Title

Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty?

Permalink

https://escholarship.org/uc/item/3pd697z8

Journal

Clinical Orthopaedics and Related Research®, 477(2)

ISSN

0009-921X

Authors

Harris, Alex HS
Kuo, Alfred C
Weng, Yingjie
et al.

Publication Date

2019-02-01

DOI

10.1097/corr.0000000000000601

Peer reviewed

Clin Orthop Relat Res (2019) 477:452-460
DOI 10.1097/CORR.0000000000000601

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

Clinical Research

Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty?

Alex H. S. Harris PhD, Alfred C. Kuo MD, PhD, Yingjie Weng MS, Amber W. Trickey PhD, Thomas Bowe PhD, Nicholas J. Giori MD, PhD

Received: 21 June 2018 / Accepted: 19 November 2018 / Published online: 16 January 2019

Abstract

Background Existing universal and procedure-specific surgical risk prediction models of death and major complications after elective total joint arthroplasty (TJA) have limitations including poor transparency, poor to modest accuracy, and insufficient validation to establish performance across diverse settings. Thus, the need remains for accurate and validated prediction models for use in preoperative management, informed consent, shared decision-making, and risk adjustment for reimbursement.

Questions/purposes The purpose of this study was to use machine learning methods and large national databases to develop and validate (both internally and externally) parsimonious risk-prediction models for mortality and complications after TJA.

Methods Preoperative demographic and clinical variables from all 107,792 nonemergent primary THAs and TKAs in the 2013 to 2014 American College of Surgeons-National Surgical Quality Improvement Program (ACS-NSQIP) were evaluated as predictors of 30-day death and major complications. The NSQIP database was chosen for its high-quality data on important outcomes and rich characterization of preoperative demographic and clinical predictors for demographically and geographically diverse patients. Least absolute shrinkage and selection operator (LASSO) regression, a type of machine learning that optimizes accuracy and parsimony, was used for model development. Tenfold validation was used to produce C-statistics, a measure of how well models discriminate patients who experience an outcome from those who do not. External validation, which evaluates the generalizability of the models to new data sources and patient groups, was accomplished using data from the Veterans Affairs Surgical Quality Improvement Program (VASQIP). Models previously developed from VASQIP data were also externally validated using NSQIP data to examine the generalizability of their performance with a different group of patients outside the VASQIP context.

This work was funded by grants from the VA HSR&D Service (I01-HAX002314-01A1; RCS14-232; CDA 13-279; AHSH).
Each author certifies that his or her institution determined that this investigation is not human subjects research and that all investigations were conducted in conformity with ethical principles of research.
This work was performed at the VA Palo Alto Healthcare System, Center for Innovation to Implementation, Palo Alto, CA, USA, and the Stanford–Surgical Policy Improvement Research and Education Center, Stanford, CA, USA.
The views expressed do not reflect those of the US Department of Veterans Affairs or other institutions.

A. H. S. Harris, T. Bowe, N. J. Giori Center for Innovation to Implementation, VA Palo Alto Healthcare System, Palo Alto, CA, USA

A. C. Kuo San Francisco Veterans Affairs Medical Center, University of California, San Francisco, CA, USA

A H. S. Harris, Y. Weng, A. W. Trickey Stanford–Surgical Policy Improvement Research and Education Center, Stanford, CA, USA

N. J. Giori Department of Orthopedic Surgery, Stanford University School of Medicine, Stanford, CA, USA

A. H. S. Harris (✉), VA Palo Alto Healthcare System Center for Innovation to Implementation 795 Willow Road (152-MPD) Menlo Park, CA 94025, USA email: Alexander.Harris2@va.gov

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.

Results The models, developed using LASSO regression with diverse clinical (for example, American Society of Anesthesiologists classification, comorbidities) and demographic (for example, age, gender) inputs, had good accuracy in terms of discriminating the likelihood a patient would experience, within 30 days of arthroplasty, a renal complication (C-statistic, 0.78; 95% confidence interval [CI], 0.76-0.80), death (0.73; 95% CI, 0.70-0.76), or a cardiac complication (0.73; 95% CI, 0.71-0.75) from one who would not. By contrast, the models demonstrated poor accuracy for venous thromboembolism (C-statistic, 0.61; 95% CI, 0.60-0.62) and any complication (C-statistic, 0.64; 95% CI, 0.63-0.65). External validation of the NSQIP-derived models using VASQIP data found them to be robust in terms of predictions about mortality and cardiac complications, but not for predicting renal complications. Models previously developed with VASQIP data had poor accuracy when externally validated with NSQIP data, suggesting they should not be used outside the context of the Veterans Health Administration.

Conclusions Moderately accurate predictive models of 30-day mortality and cardiac complications after elective primary TJA were developed as well as internally and externally validated. To our knowledge, these are the most accurate and rigorously validated TJA-specific prediction models currently available (http://med.stanford.edu/s-spire/Resources/clinical-tools-.html). Methods to improve these models, including the addition of nonstandard inputs such as natural language processing of preoperative clinical progress notes or radiographs, should be pursued as should the development and validation of models to predict longer term improvements in pain and function.

Level of Evidence Level III, diagnostic study.

Introduction

Substantial effort has gone into developing prediction models of total joint arthroplasty (TJA) outcomes for use in preoperative management, informed consent, shared decision-making, and risk-adjusting reimbursement programs [5-7, 13, 18, 28]. However, despite these efforts, no accurate and internally and externally validated risk prediction models for short-term outcomes of TJA currently exist [18]. The American College of Surgeons-National Surgical Quality Improvement Program (ACS-NSQIP) universal surgical risk calculator has good overall accuracy averaged across procedures [3], but studies of its accuracy for specific procedures, including TJA, are limited to single-site studies and have had fair to poor results [7, 26]. Multisite validation studies in more diverse contexts are needed to better evaluate the performance of the ACS-NSQIP calculator for TJA applications. Other universal risk prediction models that recently were developed using ACS-NSQIP data have excellent to good accuracy across specialties and when internally validated with a sample of orthopaedic procedures (C-statistics: 0.93 and 0.81 for 30-day mortality and morbidity, respectively) [19, 20]. However, these models have not been internally or externally validated for elective TJA, which likely has little or no variability on several key predictors (such as emergency operation, primary surgeon specialty, ventilator-dependent, work relative value unit). Thus, it is currently unknown if these universal risk models retain their reported accuracy when applied to patients undergoing elective TJA.

A recent review of TJA-specific preoperative risk prediction models noted they too have serious limitations [18], most importantly poor or unknown performance on internal or external validation. No model coefficients or accuracy metrics were originally reported for the American Joint Replacement Registry Risk Calculator (https://teamwork.aaos.org/ajrr/SitePages/Risk%20Calculator.aspx) [4], which estimates risk for 90-day mortality and 2-year prosthetic joint infection. We recently conducted an external validation study of the calculator with a sample of Medicare-eligible patients from the Veterans Health Administration (VHA) and found very poor accuracy for 30-mortality (C-statistic = 0.62) [13]. Other TJA-specific preoperative risk prediction models also have poor accuracy [21, 23] or cannot be used preoperatively because they include as predictors intraoperative or index stay characteristics (for example, lowest intraoperative heart rate) [2, 21, 27]. Established comorbidity indices such as the Charlson and Elixhauser have been found to be predictive of 90-day and 1-year mortality after elective primary TJA [15]. However, these results have not been internally or externally validated nor have these indices been associated with other TJA outcomes.

Members of our team recently developed and internally validated prediction models for 30-day mortality and complications after TJA for VHA patients with osteoarthritis (OA) [12]. Over 70,000 patients diagnosed with OA who received primary TJA in the VHA were included. Accuracy of the models was highest for cardiac complications (C-statistic, 0.75; 95% confidence interval [CI], 0.71-0.79) and 30-day mortality (C-statistic, 0.73; 95% CI, 0.66-0.79) [12]. Although the accuracies of the cardiac complications and mortality models are currently the highest reported for TJA-specific predictions, the generalizability of these models to samples outside the VHA, where patients and clinical context are not representative, are unknown.

Thus, in this study, we sought to develop and validate (both internally and externally) accurate prediction models for mortality and major complications after elective TJA that can be used to inform preoperative discussions and decisions in diverse healthcare settings. In this study, we used a machine learning regression strategy, least absolute shrinkage and selection operator (LASSO) regression, to develop and internally validate prediction models derived from elective TJAs represented in ACS-NSQIP data. To assess the generalizability of these models outside of the context of hospitals participating in ACS-NSQIP, we also conducted external validation using data from the Veterans Affairs Surgical Quality Improvement Program (VASQIP) where the patient demographics and clinical profiles are different. Finally, to assess the generalizability of models we previously developed and published from VASQIP data [12], we conducted external validation using ACS-NSQIP data.

Patients and Methods

Definitions

Machine learning is an umbrella term that refers to diverse methods for classification, in this case patients who do or do not experience death or a complication after TJA. Machine learning algorithms iteratively "learn" from patterns in data to determine the best rules for classifying new observations and assigning probabilities of those classifications being correct. LASSO regression is a machine learning method to select which variables are most important to include in a prediction model and how much weight to assign them. The goal is to optimize accuracy with the fewest number of predictors. In this context, minimizing the number of predictors is essential for ease of use and future implementation. An accessible demystification of machine learning can be found here: https://hackernoon.com/machine-learning-is-the-emperor-wearing-clothes-59933d12a3cc.

As we recently described in more detail [12], discrimination, calibration, and internal/external validation are terms that are foundational to assessing the performance of predictive models.

Discrimination, quantified by the C-statistic, is the ability of a model to distinguish patients who experience an outcome of interest from those who do not. Discrimination is the probability that a patient who experienced the outcome has a higher predicted probability than a randomly selected patient who did not experience the outcome [22]. In general, C-statistics can be interpreted as excellent (0.9–1), good (0.8–0.89), fair (0.7–0.79), poor (0.6–0.69), or fail/no discriminatory capacity (0.5–0.59) [9, 14].

Calibration compares predicted and observed outcomes across the entire range of the data and is typically visually assessed by plotting observed versus predicted outcomes. Calibration is considered good if prevalence of outcomes tracks monotonically with the predicted probabilities from the model.

Models can be overfit, meaning that the reported accuracy only applies to the data used to develop them, but not to new observations. Validation is essential in assessing model accuracy when applied to new patients from the same data source used to develop the models (internal validation) and when applied to data from different contexts or times (external validation). Internal and external validation is critical to understanding real-world model accuracy and generalizability.

Study Design and Setting

Data Sources

As a result of our focus on developing tools for informing decisions before elective surgery, all nonemergent primary THAs and TKAs included in the 2013 and 2014 ACS-NSQIP data were used for model development and internal validation (N = 107,792). External validation of the ACS-NSQIP-based models was accomplished using all non-emergent primary TJAs represented in 2005 to 2013 VASQIP data (N =70,569). The methodologies for data collection and limitation of the database are described elsewhere [1, 8, 16, 17]. Definitions of model inputs and outcomes are described in the data dictionary included in the ACS-NSQIP 2014 PUF User Guide (https://www.facs.org/∼/media/files/quality%20programs/nsqip/nsqip_puf_userguide_2014.ashx). The ACS-NSQIP database was chosen for this project because it contains high-quality data on important outcomes and rich characterization of candidate preoperative demographic and clinical predictors for a very large and geographically diverse sample of patients. Although the ACS-NSQIP data are demographically similar to the general US adult surgical population, participating hospitals (435 in 2013, 517 in 2014) need to have the infrastructure for participation and therefore overrepresent larger teaching facilities and practices that have a quality improvement infrastructure [1].

Candidate Predictors

All variables included in ACS-NSQIP data that might be known during preoperative decision-making were included as candidate predictors (Table 1). Preoperative laboratory values, although available in ACS-NSQIP data, were not considered because of substantial missing data, concerns that these data were not missing at random thereby precluding multiple imputation methods, and empiric work showing that they do not meaningfully improve the accuracy of risk models [19]. Missing data were minimal (< 1%; Table 1) on all predictor variables except race-ethnicity, which was unknown or missing for 14.8% of TJAs. For this and other categorical variables with any missing data, we

Table 1. Candidate predictor variables for 107,792 elective total joint arthroplasty cases

Variable	Number/ mean	Percent/SD
Hip procedure	41,973	38.9%
Knee procedure	65,819	61.1%
Gender		
Female	64,039	59.4%
Male	43,753	40.6%
Race-ethnicity		
White	78,048	72.4%
Black	7428	6.9%
Hispanic	3388	3.1%
Asian or Pacific Islander	2401	2.2%
Native American or Alaska Native	599	0.6%
Unknown/missing	15,928	14.8%
Age (years; mean, SD)	65.7	10.47
BMI (kg/m^2; mean, SD)	31.8	7.24
Underweight (< 18.5 kg/m^2)	495	0.46%
Normal (18.5 to < 25 kg/m^2)	14,828	13.76%
Overweight (25 to < 30 kg/m^2)	31,782	29.48%
Obese (30 to < 40 kg/m^2)	47,031	43.63%
Very obese (> 40 kg/m^2)	13,241	12.28%
BMI missing	415	0.39%
ASA class		
I: No disturbance	3252	3.0%
II: Mild disturbance	56,563	52.5%
III: Severe disturbance	46,009	42.7%
IV: Life-threatening	1851	1.7%
V: Moribund	4	0.0%
Missing	113	0.1%
Functionally health status		
Dependent	1662	1.5%
Independent	105,339	97.7%
Missing	791	0.7%
Medication and treatment		
Steroids	4172	3.9%
Hypertension	66,403	61.6%
Dialysis	210	0.2%
Renal failure	32	0.0%
History of diseases and conditions		
CHF	286	0.3%
COPD	4023	3.7%
Dyspnea		
At rest	190	0.2%
Moderate exertion	5328	4.9%
None	102,274	94.9%
Pneumonia	19	0.0%
Smoking	11,335	10.5%
> 10% loss of body weight	170	0.2%
Disseminated cancer	200	0.2%
Open wound/wound infection	315	0.3%
Diabetes	16,594	15.4%
Sepsis (48 hours before surgery)	243	0.2%
Bleeding disorders	2736	2.5%

BMI = body mass index; ASA = American Society of Anesthesiologists; CHF = congestive heart failure; COPD = chronic obstructive pulmonary disease.

included "missing" as a category so that all observations could be used.

Outcomes

Prediction models were developed for 30-day mortality, cardiac complications, central nervous system-cardiovascular system complications, respiratory complications, surgical wound infection, deep incisional surgical site infection, sepsis, return to the operating room, renal complications (failure or insufficiency), venous thromboembolism (deep vein thrombosis + pulmonary embolism), and the occurrence of any of the aforementioned complications, all as defined by the ACS-NSQIP [8].

Model Development and Internal Validation

The primary model development strategy was LASSO regression [10, 25]. LASSO regression is an iterative machine learning approach that optimizes accuracy and simplicity. LASSO regression is especially helpful when candidate predictors are highly correlated, which is the case in this context. Once the model tuning (complexity) parameter is selected using a portion of the data, a 10-fold bootstrap validation process iteratively estimates model parameters using 9/10th of the data and estimates prediction error and other performance metrics by applying the model to the held-aside data. This process is repeated 10 times and the final model and validated performance metrics are determined by pooling across these analyses [24]. The LASSO coefficients can be used to calculate a patient's risk score by multiplying the patient's values (for example, 1 = factor present, 0 = factor absent for binary variables) by the coefficients and summing the products. The risk score can then be translated to a predicted probability of an adverse event (AE) with the formula Prob(AE) = exp(score)/(1 + exp[score]).

External Validation

To assess the generalizability of these models outside of the context of hospitals participating in ACS-NSQIP, we applied the ACS-NSQIP-derived models to a sample of 70,569 VHA primary TJAs represented in 2005 to 2013 VASQIP data. Secondarily, we applied the NSQIP data to our previously published models, which were developed and internally validated with VASQIP data [12]. Aspects of this investigation related to ACS-NSQIP data were determined by the Stanford institutional review board to be nonhuman subject research as a result of the use of publicly available and deidentified data. Analysis of the VASQIP data was approved by the VHA central institutional review board.

Results

Internal Validation of the NSQIP TJA Model

Using diverse demographic and clinical predictor variables (Table 1), the LASSO regression models had good accuracy in terms of discriminating the likelihood a patient would develop a renal complication (C-statistic, 0.78; 95% CI, 0.76-0.80), die within 30 days of arthroplasty (0.73; 95% CI, 0.70-0.76), or experience a cardiac complication (0.73; 95% CI, 0.71-0.75) from one who would not. The frequency and incidence of each outcome as well as the C-statistics from the bootstrapped internal validation of the models were calculated (Table 2). A simple calculator that uses the coefficients (Table 3) from the three most accurate models to calculate risk probabilities for specific patients can be accessed here: http://med.stanford.edu/s-spire/Resources/clinical-tools-.html.

For the ACS-NSQIP sample, the mortality model predicted risk of death between 0.066% and 27.06% (that is, predicted probabilities from 0.00066 to 0.27060). Ideally, actual risk increases monotonically as model-derived predicted probabilities increase (calibration). For the mortality model, most of the observed events occur within the highest decile of predicted probability (Fig. 1). This fact should guide interpretation and use of the models. Notably, observed risk of death does not appear to increase linearly throughout the range of predicted probabilities, but jumps substantially in the highest decile of predicted probability (range, 0.0015-0.27060).

The predicted risks of cardiac complications ranged from 0.03% to 13.74% (that is, predicted probabilities from 0.0003 to 0.1374). Observed risk of cardiac complications increases slowly and linearly throughout most of the range of predicted probabilities (Fig. 2), but jumps within the highest decile of predicted probability, which in this case ranged from 0.0043 to 0.1374 (that is, 0.43% to 13.74%).

External Validation of the NSQIP Model Using VASQIP Data

External validation of the NSQIP-derived models using VASQIP data found them to be robust in terms of predictions about mortality (C-statistic, 0.69; 95% CI, 0.66-0.74) and cardiac complications (C-statistic, 0.72; 95% CI, 0.68-0.75), but not for predicting renal complications (C-statistic, 0.60; 95% CI, 0.57-0.63). Thus, two of the three NSQIP-derived models retain almost all of their predictive validity when applied to new data from a sample of patients with very different characteristics receiving TJA in different healthcare contexts.

Table 2. Outcome event rates and internal crossvalidation C-statistics and confidence intervals

Complication/outcome	Events	Incidence	Mean C-statistic	95% CI Lower	limits Upper
Renal	172	0.0016	0.777	0.758	0.796
Death	137	0.0013	0.733	0.704	0.762
Cardiac	290	0.0029	0.730	0.711	0.750
CNS-CVA	92	0.0008	0.696	0.669	0.724
Sepsis	305	0.0028	0.692	0.680	0.704
Wound infection	544	0.0050	0.664	0.652	0.676
Deep incisional SSI	234	0.0022	0.661	0.643	0.679
Respiratory	1042	0.0097	0.639	0.627	0.652
Return to OR	1611	0.0149	0.650	0.641	0.659
VTE (DVT + PE)	1236	0.0115	0.613	0.608	0.617
Any complication	3768	0.0350	0.638	0.630	0.646

CI = confidence interval; CNS-CVA = central nervous system-cerebrovascular accident; SSI = surgical site infection; OR = operating room; VTE = venous thromboembolism; DVT = deep vein thrombosis; PE = pulmonary embolism.

Table 3. LASSO regression coefficients for three outcomes

Variable	Death	Cardiac	Renal
(Intercept)	-7.3590	-8.3159	-6.8686
Surgery (hip)			
Age	0.0079	0.0363	
Male gender		0.0298	0.2350
Race-ethnicity			
Hispanic			
Black			0.3536
Asian or Pacific Islander			
Native American or Alaska Native			
BMI category			
Very obese			0.1899
ASA class			
II: Mild disturbance	-0.0572	-0.3498	-0.4563
III: Severe disturbance			
IV: Life-threatening	1.5325	0.6845	1.2534
Functional health status Dependent	1.2119		
Preoperative conditions			
Bleeding disorders	0.5524		0.8865
Diabetes			0.1446
Dialysis	1.6620	1.5366	
Disseminated cancer	0.4977		
Dyspnea at rest			
Dyspnea moderate exertion		0.1021	0.3020
History of CHF		0.8679	0.2381
History of COPD			0.4171
Hypertension medication		0.0415	0.4054
Sepsis (48 hours surgery)	2.4311		
History of renal failure			0.5478
Pneumonia		2.6248	
Smoking status			
Steroids medications			
Open wound/wound infection			
> 10% loss of body weight	1.5788	1.3607	0.0439

LASSO = least absolute shrinkage and selection operator; BMI = body mass index; ASA = American Society of Anesthesiologists; CHF = congestive heart failure; COPD = chronic obstructive pulmonary disease.

External Validation of VASQIP-derived TJA Models Using NSQIP Data

The mortality and cardiac complication models previously developed with VASQIP data [12] had poor discrimination when externally validated using the ACS-NSQIP data (C-statistics, 0.61; 95% CI, 0.47-0.74 and 0.61; 95% CI, 0.51-0.67, respectively). These results signify a substantial reduction of model accuracy for the models developed with VHA data when applied outside of the VHA context.

Discussion

To address the substantial limitations of existing universal and procedure-specific surgical risk prediction models [3, 5, 18, 20], we sought to develop and validate accurate models of death and major complications after elective primary TJA. The ability to accurately predict these outcomes could improve the quality of preoperative management, informed consent, shared decision-making, and risk adjustment for reimbursement. We were able to use ACS-NSQIP data to develop and validate fairly accurate predictive models of 30-day mortality and cardiac complications after elective primary TJA. To our knowledge, these are currently the most accurate, rigorously validated, and broadly generalizable TJA prediction models available (http://med.stanford.edu/s-spire/Resources/clinical-tools-.html). Furthermore, our mortality and cardiac complication models require fewer than half of the patient variables that the ACS-NSQIP model requires [3]. Using the model coefficients and performance metrics specific to elective TJA that we report here, other researchers can test these models on new samples to assess generalizability for different TJA patient populations. Our results may also suggest that the higher reported accuracy of the generic ACS-NSQIP models across procedures and specialties [3], and even generic orthopaedic models [19], may be overly optimistic for elective TJA.

Several issues and limitations should be considered in interpreting these results. Although the ACS-NSQIP Participant Use Data File contains high-quality data on important outcomes and preoperative predictors for a very large and geographically diverse sample of patients, it is limited to sampled TJAs from participating hospitals, therefore overrepresenting larger teaching facilities and practices that have quality improvement infrastructures. Validating these models with TJAs from nonparticipating, smaller practices is critical before assuming they generalize to those contexts. Also, these data do not contain complete information on comorbidities or other patient and setting factors that may be predictive of outcomes. The future inclusion of these currently unavailable data, for example comorbidity severity or facility complication rates, might improve accuracy of the models.

Although the ACS-NSQIP-derived models performed almost as well when applied to VHA data, our previously published VHA-derived models [12] suffered a loss of discrimination when applied to ACS-NSQIP TJAs. These results highlight the importance of conducting external validation to understand the generalizability of reported model performance metrics in new contexts, especially

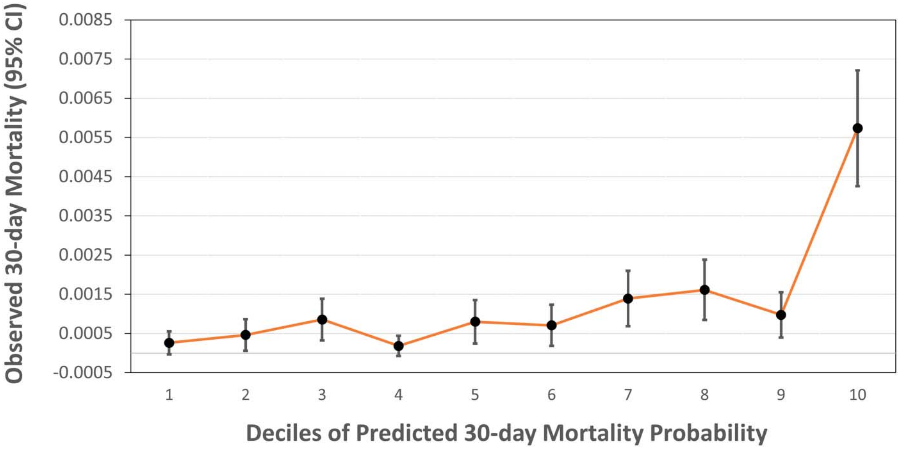

Fig. 1 Calibration plot reveals only patients in the top decile of predicted risk have substantially higher 30-day mortality.

when the underlying populations and healthcare delivery system are notably different.

Similar to our previous work using VHA data, we were unable to produce accurate ACS-NSQIP-derived models for other complications such as return to the operating room or deep infection, again highlighting the difficulty of predicting rare outcomes with preoperative patient data [12]. Several factors likely contribute to this difficulty. First, in contrast to patients undergoing emergency surgery, patients undergoing elective TJA have already been screened for risk by the operating surgeons. Therefore, the variability in risk that is needed to build predictive models is lower. Second, many of the predictor variables are dichotomized (for example, diabetes: yes/no), rather than indicative of severity, which might reduce accuracy [22]. Third, it is likely that some important complications most commonly arise from situations that occur intraoperatively or postoperatively. For example, some cases of periprosthetic infection may occur from an undetected breach of sterility at the time of the operation that may have little to do with the preoperative patient characteristics used to build current predictive models. Occurrences during the intraoperative or postoperative course of the patient's care will differentially affect certain postoperative complications, and it is unrealistic to expect that all complications are predictable with preoperatively available data. There is likely an upper bound to the predictability of various complications based solely on preoperatively available data. Where this upper bound resides for each complication of interest is unknown. Fourth, there are preoperatively available data that are not easily incorporated into these models. Complex anatomy that will increase the length and complexity of the surgery is preoperatively recognizable to the surgeon, but will not be factored with current predictive

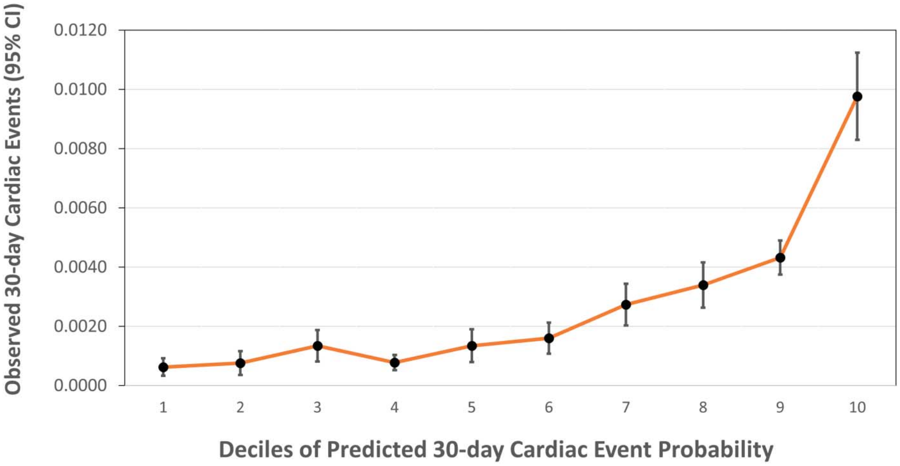

Fig. 2 Calibration plot reveals only patients in the top decile of predicted risk have a substantially higher rate of 30-day cardiac complications.

models. Methods to improve these models, including the addition of nonstandard inputs such as natural language processing of preoperative clinical progress notes or radiographs, should be pursued as should the development and validation of models to predict longer term improvements in pain and function.

In conclusion, we were able to develop as well as internally and externally validate fairly accurate models of 30-day mortality and cardiac complications for elective primary TJA. Accurate models for other complications and outcomes such as longer term improvements in pain and functioning are critically needed for myriad purposes, but do not yet exist. It remains unknown if these predictive models provide information that is not already known to surgeons or patients. Research is needed to evaluate the effects of specific applications of predictive models (for example, informed consent, shared decision-making, or risk stratification) in terms of their impact on short- and long-term patient outcomes and patient satisfaction [11].

Acknowledgments We thank Qian Ding MS, from the S-SPIRE Center in the Stanford Department of Surgery for her skillful help with analyses for the revision of the manuscript.

References

1. American College of Surgeons-National Surgical Quality Improvement Program. User Guide for the 2014 ACS NSQIP Participant Use Data File (PUF). 2015. Available at: https://www.facs.org/~/media/files/quality%20programs/nsqip/nsqip_puf_userguide_2014.ashx. Accessed October 5, 2018.
2. Berbari EF, Osmon DR, Lahr B, Eckel-Passow JE, Tsaras G, Hanssen AD, Mabry T, Steckelberg J, Thompson R. The Mayo prosthetic joint infection risk score: implication for surgical site infection reporting and risk stratification. *Infect Control Hosp Epidemiol*. 2012;33:774-781.
3. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, Cohen ME. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg*. 2013;217:833-842.e1-3.
4. Bozic KJ, Lau E, Kurtz S, Ong K, Rubash H, Vail TP, Berry DJ. Patient-related risk factors for periprosthetic joint infection and postoperative mortality following total hip arthroplasty in Medicare patients. *J Bone Joint Surg Am*. 2012;94:794-800.
5. Bozic KJ, Ong K, Lau E, Berry DJ, Vail TP, Kurtz SM, Rubash HE. Estimating risk in Medicare patients with THA: an electronic risk calculator for periprosthetic joint infection and mortality. *Clin Orthop Relat Res*. 2013;471:574-583.
6. Centers for Medicare & Medicaid Services. Quality Payment Program Overview. 2018. Available at: https://qpp.cms.gov/about/qpp-overview. Accessed November 16, 2018.
7. Edelstein AI, Kwasny MJ, Suleiman LI, Khakhkhar RH, Moore MA, Beal MD, Manning DW. Can the American College of Surgeons Risk Calculator predict 30-day complications after knee and hip arthroplasty? *J Arthroplasty*. 2015;30:5-10.
8. Fink AS, Campbell DA Jr, Mentzer RM Jr, Henderson WG, Daley J, Bannister J, Hur K, Khuri SF. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. *Ann Surg*. 2002;236:344-353; discussion 353-354.
9. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med*. 2003;29:1043-1051.
10. Guo P, Zeng F, Hu X, Zhang D, Zhu S, Deng Y, Hao Y. Improved variable selection algorithm using a LASSO-type penalty, with an application to assessing hepatitis B infection relevant factors in community residents. *PLoS One*. 2015;10:e0134151.
11. Harris AH. Path from predictive analytics to improved patient outcomes: a framework to guide use, implementation, and evaluation of accurate surgical predictive models. *Ann Surg*. 2017;265:461-463.
12. Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction models for 30-day mortality and complications after total knee and hip arthroplasties for Veteran Health Administration patients with osteoarthritis. *J Arthroplasty*. 2018;33:1539-1545.
13. Harris AHS, Kuo AC, Bozic KJ, Lau E, Bowe T, Gupta S, Giori NJ. American Joint Replacement Registry Risk Calculator does not predict 90-day mortality in veterans undergoing total joint replacement. *Clin Orthop Relat Res*. 2018;476:1869-1875.
14. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New York, NY, USA: John Wiley & Sons, Inc; 2000.
15. Inacio MCS, Pratt NL, Roughead EE, Graves SE. Evaluation of three co-morbidity measures to predict mortality in patients undergoing total joint arthroplasty. *Osteoarthritis Cartilage*. 2016;24:1718-1726.
16. Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, Chong V, Fabri PJ, Gibbs JO, Grover F, Hammermeister K, Irvin G 3rd, McDonald G, Passaro E Jr, Phillips L, Scamman F, Spencer J, Stremple JF. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998;228:491-507.
17. Khuri SF, Henderson WG, Daley J, Jonasson O, Jones RS, Campbell DA Jr, Fink AS, Mentzer RM Jr, Neumayer L, Hammermeister K, Mosca C, Healey N; Principal Investigators of the Patient Safety in Surgery Study. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: the Patient Safety in Surgery study. *Ann Surg*. 2008;248:329-336.
18. Manning DW, Edelstein AI, Alvi HM. Risk prediction tools for hip and knee arthroplasty. *J Am Acad Orthop Surg*. 2016;24:19-27.
19. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS): II. Parsimonious risk models for postoperative adverse outcomes addressing need for laboratory variables and surgeon specialty-specific models. *Ann Surg*. 2016;264:10-22.
20. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS): III. Accurate preoperative prediction of 8 adverse outcomes using 8 predictor variables. *Ann Surg*. 2016;264:23-31.
21. Mu Y, Edwards JR, Horan TC, Berrios-Torres SI, Fridkin SK. Improving risk-adjusted measures of surgical site infection for the national healthcare safety network. *Infect Control Hosp Epidemiol*. 2011;32:970-986.
22. Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: The C statistic. *JAMA*. 2015;314:1063-1064.

23. Romine LB, May RG, Taylor HD, Chimento GF. Accuracy and clinical utility of a peri-operative risk calculator for total knee arthroplasty. *J Arthroplasty*. 2013;28:445-448.
24. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54:774-781.
25. Tibshirani R. Regression Shrinkage and Selection via the lasso. *J R Stat Soc Series B Methodol*. 1996;58:267–288.
26. Wingert NC, Gotoff J, Parrilla E, Gotoff R, Hou L, Ghanem E. The ACS NSQIP Risk Calculator is a fair predictor of acute periprosthetic joint infection. *Clin Orthop Relat Res*. 2016;474: 1643-1648.
27. Wuerz TH, Kent DM, Malchau H, Rubash HE. A nomogram to predict major complications after hip and knee arthroplasty. *J Arthroplasty*. 2014;29:1457-1462.
28. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE). 2017 *Procedure-specific Measure Updates and Specifications Report Hospital-level Risk-standardized Complication Measure.* New Haven, CT, USA: Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE); 2017.