables. It is important to note that patients were only enrolled in the registry after deciding to have surgery. Therefore, baseline PROMs were measured before surgery but not before the decision to undergo surgery.

Statistical Analysis

For each of our four PROMs, we began by randomly splitting TJAs into two mutually exclusive sets: a training set (80% of TJAs), and a validation set (remaining 20%). Importantly, all modeling was conducted on the 80%

training set, and all testing on the 20% validation sample. A model generated using specific data is by definition fit to explain those same data well; the true test of predictive ability is whether the model performs well on data from other patients.

We considered models that incrementally added features available at each of the four time points to predict whether a patient would not achieve a 2-year MCID. That is, the before-decision models included only features available before decision; the before-surgery models included features available before decision and before surgery.

Missing features were handled differently for categorical and numerical variables. For categorical variables, we created a separate "missing" category, thereby allowing missing information to be informative. For numerical variables, we imputed the missing values to the mean among nonmissing observations in the training set. Only three categorical features had missing values for more than 10% of patients; only 12 categorical features had missing values for more than 10% of patients (see Appendices, Supplemental Digital Content 1, http://links.lww.com/CORR/A141, 2, http://links.lww.com/CORR/A142, and 3, http://links.lww.com/CORR/A143).

Given whether a patient achieves a MCID is a direct function of the baseline PROM, we further considered abbreviated models, which used as predictors only baseline PROMs and whether the patient was having their hip or knee replaced. The first set of abbreviated models used only the baseline PROM for the MCID being predicted, for example, for predicting the PCS MCIDs, we use only the PCS baseline scores. The second set of abbreviated models used all baseline PROMs. For example, for predicting PCS MCIDs, we used the PCS, MCS, KOOS JR, and HOOS JR baseline scores. By comparing these models to others, we could isolate how much of our prediction performance was driven by the baseline PROMs and, therefore, the mechanical relationship between baseline score and MCID achievement.

For each of the four PROMs and four time points, we applied three of the most popular supervised machine learning algorithms: (1) logistic least absolute shrinkage and selection operator (LASSO), (2) random forest, and (3) linear support vector machine. Supervised machine learning refers to a class of algorithms that learns a mapping of inputs to some output based on many input-output examples; this mapping can then be applied to new inputs to predict their likely outputs (or probabilities of outputs). Classical logistic regression is an example of a supervised machine learning algorithm; however, it is typically used to identify associations (such as, calculate the sign and magnitude of relevant coefficients). Here, our goal is different: namely, to generate models and then perform prediction. We specifically chose three algorithms that automatically select and weigh a subset of features among a larger pool of available variables. These algorithms often have tuning parameters associated with them (also called hyperparameters), which control technical details of how each algorithm operates.

Logistic LASSO is very similar to classical logistic regression, but with an additional tuning parameter that forces some variables to be assigned zero weight (coefficients equal to zero); the other features are included in the LASSO and assigned non-zero coefficients. This regularization tends to make models perform better on new samples (or validation samples) because, with fewer features, models are less fit to the specific training data, and therefore more generalizable. This is exactly why we chose logistic LASSO over classical logistic regression: the latter does not perform regularization, and instead includes all variables, and is therefore likely to perform much worse on new or validation data.

Random forests are a nonparametric approach (that is, no coefficients) that are likely less familiar to readers. They are based on the average predictions of many individual decision trees. For each decision tree, first a random subset of features are chosen (the exact number is based on a tuning parameter). Next, among the selected features, the feature that splits the data into cases and controls with the most discriminatory power is first included (at the base of the tree), and subsequent branches are filled out with features that split the data second best, third best, and so on, until some predefined stopping criteria are reached (controlled by more tuning parameters). After generating many thousands of these random decision trees (the exact number is yet another tuning parameter), each of which produces a prediction (and can be generated by following the branches of the tree to the relevant leaf), the predictions are finally averaged across trees to reach a forest-based prediction. There are two types of random forest – classification and regression – which vary in terms of how predictions are calculated. Although typically classification is used with binary outcomes, and regression with continuous outcomes, regression random forest can also be used with binary outcomes; we used regression random forest here given evidence of superior predictive ability in binary orthopaedic outcomes [8, 29].

A support vector machine (SVM) assigns individuals to cases and controls based on their relative distances from each other according to their features. The algorithm is easy to visualize with only two features (two dimensions) and a linear kernel – draw a line that separates the cases and controls such that cases and controls, on average, are separated the most. With more features, this line (support vector) becomes a higher-dimensional separating hyperplane that maximizes the separation between that hyperplane and the data points. With other kernels, this support vector can be a nonlinear or curved. As with logistic LASSO, regularization is often included via a tuning parameter to control the number of features considered by the algorithm. While we included regularization, we only used a linear kernel; other kernels (for example, polynomial, sigmoid, radial basis function) were experimented with, but yielded no better predictive power and much longer computation times.

We empirically optimized the various tuning parameters of these algorithms by performing fivefold cross-validation within the 80% training sample. That is, to figure out which values of the hyperparameters tuned the models to provide the best predictions, we trained a model on a random 80% of the training set and tested it on the remaining 20% of the training set. We repeated this for all five possible permutations, averaged the results, and repeated for various potential values of the tuning parameters, noting and retaining which values performed best. For logistic LASSO and linear SVM, the only tuning parameter was the regularization parameter. For random forest, we tuned three: (1) the number of features used for each decision tree, (2) the size of the terminal nodes (that is, the minimum number of samples required to be at a leaf), and (3) the number of trees generated in each forest.

We evaluated the performance of our models by using area under the receiver operating characteristic (AUROC) statistics on the 20% randomly selected validation set of patients not used in model generation. AUROCs theoretically range from 0.50 (no better than a coin flip) to 1.00 (perfect prediction) and can be interpreted as the probability that two random patients, one who achieved the MCID and one who did not, would be correctly ranked by the model in terms of their predicted probabilities. We calculated 95% confidence intervals (CIs) for the AUROCs via bootstrapping (that is, we calculated the AUROC on a random subset of validation sample patients, repeated that 1000 times with different random subsets, sorted the 1000 estimates from smallest to biggest, and cited the 25th and 975th as the 95% CI).

For each PROM, for our best-performing full logistic LASSO models (and in the case of ties, most parsimonious, that is, the fewest features), as well as for our best-performing abbreviated logistic LASSO models, we report all model parameters (including coefficients, intercepts, and how we imputed and standardized each continuous feature). Although these models cannot be used off-the-shelf without proper testing, our hope, beyond providing methodological guidance, is to facilitate such validation work. We also report calibration tables and plots for these models, which indicate whether the probabilities predicted for our validation samples actually reflect observed outcomes (for example, does a predicted probability of 50% really translate to equal likelihoods of achieving an MCID or not).

We also compared machine learning model performances with dummy models. These dummy models used very simple heuristics for prediction, for example, they always guessed "achieve MCID" or guessed randomly. The idea is that it is possible that a very simple heuristic might outperform our machine learning models, and we wanted to make sure such a simple, powerful heuristic does not exist. We only report the best-performing dummy models.

Finally, we also detailed the top five features ranked by each model (for each PROM and time point) as most predictive. Although these top predictors cannot be interpreted causally nor as unbiased associations, they are useful to report to get a direct sense of how the models are working and what is driving their predictions. We focused on logistic LASSO before-surgery models for three reasons: (1) our three machine learning models performed nearly identically for a given PROM and time point, (2) logistic LASSO is similar to standard logistic regression and is therefore likely to be most familiar and easily understood (and it is straightforward to report its coefficient estimates), and (3) our before-surgery models performed just as well as models that incorporated additional information available during hospitalization and after discharge. We evaluated feature importance based on the magnitude of features' coefficients.

Results

Could Machine Learning Models Predict Patients Who Would Not Achieve a MCID?

The three machine learning algorithms performed in the poor-to-good range, with AUROCs ranging from 0.60 to 0.89. They performed virtually identically to each other for a given time point and PROM, such as the SF-36 PCS and MCS (see Appendix, Supplemental Digital Content 4, http://links.lww.com/CORR/A144) and HOOS JR and KOOS JR (see Appendix, Supplemental Digital Content 5, http://links.lww.com/CORR/A145). All machine learning models performed better than dummy models, which never achieved AUROCs higher than 0.52.

How Does Predictive Ability Change as More Information is Included?

Including more information available from "before decision" to "before surgery" improved predictive power dramatically (fair-to-good range), but adding further information available before discharge and after discharge yielded no further improvements. The abbreviated models performed worse than the full before-surgery models, but not by much (particularly for SF-36 MCS). AUROCs for predicting PCS scores at the four time points (before decision, before surgery, before discharge, and after discharge) for the logistic LASSO models were: 0.69, 0.78, 0.78, and 0.78, respectively (Fig. 1A). For MCS scores these were: 0.63, 0.89, 0.89, and 0.88 (Fig. 1B); for

HOOS JR they were: 0.67, 0.78, 0.77, and 0.77 (Fig. 1C); and for KOOS JR, they were: 0.61, 0.75, 0.75, and 0.75 (Fig. 1D).

Given the fact that the before-surgery models performed best (and had the fewest features among ties), we report the full parameters for each of these four PROMs MCIDs prediction models: SF-36 PCS (see Appendix, Supplemental Digital Content 6, http://links.lww.com/CORR/A146), SF-36 MCS (see Appendix, Supplemental Digital Content 7, http://links.lww.com/CORR/A147), HOOS JR (see Appendix, Supplemental Digital Content 8, http://links.lww.com/CORR/A148), and KOOS JR (see Appendix, Supplemental Digital Content 9, http://links.lww.com/CORR/A149). The associated calibration tables (see Appendix, Supplemental Digital Content 10, http://links.lww.com/CORR/A150) and figures (see Figure, Supplemental Digital Content 11, http://links.lww.com/CORR/A151) indicate that the probabilities reported by the before-surgery models have good, albeit imperfect face validity. The SF-36 MCS MCID model is nearly perfectly calibrated. The other models are not as well calibrated, although reported probabilities are indeed monotonically related to their actual incidences in the validation sample. In practice, other facilities wanting to use our models would first want to test them on their own validation sample (for acceptable AUROC), and then use their validation sample to create their own calibration tables.

The first abbreviated model achieved AUROCs of 0.65, 0.88, 0.68, and 0.69 (for the PCS, MCS, HOOS JR, and KOOS JR, respectively) (Fig. 1A-D). The second abbreviated model achieved AUROCs of 0.71, 0.88, 0.72, and 0.71 (Fig. 1A-D). Given the fact that the second abbreviated model performed better than the first, we report their full model parameters for each of the four PROMs MCIDs (see Appendix, Supplemental Digital Content 12, http://links.lww.com/CORR/A152). The associated calibration

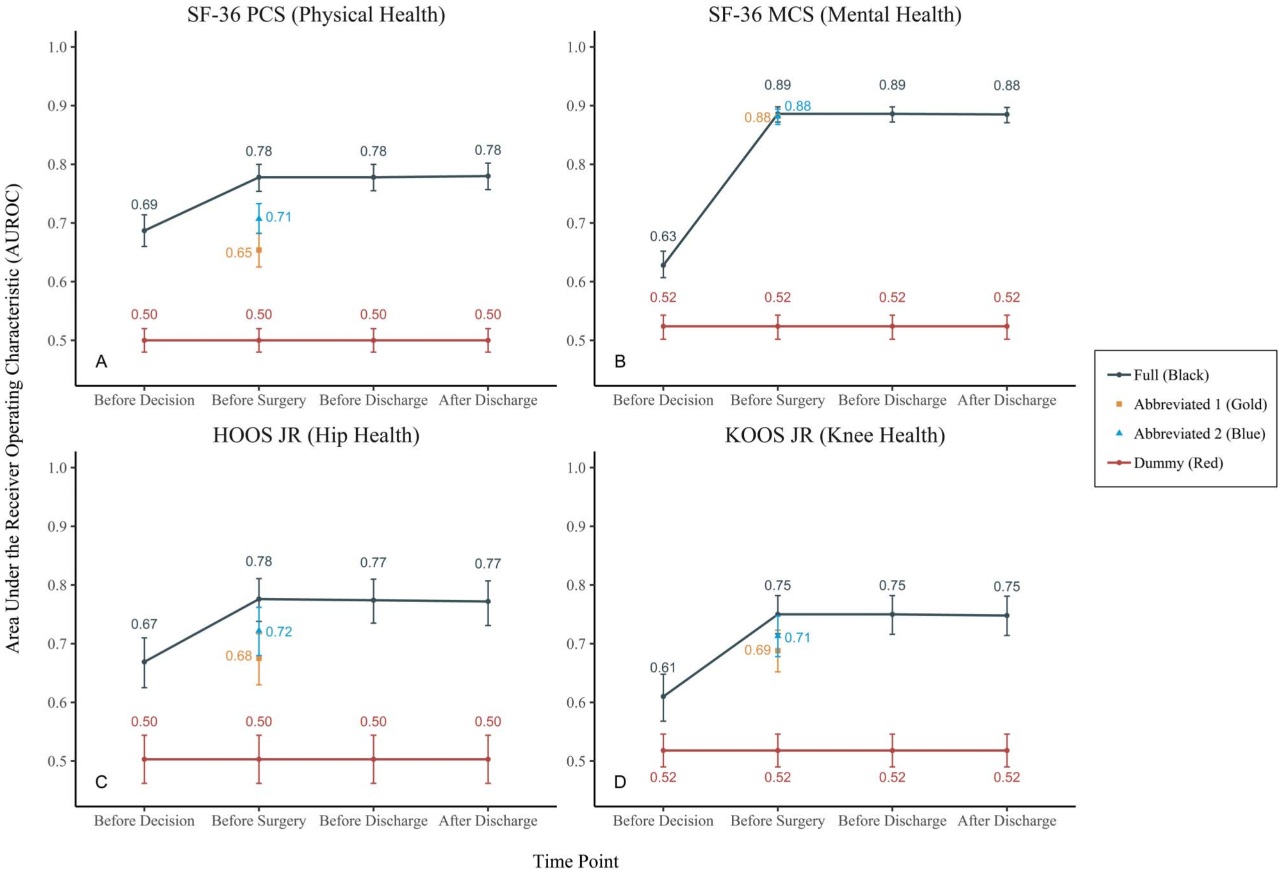

Fig 1 A-D. Area under the receiver operating characteristic (AUROC) statistics with 95% confidence intervals for full logistic least absolute shrinkage and selection operator (LASSO) models (at four time points: before decision, before surgery, before discharge, and after discharge, in black), abbreviated models 1 (before surgery, in gold), abbreviated model 2 (before surgery, in blue), and best-performing dummy models (four time points, in red), predicting: (**A**) SF-36 physical component score (PCS) minimally clinically important differences (MCIDs), (**B**) SF-36 mental component score (MCS) MCIDs, (**C**) Hip Disability and Osteoarthritis Outcome Score for joint replacement (HOOS JR) MCIDs, and (**D**) Knee Disability and Osteoarthritis Outcome Score for joint replacement (KOOS JR) MCIDs.

tables (see Appendix, Supplemental Digital Content 13, http://links.lww.com/CORR/A153) and figures (see Figure, Supplemental Digital Content 14, http://links.lww.com/CORR/A154) indicate the probabilities reported by the models are also monotonically related to their actual incidences in the validation sample.

Which Features Drive Predictive Power?

For each before-surgery logistic LASSO model, the associated baseline PROM score was always the first or second most predictive feature (for example, the baseline SF-36 PCS score is most predictive of whether a patient will achieve an SF-36 PCS MCID) (Table 4). For the PCS, the next top four features were: whether the patient was undergoing unilateral revision, self-reported back pain, the type of surgery (knee versus hip), and having one particular surgeon. For the MCS, the next top four features were: self-reported depression, the baseline SF-36 role emotional score, the baseline SF-36 social functioning score, and whether the secondary payor was Medicaid. For the HOOS JR, the top feature was whether the patient was undergoing unilateral revision, followed (very closely) by the baseline HOOS JR score, whether the secondary payor was commercial, whether the secondary payor was Blue Cross, and the baseline WOMAC stiffness score. Finally, for KOOS JR, the next top four features (after baseline KOOS JR score) were: whether the patient was undergoing unilateral revision, whether the primary payor was commercial, whether the patient reported their race as white, and whether the patient underwent bilateral surgery.

Discussion

Surgeons and healthcare systems are increasingly using PROMs to measure quality and value in TJA [10, 34]. However, it is not enough to know whether a patient's score on a PROM improved; what really matters is whether the patient's long-term perception of improvement was enough to consider the intervention worthwhile [27, 30]. The concept of the MCID captures this, by defining the smallest change in a PROM that a patient considers beneficial [3, 4, 15, 25, 37]. It is well documented that some 10% to 30% of TJA patients do not achieve a "good" outcome (such as, satisfaction [7], lack pain or function [33, 48], MCID [28]). The rapidly growing field of machine learning has also seen new applications in predicting medically relevant outcomes [9, 11, 12, 13, 16–19, 21, 24, 36, 38, 40, 44, 46, 47]. Analogously, we believe machine learning holds the potential to help clinicians identify, in advance, patients who are less likely to achieve meaningful improvements and, therefore, inform presurgical discussions of likely outcomes of TJA.

This study has several limitations. First, the registry data used was from a single, American, high-volume specialty hospital from 2007 to 2012. It is unclear the extent to which these models are externally valid to other hospitals and time periods. We have reported all parameters of our best-performing full logistic LASSO model so that other facilities can test them. However, these models include a large number of features, making their use practically challenging; it is also unlikely that other facilities would collect the exact same features. Therefore, we also report all parameters of our best-performing abbreviated logistic LASSO models, which only include baseline PROMs and surgery

Table 4. Top five features for each PROM for before surgery logistic LASSO models

Feature	SF-36 PCS		SF-36 MCS		HOOS JR		KOOS JR	
rank	Feature	Coefficient	Feature	Coefficient	Feature	Coefficient	Feature	Coefficient
1	Baseline SF-36 PCS	1.270	Baseline SF-36 MCS	2.500	Unilateral revision	1.112	Baseline KOOS JR	1.132
2	Unilateral revision	0.756	Depression	1.102	Baseline HOOS JR	1.110	Unilateral revision	1.131
3	Back pain	0.465	Baseline SF-36 role emotional	0.708	Secondary payor: commercial	-0.433	Primary payor: commercial	0.635
4	Knee patient (not hip patient)	0.449	Baseline SF-36 social functioning	0.401	Secondary payor: Blue Cross	-0.355	Race: white	-0.479
5	Surgeon #105	-0.425	Secondary payor: Medicaid	0.329	Baseline WOMAC stiffness score	-0.334	Bilateral	-0.313

The outcome was coded so that 0 = (achieve MCID) and 1 = (did not achieve MCID); depression and back pain measured with a self-reported health problems survey asking, "Do you have this problem?" with "Yes" or "No" as possible responses; MCID = minimally clinically important difference; LASSO = least absolute shrinkage and selection operator; PROM = patient-reported outcome measure; PCS = physical component score; MCS = mental component score; HOOS JR = Hip Disability and Osteoarthritis Outcome Score for joint replacement; KOOS JR = Knee Disability and Osteoarthritis Outcome Score for joint replacement.

type as features. It is also important to note that everyone in our analyses had surgery. Using these tools for presurgical decision support is therefore probably most appropriate among patients who would otherwise have surgery but might reconsider given (in part) their likelihood of not achieving a MCID.

Second, it is worth emphasizing the limitations on how our analyses can be interpreted. Correlation is not causation. However, this adage must be taken a step further when a large number of correlated features are included; this multicollinearity breaks one of the assumptions guaranteeing that logistic regression produces unbiased coefficients. Because we are focused on the combined predictive power of all these variables, it is acceptable to break this assumption. We can look at these coefficients to get a sense of what is mechanically driving prediction, but we cannot read too much into the sign nor magnitude of those (potentially biased) coefficients. For example, the SF-36 PCS MCID model ranks having one particular surgeon as the fifth most important predictor. We cannot look at the associated coefficient's sign and say that this surgeon's patients are more or less likely to achieve an MCID.

Third, there may exist selection bias given that only 44.3% of patients undergoing TJA in the relevant time period joined the registry, and among those, 71.5% were included in our analyses. Included patients were indeed more likely to be male, have lower BMI, have a shorter length of stay, and were healthier. It is therefore possible that our models would not work as well on sicker patients. However, these variables overall explained less than 1.2% of the variation in inclusion. Further testing on less healthy samples would ameliorate this concern. Missing predictor data was far less of an issue. There were also a small percentage of patients not included because their baseline HOOS JR scores were so high that it was mathematically impossible to reach an MCID; these patients were more likely to be female, have lower baseline expectations, more diagnoses, and higher baseline SF-36 PCS scores. Given that our sample is already tilted toward healthier patients, excluding these healthier patients likely helped mitigate selection bias.

Fourth, the predicted probabilities produced by these models were not perfectly calibrated in our validation sample (except for the SF-36 MCS), although all calibration plots were upward sloping as expected. Someone applying these models to a new sample would still need to first assemble, impute, and scale the relevant validation data; calculate uncalibrated probabilities with the models; test for acceptable AUROC; and then generate their own sample-specific calibration plots. This highlights that machine learning is rarely a simple matter of applying existing models. Fifth, we restricted our attention to MCIDs for four particular PROMs after TJA. There are many other PROMs and orthopaedic procedures for which similar analyses would be useful; we chose PROMs that have generally been adopted globally in arthroplasty research. Similarly, there are many machine learning algorithms; we picked three of the most popular. Sixth, there are additional features not included in the registry that might be powerful predictors, for example, unstructured image and text data such as radiology images or reports, or medications taken during the inpatient stay; these were not part of our registry.

The three machine learning algorithms performed in the poor-to-good range, with AUROCs ranging from 0.60 to 0.89. This performance is in line with prior studies that attempted to predict orthopaedic-related postsurgical PROMs [5, 6, 23, 26, 31] (AUROCs ranging from 0.64 to 0.83). However, few existing studies in orthopaedics have used machine learning algorithms beyond classic logistic regression or a proper validation sample to test predictive ability, although there are exceptions [14, 22, 26, 39]. In unreported analyses, we reproduced our four full models for each PROM using classical logistic regression instead of LASSO. Indeed, with so many features, the models were overfit to the training data, and performed substantially worse on the validation sample. We are aware of no previous studies that have attempted to predict postsurgical PROMs MCID achievement with machine learning algorithms and a proper validation sample among TJA patients. Extending these analyses to other orthopaedic treatments and PROMs should be the subject of future research.

Including more information available moving from before decision to before surgery (such as baseline PROMs) improved predictive power, but adding further information from the hospitalization yielded no further improvements. Moreover, the abbreviated models, which included just baseline PROMs, performed worse than the full before-surgery models, but not by much. Most existing orthopaedic research does not compare predictive power across information available at different points along the timeline of care; most simply consider some single set of features [5, 6, 7, 14, 20, 22, 23, 26, 31, 32, 35, 39, 41, 42]. Baker et al. [2] compared the ability of preoperative versus postoperative information to predict patient satisfaction after knee replacement; they found that postoperative features were more predictive. It is possible that postsurgical PROMs collected sooner after surgery could add predictive value, but these would not be useful for presurgical decision support. Baker et al. [1] compared patient and surgical features' relative ability to predict PROM improvements; they found that patient factors were more important, particularly preoperative PROM scores and general health status. Overall, collecting baseline PROMs before the decision to have surgery (that is, moving up collection) would facilitate these sorts of predictions for presurgical decision support. Future studies might more carefully ascertain the best time to measure baseline PROMs before the surgical decision. Considering additional data before the decision might also be informative (for example, unstructured notes or radiographs). Finally, whether additional information collected during the hospitalization (such as, inpatient medications) might be additionally predictive should be explored.

Each before-surgery logistic LASSO model directly ranked the relevant baseline PROM score as the first or second most-predictive feature. The degree to which the baseline PROMs drove prediction varied by PROM: the MCS predictions were almost entirely driven by baseline MCS scores; the PCS, HOOS JR, and KOOS JR predictions were somewhat driven by their own baseline scores, but also by other features, although which exact variables were most important differed by PROM. The power of baseline PROMs to predict postsurgical PROMs is consistent with previous work [1, 7, 20, 26, 31, 32, 35, 41, 42]. Fewer such studies directly tried to predict MCIDs [5, 6, 23], although these studies considered only baseline PROMs as predictors, and did not test on a validation sample. This again highlights the importance of collecting baseline PROMs early enough to help identify patients who might be at risk of not achieving an MCID.

In conclusion, machine learning has the potential to improve clinical decision-making and patient care by informing presurgical discussions of likely outcomes from TJA. Patients' own perceptions of the benefit of surgery should be placed at the center of such evaluations; MCIDs facilitate exactly that. Supervised machine learning algorithms using presurgical registry data can predict, with fair-to-good predictive ability, 2-year postsurgical MCIDs for general and joint-specific health PROMs. The largest gains in predictive power were from incorporating information available before surgery, namely baseline PROM scores; registry information from the hospitalization provided negligible improvement. Although we report all parameters of our best-performing models, they cannot simply be applied off-the-shelf in new settings without proper testing (such as the acceptable AUROC and sample-specific calibration). Indeed, machine learning is rarely, in practice, a simple matter of applying existing models, though it is clear from our analyses that they hold much promise for orthopaedic outcomes.

References

1. Baker PN, Deehan DJ, Lees D, Jameson S, Avery PJ, Gregg PJ, Reed MR. The effect of surgical factors on early patient-reported outcome measures (PROMS) following total knee replacement. *J Bone Joint. Surg Br.* 2012;94:1058–1066.
2. Baker PN, Rushton S, Jameson SS, Reed M, Gregg P, Deehan DJ. Patient satisfaction with total knee replacement cannot be predicted from pre-operative variables alone: a cohort study from the national joint registry for England and Wales. *Bone Joint. J.* 2013;95B:1359–1365.
3. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, Price AJ. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol.* 2015;68:73–79.
4. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol*. 2002;14:109–114.
5. Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ. John Charnley Award: preoperative patient-reported outcome measures predict clinically meaningful improvement in function after THA. *Clin Orthop Relat Res*. 2016;474:321–329.
6. Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ. Can preoperative patient-reported outcome measures be used to predict meaningful improvement in function after TKA? *Clin Orthop Relat Res*. 2017;475:149–157.
7. Bourne RB, Chesworth BM, Davis AM, Mahomed NN, Charron KDJ. Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? *Clin Orthop Relat Res*. 2010;468:57–63.
8. Cafri G, Li L, Paxton EW, Fan J. Predicting risk for adverse health events using random forest. *J Appl Stat*. 2018;45:2279–2294.
9. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE, Kratz A, Wefers AK, Huang K, Pajtler KW, Schweizer L, Stichel D, Olar A, Engel NW, Lindenberg K, Harter PN, Braczynski AK, Plate KH, Dohmen H, Garvalov BK, Coras R, Hölsken A, Hewer E, Bewerunge-Hudler M, Schick M, Fischer R, Beschorner R, Schittenhelm J, Staszewski O, Wani K, Varlet P, Pages M, Temming P, Lohmann D, Selt F, Witt H, Milde T, Witt O, Aronica E, Giangaspero F, Rushing E, Scheurlen W, Geisenberger C, Rodriguez FJ, Becker A, Preusser M, Haberler C, Bjerkvig R, Cryan J, Farrell M, Deckert M, Hench J, Frank S, Serrano J, Kannan K, Tsirigos A, Brück W, Hofer S, Brehmer S, Seiz-Rosenhagen M, Hänggi D, Hans V, Rozsnoki S, Hansford JR, Kohlhof P, Kristensen BW, Lechner M, Lopes B, Mawrin C, Ketter R, Kulozik A, Khatib Z, Heppner F, Koch A, Jouvet A, Keohane C, Mühleisen H, Mueller W, Pohl U, Prinz M, Benner A, Zapatka M, Gottardo NG, Driever PH, Kramm CM, Müller HL, Rutkowski S, Von Hoff K, Frühwald MC, Gnekow A, Fleischhack G, Tippelt S, Calaminus G, Monoranu CM, Perry A, Jones C, Jacques TS, Radlwimmer B, Gessi M, Pietsch T, Schramm J, Schackert G, Westphal M, Reifenberger G, Wesseling P, Weller M, Collins VP, Blümcke I, Bendszus M, Debus J, Huang A, Jabado N, Northcott PA, Paulus W, Gajjar A, Robinson GW, Taylor MD, Jaunmuktane Z, Ryzhova M, Platten M, Unterberg A, Wick W, Karajannis MA, Mittelbronn M, Acker T, Hartmann C, Aldape K, Schüller U, Buslei R, Lichter P, Kool M, Herold-Mende C, Ellison DW, Hasselblatt M, Snuderl M, Brandner S, Korshunov A, Von Deimling A, Pfister SM. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555:469–474.
10. Centers for Medicare & Medicaid Services. Overview of CJR quality measures, composite quality score, and pay-for-performance methodology. 2017. Available at: https://innovation.cms.gov/Files/x/cjr-qualsup.pdf. Accessed August 31, 2018.
11. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc*. 2017;24:361–370.
12. Ehlers AP, Roy SB, Khor S, Mandagani P, Maria M, Alfonso-Cristancho R, Flum DR. Improved risk prediction following surgery using machine learning algorithms. *EGEMS (Wash DC)*. 2017;5:3.
13. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A,

Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin H-J, Heng P-A, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang Y-W, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318:2199–2210.

14. Endo A, Baer HJ, Nagao M, Weaver MJ. Prediction model of in-hospital mortality after hip fracture surgery. *J Orthop Trauma*. 2018;32:34–38.
15. Escobar A, Quintana JM, Bilbao A, Aróstegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartilage.* 2007;15:273–280.
16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
17. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform.* 2015;56:229–238.
18. Giannini HM, Chivers C, Draugelis M, Hanish A, Fuchs B, Donnelly P, Lynch M, Meadows L, Parker SJ, Schweickert WD, Mikkelsen ME, Fishman N, Hansen C, Umscheid C. Development and implementation of a machine-learning algorithm for early identification of sepsis in a multi-hospital academic healthcare system. *Am J Respir Crit Care Med*. 2017;195:A7015.
19. 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. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
20. Hamilton DF, Lane J V, Gaston P, Patton JT, MacDonald D, Simpson AHRW, Howie CR. What determines patient satisfaction with surgery? A prospective cohort study of 4709 patients following total joint replacement. *BMJ Open*. 2013;3:e002525.
21. Hao S, Wang Y, Jin B, Shin AY, Zhu C, Huang M, Zheng L, Luo J, Hu Z, Fu C, Dai D, Wang Y, Culver DS, Alfreds ST, Rogow T, Stearns F, Sylvester KG, Widen E, Ling XB. Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine healthcare information exchange. *PLoS One*. 2015;10:e0140271.
22. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty. *Clin Orthop Relat Res*. 2019;477:452-460.
23. Ho B, Houck JR, Flemister AS, Ketz J, Oh I, Digiovanni BF. Preoperative PROMIS scores predict postoperative success in foot and ankle patients. *Foot Ankle Int*. 2016;37:911–918.
24. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One*. 2017;12:e0174708.
25. Keurentjes JC, Van Tol FR, Fiocco M, Schoones JW, Nelissen RG. Minimal clinically important differences in health-related quality of life after total hip or knee replacement: A systematic review. *Bone Joint Res*. 2012;1:71–77.
26. Khor S, Lavallee D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, Lu D, Alex Mohit A, Oskouian RJ, Roh JR, Shonnard N, Dagal A, Flum DR. Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. *JAMA Surg*. 2018;153:634–642.
27. Leopold SS, Porcher R. Editorial: the minimum clinically important difference—the least we can do. *Clin Orthop Relat Res*. 2017;475:929–932.
28. Lyman S, Lee YY, McLawhorn AS, Islam W, MacLean C. What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? *Clin Orthop Relat Res*. 2018;476:2432–2441.
29. Malley J, Kruppa J, Dasgupta A, Malley K, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med*. 2012;51:74–81.
30. Maltenfort M, Díaz-Ledezma C. Statistics in brief: minimum clinically important difference—availability of reliable estimates. *Clin Orthop Relat Res*. 2017;475:933–946.
31. McGirt M, Bydon M, Archer K, Devin C, Chotai S, Parker S, Nian H, Harrell FJ, Speroff T, Dittus R, Philips S, Shaffrey C, Foley K, Asher A. An analysis from the quality outcomes database, part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: predicting likely individual patient outcomes for shared decision-making. *J Neurosurg Spine*. 2017;27:357–369.
32. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ. Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. *Neurosurg Focus*. 2015;39:E13.
33. Murray D, Frost S. Pain in the assessment of total knee replacement. *J Bone Joint. Surg Br*. 1998;80:426–431.
34. National Health Service England. National patient reported outcome measures (PROMs) programme consultation. 2016. Available at: https://www.engage.england.nhs.uk/consultation/proms-programme/. Accessed August 31, 2018.
35. Nemes S, Rolfson O, Garellick G. Development and validation of a shared decision-making instrument for health-related quality of life one year after total hip replacement based on quality registries data. *J Eval Clin Pract.* 2018;24:13–21.
36. Poplin R, Varadarajan A V., Blumer K, Liu Y, McConnell M V., Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
37. Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthritis Cartilage.* 2005;13:1076–1083.
38. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.
39. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, Patterson BM, Krebs VE. Development and validation of a machine-learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. *J Arthroplasty.* [Published online ahead of print Decebmer 27, 2018]. DOI: 10.1016/j.arth.2018.12.030.
40. Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality

utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *J Gen Intern Med*. 2018;33:921–928.

41. Sanchez-Santos MT, Garriga C, Judge A, Batra RN, Price AJ, Liddle AD, Javaid MK, Cooper C, Murray DW, Arden NK. Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery. *Sci Rep*. 2018;8:3381.
42. Scott CEH, Howie CR, MacDonald D, Biant LC. Predicting dissatisfaction following total knee replacement: a prospective study of 1217 patients. *J Bone Joint Surg Br*. 2010;92:1253–1258.
43. SooHoo NF, Li Z, Chenok KE, Bozic KJ. Responsiveness of patient reported outcome measures in total joint arthroplasty patients. *J Arthroplasty.* 2015;30:176–191.
44. Tong L, Erdmann C, Daldalian M, Li J, Esposito T. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. *BMC Med Res Methodol*. 2016;16:26.
45. Van S, Van Der Straeten C, Arnout N, Deprez P, Van Damme G, Victor J. A new prediction model for patient satisfaction after total knee arthroplasty. *J Arthroplasty.* 2016;31:2660–2667.
46. Vovsha I, Salleb-Aouissi A, Raja A, Koch T, Rybchuk A, Radeva A, Rajan A, Huang Y, Diab H, Tomar A, Wapner R. Using kernel methods and model selection for prediction of preterm birth. *Proc Mach Learn Healthc.* 2016. Available at: http://proceedings.mlr.press/v56/Vovsha16.html. Accessed August 31, 2018.
47. Warner JL, Zhang P, Liu J, Alterovitz G. Classification of hospital acquired complications using temporal clinical information from a large electronic health record. *J Biomed Inf.* 2016;59: 209–217.
48. Wylde V, Blom AW, Whitehouse SL, Taylor AH, Pattison GT, Bannister GC. Patient-reported outcomes after total hip and knee arthroplasty: comparison of midterm results. *J Arthroplasty.* 2009;24:210–216.